ECONOMETRICS

Bruce E. Hansen

University of Wisconsin

www.ssc.wisc.edu/~bhansen

This Revision: January 10, 2010
Comments Welcome

Contents

Preface

This book is intended to serve as the textbook for a first-year graduate course in econometrics. It can be used as a stand-alone text, or be used as a supplement to another text.

Students are assumed to have an understanding of multivariate calculus, probability theory, linear algebra, and mathematical statistics. A prior course in undergraduate econometrics would be helpful, but not required.

For reference, some of the basic tools of matrix algebra, probability, and statistics are reviewed in the Appendix.

For students wishing to deepen their knowledge of matrix algebra in relation to their study of econometrics, I recommend *Matrix Algebra* by Abadir and Magnus (2005).

An excellent introduction to probability and statistics is *Statistical Inference* by Casella and Berger (2002). For those wanting a deeper foundation in probability, I recommend Ash (1972) or Billingsley (1995). For more advanced statistical theory, I recommend Lehmann and Casella (1998), van der Vaart (1998), Shao (2003), and Lehmann and Romano (2005).

For further study in econometrics beyond this text, I recommend Davidson (1994) for asymptotic theory, Hamilton (1994) for time-series methods, Wooldridge (2002) for panel data and discrete response models, and Li and Racine (2007) for nonparametrics and semiparametric econometrics. Beyond these texts, the *Handbook of Econometrics* series provides advanced summaries of contemporary econometric methods and theory.

As this is a manuscript in progress, some parts are quite incomplete, in particular the later sections of the manuscript. Hopefully one day these sections will be fleshed out and completed in more detail.

Chapter 1

Introduction

1.1 What is Econometrics?

The term "econometrics" is believed to have been crafted by Ragnar Frisch (1895-1973) of Norway, one of the three principle founders of the Econometric Society, first editor of the journal *Econometrica*, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of *Econometrica* for an explanation of the discipline.

> A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems...."
>
> But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

This definition remains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data.

Within the field of econometrics there are sub-divisions and specializations. Econometric theory concerns the development of tools and methods, and the study of the properties of econometric methods. Applied econometrics is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal

paper "The probability approach in econometrics", *Econometrica* (1944). Haavelmo argued that quantitative economic models must necessarily be probability models (by which today we would mean *stochastic*). Deterministic models are blatently inconsistent with observed economic quantities, and it is incohorent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that the best way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Haavelmo's probability approach was quickly embraced by the economics profession. Today no quantitative work in economics shuns its fundamental vision.

While all economists embrace the probability approach, there has been some evolution in its implementation.

The **structural approach** is the closest to Haavelmo's original idea. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as "taking their model seriously." The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation. In this case, how should we interpret structural econometric analysis? The **quasi-structural approach** to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Closely related is the **semiparametric approach**. A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least-squares and the Generalized Method of Moments. The semiparametric approach dominates contemporary econometrics, and is the main focus of this textbook.

Another branch of quantitative structural economics is the **calibration approach.** Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics as inappropriate for approximate models, and instead selects parameters by matching model and data moments using non-statistical *ad hoc*[1] methods.

1.3 Econometric Terms and Notation

In a typical application, an econometrician has a set of repeated measurements on a set of variables. For example, in an labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the **data, dataset**, or **sample**.

We use the term **observations** to refer to the distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region. An individual observation could also be a measurement at a point in time, such as quarterly GDP or a daily interest rate.

Economists typically denote variables by the italized roman characters y, x, and/or z. The convention in econometrics is to use the character y to denote the variable to be explained, while

[1] *Ad hoc* means "for this purpose" – a method designed for a specific problem – and not based on a generalizable principle.

the characters x and z are used to denote the conditioning (explaining) variables.

Following mathematical convention, real numbers (elements of the real line $\mathbb{R}$) are written using lower case italics such as y, and vectors (elements of $\mathbb{R}^k$) by lower case bold italics such as $\boldsymbol{x}$, e.g.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Upper case bold italics such as $\boldsymbol{X}$ will be used for matrices.

We typically denote the number of observations by the natural number n, and subscript the variables by the index i to denote the individual observation, e.g. y_i, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. In some contexts we use indices other than i, such as in time-series applications where the index t is common, and in panel studies we typically use the double index it to refer to individual i at a time period t.

The i'th observation is the set $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$.

It is proper mathematical practice to use upper case X for random variables and lower case x for realizations or specific values. This practice is not commonly followed in econometrics because instead we use upper case to denote matrices. Thus the notation y_i will in some places refer to a random variable, and in other places a specific realization. Hopefully there will be no confusion as the use should be evident from the context.

As we mentioned before, ideally each observation consists of a set of measurements on the list of variables. In practice it is common to find that some variables are not measured for some observations, and in these cases we describe these variables or observations as **unobserved** or **missing**.

We typically use Greek letters such as β, θ and σ^2 to denote unknown parameters of an econometric model, and will use boldface, e.g. $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$, when these are vector-valued. Estimates are typically denoted by putting a hat "^", tilde "~" or bar "-" over the corresponding letter, e.g. $\hat{\beta}$ and $\tilde{\beta}$ are estimates of β.

The covariance matrix of an econometric estimator will typically be written using the capital boldface $\boldsymbol{V}$, often with a subscript to denote the estimator, e.g. $\boldsymbol{V}_{\hat{\beta}} = \text{var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)$ as the covariance matrix for $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. Hopefully without causing confusion, we will use the notation $\boldsymbol{V}_{\boldsymbol{\beta}}$ to denote the asymptotic covariance matrix of $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ (the variance of the asymptotic distribution). Estimates will be denoted by appending hats or tildes, e.g. $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{V}_{\boldsymbol{\beta}}$.

1.4 Observational Data

A common econometric question is to quantify the impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path after they mature and enter the labor force. The differences between the groups would be direct measurements of the effects of different levels of education. However, experiments such as this would be widely condemned as immoral! Consequently, we see few non-laboratory experimental data sets in economics.

Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage. With such data we can measure the joint distribution of these variables, and assess the joint dependence. But from observational data it is difficult to infer causality, as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices as well as their educational level. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distibution alone may not be able to distinguish between these explanations.

Most economic data sets are observational, not experimental. This means that all variables must be treated as random and possibly jointly determined.

This discussion means that it is difficult to infer causality from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will return to a discussion of some of these issues in Chapter 11.

1.5 Standard Data Structures

There are three major types of economic data sets: cross-sectional, time-series, and panel. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms or other economic agents. In many contemporary econometric cross-section studies the sample size n is quite large. It is conventional to assume that cross-sectional observations are mutually independent. Most of this text is devoted to the study of cross-section data.

Time-series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence so the random sampling assumption is inappropriate. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly) so the sample size can be much smaller than in typical cross-section studies. The exception is financial data where data are available at a high frequency (weekly, data, hourly, or tick-by-tick) so sample sizes can be quite large.

Panel data combines elements of cross-section and time-series. These data sets consist of a set of individuals (typically persons, households, or corporations) surveyed repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but a given individual's observations are mutually dependent. This is a modified random sampling environment.

Data Structures

- Cross-section
- Time-series
- Panel

Some contemporary econometric applications combine elements of cross-section, time-series, and panel data modeling. These include models of spatial correlation and clustering.

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the i'th observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ is independent of the j'th observation $(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$ for $i \neq j$. (Sometimes the label "independent" is misconstrued. It is a statement about the relationship between observations i and j, not a statement about the relationship between y_i and $\boldsymbol{x}_i$ and/or $\boldsymbol{z}_i$.)

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case we say that the data are **independent and identically distributed** or **iid**. We call this a **random sample**. For most of this text we will assume that our observations come from a random sample.

Definition 1.5.1 *The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are a **random sample** if they are mutually independent and identically distributed (**iid**) across $i = 1, ..., n$.*

In the random sampling framework, we think of an individual observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ as a realization from a joint probability distribution $F(y, \boldsymbol{x}, \boldsymbol{z})$ which can call the **population**. This "population" is infinitely large. This abstraction can be a source of confusion as it does not correspond to a physical population in the real world. The distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample. The *assumption* of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

The random sampling framework was a major intellectural breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as they were viewed as inappropraite. The random sampling framework enabled economic samples to be viewed as homogenous and random, a necessary precondition for the application of statistical methods.

1.6 Sources for Economic Data

Fortunately for economists, the the internet provides a convenient forum for dissemination of economic data. Many large-scale economic datasets are available without charge from governmental agencies. An excellent starting point is the Resources for Economists Data Links, available at `rfe.org`. From this site you can find almost every publically available economic data set. Some specific data sources of interest include

- Bureau of Labor Statistics
- US Census
- Current Population Survey
- Survey of Income and Program Participation
- Panel Study of Income Dynamics
- Federal Reserve System (Board of Governors and regional banks)
- National Bureau of Economic Research

- U.S. Bureau of Economic Analysis
- CompuStat
- International Financial Statistics

Another good source of data is from authors of published empirical studies. Most journals in economics require authors of published papers to make their datasets generally available. For example, in its instructions for submission, *Econometrica* states:

> *Econometrica* has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication and some limited sensitivity analysis.

The *American Economic Review* states:

> All data used in analysis must be made available to any researcher for purposes of replication.

The *Journal of Political Economy* states:

> It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.

If you are interested in using the data from a published paper, first check the journal's website, as many journals archive data and replication programs online. Second, check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs. If these investigations fail, email the author(s), politely requesting the data. You may need to be persistent.

As a matter of professional etiquette, all authors absolutely have the obligation to make their data and programs available. Unfortunately, many fail to do so, and typically for poor reasons. The irony of the situation is that it is typically in the best interests of a scholar to make as much of their work (including all data and programs) freely available, as this only increases the likelihood of their work being cited and having an impact.

Keep this in mind as you start your own empirical project. Remember that as part of your end product, you will need (and want) to provide all data and programs to the community of scholars. The greatest form of flattery is to learn that another scholar has read your paper, wants to extend your work, or wants to use your empirical methods. In addition, public openness provides a healthy incentive for transparency and integrity in empirical analysis.

1.7 Econometric Software

Economists use a variety of econometric, statistical, and programming software.

STATA (www.stata.com) is a powerful statistical program with a broad set of pre-programmed econometric and statistical tools. It is quite popular among economists, and is continuously being updated with new methods. It is an excellent package for most econometric analysis, but is limited when you want to use new or less-common econometric methods which have not yet been programed.

GAUSS (www.aptech.com), MATLAB (www.mathworks.com), and Ox (www.oxmetrics.net) are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programed in these languages and are available on the web. The advantage of these packages is that you are in complete control of your analysis, and it is

easier to program new methods than in STATA. Some disadvantages are that you have to do much of the programming yourself, programming complicated procedures takes significant time, and programming errors are hard to prevent and difficult to detect and eliminate.

R (www.r-project.org) is an integrated suite of statistical and graphical software that is flexible, open source, and best of all, free!

For highly-intensive computational tasks, some economists write their programs in a standard programming language such as Fortran or C. This can lead to major gains in computational speed, at the cost of increased time in programming and debugging.

As these different packages have distinct advantages, many empirical economists end up using more than one package. As a student of econometrics, you will learn at least one of these packages, and probably more than one.

1.8 Reading the Manuscript

Chapters 2 through 7 deal with the core linear regression and projection models. Chapter 8 introduces the bootstrap. Chapters 9 through 11 deal with the Generalized Method of Moments, empirical likelihood and endogeneity. Chapters 12 and 13 cover time series, and Chapters 14, 15 and 16 cover limited dependent variables, panel data, and nonparametrics. Reviews of matrix algebra, probability theory, asymptotic theory, maximum likelihood, and numerical optimization can be found in the appendix.

Chapter 2

Regression and Projection

2.1 Introduction

The most commonly applied econometric tool is least-squares estimation, also known as **regression**. As we will see, least-squares is a tool to estimate an approximate conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).

In this chapter we abstract from estimation, and focus on the probabilistic foundation of the regression model and its projection approximation.

2.2 Notation

We let y denote the dependent variable and let $(x_1, x_2, ..., x_k)$ denote the k regressors. Throughout this section we maintain the assumption that the variables are stochastic.

Assumption 2.2.1 *$(y, x_1, x_2, ..., x_k)$ is a random vector with a joint probability distribution such that*

1. $\mathbb{E}y^2 < \infty$.
2. $\mathbb{E}x_j^2 < \infty$ *for* $j = 1, ..., k$.

The finite second moment conditions imposed in Assumption 2.2.1.1 and 2.2.1.2 imply that the variables have finite means and variances.

It is convenient to write the set of regressors as a vector in $\mathbb{R}^k$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}. \tag{2.1}$$

For most of our analysis it is unimportant whether the regressors $\boldsymbol{x}$ come from continuous or discrete distributions. As an example of a discrete variable, many regressors in econometric applications are binary, taking on only the values 0 and 1, and are called **dummy variables**.

For some purposes, the same is true about the dependent variable – it could be continuous or discrete. But when the dependent variable is discrete we typically use specific models and techniques built for this purpose (see Chapter 14).

2.3 Conditional Mean

To study how the distribution of y varies with the variables $\boldsymbol{x}$ in the population, we start with $f(y \mid \boldsymbol{x})$, the conditional density of y given $\boldsymbol{x}$.

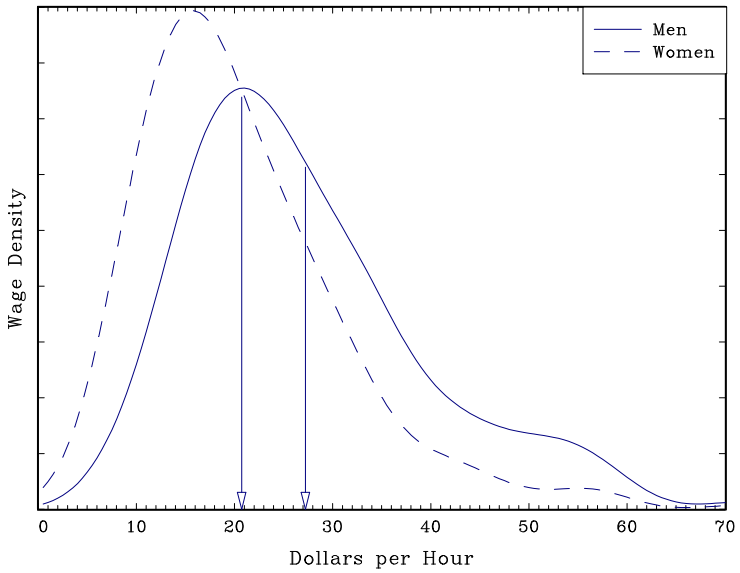

Figure 2.1: Wage Densities for White College Grads with 10-15 Years Work Experience

To illustrate, Figure 2.1 displays the density[1] of hourly wages for men and women, from the population of white non-military wage earners in the U.S. with a college degree and 10-15 years of potential work experience. These are conditional density functions – the density of hourly wages conditional on race, gender, education and experience. The two density curves show the effect of gender on the distribution of wages, holding the other variables constant.

While it is easy to observe that the two densities are unequal, it is useful to have numerical measures of the difference. An important summary measure is the **conditional mean**[2]

$$m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x}) = \int_{-\infty}^{\infty} y f(y \mid \boldsymbol{x})\, dy. \tag{2.2}$$

The function $m(\boldsymbol{x})$ varies with the vector $\boldsymbol{x}$ and is thus a function from $\mathbb{R}^k$ to $\mathbb{R}$. The conditional mean $m(\boldsymbol{x})$ is sometimes called the **regression function**. In general, $m(\boldsymbol{x})$ can have arbitrary shape, although in some cases an economic model may dictate a specific shape restriction (such as monotonicity) or a specific functional form (such as linearity). The regression function $m(\boldsymbol{x})$ is defined for values of $\boldsymbol{x}$ in the support[3] of $\boldsymbol{x}$. Thus when $\boldsymbol{x}$ has a discrete distribution then $m(\boldsymbol{x})$ is defined for those values of $\boldsymbol{x}$ with positive probability. When $\boldsymbol{x}$ has a continuous distribution with density $f_{\boldsymbol{x}}(\boldsymbol{x})$ then $m(\boldsymbol{x})$ is defined for those values of $\boldsymbol{x}$ for which $f_{\boldsymbol{x}}(\boldsymbol{x}) > 0$.

In the example presented in Figure 2.1, the mean wage for men is \$27.22, and that for women is \$20.73. These are indicated in Figure 2.1 by the arrows drawn to the x-axis. These values are the conditional means of U.S. wages in 2004 (conditional on gender, and conditional for white non-military wages earners with a college degree and 10-15 years of work experience).

Take a closer look at the density functions displayed in Figure 2.1. You can see that the right tail of the density is much thicker than the left tail. These are asymmetric (skewed) densities,

[1] These are nonparametric density estimates using a normal kernel with the bandwidth selected by cross-validation. See Chapter 16. The data are from the 2004 Current Population Survey.

[2] The conditional mean exists if $\mathbb{E}|y| < \infty$. For a rigorous definition see Section 2.16.

[3] The support of a random vector $\boldsymbol{x}$ is the closed set of points for which its distribution $F(\boldsymbol{x})$ is increasing in all elements of $\boldsymbol{x}$.

which is a common feature of many economic variables. When a distribution is skewed, the mean is not necessarily a good summary of the central tendency. In this context it is often convenient to transform the data by taking the (natural) logarithm[4]. Figure 2.2 shows the density of log hourly wages for the same population, with mean log hourly wages (3.21 and 2.91, respectively) drawn in with the arrows. The difference between the mean log wage of men and women is 0.30, which implies a 30% average wage difference for this population. The difference in the mean log wage is a more robust measure of the typical wage gap than the difference in the untransformed wage means. For this reason, wage regressions typically use log wages as a dependent variable rather than the level of wages.

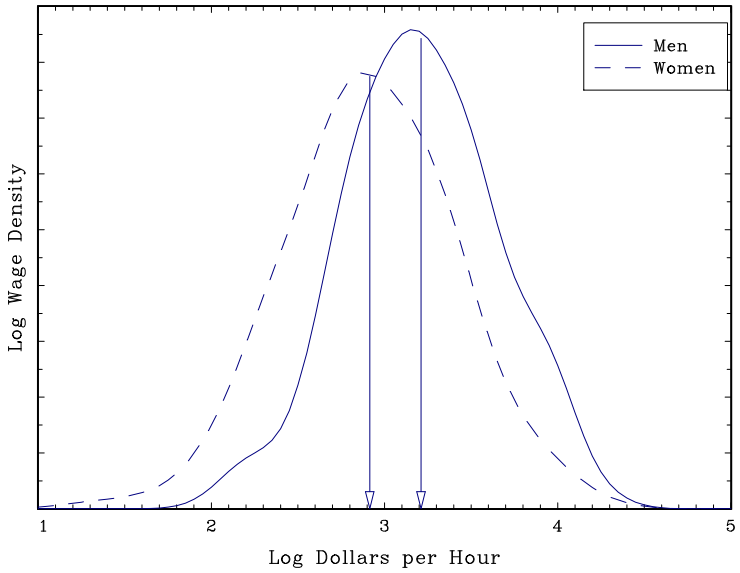

Figure 2.2: Log Wage Densities for White College Grads with 10-15 Years Work Experience

The comparisons in Figures 2.1 and 2.2 are facilitated by the fact that the control variable (gender) is binary. When the distribution of the control variable takes on multiple values or is continuous, then comparisons become more complicated. To illustrate, Figure 2.3 displays a scatter plot[5] of log wages against education levels. Assuming for simplicity that this is the true joint distribution, the solid line displays the conditional expectation of log wages varying with education. The conditional expectation function is close to linear; the dashed line is a linear projection approximation which will be discussed in Section 2.9. The main point to be learned from Figure 2.3 is that the conditional expectation is a useful summary of the central tendency of the conditional distribution when the control variable takes multiple values. Of particular interest to graduate students may be the observation that difference between a B.A. and a Ph.D. degree in mean log hourly wages is 0.36, implying an average 36% difference in wage levels.

As another example, Figure 2.4 displays the conditional mean[6] of log hourly wages as a function of labor market experience. The solid line is the conditional mean. We see that the conditional mean is strongly non-linear and non-monotonic. The main lesson to be learned at this point from Figure 2.4 is that conditional expectations can be quite non-linear.

[4] Mathematically, this is equivalent to measuring the central tendency by the conditional geometric mean $\exp\left(\mathbb{E}\left(\log y \mid \boldsymbol{x}\right)\right)$. For example, the conditional geometric means for the densities in Figure 2.1 are \$24.78 and \$18.36, respectively.

[5] White non-military male wage earners with 10-15 years of potential work experience.

[6] In the population of white non-military male wage earners with 12 years of education.

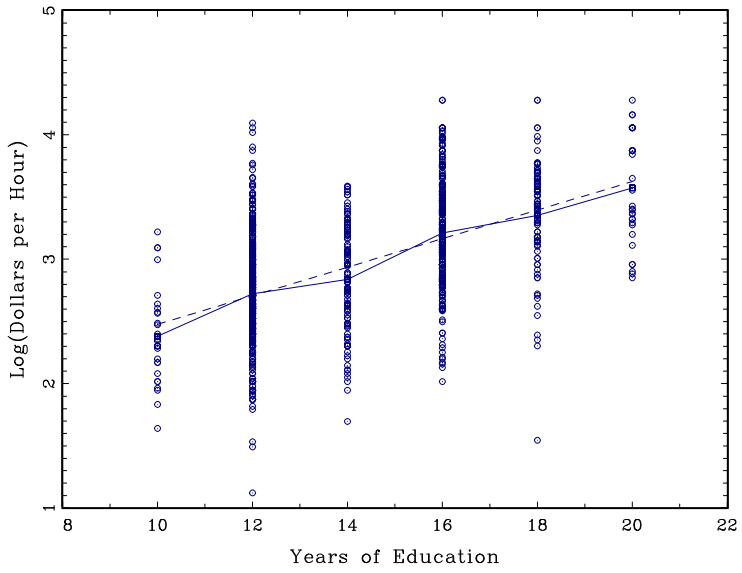

Figure 2.3: Scatter Plot and Conditional Mean of Log Wages Given Education

2.4 Regression Error

The regression error e is defined as the difference between y and its conditional mean (2.2) evaluated at the random vector $\boldsymbol{x}$:

$$e = y - m(\boldsymbol{x}).$$

By construction, this yields the formula

$$y = m(\boldsymbol{x}) + e. \tag{2.3}$$

It is useful to understand that the regression error is derived from the joint distribution of $(y, \boldsymbol{x})$, and so its properties are derived from this construction. We now discuss some of these properties.

Theorem 2.4.1 ***Properties of the regression error*** e.
Under Assumption 2.2.1,

1. $\mathbb{E}(e \mid \boldsymbol{x}) = 0$.
2. $\mathbb{E}(e) = 0$.
3. $\mathbb{E}(h(\boldsymbol{x})e) = 0$ *for any function* $h(\cdot)$ *such that* $\mathbb{E}h(\boldsymbol{x})^2 < \infty$
4. $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$.

Proof of Theorem 2.4.1.1:

By the definition of e and the linearity of conditional expectations,

$$\begin{aligned} \mathbb{E}(e \mid \boldsymbol{x}) &= \mathbb{E}((y - m(\boldsymbol{x})) \mid \boldsymbol{x}) \\ &= \mathbb{E}(y \mid \boldsymbol{x}) - \mathbb{E}(m(\boldsymbol{x}) \mid \boldsymbol{x}) \\ &= m(\boldsymbol{x}) - m(\boldsymbol{x}) = 0. \end{aligned}$$

Proofs of the remaining parts of Theorem 2.4.1.1 are left to Exercise 2.1.

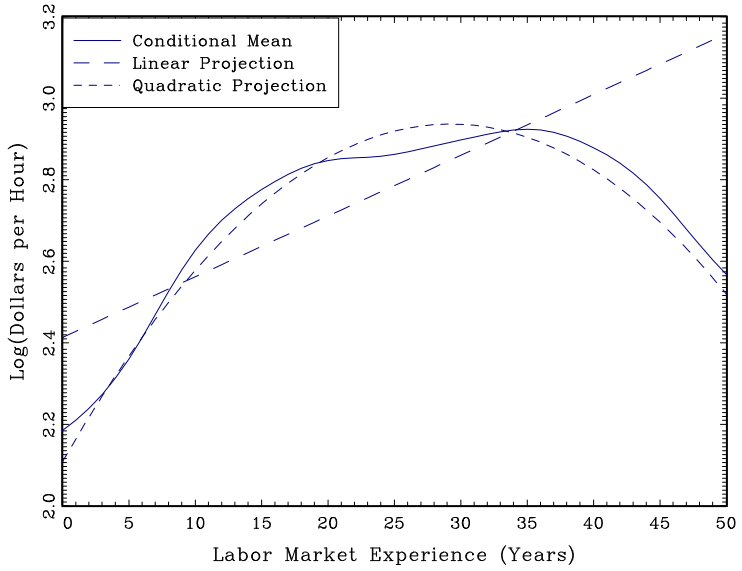

Figure 2.4: Log Hourly Wage as a Function of Experience

The equations

$$\begin{aligned} y &= m(\boldsymbol{x}) + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0. \end{aligned}$$

are often stated jointly as the regression framework. It is important to understand that this is a framework, not a model, because no restrictions have been placed on the joint distribution of the data. These equations hold true by definition. A regression model imposes further restrictions on the permissible class of regression functions $m\left(\boldsymbol{x}\right)$.

The condition $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$ is the key implication of the conditional mean model. This equation is sometimes called a conditional mean restriction, since the conditional mean is restricted to equal zero. The property is also sometimes called **mean independence**, for the conditional mean of e is 0 and thus independent of $\boldsymbol{x}$. It is quite important to understand, however, that it does not imply that the distribution of e is independent of $\boldsymbol{x}$. Sometimes the assumption "e is independent of $\boldsymbol{x}$" is added as a convenient simplification, but it is not generic feature of regression. Typically and generally, e and $\boldsymbol{x}$ are jointly dependent, even though the conditional mean of e is zero.

As a simple example, suppose that $y = xu$ where x and u are independent and $\mathbb{E}u = 1$. Then $\mathbb{E}\left(y \mid x\right) = x$ so the regression equation is $y = x + e$ where $e = x(u - 1)$. Yet e is not independent of x, even though $\mathbb{E}\left(e \mid x\right) = 0$.

2.5 Best Predictor

Given a realized value of $\boldsymbol{x}$, we can view $m(\boldsymbol{x})$ as a predictor or forecast of y. The prediction error is $e = y - m(\boldsymbol{x})$, which is random. A non-stochastic measure of the magnitude of the prediction error is the expectation of the squared error, or mean squared error

$$\mathbb{E}\left(y - m\left(\boldsymbol{x}\right)\right)^2 = \mathbb{E}e^2 \equiv \sigma^2. \tag{2.4}$$

The parameter σ^2 is also known as the variance of the regression error.

It turns out that the the conditional mean is a good predictor of y in the sense that it has the lowest mean squared error among all predictors. This holds regardless of the joint distribution of $(y, \boldsymbol{x})$. We state this formally in the following result.

Theorem 2.5.1 ***Conditional Mean as Best Predictor***
Let $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ be the conditional mean and let $g(\boldsymbol{x})$ be any other predictor of y given $\boldsymbol{x}$. Under Assumption 2.2.1,

$$\mathbb{E}(y - g(\boldsymbol{x}))^2 \geq \mathbb{E}(y - m(\boldsymbol{x}))^2.$$

Proof of Theorem 2.5.1: Since $y = m(\boldsymbol{x}) + e$, the mean squared error using $g(\boldsymbol{x})$ is

$$\begin{aligned}
\mathbb{E}(y - g(\boldsymbol{x}))^2 &= \mathbb{E}(e + m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&= \mathbb{E}e^2 + 2\mathbb{E}(e(m(\boldsymbol{x}) - g(\boldsymbol{x}))) + \mathbb{E}(m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&= \mathbb{E}e^2 + \mathbb{E}(m(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \\
&\geq \mathbb{E}e^2 = \mathbb{E}(y - m(\boldsymbol{x}))^2
\end{aligned}$$

where the third equality uses Theorem 2.4.1.3. The right-hand-side after the third equality is minimized by setting $g(\boldsymbol{x}) = m(\boldsymbol{x})$, yielding the final inequality.

2.6 Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution, it does not provide information about the spread of the distribution. A common measure of the dispersion is the **conditional variance**.

Definition 2.6.1 *The **conditional variance** of y given $\boldsymbol{x}$ is*

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= \operatorname{var}(y \mid \boldsymbol{x}) \\
&= \mathbb{E}\left(y^2 \mid \boldsymbol{x}\right) - (\mathbb{E}(y \mid \boldsymbol{x}))^2 \\
&= \mathbb{E}\left((y - \mathbb{E}(y \mid \boldsymbol{x}))^2 \mid \boldsymbol{x}\right) \\
&= \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right)
\end{aligned}$$

Generally, $\sigma^2(\boldsymbol{x})$ is a non-trivial function of $\boldsymbol{x}$ and can take any form subject to the restriction that it is non-negative. The **conditional standard deviation** is its square root $\sigma(\boldsymbol{x}) = \sqrt{\sigma^2(\boldsymbol{x})}$. One way to think about $\sigma^2(\boldsymbol{x})$ is that it is the conditional mean of e^2 given $\boldsymbol{x}$.

As an example of how the conditional variance depends on observables, compare the conditional wage densities for men and women displayed in Figure 2.1. The difference between the densities is not just a location shift, but is also a difference in spread. Specifically, we can see that the density for men's wages is somewhat more spread out than that for women, while the density for women's wages is somewhat more peaked. Indeed, the conditional standard deviation for men's wages is 12.1 and that for women is 10.5. So while men have higher average wages, they are also somewhat more dispersed.

Many econometric studies focus on the conditional mean $m(\boldsymbol{x})$ and either ignore the conditional variance $\sigma^2(\boldsymbol{x})$, treat it as a constant $\sigma^2(\boldsymbol{x}) = \sigma^2$, or treat it as a nuisance parameter (a parameter not of primary interest). This may be unfortunate as dispersion is relevant to many economic topics, including income and wealth distribution, economic inequality, and price dispersion.

The perverse consequences of a narrow-minded focus on the mean has been parodied in a classic joke:

An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, "On average I feel just fine."

Clearly, the economist in question ignored variance!

2.7 Homoskedasticity and Heteroskedasticity

An important special case obtains when the conditional variance of the regression error $\sigma^2(\boldsymbol{x})$ is a constant and independent of $\boldsymbol{x}$. This is called **homoskedasticity**.

Definition 2.7.1 *The error is* ***homoskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$ *does not depend on* $\boldsymbol{x}$.

In the general case where $\sigma^2(\boldsymbol{x})$ depends on $\boldsymbol{x}$ we say that the error e is **heteroskedastic**.

Definition 2.7.2 *The error is* ***heteroskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2(\boldsymbol{x})$ *depends on* $\boldsymbol{x}$.

Even when the error is heteroskedastic we still define the unconditional variance σ^2 of the error e as in (2.4). It may be helpful to notice that by using iterated expectations the unconditional variance can be written as the expected conditional error variance

$$\sigma^2 = \mathbb{E}\left(e^2\right) = \mathbb{E}\left(\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right)\right) = \mathbb{E}\left(\sigma^2(\boldsymbol{x})\right).$$

Thus σ^2 is well-defined whether or not the error is homoskedastic or heteroskedastic.

Some older or introductory textbooks describe heteroskedasticity as the case where "the variance of e varies across observations". This is a poor and confusing definition. It is more constructive to understand that heteroskedasticity is the case where the conditional variance $\sigma^2(\boldsymbol{x})$ depends on the variables $\boldsymbol{x}$. (Once again, recall Figure 2.1 and how the variance of wages varies between men and women.)

Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification, and describe heteroskedasticity as an exception or deviance. This description has influenced many generations of economists, but it is unfortunately backwards. The correct view is that heteroskedasticity is generic and "standard", while homoskedasticity is unusual and exceptional. The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.

In apparent contraction to the above statement, we will still frequently impose the homoskedasticity assumption when making theoretical investigations into the properties of regression techniques. The reason is that in many cases homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning. It should always be remembered, however, that homoskedasticity is never imposed because it is believed to be a correct feature of an empirical regression, but rather because of its simplicity.

2.8 Linear Regression

An important special case of (2.3) is when the conditional mean function $m(\boldsymbol{x})$ is linear in $\boldsymbol{x}$ (or linear in functions of $\boldsymbol{x}$). In this case we can write the mean equation as

$$m(\boldsymbol{x}) = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k.$$

Notationally it is convenient to write this as a simple function of the vector $\boldsymbol{x}$. An easy way to do so is to augment the regressor vector $\boldsymbol{x}$ by listing the number "1" as an element. We call this the "constant" and the corresponding coefficient is called the "intercept". Equivalently, assuming that the first element[7] of the vector $\boldsymbol{x}$ is the intercept, then $x_1 = 1$. Thus (2.1) has been redefined as the $k \times 1$ vector

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}. \tag{2.5}$$

With this redefinition, then the mean equation is

$$\begin{aligned} m(\boldsymbol{x}) &= x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k \\ &= \boldsymbol{x}'\boldsymbol{\beta} \end{aligned} \tag{2.6}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{2.7}$$

is a $k \times 1$ coefficient vector. This is called the linear regression model.

Linear Regression

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}(e \mid \boldsymbol{x}) &= 0 \end{aligned}$$

If in addition the error is homoskedastic, we call this the homoskedastic linear regression model.

Homoskedastic Linear Regression

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}(e \mid \boldsymbol{x}) &= 0 \\ \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) &= \sigma^2 \end{aligned}$$

[7] The order doesn't matter. It could be any element.

2.9 Best Linear Predictor

While the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is the best predictor of y among all functions of $\boldsymbol{x}$, its functional form is typically unknown. In particular, the linear equation of the previous section is empirically unlikely to be accurate. In practice it is more realistic to view the linear specification (2.6) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.5.1 showed that the conditional mean $m(\boldsymbol{x})$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define a linear approximation to the conditional mean function as the linear function with the lowest mean squared error among all linear predictors.

To be precise, a linear predictor for y given $\boldsymbol{x}$ is $\boldsymbol{x}'\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. The mean squared error of this predictor is

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2.$$

The **best linear predictor** of y given $\boldsymbol{x}$ is defined by finding the vector $\boldsymbol{\beta}$ which minimizes $S(\boldsymbol{\beta})$.

Definition 2.9.1 *The **Best Linear Predictor** of y given $\boldsymbol{x}$ is $\boldsymbol{x}'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ minimizes the mean squared error*

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2.$$

The minimizer

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\operatorname{argmin}}\, S(\boldsymbol{\beta}) \tag{2.8}$$

*is called the **Linear Projection Coefficient**.*

The quadratic structure of $S(\boldsymbol{\beta})$ means that we can solve explicitly for $\boldsymbol{\beta}$. The mean squared prediction error can be written out as a quadratic function of $\boldsymbol{\beta}$:

$$S(\boldsymbol{\beta}) = \mathbb{E}y^2 - 2\boldsymbol{\beta}'\mathbb{E}(\boldsymbol{x}y) + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}$$

The first-order condition for minimization (from Appendix A.9) is

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\mathbb{E}(\boldsymbol{x}y) + 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}. \tag{2.9}$$

This has a unique solution under the following condition.

Assumption 2.9.1 $\boldsymbol{Q} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ *is invertible.*

The matrix $\boldsymbol{Q}$ is sometimes called the **design matrix**, as in experimental settings the researcher is able to control $\boldsymbol{Q}$ by manipulating the distribution of the regressors $\boldsymbol{x}$.

Rewriting (2.9) as

$$2\mathbb{E}(\boldsymbol{x}y) = 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}$$

dividing by 2, and then inverting the $k \times k$ matrix $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$, we obtain the solution for $\boldsymbol{\beta}$.

Theorem 2.9.1 ***Linear Projection Coefficient***
Under Assumptions 2.2.1 and 2.9.1, the linear projection coefficient equals

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right). \tag{2.10}$$

It is worth taking the time to understand the notation involved in the expression (2.10). $\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ is a $k \times k$ matrix and $\mathbb{E}\left(\boldsymbol{x}y\right)$ is a $k \times 1$ column vector. Therefore, alternative expressions such as $\frac{\mathbb{E}(\boldsymbol{x}y)}{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')}$ or $\mathbb{E}\left(\boldsymbol{x}y\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}$ are incoherent and incorrect.

Given the definition of $\boldsymbol{\beta}$ in (2.10), $\boldsymbol{x}'\boldsymbol{\beta}$ is the best linear predictor for y. The **projection error** is

$$e = y - \boldsymbol{x}'\boldsymbol{\beta}. \tag{2.11}$$

The error e from the linear prediction equation is equal to the error from the regression equation when (and only when) the conditional mean is linear in $\boldsymbol{x}$, otherwise they are distinct.

Rewriting, we obtain a decomposition of y into linear predictor and error

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e. \tag{2.12}$$

This completes the derivation of the model. We call $\boldsymbol{x}'\boldsymbol{\beta}$ the best linear predictor of y given $\boldsymbol{x}$**,** or the linear projection of y onto $\boldsymbol{x}$. In general we call equation (2.12) the **linear projection model**.

The following are important properties of the model.

Theorem 2.9.2 ***Properties of Linear Projection Model***
Under Assumptions 2.2.1 and 2.9.1, then (2.11) and (2.12) exist and are unique,

$$\sigma^2 = \mathbb{E}\left(e^2\right) < \infty, \tag{2.13}$$

and

$$\mathbb{E}\left(\boldsymbol{x}e\right) = \mathbf{0}. \tag{2.14}$$

A complete proof of Theorem 2.9.1 is presented below.

We have shown that under mild regularity conditions, for any pair $(y, \boldsymbol{x})$ we can define a linear equation (2.12) with the properties listed in Theorem 2.9.1. No additional assumptions are required. Thus the linear model (2.12) exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation (2.12) is defined as the best linear predictor. In contrast, in many economic models the parameter $\boldsymbol{\beta}$ may be defined within the model. In this case (2.10) may not hold and the implications of Theorem 2.9.1 may be false. These structural models require alternative estimation methods, and are discussed in Chapter 11.

Linear Projection Model

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e. \\ \mathbb{E}\left(\boldsymbol{x}e\right) &= \mathbf{0} \\ \boldsymbol{\beta} &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right) \end{aligned}$$

Equation (2.14) is a set of k equations, one for each regressor. In other words, (2.14) is equivalent to

$$\mathbb{E}\left(\boldsymbol{x}_j e\right) = 0 \tag{2.15}$$

for $j = 1, ..., k$. As in (2.5), the regressor vector $\boldsymbol{x}$ typically contains a constant, e.g. $x_1 = 1$. In this case (2.15) for $j = 1$ is the same as

$$\mathbb{E}\left(e\right) = 0. \tag{2.16}$$

Thus the projection error has a mean of zero when the regression contains a constant. (When $\boldsymbol{x}$ does not have a constant, this is not guarenteed. As it is desireable for e to have a zero mean, this is a good reason to always include a constant in any regression.)

It is also useful to observe that since $\operatorname{cov}(\boldsymbol{x}_j, e) = \mathbb{E}\left(\boldsymbol{x}_j e\right) - \mathbb{E}\left(\boldsymbol{x}_j\right)\mathbb{E}\left(e\right)$, then (2.15)-(2.16) together imply that the variables $\boldsymbol{x}_j$ and e are uncorrelated.

Invertibility and Identification

The vector (2.10) exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ is invertible. Observe that for any non-zero $\boldsymbol{\alpha} \in \mathbb{R}^k$,

$$\boldsymbol{\alpha}'\boldsymbol{Q}\boldsymbol{\alpha} = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\alpha}\right) = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 \geq 0$$

so $\boldsymbol{Q}$ by construction is positive semi-definite. It is invertible if and only if it is positive definite, which requires that for all non-zero $\boldsymbol{\alpha}$, $\mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 > 0$. Equivalently, there cannot exist a non-zero vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\boldsymbol{x} = 0$ identically. This occurs when redundant variables are included in $\boldsymbol{x}$. In order for $\boldsymbol{\beta}$ to be uniquely defined, this situation must be excluded.

Theorem 2.9.1 shows that the linear projection coefficient $\boldsymbol{\beta}$ is **identified** (uniquely determined) under Assumptions 2.2.1 and 2.9.1. The key is invertibility of $\boldsymbol{Q}$. Otherwise, there is no unique solution to the equation

$$\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta} = \mathbb{E}\left(\boldsymbol{x}y\right). \tag{2.17}$$

When $\boldsymbol{Q}$ is not invertible there are multiple solutions to (2.17), all of which yield an equivalent best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$. In this case the coefficient $\boldsymbol{\beta}$ is **not identified** as it does not have a unique value. Even so, the best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$ still identified. One solution is to set

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-}\mathbb{E}\left(\boldsymbol{x}y\right)$$

where $\boldsymbol{A}^{-}$ denotes the generalized inverse of $\boldsymbol{A}$ (see Appendix A.5).

Proof of Theorem 2.9.1

We first show that the moments $\mathbb{E}\left(\boldsymbol{x}\boldsymbol{y}\right)$ and $\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are finite and well defined. First, it is useful to note that Assumption 2.2.1 implies that

$$\mathbb{E}\left\|\boldsymbol{x}\right\|^2 = \mathbb{E}\left(\boldsymbol{x}'\boldsymbol{x}\right) = \sum_{j=1}^{k}\mathbb{E}x_j^2 < \infty. \tag{2.18}$$

Note that for $j = 1, ..., k$, by the Cauchy-Schwarz Inequality (C.3) and Assumption 2.2.1

$$\mathbb{E}\left|x_j y\right| \leq \left(\mathbb{E}x_j^2\right)^{1/2}\left(\mathbb{E}y^2\right)^{1/2} < \infty.$$

Thus the elements in the vector $\mathbb{E}(\boldsymbol{xy})$ are well defined and finite. Next, note that the jl'th element of $\mathbb{E}(\boldsymbol{xx}')$ is $\mathbb{E}(x_j x_l)$. Observe that

$$\mathbb{E}|x_j x_l| \leq \left(\mathbb{E}\boldsymbol{x}_j^2\right)^{1/2}\left(\mathbb{E}\boldsymbol{x}_l^2\right)^{1/2} < \infty.$$

Thus all elements of the matrix $\mathbb{E}(\boldsymbol{xx}')$ are finite.

Equation (2.10) states that $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{xx}'))^{-1}\mathbb{E}(\boldsymbol{xy})$ which is well defined since $(\mathbb{E}(\boldsymbol{xx}'))^{-1}$ exists under Assumption 2.9.1. It follows that $e = y - \boldsymbol{x}'\boldsymbol{\beta}$ as defined in (2.11) is also well defined.

Note the Schwarz Inequality (A.6) implies $(\boldsymbol{x}'\boldsymbol{\beta})^2 \leq \|\boldsymbol{x}\|^2\|\boldsymbol{\beta}\|^2$ and therefore combined with (2.18) we see that

$$\mathbb{E}\left(\boldsymbol{x}'\boldsymbol{\beta}\right)^2 \leq \mathbb{E}\|\boldsymbol{x}\|^2\|\boldsymbol{\beta}\|^2 < \infty. \tag{2.19}$$

Using Minkowski's Inequality (C.5), Assumption 2.2.1, and (2.19) we find

$$\begin{aligned}\left(\mathbb{E}\left(e^2\right)\right)^{1/2} &= \left(\mathbb{E}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right)^{1/2} \\ &\leq \left(\mathbb{E}y^2\right)^{1/2} + \left(\mathbb{E}\left(\boldsymbol{x}'\boldsymbol{\beta}\right)^2\right)^{1/2} \\ &< \infty\end{aligned}$$

establishing (2.13).

An application of the Cauchy-Schwarz Inequality (C.3) shows that for any j

$$\mathbb{E}|x_j e| \leq \left(\mathbb{E}\boldsymbol{x}_j^2\right)^{1/2}\left(\mathbb{E}e^2\right)^{1/2} < \infty$$

and therefore the elements in the vector $\mathbb{E}(\boldsymbol{x}e)$ are well defined and finite.

Using the definitions (2.11) and (2.10), and the matrix properties that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ and $\boldsymbol{I}\boldsymbol{a} = \boldsymbol{a}$,

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}e) &= \mathbb{E}\left(\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\right) \\ &= \mathbb{E}(\boldsymbol{xy}) - \mathbb{E}\left(\boldsymbol{xx}'\right)\left(\mathbb{E}\left(\boldsymbol{xx}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{xy}) = \boldsymbol{0}\end{aligned}$$

completing the proof.

2.10 Regression Coefficients

Sometimes it is useful to separate the intercept from the other regressors, and write the regression equation in the format

$$y = \alpha + \boldsymbol{x}'\boldsymbol{\beta} + e \tag{2.20}$$

where α is the intercept and $\boldsymbol{x}$ does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}y = \mathbb{E}\alpha + \mathbb{E}\boldsymbol{x}'\boldsymbol{\beta} + \mathbb{E}e$$

or

$$\mu_y = \alpha + \mu_x'\boldsymbol{\beta}$$

where $\mu_y = \mathbb{E}y$ and $\mu_x = \mathbb{E}\boldsymbol{x}$, since $\mathbb{E}(e) = 0$ from (2.16). Rearranging, we find

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta}.$$

Subtracting this equation from (2.20) we find

$$y - \mu_y = \left(\boldsymbol{x} - \mu_x\right)'\boldsymbol{\beta} + e, \tag{2.21}$$

a linear equation between the centered variables $y - \mu_y$ and $\boldsymbol{x} - \mu_x$. (They are centered at their means, or equivalently are mean-zero random variables.) Because $\boldsymbol{x} - \mu_x$ is uncorrelated with e, (2.21) is also a linear projection, thus by the formula for the linear projection model,

$$\begin{aligned}\boldsymbol{\beta} &= \left(\mathbb{E}\left(\left(\boldsymbol{x}-\mu_x\right)\left(\boldsymbol{x}-\mu_x\right)'\right)\right)^{-1}\mathbb{E}\left(\left(\boldsymbol{x}-\mu_x\right)\left(y-\mu_y\right)\right) \\ &= \operatorname{cov}\left(\boldsymbol{x},\boldsymbol{x}\right)^{-1}\operatorname{cov}\left(\boldsymbol{x},y\right)\end{aligned}$$

a function only of the covariances[8] of $\boldsymbol{x}$ and y.

Theorem 2.10.1 *In the linear projection model*

$$y = \alpha + \boldsymbol{x}'\boldsymbol{\beta} + e,$$

then

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta} \tag{2.22}$$

and

$$\boldsymbol{\beta} = \operatorname{cov}\left(\boldsymbol{x},\boldsymbol{x}\right)^{-1}\operatorname{cov}\left(\boldsymbol{x},y\right). \tag{2.23}$$

2.11 Best Linear Approximation

There are alternative ways we could construct a linear approximation $\boldsymbol{x}'\boldsymbol{\beta}$ to the conditional mean $m(\boldsymbol{x})$. In this section we show that one natural approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of $\boldsymbol{x}'\boldsymbol{\beta}$ to $m(\boldsymbol{x})$ as the expected squared difference between $\boldsymbol{x}'\boldsymbol{\beta}$ and the conditional mean $m(\boldsymbol{x})$

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2. \tag{2.24}$$

The function $d(\boldsymbol{\beta})$ is a measure of the deviation of $\boldsymbol{x}'\boldsymbol{\beta}$ from $m(\boldsymbol{x})$. If the two functions are identical then $d(\boldsymbol{\beta}) = 0$, otherwise $d(\boldsymbol{\beta}) > 0$. We can also view the mean-square difference $d(\boldsymbol{\beta})$ as a density-weighted average of the function $\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2$.

We can then define the best linear approximation to the conditional $m(\boldsymbol{x})$ as the function $\boldsymbol{x}'\boldsymbol{\beta}$ obtained by selecting $\boldsymbol{\beta}$ to minimize $d(\boldsymbol{\beta})$:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}}\, d(\boldsymbol{\beta}). \tag{2.25}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.8) selects $\boldsymbol{\beta}$ to minimize the expected squared prediction error, while the best linear approximation (2.25) selects $\boldsymbol{\beta}$ to minimize the expected squared approximation error.

Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.9) plus an application of conditional expectations we can find that

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}m(\boldsymbol{x})\right) \tag{2.26}$$

$$= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right) \tag{2.27}$$

(see Exercise 2.14). Thus (2.25) equals (2.8). We conclude that the definition (2.25) can be viewed as an alternative motivation for the linear projection coefficient.

[8] The **covariance matrix** between vectors x and z is $\operatorname{cov}\left(\boldsymbol{x},\boldsymbol{z}\right) = \mathbb{E}\left(\left(\boldsymbol{x}-\mathbb{E}\boldsymbol{x}\right)\left(\boldsymbol{z}-\mathbb{E}\boldsymbol{z}\right)'\right)$. We call $\operatorname{cov}\left(\boldsymbol{x},\boldsymbol{x}\right)$ the covariance matrix of $\boldsymbol{x}$.

2.12 Normal Regression

Suppose the variables $(y, \boldsymbol{x})$ are jointly normally distributed. Consider the best linear predictor of y on $\boldsymbol{x}$

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e. \\ \boldsymbol{\beta} &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right). \end{aligned}$$

Since the error e is a linear transformation of the normal vector $(y, \boldsymbol{x})$, it follows that $(e, \boldsymbol{x})$ is jointly normal, and since they are jointly normal and uncorrelated (since $\mathbb{E}(\boldsymbol{x}e) = 0$) they are also independent (see Appendix B.9). Independence implies that

$$\mathbb{E}(e \mid \boldsymbol{x}) = \mathbb{E}(e) = 0$$

and

$$\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \mathbb{E}\left(e^2\right) = \sigma^2$$

which are properties of a homoskedastic linear conditional regression.

We have shown that when $(y, \boldsymbol{x})$ are jointly normally, they satisfy a normal linear regression

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$

where

$$e \sim N(0, \sigma^2)$$

is independent of $\boldsymbol{x}$.

This is an alternative (and traditional) motivation for the linear regression model. This motivation has limited merit in econometric applications since economic data is typically non-normal.

2.13 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton published in 1886, where he examined the joint distribution of the stature (height) of parents and children. Effectively, he was estimating the conditional mean of children's height given their parents height. Galton discovered that this conditional mean was approximately linear with a slope of 2/3. This implies that *on average* a child's height is more mediocre than his or her parent's height. Galton called this phenomenon **regression to the mean**, and the label **regression** has stuck to this day to describe most conditional relationships.

One of Galton's fundamental insights was to recognize that if the marginal distributions of y and x are the same (e.g. the heights of children and parents in a stable environment) then the regression slope in a linear projection is always less than one.

To be more precise, take the simple regression

$$y = \alpha + x\beta + e \tag{2.28}$$

where y equals the height of the child and x equals the height of the parent. Assume that y and x have the same mean, so that $\mu_y = \mu_x = \mu$. Then from (2.22)

$$\alpha = (1 - \beta)\,\mu$$

so we can write the conditional mean of (2.28) as

$$\mathbb{E}(y \mid \boldsymbol{x}) = (1 - \beta)\,\mu + x\beta.$$

This shows that the expected height of the child is a weighted average of the population average height μ and the parents height x, with the weight equal to the regression slope β. When the height

distribution is stable across generations, so that $\text{var}(y) = \text{var}(x)$, then this slope is the simple correlation of y and x. Using (2.23)

$$\beta = \frac{\text{cov}(\boldsymbol{x}, y)}{\text{var}(x)} = \text{corr}(x, y).$$

By the properties of correlation (e.g. equation (B.7) in the Appendix), $-1 \leq \text{corr}(x, y) \leq 1$, with $\text{corr}(x, y) = 1$ only in the degenerate case $y = x$. Thus if we exclude degeneracy, β is strictly less than 1.

This means that on average a child's height is more mediocre (closer to the population average) than the parent's.

Sir Francis Galton

Sir Francis Galton (1822-1911) of England was one of the leading figures in late 19th century statistics. In addition to inventing the concept of regression, he is credited with introducing the concepts of correlation, the standard deviation, and the bivariate normal distribution. His work on heredity made a significant intellectual advance by examing the joint distributions of observables, allowing the application of the tools of mathematical statistics to the social sciences.

A common error – known as the **regression fallacy** – is to infer from $\beta < 1$ that the population is **converging**[9]. This is a fallacy because we have shown that under the assumption of constant (e.g. stable, non-converging) means and variances, the slope coefficient *must be* less than one. It cannot be anything else. A slope less than one does not imply that the variance of y is less than than the variance of x.

Another way of seeing this is to examine the conditions for convergence in the context of equation (2.28). Since x and e are uncorrelated, it follows that

$$\text{var}(y) = \beta^2 \text{var}(x) + \text{var}(e).$$

Then $\text{var}(y) < \text{var}(x)$ if and only if

$$\beta^2 < 1 - \frac{\text{var}(e)}{\text{var}(x)}$$

which is not implied by the simple condition $|\beta| < 1$.

The regression fallacy arises in related empirical situations. Suppose you sort families into groups by the heights of the parents, and then plot the average heights of each subsequent generation over time. If the population is stable, the regression property implies that the plots lines will converge – children's height will be more average than their parents. The regression fallacy is to incorrectly conclude that the population is converging. The message is that such plots are misleading for inferences about convergence.

The regression fallacy is subtle. It is easy for intelligent economists to succumb to its temptation. A famous example is *The Triumph of Mediocrity in Business* by Horace Secrist, published in 1933. In this book, Secrist carefully and with great detail documented that in a sample of department stores over 1920-1930, when he divided the stores into groups based on 1920-1921 profits, and plotted the average profits of these groups for the subsequent 10 years, he found clear and persuasive evidence for convergence "toward mediocrity". Of course, there was no discovery – regression to the mean is a necessary feature of stable distributions.

[9] A population is **converging** if its variance is declining towards zero.

2.14 Reverse Regression

Galton noticed another interesting feature of the bivariate distribution. There is nothing special about a regression of y on x. We can also regress x on y. (In his heredity example this is the best linear predictor of the height of parents given the height of their children.) This regression takes the form

$$x = \alpha^* + y\beta^* + e^*. \tag{2.29}$$

This is sometimes called the **reverse regression**. In this equation, the coefficients α^*, β^* and error e^* are defined by linear projection. In a stable population we find that

$$\beta^* = \operatorname{corr}(x, y) = \beta$$

$$\alpha^* = (1 - \beta)\,\mu = \alpha$$

which are exactly the same as in the regression of y on x! The intercept and slope have exactly the same values in the forward and reverse regression!

While this algebraic discovery is quite simple, it is counter-intuitive. Instead, a common yet mistaken guess for the form of the reverse regression is to take the regression (2.28), divide through by β and rewrite to find the equation

$$x = -\frac{\alpha}{\beta} + y\frac{1}{\beta} - \frac{1}{\beta}e \tag{2.30}$$

suggesting that the regression of x on y should have a slope coefficient of $1/\beta$ instead of β, and intercept of -α/β rather than α. What went wrong? Equation (2.30) is perfectly valid, because it is a simple manipulation of the valid equation (2.28). The trouble is that (2.30) is not a regression equation. Inverting a regression does not yield a regression. Instead, (2.29) is a valid regression, not (2.30).

In any event, Galton's finding was that when the variables are standardized, the slope in both regressions (y on x, and x and y) equals the correlation, and both equations exhibit regression to the mean. It is not a causal relation, but a natural feature of all joint distributions.

2.15 Limitations of the Best Linear Predictor

Let's compare the linear projection and linear regression models.

From Theorem 2.4.1.4 we know that the regression error has the property $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$. Thus a linear regression is a linear projection. However, the converse is not true as the projection error does not necessarily satisfy $\mathbb{E}(e \mid \boldsymbol{x}) = 0$.

To see this in a simple example, suppose we take a normally distributed random variable $x \sim N(0, 1)$ and set $y = x^2$. Note that y is a deterministic function of x! Now consider the linear projection of y on x and an intercept. The intercept and slope may be calculated as

$$\begin{aligned}
\begin{pmatrix} \alpha \\ \beta \end{pmatrix} &= \begin{pmatrix} 1 & \mathbb{E}(x) \\ \mathbb{E}(x) & \mathbb{E}\left(x^2\right) \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}(y) \\ \mathbb{E}(xy) \end{pmatrix} \\
&= \begin{pmatrix} 1 & \mathbb{E}(x) \\ \mathbb{E}(x) & \mathbb{E}\left(x^2\right) \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}\left(x^2\right) \\ \mathbb{E}\left(x^3\right) \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ 0 \end{pmatrix}
\end{aligned}$$

Thus the linear projection equation takes the form

$$y = \alpha + x\beta + e$$

where $\alpha = 1$, $\beta = 0$ and $e = x^2 - 1$. Observe that $\mathbb{E}(e) = \mathbb{E}\left(x^2\right) - 1 = 0$ and $\mathbb{E}(xe) = \mathbb{E}\left(x^3\right) - \mathbb{E}(e) = 0$, yet $\mathbb{E}(e \mid x) = x^2 - 1 \neq 0$. In this simple example e is a deterministic function of x, yet e and x are uncorrelated! The point is that a projection error need not be a regression error.

Return for a moment to the joint distributions displayed in Figures 2.3 and 2.4. In these figures, the solid lines are the conditional means and the straight dashed lines are the linear projections. In Figure 2.3 (the conditional mean of log hourly wages as a function of education) the conditional mean and linear projection are quite close to one another. In this example the linear predictor is a close approximation to the conditional mean. However, in Figure 2.4 (the conditional mean of log hourly wages as a function of labor market experience) the conditional mean is quite nonlinear, so the linear projection is a poor approximation. It over-predicts wages for young and old workers, and under-predicts for the rest. Most importantly, it misses the strong downturn in expected wages for those above 35 years work experience (equivalently, for those over 53 in age).

This defect in the best linear predictor can be partially corrected through a careful selection of regressors. In the example of Figure 2.4, we can augment the regressor vector $\boldsymbol{x}$ to include both *experience* and *experience*^{2}. The best linear predictor of log wages given these two variables can be called a quadratic projection, since the resulting function is quadratic in *experience*. Other than the redefinition of the regressor vector, there are no changes in our methods or analysis. In Figure 2.4 we display as well the quadratic projection. In this example it is a much better approximation to the conditional mean than the linear projection.

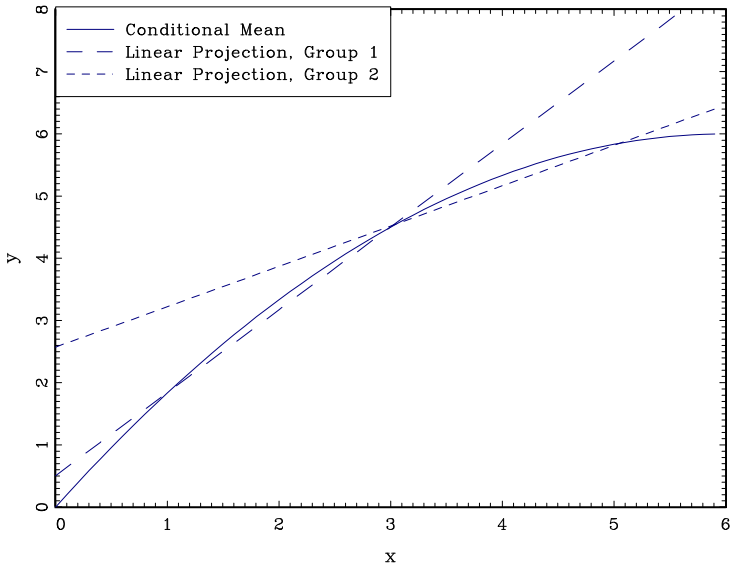

Figure 2.5: Conditional Mean and Two Linear Projections

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is non-linear. We illustrate the issue in Figure 2.5 for a constructed[10] joint distribution of y and x. The solid line is the non-linear conditional mean of y given x. The data are divided in two – Group 1 and Group 2 – which have different marginal distributions for the regressor x, and Group 1 has a lower mean value of x than Group 2. The separate linear projections of y on x for these two groups are displayed in the Figure by the dashed lines. These two projections are distinct approximations to the conditional mean. A defect with linear projection is that it leads to the incorrect conclusion that the effect of x on y is different for individuals in the two Groups. This conclusion is incorrect because in fact there is no difference in the conditional mean function. The apparant difference is a by-product of a linear approximation

[10] The x in Group 1 are $\mathrm{N}(2,1)$ and those in Group 2 are $\mathrm{N}(4,1)$, and the conditional distriubtion of y given x is $\mathrm{N}(m(x),1)$ where $m(x) = 2x - x^2/6$.

to a non-linear mean, combined with different marginal distributions for the conditioning variables.

2.16 Identification of the Conditional Mean

When a parameter is uniquely determined by the distribution of the observable variables, we say that the parameter is **identified**. Typically, identification only holds under a set of restrictions, and an identification theorem carefully describes a set of such conditions which are sufficient for identification. Identification is a necessary pre-condition for estimation.

For example, consider the unconditional mean $\mu = \mathbb{E}y$. It is well defined and unique for all distributions for which $\mathbb{E}|y| < \infty$. Thus the mean μ is identified from the distribution of y under the restriction $\mathbb{E}|y| < \infty$. Unless $\mathbb{E}|y| < \infty$, it is meaningless to attempt to estimate $\mathbb{E}y$.

As another example, consider the ratio of means $\theta = \mu_1/\mu_2$ where $\mu_1 = \mathbb{E}y_1$ and $\mu_2 = \mathbb{E}y_2$. It is well defined when μ_1 and μ_2 are both finite and $\mu_2 \neq 0$, but if $\mu_2 = 0$ then θ is undefined. Thus θ is identified from the distribution of (y_1, y_2) under the restrictions $\mathbb{E}|y_1| < \infty$, $\mathbb{E}|y_2| < \infty$, and $\mathbb{E}y_2 \neq 0$. Unless these conditions hold, it is meaningless to estimate θ.

Now consider the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$. Under which conditions is $m(\boldsymbol{x})$ defined and unique? The answer is provided in the following deep result from probability theory, which establishes the existence of the conditional mean.

Theorem 2.16.1 ***Existence of the Conditional Mean***
If $\mathbb{E}|y| < \infty$ then there exists a function $m(\boldsymbol{x})$ such that for all measurable sets $\mathcal{X}$

$$\mathbb{E}(1(\boldsymbol{x} \in \mathcal{X})\, y) = \mathbb{E}(1(\boldsymbol{x} \in \mathcal{X})\, m(\boldsymbol{x})). \tag{2.31}$$

The function $m(\boldsymbol{x})$ is almost everywhere unique, in the sense that if $h(\boldsymbol{x})$ satisfies (2.31), then there is a set S^* such that $\mathbb{P}(S^*) = 1$ and $m(\boldsymbol{x}) = h(\boldsymbol{x})$ for $\boldsymbol{x} \in S^*$. The function $m(\boldsymbol{x})$ is called the conditional mean and is written $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$.

See, for example, Ash (1972), Theorem 6.3.3.

The function $m(\boldsymbol{x})$ defined by (2.31) specializes to (2.2) when $(y, \boldsymbol{x})$ have a joint density.

Theorem 2.16.1 shows that the conditional mean function $m(\boldsymbol{x})$ exists and is almost everywhere unique, and is thus is identified.

Theorem 2.16.2 ***Identification of the Conditional Mean***
If $\mathbb{E}|y| < \infty$, the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is identified for $\boldsymbol{x} \in S^*$ where $\mathbb{P}(S^*) = 1$.

Exercises

Exercise 2.1 Prove parts 2, 3 and 4 of Theorem 2.4.1.

Exercise 2.2 Suppose that the random variables y and x only take the values 0 and 1, and have the following joint probability distribution

	$x=0$	$x=1$
$y=0$	.1	.2
$y=1$	.4	.3

Find $\mathbb{E}(y \mid x)$, $\mathbb{E}\left(y^2 \mid x\right)$ and $\operatorname{var}(y \mid x)$ for $x=0$ and $x=1$.

Exercise 2.3 Show that $\sigma^2(\boldsymbol{x})$ is the best predictor of e^2 given $\boldsymbol{x}$:

(a) Write down the mean-squared error of a predictor $h(\boldsymbol{x})$ for e^2.

(b) What does it mean to be predicting e^2?

(c) Show that $\sigma^2(\boldsymbol{x})$ minimizes the mean-squared error and is thus the best predictor.

Exercise 2.4 Use $y = m(\boldsymbol{x}) + e$ to show that

$$\operatorname{var}(y) = \operatorname{var}(m(\boldsymbol{x})) + \sigma^2$$

Exercise 2.5 Suppose that y is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of y given $\boldsymbol{x}$ is Poisson:

$$\mathbb{P}(y = j \mid \boldsymbol{x}) = \frac{\exp(-\boldsymbol{x}'\boldsymbol{\beta})(\boldsymbol{x}'\boldsymbol{\beta})^j}{j!}, \qquad j = 0, 1, 2, ...$$

Compute $\mathbb{E}(y \mid \boldsymbol{x})$ and $\operatorname{var}(y \mid \boldsymbol{x})$. Does this justify a linear regression model of the form $y = \boldsymbol{x}'\boldsymbol{\beta} + e$?

Hint: If $\mathbb{P}(y = j) = \frac{\exp(-\lambda)\lambda^j}{j!}$, then $\mathbb{E}y = \lambda$ and $\operatorname{var}(y) = \lambda$.

Exercise 2.6 Let x and y have the joint density $f(x, y) = \frac{3}{2}\left(x^2 + y^2\right)$ on $0 \le x \le 1$, $0 \le y \le 1$. Compute the coefficients of the best linear predictor $y = \alpha + \beta x + e$. Compute the conditional mean $m(x) = \mathbb{E}(y \mid x)$. Are the best linear predictor and conditional mean different?

Exercise 2.7 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}(e \mid x) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.8 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}(xe) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.9 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}(e \mid \boldsymbol{x}) = 0$, then e is independent of $\boldsymbol{x}$.

Exercise 2.10 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$, then $\mathbb{E}(e \mid \boldsymbol{x}) = 0$.

Exercise 2.11 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$, $\mathbb{E}(e \mid \boldsymbol{x}) = 0$, and $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$, a constant, then e is independent of $\boldsymbol{x}$.

Exercise 2.12 Let x be a random variable with $\mu = \mathbb{E}x$ and $\sigma^2 = \operatorname{var}(x)$. Define

$$g\left(x \mid \mu, \sigma^2\right) = \begin{pmatrix} x - \mu \\ (x-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Show that $\mathbb{E}g(x \mid m, s) = 0$ if and only if $m = \mu$ and $s = \sigma^2$.

Exercise 2.13 Suppose that

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x_2 \\ x_3 \end{pmatrix}$$

and $x_3 = \alpha_1 + \alpha_2 x_2$ is a linear function of x_2.

(a) Show that $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ is not invertible.

(b) Use a linear transformation of $\boldsymbol{x}$ to find an expression for the best linear predictor of y given $\boldsymbol{x}$. (Be explicit, do not just use the generalized inverse formula.)

Exercise 2.14 Show (2.26)-(2.27), namely that for

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2$$

then

$$\begin{aligned} \boldsymbol{\beta} &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} d(\boldsymbol{\beta}) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} y\right). \end{aligned}$$

Hint: To show $\mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) = \mathbb{E}\left(\boldsymbol{x} y\right)$ use the law of iterated expectations.

Chapter 3

The Algebra of Least Squares

3.1 Introduction

In this chapter we introduce the popular least-squares estimator. Most of the discussion will be algebraic, with questions of distribution and inference defered to later chapters.

3.2 Least Squares Estimator

In Section 2.9 we derived and discussed the best linear predictor of y given $\boldsymbol{x}$ for a pair of random variables $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^k$, and called this the linear projection model. Applied to observations from a random sample with observations $(y_i, \boldsymbol{x}_i : i = 1, ..., n)$ this model takes the form

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{3.1}$$

where $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\operatorname{argmin}} S(\boldsymbol{\beta}), \tag{3.2}$$

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2, \tag{3.3}$$

and

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x}y\right). \tag{3.4}$$

When a parameter is defined as the minimizer of a function as in (3.2), a standard approach to estimation is to construct an empirical analog of the function, and define the estimator of the parameter as the minimizer of the empirical function.

The empirical analog of the expected squared error (3.3) is the sample average squared error

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \frac{1}{n} SSE_n(\boldsymbol{\beta}) \end{aligned} \tag{3.5}$$

where

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

is called the sum-of-squared-errors function.

An estimator for $\boldsymbol{\beta}$ is the minimizer of (3.5):

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\operatorname{argmin}} S_n(\boldsymbol{\beta}).$$

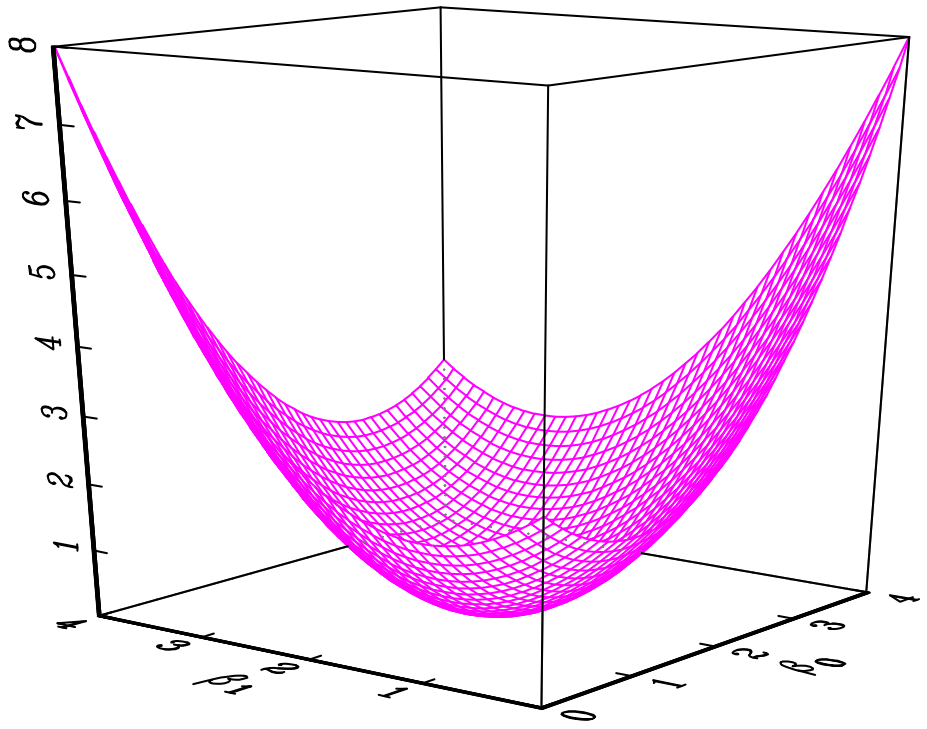

Figure 3.1: Sum-of-Squared Errors Function

Alternatively, as $S_n(\boldsymbol{\beta})$ is a scale multiple of $SSE_n(\boldsymbol{\beta})$, we may equivalently define $\hat{\boldsymbol{\beta}}$ as the minimizer of $SSE_n(\boldsymbol{\beta})$. Hence $\hat{\boldsymbol{\beta}}$ is commonly called the **least-squares estimator** of $\boldsymbol{\beta}$.

To visualize the quadratic function $S_n(\boldsymbol{\beta})$, Figure 3.1 displays an example sum-of-squared errors function $SSE_n(\boldsymbol{\beta})$ for the case $k = 2$. The least-squares estimator $\hat{\boldsymbol{\beta}}$ is the the pair $(\hat{\beta}_1, \hat{\beta}_2)$ minimizing this function.

3.3 Solving for Least Squares

To solve for $\hat{\boldsymbol{\beta}}$, expand the SSE function to find

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i^2 - 2\boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i y_i + \boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}$$

which is quadratic in the vector argument $\boldsymbol{\beta}$. The first-order-condition for minimization of $SSE_n(\boldsymbol{\beta})$ is

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} S_n(\hat{\boldsymbol{\beta}}) = -2 \sum_{i=1}^{n} \boldsymbol{x}_i y_i + 2 \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}. \tag{3.6}$$

By inverting the $k \times k$ matrix $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ we find an explicit formula for the least-squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i \right). \tag{3.7}$$

This is the natural estimator of the best linear prediction coefficient $\boldsymbol{\beta}$ defined in (3.2), and can also be called the linear projection estimator.

Early Use of Matrices

The earliest known treatment of the use of matrix methods to solve simultaneous systems is found in Chapter 8 of the Chinese text *The Nine Chapters on the Mathematical Art*, written by several generations of scholars from the 10th to 2nd century BCE.

Alternatively, equation (3.4) writes the projection coefficient $\boldsymbol{\beta}$ as an explicit function of the population moments $\mathbb{E}(\boldsymbol{x}_i y_i)$ and $\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')$. Their moment estimators are the sample moments

$$\begin{aligned} \widehat{\mathbb{E}}(\boldsymbol{x}_i y_i) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i \\ \widehat{\mathbb{E}}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right) &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'. \end{aligned}$$

The moment estimator of $\boldsymbol{\beta}$ replaces the population moments in (3.4) with the sample moments:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\widehat{\mathbb{E}}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right)\right)^{-1} \widehat{\mathbb{E}}(\boldsymbol{x}_i y_i) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\ &= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \end{aligned}$$

which is identical with (3.7).

Least Squares Estimation

Definition 3.3.1 *The* ***least-squares estimator*** $\hat{\boldsymbol{\beta}}$ *is*

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^k} S_n(\boldsymbol{\beta})$$

where

$$S_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

and has the solution

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right).$$

To illustrate least-squares estimation in practice, consider the data used to generate Figure 2.3. These are white male wage earners from the March 2004 Current Population Survey, excluding

military, with 10-15 years of potential work experience. This sample has 988 observations. Let y_i be log wages and $\boldsymbol{x}_i$ be an intercept and years of education. Then

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i = \begin{pmatrix} 2.951 \\ 42.405 \end{pmatrix}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' = \begin{pmatrix} 1 & 14.136 \\ 14.136 & 205.826 \end{pmatrix}.$$

Thus

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{pmatrix} 1 & 14.136 \\ 14.136 & 205.826 \end{pmatrix}^{-1}\begin{pmatrix} 2.951 \\ 42.405 \end{pmatrix} \\ &= \begin{pmatrix} 1.33 \\ 0.115 \end{pmatrix}. \end{aligned} \tag{3.8}$$

We often write the estimated equation using the format

$$\widehat{\log(Wage)} = 1.33 + 0.115\ education. \tag{3.9}$$

An interpretation of the estimated equation is that each year of education is associated with an 11% increase in mean wages.

Equation (3.9) is called a bivariate regression as there are only two variables. A multivariate regression has two or more regressors, and allows a more detailed investigation. Let's redo the example, but now including all levels of experience. This expanded sample includes 6578 observations. Including as regressors years of experience and its square ($experience^2/100$) (we divide by 100 to simplify reporting), we obtain the estimates

$$\widehat{\log(Wage)} = 0.959 + 0.100\ education + 0.053\ experience - 0.095\ experience^2/100. \tag{3.10}$$

These estimates suggest a 10% increase in mean wages per year of education.

Adrien-Marie Legendre

The method of least-squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least-squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.1) is a set of n equations with k unknowns. As the equations cannot be solved exactly, Legendre's goal was to select $\boldsymbol{\beta}$ to make the set of errors as small as possible. He proposed the sum of squared error criterion, and derived the algebraic solution presented above. As he noted, the first-order conditions (3.6) is a system of k equations with k unknowns, which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

3.4 Least Squares Residuals

As a by-product of estimation, we define the **fitted** or **predicted value**

$$\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$$

and the **residual**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}. \tag{3.11}$$

Note that $y_i = \hat{y}_i + \hat{e}_i$. We make a distinction between the **error** e_i and the **residual** $\hat{e}_i$. The error e_i is unobservable while the residual $\hat{e}_i$ is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

Equation (3.6) implies that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i = \mathbf{0}. \tag{3.12}$$

To see this by a direct calculation, using (3.11) and (3.7),

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i y_i \\
&= 0.
\end{aligned}$$

When $\boldsymbol{x}_i$ contains a constant, an implication of (3.12) is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0.$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

Given the residuals, we can construct an estimator for σ^2 as defined in (2.13):

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2. \tag{3.13}$$

3.5 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The linear equation (2.12) is a system of n equations, one for each observation. We can stack these n equations together as

$$\begin{aligned}
y_1 &= \boldsymbol{x}_1'\boldsymbol{\beta} + e_1 \\
y_2 &= \boldsymbol{x}_2'\boldsymbol{\beta} + e_2 \\
&\vdots \\
y_n &= \boldsymbol{x}_n'\boldsymbol{\beta} + e_n.
\end{aligned}$$

Now define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that $\boldsymbol{y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors, and $\boldsymbol{X}$ is an $n \times k$ matrix. Then the system of n equations can be compactly written in the single equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Sample sums can also be written in matrix notation. For example

$$\begin{aligned} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' &= \boldsymbol{X}'\boldsymbol{X} \\ \sum_{i=1}^{n} \boldsymbol{x}_i y_i &= \boldsymbol{X}'\boldsymbol{y}. \end{aligned}$$

Therefore

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y}\right). \tag{3.14}$$

Using matrix notation we have simple expressions for most estimators. This is particularly convenient for computer programming, as most languages allow matrix notation and manipulation.

Important Matrix Expressions

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y}\right) \\ \hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ \hat{\sigma}^2 &= n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}. \end{aligned}$$

3.6 Projection Matrices

Define the matrices

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

and

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ &= \boldsymbol{I}_n - \boldsymbol{P} \end{aligned}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. $\boldsymbol{P}$ and $\boldsymbol{M}$ are called **projection matrices** due to the property that for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the **range space** of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}$$

and

$$\boldsymbol{M}\boldsymbol{Z} = \left(\boldsymbol{I}_n - \boldsymbol{P}\right)\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{Z} = \boldsymbol{0}.$$

As an important example of this property, partition the matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ so that

$$\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2].$$

Then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$. It follows that $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{M}\boldsymbol{P} = \boldsymbol{0}$, so $\boldsymbol{M}$ and $\boldsymbol{P}$ are orthogonal.

The matrices $\boldsymbol{P}$ and $\boldsymbol{M}$ are **symmetric** and **idempotent**[1]. To see that $\boldsymbol{P}$ is symmetric,

$$\begin{aligned}
\boldsymbol{P}' &= \left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)' \\
&= \left(\boldsymbol{X}'\right)'\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{X}\right)' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}\right)'\left(\boldsymbol{X}'\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

To establish that it is idempotent,

$$\begin{aligned}
\boldsymbol{P}\boldsymbol{P} &= \left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right) \\
&= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

Similarly,

$$\boldsymbol{M}' = \left(\boldsymbol{I}_n - \boldsymbol{P}\right)' = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{M}$$

and

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M} &= \boldsymbol{M}\left(\boldsymbol{I}_n - \boldsymbol{P}\right) \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{P} \\
&= \boldsymbol{M},
\end{aligned}$$

since $\boldsymbol{M}\boldsymbol{P} = \boldsymbol{0}$.

Another useful property is that

$$\operatorname{tr}\boldsymbol{P} = k \tag{3.15}$$

$$\operatorname{tr}\boldsymbol{M} = n - k \tag{3.16}$$

(See Appendix A.4 for definition and properties of the trace operator.) To show (3.15) and (3.16),

$$\begin{aligned}
\operatorname{tr}\boldsymbol{P} &= \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right) \\
&= \operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right) \\
&= \operatorname{tr}\left(\boldsymbol{I}_k\right) \\
&= k,
\end{aligned}$$

and

$$\operatorname{tr}\boldsymbol{M} = \operatorname{tr}\left(\boldsymbol{I}_n - \boldsymbol{P}\right) = \operatorname{tr}\left(\boldsymbol{I}_n\right) - \operatorname{tr}\left(\boldsymbol{P}\right) = n - k.$$

[1] A matrix $\boldsymbol{P}$ is symmetric if $\boldsymbol{P}' = \boldsymbol{P}$. A matrix $\boldsymbol{P}$ is idempotent if $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. See Appendix A.8.

Given the definitions of $\boldsymbol{P}$ and $\boldsymbol{M}$, observe that

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

and

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}. \tag{3.17}$$

Furthermore, since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$, then

$$\hat{\boldsymbol{e}} = \boldsymbol{M}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right) = \boldsymbol{M}\boldsymbol{e}. \tag{3.18}$$

Another way of writing (3.17) is

$$\boldsymbol{y} = \left(\boldsymbol{P} + \boldsymbol{M}\right)\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}.$$

This decomposition is **orthogonal**, that is

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = \left(\boldsymbol{P}\boldsymbol{y}\right)'\left(\boldsymbol{M}\boldsymbol{y}\right) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{y} = 0.$$

The projection matrix $\boldsymbol{P}$ is also known as the **hat matrix** due to the equation $\hat{\boldsymbol{y}} = \boldsymbol{P}\boldsymbol{y}$. The i'th diagonal element of $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is

$$h_{ii} = \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \tag{3.19}$$

which is called the **leverage** of the i'th observation. The h_{ii} take values in $[0, 1]$ and sum to k

$$\sum_{i=1}^{n} h_{ii} = k \tag{3.20}$$

(See Exercise 3.6).

3.7 Residual Regression

Partition

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix}.$$

Then the regression model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}. \tag{3.21}$$

Observe that the OLS estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ can be obtained by regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$. OLS estimation can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}} \tag{3.22}$$

Suppose that we are primarily interested in $\boldsymbol{\beta}_2$, not in $\boldsymbol{\beta}_1$, and we want to obtain the OLS subcomponent $\hat{\boldsymbol{\beta}}_2$. In this section we derive an alternative expression for $\hat{\boldsymbol{\beta}}_2$ which does not involve estimation of the full model.

Define

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'.$$

Recalling the definition $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, observe that $\boldsymbol{X}_1'\boldsymbol{M}_1 = \boldsymbol{0}$ and thus

$$\boldsymbol{M}_1\boldsymbol{M} = \boldsymbol{M} - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M} = \boldsymbol{M}.$$

It follows that

$$\boldsymbol{M}_1\hat{\boldsymbol{e}} = \boldsymbol{M}_1\boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{e}}.$$

Using this result, if we premultiply (3.22) by $\boldsymbol{M}_1$ we obtain

$$\begin{aligned} \boldsymbol{M}_1\boldsymbol{y} &= \boldsymbol{M}_1\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{M}_1\hat{\boldsymbol{e}} \\ &= \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}} \end{aligned} \tag{3.23}$$

the second equality since $\boldsymbol{M}_1\boldsymbol{X}_1 = \boldsymbol{0}$. Premultiplying by $\boldsymbol{X}_2'$ and recalling that $\boldsymbol{X}_2'\hat{\boldsymbol{e}} = \boldsymbol{0}$, we obtain

$$\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y} = X_2'\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\hat{\boldsymbol{e}} = \boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Solving,

$$\hat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right)$$

an alternative expression for $\hat{\boldsymbol{\beta}}_2$.

Now, define

$$\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2 \tag{3.24}$$

$$\tilde{\boldsymbol{y}} = \boldsymbol{M}_1\boldsymbol{y}, \tag{3.25}$$

the least-squares residuals from the regression of $\boldsymbol{X}_2$ and $\boldsymbol{y}$, respectively, on the matrix $\boldsymbol{X}_1$ only. Since the matrix $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{X}}_2\right)^{-1}\left(\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{y}}\right). \end{aligned}$$

This shows that $\hat{\boldsymbol{\beta}}_2$ can be calculated by the OLS regression of $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$. This technique is called **residual regression**.

Furthermore, using the definitions (3.24) and (3.25), expression (3.23) can be equivalently written as

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}_2\hat{\boldsymbol{\beta}}_2 + \hat{e}.$$

Since $\hat{\boldsymbol{\beta}}_2$ is precisely the OLS coefficient from a regression of $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$, this shows that the residual vector from this regression is $\hat{\boldsymbol{e}}$, numerically the same residual vector as from the joint regression (3.22). We have proven the following theorem.

Theorem 3.7.1 ***Frisch-Waugh-Lovell***
In the model (3.21), the OLS estimator of $\boldsymbol{\beta}_2$ and the OLS residuals $\hat{\boldsymbol{e}}$ may be equivalently computed by either the OLS regression (3.22) or via the following algorithm:

1. *Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{y}}$;*
2. *Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{X}}_2$;*
3. *Regress $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$, obtain OLS estimates $\hat{\boldsymbol{\beta}}_2$ and residuals $\hat{\boldsymbol{e}}$.*

In some contexts, the FWL theorem can be used to speed computation, but in most cases there is little computational advantage to using the two-step algorithm. Rather, the primary use is theoretical.

A common application of the FWL theorem, which you may have seen in an introductory econometrics course, is the demeaning formula for regression. Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 = \boldsymbol{\iota}$ is a vector of ones, and $\boldsymbol{X}_2$ is the vector of observed regressors. In this case,

$$\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{\iota}\left(\boldsymbol{\iota}'\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}'.$$

Observe that

$$\begin{aligned}\tilde{\boldsymbol{X}}_2 &= \boldsymbol{M}_1\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \boldsymbol{\iota}\left(\boldsymbol{\iota}'\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}'\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2\end{aligned}$$

and

$$\begin{aligned}\tilde{\boldsymbol{y}} &= \boldsymbol{M}_1\boldsymbol{y} \\ &= \boldsymbol{y} - \boldsymbol{\iota}\left(\boldsymbol{\iota}'\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}'\boldsymbol{y} \\ &= \boldsymbol{y} - \overline{\boldsymbol{y}},\end{aligned}$$

which are "demeaned". The FWL theorem says that $\hat{\boldsymbol{\beta}}_2$ is the OLS estimate from a regression of $y_i - \overline{\boldsymbol{y}}$ on $\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2$:

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(y_i - \overline{\boldsymbol{y}}\right)\right).$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

Ragnar Frisch

Ragnar Frisch (1895-1973) was co-winner with Jan Tinbergen of the first Nobel Memorial Prize in Economic Sciences in 1969 for their work in developing and applying dynamic models for the analysis of economic problems. Frisch made a number of foundational contributions to modern economics beyond the Frisch-Waugh-Lovell Theorem, including formalizing consumer theory, production theory, and business cycle theory.

3.8 Prediction Errors

The least-squares residual $\hat{e}_i$ are not true prediction errors, as they are constructed based on the full sample including y_i. A proper prediction for y_i should be based on estimates constructed only using the other observations. We can do this by defining the **leave-one-out** OLS estimator of $\boldsymbol{\beta}$ as that obtained from the sample *excluding* the i'th observation:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j y_j\right) \\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{y}_{(-i)} \qquad (3.26)\end{aligned}$$

where $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{y}_{(-i)}$ are the data matrices omitting the i'th row. The leave-one-out predicted value for y_i is

$$\tilde{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(-i)},$$

and the **leave-one-out residual** or **prediction error** is

$$\tilde{e}_i = y_i - \tilde{y}_i.$$

A convenient alternative expression for $\hat{\boldsymbol{\beta}}_{(-i)}$ (derived below) is

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \tag{3.27}$$

where h_{ii} are the leverage values as defined in (3.19).

Using (3.27) we can simplify the expression for the prediction error:

$$\begin{aligned} \tilde{e}_i &= y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{(-i)} \\ &= y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + (1 - h_{ii})^{-1} \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \\ &= \hat{e}_i + (1 - h_{ii})^{-1} h_{ii} \hat{e}_i \\ &= (1 - h_{ii})^{-1} \hat{e}_i. \end{aligned} \tag{3.28}$$

A convenient feature of this expression is that it shows that computation of $\tilde{e}_i$ is based on a simple linear operation, and does not really require n separate estimations.

One use of the prediction errors is to estimate the out-of-sample mean squared error

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \tilde{e}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (1 - h_{ii})^{-2} \hat{e}_i^2. \end{aligned}$$

This is also known as the **mean squared prediction error**. Its square root $\tilde{\sigma} = \sqrt{\tilde{\sigma}^2}$ is the **prediction standard error**.

Proof of Equation (3.27). The Sherman–Morrison formula (A.2) from Appendix A.5 states that for nonsingular $\boldsymbol{A}$ and vector $\boldsymbol{b}$

$$\left(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1 - \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

This implies

$$\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + (1 - h_i)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

and thus

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{(-i)} &= \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i \\ &\quad + (1 - h_i)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i + (1 - h_i)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} - h_i y_i\right) \\ &= \hat{\boldsymbol{\beta}} - (1 - h_i)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left((1 - h_i)\, y_i - \boldsymbol{x}_i'\hat{\beta} + h_i y_i\right) \\ &= \hat{\boldsymbol{\beta}} - (1 - h_i)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \end{aligned}$$

the third equality making the substitutions $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$ and $h_i = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$, and the remainder collecting terms.

3.9 Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of **influential observations**, sometimes called **outliers**. We say that observation i is influential if its omission from the sample induces a substantial change in a parameter of interest. From (3.27)-(3.28) we know that

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} &= (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \tilde{e}_i.\end{aligned}$$

By direct calculation of this quantity for each observation i, we can directly discover if a specific observation i is influential for a coefficient estimate of interest.

For a more general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\begin{aligned}\hat{y}_i - \tilde{y}_i &= \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(-i)} \\ &= \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \tilde{e}_i \\ &= h_{ii}\tilde{e}_i\end{aligned}$$

which is a simple function of the leverage values h_{ii} and prediction errors $\tilde{e}_i$. Observation i is influential for the predicted value if $|h_{ii}\tilde{e}_i|$ is large, which requires that both h_{ii} and $|\tilde{e}_i|$ are large.

One way to think about this is that a large leverage value h_{ii} gives the potential for observation i to be influential. A large h_{ii} means that observation i is unusual in the sense that the regressor $\boldsymbol{x}_i$ is far from its sample mean. We call this observation with large h_{ii} **a leverage point**. A leverage point is not necessarily influential as this also requires that the prediction error $\tilde{e}_i$ is large.

To determine if any individual observations are influential in this sense, a useful summary statistic is

$$\textit{Influence} = \max_{1\leq i\leq n} \frac{|\hat{y}_i - \tilde{y}_i|}{\tilde{\sigma}} = \max_{1\leq i\leq n} \frac{h_{ii}\,|\tilde{e}_i|}{\tilde{\sigma}}$$

which scales the maximum change in predicted values by the prediction standard error. If *Influence* is large, it may be useful to examine the corresponding observation or observations. (As this is an informal comparison there is no magic threshold, so judgement must be employed.)

If an observation is determined to be influential, what should be done? Certainly, the recorded values for the observations should be examined. It is quite possible that there is a data error, and this is a common cause of influential observations. If there is an error, you should scrutinize all observations more carefully, as it would seem unlikely that data error would be confined to a single observation. If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. When this is done it is proper empirical practice to document such choices. (It is useful to keep the source data in its original form, a revised data file after cleaning, and a record describing the revision process. This is especially useful when revising empirical work at a later date.)

It is also possible that an observation is correctly measured, but unusual and influential. In this case it is unclear how to proceed. Some researchers will try to alter the specification to properly model the influential observation. Other researchers will delete the observation from the sample. The motivation for this choice is to prevent the results from being skewed or determined by individual observations, but this practice is viewed skeptically by many researchers, who believe it reduces the integrity of reported empirical results.

3.10 Measures of Fit

When a least-squares regression is reported in applied economics, it is common to see a reported summary measure of fit, measuring how well the regressors explain the observed variation in the dependent variable.

Some common summary measures are based on scaled or transformed estimates of the mean-squared error σ^2. These include the **sum of squared errors** $\sum_{i=1}^n \hat{e}_i^2$, the mean squared error of sample variance $n^{-1}\sum_{i=1}^n \hat{e}_i^2 = \hat{\sigma}^2$, and the root mean squared error $\sqrt{n^{-1}\sum_{i=1}^n \hat{e}_i^2}$ (sometimes called the **standard error of the regression**), and the mean prediction error $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$.

A related and commonly reported statistic is the **coefficient of determination** or **R-squared**:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$$

where

$$\hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$$

is the sample variance of y_i. R^2 can be viewed as an estimator of the population parameter

$$\rho^2 = \frac{\text{var}(\boldsymbol{x}_i'\boldsymbol{\beta})}{\text{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}$$

where $\sigma_y^2 = \text{var}(y_i)$. A high ρ^2 or R^2 means that forecasts of y using $\boldsymbol{x}'\boldsymbol{\beta}$ or $\boldsymbol{x}'\widehat{\boldsymbol{\beta}}$ will be quite accurate relative to the unconditional mean. In this sense R^2 can be a useful summary measure for an out-of-sample forecast or policy experiment.

An alternative estimator of ρ^2 proposed by Theil called **R-bar-squared** or **adjusted** R^2 is

$$\overline{R}^2 = 1 - \frac{(n-1)\sum_{i=1}^n \hat{e}_i^2}{(n-k)\sum_{i=1}^n (y_i - \overline{y})^2}.$$

Theil's estimator $\overline{R}^2$ is better estimator of ρ^2 than the unadjusted estimator R^2 because it can be expressed as a ratio of bias-corrected variance estimates.

Unfortunately, the frequent reporting of R^2 and $\overline{R}^2$ seems to have led to exaggerated beliefs regarding their usefulness. One mistaken belief is that R^2 is a measure of "fit". This belief is incorrect, as an incorrectly specified model can still have a reasonably high R^2. For example, suppose the truth is that $x_i \sim N(0,1)$ and $y_i = \beta x_i + x_i^2$. If we regress y_i on x_i (incorrectly omitting x_i^2), the best linear predictor is $y_i = 1 + \beta x_i + e_i$ where $e_i = x_i^2 - 1$. This is a misspecified regression, as the true relationship is deterministic! You can also calculate that the population $\rho^2 = \beta/(2+\beta)$ which can be arbitrarily close to 1 if β is large. For example, if $\beta = 8$, then $R^2 \simeq \rho^2 = .8$, or if $\beta = 18$ then $R^2 \simeq \rho^2 = .9$. This example shows that a regression with a high R^2 can actually have poor fit.

Another mistaken belief is that a high R^2 is important in order to justify interpretation of the regression coefficients. This is mistaken as there is no known association between the level of R^2 and the "correctness" of a regression, the accuracy of the coefficient estimates, or the validity of statistical inferences based on the estimated regression. In contrast, even if the R^2 is quite small, accurate estimates of regression coefficients is quite possible when sample sizes are large.

The bottom line is that while R^2 and $\overline{R}^2$ have appropriate uses, their usefulness should not be exaggerated.

Henri Theil

Henri Theil (1924-2000) of Holland invented $\overline{R}^2$ and two-stage least squares, both of which are routinely seen in applied econometrics. He also wrote an early and influential advanced textbook on econometrics (Theil, 1971).

3.11 Normal Regression Model

The normal regression model is the linear regression model under the restriction that the error e_i is independent of $\boldsymbol{x}_i$ and has the distribution $\mathrm{N}\left(0, \sigma^2\right)$. We can write this as

$$e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right).$$

This assumption implies

$$y_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right).$$

Normal regression is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory.

The log-likelihood function for the normal regression model is

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right)\right) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} SSE_n(\boldsymbol{\beta}). \end{aligned}$$

The maximum likelihood estimator (MLE) $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ maximize $\log L(\boldsymbol{\beta}, \sigma^2)$. Since the latter is a function of $\boldsymbol{\beta}$ only through the sum of squared errors $SSE_n(\boldsymbol{\beta})$, maximizing the likelihood is identical to minimizing $SSE_n(\boldsymbol{\beta})$. Hence

$$\hat{\boldsymbol{\beta}}_{mle} = \hat{\boldsymbol{\beta}}_{ols},$$

the MLE for $\boldsymbol{\beta}$ equals the OLS estimator. Due to this equivalence, the least squares estimator $\hat{\boldsymbol{\beta}}$ is also known as the MLE.

We can also find the MLE for σ^2. Plugging $\hat{\boldsymbol{\beta}}$ into the log-likelihood we obtain

$$\log L\left(\hat{\boldsymbol{\beta}}, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\hat{e}_i^2.$$

Maximization with respect to σ^2 yields the first-order condition

$$\frac{\partial}{\partial\sigma^2}\log L\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2}\sum_{i=1}^{n}\hat{e}_i^2 = 0.$$

Solving for $\hat{\sigma}^2$ yields the MLE for σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$$

which is the same as the moment estimator (3.13).

It may seem surprising that the MLE $\hat{\boldsymbol{\beta}}$ is numerically equal to the OLS estimator, despite emerging from quite different motivations. It is not completely accidental. The least-squares estimator minimizes a particular sample loss function – the sum of squared error criterion – and most loss functions are equivalent to the likelihood of a specific parametric distribution, in this case the normal regression model. In this sense it is not surprising that the least-squares estimator can be motivated as either the minimizer of a sample loss function or as the maximizer of a likelihood function.

Carl Friedrich Gauss

The mathematician Carl Friedrich Gauss (1777-1855) proposed the normal regression model, and derived the least squares estimator as the maximum likelihood estimator for this model. He claimed to have discovered the method in 1795 at the age of eighteen, but did not publish the result until 1809. Interest in Gauss's approach was reinforced by Laplace's simultaneous discovery of the central limit theorem, which provided a justification for viewing random disturbances as approximately normal.

Exercises

Exercise 3.1 Let y be a random variable with $\mu = \mathbb{E}y$ and $\sigma^2 = \text{var}(y)$. Define

$$g\left(y, \mu, \sigma^2\right) = \begin{pmatrix} y - \mu \\ (y-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let $(\hat{\mu}, \hat{\sigma}^2)$ be the values such that $\overline{g}_n(\hat{\mu}, \hat{\sigma}^2) = \mathbf{0}$ where $\overline{g}_n(m, s) = n^{-1} \sum_{i=1}^n g\left(y_i, m, s\right)$. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{y}$ on the $n \times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\hat{\boldsymbol{e}}$.

Exercise 3.4 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.5 Let $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{y}}$ on $\boldsymbol{X}$.

Exercise 3.6 Show ()3.20), that h_{ii} in (3.19) sum to k. (Hint: Use (3.15).)

Exercise 3.7 A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be vectors of 1's and 0's, with the i'th element of $\boldsymbol{d}_1$ equaling 1 and that of $\boldsymbol{d}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider the three regressions

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.29}$$

$$\boldsymbol{y} = \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.30}$$

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\phi + \boldsymbol{e} \tag{3.31}$$

Can all three regressions (3.29), (3.30), and (3.31) be estimated by OLS? Explain if not.

(a) Compare regressions (3.30) and (3.31). Is one more general than the other? Explain the relationship between the parameters in (3.30) and (3.31).

(b) Compute $\boldsymbol{\iota}'\boldsymbol{d}_1$ and $\boldsymbol{\iota}'\boldsymbol{d}_2$, where $\boldsymbol{\iota}$ is an $n \times 1$ is a vector of ones.

(c) Letting $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)'$, write equation (3.30) as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + e$. Consider the assumption $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Is there any content to this assumption in this setting?

Exercise 3.8 Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\gamma}_1 + \boldsymbol{d}_2\hat{\gamma}_2 + \hat{\boldsymbol{u}},$$

show that $\hat{\gamma}_1$ is sample mean of the dependent variable among the men of the sample ($\overline{y}_1$), and that $\hat{\gamma}_2$ is the sample mean among the women ($\overline{y}_2$).

(b) Describe in words the transformations

$$\begin{aligned} \boldsymbol{y}^* &= \boldsymbol{y} - \boldsymbol{d}_1\overline{y}_1 - \boldsymbol{d}_2\overline{y}_2 \\ \boldsymbol{X}^* &= \boldsymbol{X} - \boldsymbol{d}_1\overline{\boldsymbol{X}}_1 - \boldsymbol{d}_2\overline{\boldsymbol{X}}_2. \end{aligned}$$

(c) Compare $\tilde{\boldsymbol{\beta}}$ from the OLS regresion

$$\boldsymbol{y}^* = \boldsymbol{X}^*\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{e}}$$

with $\hat{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\alpha}_1 + \boldsymbol{d}_2\hat{\alpha}_2 + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}.$$

Exercise 3.9 Let $\hat{\boldsymbol{\beta}}_n = \left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$ denote the OLS estimate when $\boldsymbol{y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ becomes available. Prove that the OLS estimate computed using this additional observation is

$$\hat{\boldsymbol{\beta}}_{n+1} = \hat{\boldsymbol{\beta}}_n + \frac{1}{1 + \boldsymbol{x}_{n+1}'\left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_{n+1}}\left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_{n+1}\left(y_{n+1} - \boldsymbol{x}_{n+1}'\hat{\boldsymbol{\beta}}_n\right).$$

Exercise 3.10 Prove that R^2 is the square of the simple correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Exercise 3.11 The data file `cps85.dat` contains a random sample of 528 individuals from the 1985 Current Population Survey by the U.S. Census Bureau. The file contains observations on nine variables, listed in the file `cps85.pdf`.

V1 = education (in years)
V2 = region of residence (coded 1 if South, 0 otherwise)
V3 = (coded 1 if nonwhite and non-Hispanic, 0 otherwise)
V4 = (coded 1 if Hispanic, 0 otherwise)
V5 = gender (coded 1 if female, 0 otherwise)
V6 = marital status (coded 1 if married, 0 otherwise)
V7 = potential labor market experience (in years)
V8 = union status (coded 1 if in union job, 0 otherwise)
V9 = hourly wage (in dollars)

Estimate a regression of wage y_i on education x_{1i}, experience x_{2i}, and experienced-squared $x_{3i} = x_{2i}^2$ (and a constant). Report the OLS estimates.

Let $\hat{e}_i$ be the OLS residual and $\hat{y}_i$ the predicted value from the regression. Numerically calculate the following:

(a) $\sum_{i=1}^n \hat{e}_i$

(b) $\sum_{i=1}^n x_{1i}\hat{e}_i$

(c) $\sum_{i=1}^n x_{2i}\hat{e}_i$

(d) $\sum_{i=1}^n x_{1i}^2\hat{e}_i$

(e) $\sum_{i=1}^n x_{2i}^2\hat{e}_i$

(f) $\sum_{i=1}^n \hat{y}_i\hat{e}_i$

(g) $\sum_{i=1}^n \hat{e}_i^2$

(h) R^2

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.12 Using the data from the previous problem, restimate the slope on education using the residual regression approach. Regress y_i on $(1, x_{2i}, x_{2i}^2)$, regress x_{1i} on $(1, x_{2i}, x_{2i}^2)$, and regress the residuals on the residuals. Report the estimate from this regression. Does it equal the value from the first OLS regression? Explain.

In the second-stage residual regression, (the regression of the residuals on the residuals), calculate the equation R^2 and sum of squared errors. Do they equal the values from the initial OLS regression? Explain.

Chapter 4

Least Squares Regression

4.1 Introduction

In this chapter we investigate some finite-sample properties of least-squares applied to a random sample in the the linear regression model. Throughout this chapter we maintain the following.

Assumption 4.1.1 ***Linear Regression Model***
The observations $(y_i, \boldsymbol{x}_i)$ *come from a random sample and satisfy the linear regression equation*

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{4.1}$$
$$\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}y_i^2 < \infty$$

and

$$\mathbb{E}x_{ji}^2 < \infty$$

for $j = 1, ..., k$, *and an invertible design matrix*

$$\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) > 0.$$

We will consider both the general case of heteroskedastic regression, where the conditional variance

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma_i^2$$

is unrestricted, and the specialized case of homoskedastic regression, where the conditional variance is constant. In the latter case we add the following assumption.

Assumption 4.1.2 ***Homoskedastic Linear Regression Model***
In addition to Assumption 4.1.1,

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma^2 \tag{4.3}$$

is independent of x_i.

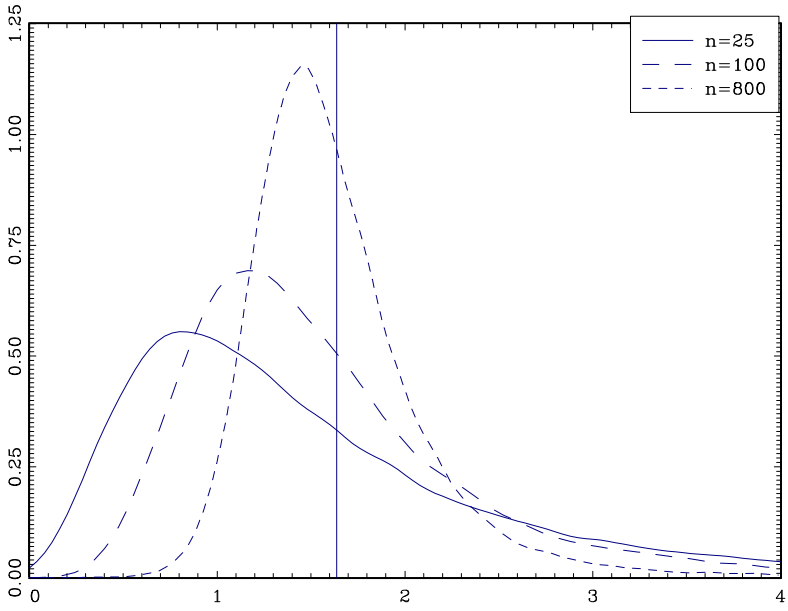

Figure 4.1: Sampling Density of $\hat{\beta}$

4.2 Sampling Distribution

The least-squares estimator is random, since it is a function of random data, and therefore has a sampling distribution. In general, its distribution is a complicated function of the joint distribution of $(y_i, \boldsymbol{x}_i)$ and the sample size n.

To illustrate the possibilities in one example, let y_i and x_i be drawn from the joint density

$$f(x,y) = \frac{1}{2\pi xy} \exp\left(-\frac{1}{2}\left(\log y - \log x\right)^2\right) \exp\left(-\frac{1}{2}\left(\log x\right)^2\right)$$

and let $\hat{\beta}$ be the slope coefficient estimate from a bivariate regression on observations from this joint density. Using simulation methods, the density function of $\hat{\beta}$ was computed and plotted in Figure 4.1 for sample sizes of $n = 25$, $n = 100$ and $n = 800$. The vertical line marks the true value of the projection coefficient.

From the figure we can see that the density functions are dispersed and highly non-normal. As the sample size increases the density becomes more concentrated about the population coefficient. To learn about the true value of β from the sample estimate $\hat{\beta}$, we need to have a way to characterize the sampling distribution of $\hat{\beta}$. We start in the next sections by deriving the mean and variance of $\hat{\beta}$.

4.3 Mean of Least-Squares Estimator

In this section we show that the OLS estimator is unbiased in the linear regression model.

Under (4.1)-(4.2) note that

$$\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(y_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \boldsymbol{x}_i'\boldsymbol{\beta} \\ \vdots \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta}. \tag{4.4}$$

Similarly

$$\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i \mid \boldsymbol{X}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i \mid \boldsymbol{x}_i) \\ \vdots \end{pmatrix} = \mathbf{0}. \tag{4.5}$$

By (3.14), conditioning on $\boldsymbol{X}$, the linearity of expectations, (4.4), and the properties of the matrix inverse,

$$\begin{aligned} \mathbb{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

Applying the law of iterated expectations to $\mathbb{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$, we find that

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\mathbb{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right) = \boldsymbol{\beta}.$$

Another way to calculate the same result is as follows. Insert $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ into the formula (3.14) for $\hat{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right)\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{e}\right) \\ &= \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}. \end{aligned} \tag{4.6}$$

This is a useful linear decomponsition of the estimator $\hat{\boldsymbol{\beta}}$ into the true parameter $\boldsymbol{\beta}$ and the stochastic component $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$.

Using (4.6), conditioning on $\boldsymbol{X}$, and (4.5),

$$\begin{aligned} \mathbb{E}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) \\ &= \mathbf{0}. \end{aligned}$$

Using either derivation, we have shown the following theorem.

Theorem 4.3.1 ***Mean of Least-Squares Estimator***
In the linear regression model (Assumption 4.1.1)

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \tag{4.7}$$

and

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{4.8}$$

Equation (4.8) says that the estimator is unbiased, meaning that the distribution of $\hat{\boldsymbol{\beta}}$ is centered at $\boldsymbol{\beta}$. Equation (4.7) says that the estimator is conditionally unbiased, which is a stronger result. It says that $\hat{\boldsymbol{\beta}}$ is unbiased for any realization of the regressor matrix $\boldsymbol{X}$.

4.4 Variance of Least Squares Estimator

In this section we calculate the conditional variance of the OLS estimator.

For any $r \times 1$ random vector $\mathbf{Z}$ define the $r \times r$ covariance matrix

$$\begin{aligned} \operatorname{var}(\boldsymbol{Z}) &= \mathbb{E}\left(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\right)\left(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\right)' \\ &= \mathbb{E}\boldsymbol{Z}\boldsymbol{Z}' - \left(\mathbb{E}\boldsymbol{Z}\right)\left(\mathbb{E}\boldsymbol{Z}\right)' \end{aligned}$$

and for any pair $(\mathbf{Z}, \mathbf{X})$ define the conditional covariance matrix

$$\operatorname{var}(\boldsymbol{Z} \mid \boldsymbol{X}) = \mathbb{E}\left(\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z} \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z} \mid \boldsymbol{X}\right)\right)' \mid \boldsymbol{X}\right).$$

The conditional covariance matrix of the $n \times 1$ regression error $\boldsymbol{e}$ is the $n \times n$ matrix

$$\boldsymbol{D} = \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right).$$

The i'th diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2$$

while the ij'th off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left(e_i e_j \mid \boldsymbol{X}\right) = \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right)\mathbb{E}\left(e_j \mid \boldsymbol{x}_j\right) = 0.$$

where the first equality uses independence of the observations (Assumption 1.5.1) and the second is (4.2). Thus $\boldsymbol{D}$ is a diagonal matrix with i'th diagonal element σ_i^2:

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.9}$$

In the special case of the linear homoskedastic regression model (4.3), then

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2 = \sigma^2$$

and we have the simplification

$$\boldsymbol{D} = \boldsymbol{I}_n \sigma^2.$$

In general, however, $\boldsymbol{D}$ need not necessarily take this simplified form.

For any matrix $n \times r$ matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\operatorname{var}(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}) = \operatorname{var}(\boldsymbol{A}'\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}. \tag{4.10}$$

In particular, we can write $\hat{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ where $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and thus

$$\begin{aligned} \operatorname{var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \end{aligned}$$

It is useful to note that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \sigma_i^2,$$

a weighted version of $\boldsymbol{X}'\boldsymbol{X}$.

Rather than working with the variance of the unscaled estimator $\hat{\boldsymbol{\beta}}$, it will be useful to work with the conditional variance of the scaled estimator $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$

$$\begin{aligned}
\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \mid \boldsymbol{X}\right) \\
&= n \operatorname{var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \\
&= n\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.
\end{aligned}$$

This rescaling might seem rather odd, but it will help provide continuity between the finite-sample treatment of this chapter and the asymptotic treatment of later chapters. As we will see in the next chapter, $\operatorname{var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$ vanishes as n tends to infinity, yet $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ converges to a constant matrix.

In the special case of the linear homoskedastic regression model, $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, so $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}\sigma^2$, and the variance matrix simplifies to

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

Theorem 4.4.1 ***Variance of Least-Squares Estimator***
In the linear regression model (Assumption 4.1.1),

$$\begin{aligned}
\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \mid \boldsymbol{X}\right) \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
\end{aligned}$$

where $\boldsymbol{D}$ is defined in (4.9).
In the homoskedastic linear regression model (Assumption 4.1.2), the covariance matrix simplifies to

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

4.5 Gauss-Markov Theorem

Now consider the class of estimators of $\boldsymbol{\beta}$ which are linear functions of the vector $\boldsymbol{y}$, and thus can be written as

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$$

where $\boldsymbol{A}$ is an $n \times k$ function of $\boldsymbol{X}$. The least-squares estimator is the special case obtained by setting $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. What is the best choice of $\boldsymbol{A}$? The Gauss-Markov theorem, which we now present, says that the least-squares estimator is the best choice when the errors are homoskedastic, as the least-squares estimator has the smallest variance among all unbiased linear estimators.

To see this, since $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$, then for any linear estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ we have

$$\mathbb{E}\left(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta},$$

so $\tilde{\boldsymbol{\beta}}$ is unbiased if (and only if) $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. Furthermore, we saw in (4.10) that

$$\operatorname{var}\left(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{A}\sigma^2.$$

the last equality using the homoskedasticity assumption $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$. The "best" unbiased linear estimator is obtained by finding the matrix $\boldsymbol{A}$ such that $\boldsymbol{A}'\boldsymbol{A}$ is minimized in the positive definite sense.

Theorem 4.5.1 ***Gauss-Markov***

1. *In the homoskedastic linear regression model (Assumption 4.1.2), the best (minimum-variance) unbiased linear estimator is the least-squares estimator*
$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$
2. *In the linear regression model (Assumption 4.1.1), the best unbiased linear estimator is*
$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y} \tag{4.11}$$

The first part of the Gauss-Markov theorem is a limited efficiency justification for the least-squares estimator. The justification is limited because the class of models is restricted to homoskedastic linear regression and the class of potential estimators is restricted to linear unbiased estimators. This latter restriction is particularly unsatisfactory as the theorem leaves open the possibility that a non-linear or biased estimator could have lower mean squared error than the least-squares estimator.

The second part of the theorem shows that in the (heteroskedastic) linear regression model, the least-squares estimator is inefficient. Within the class of linear unbiased estimators the best estimator is (4.11) and is called the **Generalized Least Squares** (GLS) estimator. This estimator is infeasible as the matrix $\boldsymbol{D}$ is unknown. This result does not suggest a practical alternative to least-squares. We return to the issue of feasible implementation of GLS in Section 7.1.

Proof of Theorem 4.5.1.1. Let $\boldsymbol{A}$ be any $n \times k$ function of $\boldsymbol{X}$ such that $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. The variance of the least-squares estimator is $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$ and that of $\boldsymbol{A}'\boldsymbol{y}$ is $\boldsymbol{A}'\boldsymbol{A}\sigma^2$. It is sufficient to show that the difference $\boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ is positive semi-definite. Set $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Note that $\boldsymbol{X}'\boldsymbol{C} = \boldsymbol{0}$. Then we calculate that

$$\begin{aligned}
\boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} &= \left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{C} \\
&\quad + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C}
\end{aligned}$$

The matrix $\boldsymbol{C}'\boldsymbol{C}$ is positive semi-definite (see Appendix A.7) as required.

The proof of **Theorem 4.5.1.2** is left for Exercise 4.3.

4.6 Residuals

What are some properties of the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ and prediction errors $\tilde{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(-i)}$, at least in the context of the linear regression model?

Recall from (3.17) and (3.18) that we can write the residuals in vector notation as

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{e}$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is the matrix which projects on the the space orthogonal to the columns of $\boldsymbol{X}$. Using the properties of conditional expectation

$$\mathbb{E}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{0}$$

and

$$\operatorname{var}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\operatorname{\mathbf{var}}\left(\boldsymbol{e} \mid \boldsymbol{X}\right)\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \tag{4.12}$$

where $\boldsymbol{D}$ is defined in (4.9).

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2.$$

In this case (4.12) simplies to

$$\operatorname{var}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \boldsymbol{M}\sigma^2.$$

In particular, for a single observation i, we obtain

$$\operatorname{var}\left(\hat{e}_i \mid \boldsymbol{X}\right) = \mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) = \left(1 - h_{ii}\right)\sigma^2 \tag{4.13}$$

since the diagonal elements of $\boldsymbol{M}$ are $1 - h_{ii}$ as defined in (3.19). Thus the residuals are heteroskedastic even if the errors are homoskedastic.

Similarly, we can write the prediction errors $\tilde{e}_i = \left(1 - h_{ii}\right)^{-1}\hat{e}_i$ in vector notation. Set

$$\boldsymbol{M}^* = \operatorname{diag}\{\left(1 - h_{11}\right)^{-1}, .., \left(1 - h_{nn}\right)^{-1}\}$$

Then we can write the prediction errors as

$$\begin{aligned}\tilde{\boldsymbol{e}} &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{y} \\ &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{e}.\end{aligned}$$

We can calculate that

$$\mathbb{E}\left(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \boldsymbol{M}^*\boldsymbol{M}\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{0}$$

and

$$\operatorname{var}\left(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \boldsymbol{M}^*\boldsymbol{M}\operatorname{\mathbf{var}}\left(\boldsymbol{e} \mid \boldsymbol{X}\right)\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*$$

which simplifies under homoskedasticity to

$$\begin{aligned}\operatorname{var}\left(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}\right) &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}^*\sigma^2 \\ &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2.\end{aligned}$$

The variance of the i'th prediction error is then

$$\begin{aligned}\operatorname{var}\left(\tilde{e}_i \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\tilde{e}_i^2 \mid \boldsymbol{X}\right) \\ &= \left(1 - h_{ii}\right)^{-1}\left(1 - h_{ii}\right)\left(1 - h_{ii}\right)^{-1}\sigma^2 \\ &= \left(1 - h_{ii}\right)^{-1}\sigma^2.\end{aligned}$$

A residual with proper variance can be obtained by rescaling. The **studentized residuals** are

$$\bar{e}_i = (1 - h_i)^{-1/2}\, \hat{e}_i, \tag{4.14}$$

and in vector notation

$$\bar{\boldsymbol{e}} = (\bar{e}_1, ..., \bar{e}_n)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\,\boldsymbol{e}.$$

From our above calculations, under homoskedasticity,

$$\operatorname{var}(\bar{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^2$$

and

$$\operatorname{var}(\bar{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\bar{e}_i^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.15}$$

and thus these rescaled residuals have the same bias and variance as the original errors when the latter are homoskedastic.

4.7 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}e_i^2$ can be a parameter of interest, even in a heteroskedastic regression or a projection model. σ^2 measures the variation in the "unexplained" part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

and equals the MLE in the normal regression model (3.13).

In the linear regression model we can calculate the mean of $\hat{\sigma}^2$. From (3.18), the properties of projection matrices and the trace operator, observe that

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{M}\,\boldsymbol{e} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e} = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\,\boldsymbol{e}\boldsymbol{e}'\right).$$

Then

$$\begin{aligned}
\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{M}\,\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{D}\right).
\end{aligned} \tag{4.16}$$

Adding the assumption of conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, so that $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, then (4.16) simplifies to

$$\begin{aligned}
\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) \\
&= \sigma^2\left(\frac{n-k}{n}\right),
\end{aligned}$$

the final equality by (3.16). This calculation shows that $\hat{\sigma}^2$ is biased towards zero. The order of the bias depends on k/n, the ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.13). Note that

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)\sigma^2 \\ &= \left(\frac{n-k}{n}\right)\sigma^2\end{aligned}$$

using (3.20).

Since the bias takes a scale form, a classic method to obtain an unbiased estimator is by rescaling the estimator. Define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2. \tag{4.17}$$

By the above calculation,

$$\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.18}$$

so

$$\mathbb{E}\left(s^2\right) = \sigma^2$$

and the estimator s^2 is unbiased for σ^2. Consequently, s^2 is known as the "bias-corrected estimator" for σ^2 and in empirical practice s^2 is the most widely used estimator for σ^2.

Interestingly, this is not the only method to construct an unbiased estimator for σ^2. An alternative unbiased estimator can be using the studentized residuals $\bar{e}_i$ from (4.14), yielding the estimator

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\hat{e}_i^2.$$

You can show (see Exercise 4.6) that

$$\mathbb{E}\left(\bar{\sigma}^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.19}$$

and thus $\bar{\sigma}^2$ is unbiased for σ^2 (in the homoskedastic linear regression model).

When the sample sizes are large and the number of regressors small, the estimators $\hat{\sigma}^2$, s^2 and $\bar{\sigma}^2$ are likely to be close. For example, in the regression (3.10), $\hat{\sigma}$, s, and $\bar{\sigma}$ all equal 0.490. The estimators are more likely to differ when n is small and k is large.

4.8 Covariance Matrix Estimation Under Homoskedasticity

For inference, we need an estimate of the covariance matrix $\boldsymbol{V}_{\hat{\beta}}$ of the least-squares estimator. In this section we consider estimation of $\boldsymbol{V}_{\hat{\beta}}$ in the homoskedastic regression model (4.1)-(4.2)-(4.3).

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

which is known up to the unknown scale σ^2. In the previous section we discussed three estimators of σ^2. The most commonly used choice is s^2, leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}s^2. \tag{4.20}$$

Since s^2 is conditionally unbiased for σ^2, it is simple to calculate that $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ is conditionally unbiased for $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$ under the assumption of homoskedasticity:

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) \\ &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ &= \boldsymbol{V}_{\hat{\boldsymbol{\beta}}}.\end{aligned}$$

This estimator was the dominant covariance matrix estimator in applied econometrics in previous generations, and is still the default in most regression packages.

If the estimator (4.20) is used, but the regression error is heteroskedastic, it is possible for $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ to be quite biased for the correct covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. For example, suppose $k = 1$ and $\sigma_i^2 = x_i^2$ (extreme heteroskedasticity). The ratio of the true variance of the least-squares estimator to the expectation of the variance estimator is

$$\frac{\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}}{\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right)} = \frac{\frac{1}{n}\sum_{i=1}^n x_i^4}{\sigma^2\frac{1}{n}\sum_{i=1}^n x_i^2} \simeq \frac{\mathbb{E}x_i^4}{\sigma^2\mathbb{E}x_i^2} = \frac{\mathbb{E}x_i^4}{\left(\mathbb{E}x_i^2\right)^2}.$$

(Notice that we use the fact that $\sigma_i^2 = x_i^2$ implies $\sigma^2 = \mathbb{E}\sigma_i^2 = \mathbb{E}x_i^2$.) This is the kurtosis of the regressor x_i. As the kurtosis can be any number greater than one, we conclude that the bias of $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ can be arbitrarily large. While this is an extreme and constructed example, the point is that the classic covariance matrix estimator (4.20) may be quite biased when the homoskedasticity assumption fails.

4.9 Covariance Matrix Estimation Under Heteroskedasticity

In the previous section we showed that that the classic covariance matrix estimator can be highly biased if homoskedasticity fails. In this section we show how to contruct covariance matrix estimators which do not require homoskedasticity.

Recall that the general form for the covariance matrix is

$$\boldsymbol{V}_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

This depends on the unknown matrix $\boldsymbol{D}$ which we can write as

$$\begin{aligned}\boldsymbol{D} &= \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) \\ &= \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) \\ &= \mathbb{E}\left(\operatorname{diag}\left(e_1^2, ..., e_n^2\right) \mid \boldsymbol{X}\right).\end{aligned}$$

Thus $\boldsymbol{D}$ is the conditional mean of $\operatorname{diag}\left(e_1^2, ..., e_n^2\right)$, so the latter is an unbiased estimator for $\boldsymbol{D}$. Therefore, if the squared errors e_i^2 were observable, we could construct the unbiased estimator

$$\begin{aligned}\boldsymbol{V}^{ideal}_{\hat{\boldsymbol{\beta}}} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\operatorname{diag}\left(e_1^2, ..., e_n^2\right)\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

Indeed,

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{ideal} \mid \boldsymbol{X}\right) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right)\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{V}_{\hat{\boldsymbol{\beta}}}
\end{aligned}$$

verifying that $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{ideal}$ is unbiased for $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}$

Since the errors e_i^2 are unobserved, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}^{ideal}$ is not a feasible estimator. To construct a feasible estimator we can replace the errors with the least-squares residuals $\hat{e}_i$, the prediction errors $\tilde{e}_i$ or the unbiased residuals $\bar{e}_i$, e.g.

$$\begin{aligned}
\widehat{\boldsymbol{D}} &= \operatorname{diag}\left(\hat{e}_1^2, ..., \hat{e}_n^2\right), \\
\widetilde{\boldsymbol{D}} &= \operatorname{diag}\left(\tilde{e}_1^2, ..., \tilde{e}_n^2\right), \\
\overline{\boldsymbol{D}} &= \operatorname{diag}\left(\bar{e}_1^2, ..., \bar{e}_n^2\right).
\end{aligned}$$

Substituting these matrices into the formula for $\boldsymbol{V}_{\hat{\beta}}$ we obtain the estimators

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\widehat{\boldsymbol{D}}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1},
\end{aligned}$$

$$\begin{aligned}
\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\tilde{e}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1},
\end{aligned}$$

and

$$\begin{aligned}
\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\overline{\boldsymbol{D}}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\bar{e}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.
\end{aligned}$$

The estimators $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ are often called *robust, heteroskedasticity-consistent*, or *heteroskedasticity-robust* covariance matrix estimators. The estimator $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ was first developed by Eicker (1963), and introduced to econometrics by White (1980), and is sometimes called the **Eicker-White** or **White**

covariance matrix estimator[1]. The estimator $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ was introduced by Andrews (1991) based on the principle of leave-one-out cross-validation, and the estimator $\overline{\boldsymbol{V}}_{\hat{\beta}}$ was introduced by Horn, Horn and Duncan (1975) as a reduced-bias covariance matrix estimator.

In general, the bias of the estimators $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ and $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, is quite complicated, but they greatly simplify under the assumption of homoskedasticity (4.3). For example, using (4.13),

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right)\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\left(1-h_{ii}\right)\sigma^2\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'h_{ii}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&< \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&= \boldsymbol{V}_{\hat{\boldsymbol{\beta}}}.
\end{aligned}$$

The inequality $\boldsymbol{A} < \boldsymbol{B}$ when applied to matrices means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is positive definite, which holds here since $\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'h_{ii}$ is positive definite. This calculation shows that $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is biased downwards.

Similarly, (again under homoskedasticity) we can calculate that $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is biased upwards, specifically

$$\mathbb{E}\left(\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) > \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \tag{4.21}$$

while the estimator $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is unbiased

$$\mathbb{E}\left(\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.22}$$

(See Exercise 4.7

It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity, but it does give us a baseline for comparison.

We have introduced four covariance matrix estimators, $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$, $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$. Which should you use? The classic estimator $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ is typically a poor choice, as it is only valid under the unlikely homoskedasticity restriction. For this reason it is not typically used in contemporary econometric research. Of the three robust estimators, $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is the most commonly used, as it is the most straightforward and familiar. However, $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, and in particular $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, are perferred based on their improved bias. Unfortunately, standard regression packages set the classic estimator $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$ as the default. As $\widetilde{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ and $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ are simple to implement, this should not be a barrier. For example, in STATA, $\overline{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is implemented by selecting "Robust" standard errors and selecting the bias correction option "$1/(1-h)$", or using the vce(hc2) option.

4.10 Standard Errors

A variance estimator such as $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ is an estimate of the variance of the distribution of $\hat{\boldsymbol{\beta}}$. A more easily interpretable measure of spread is its square root – the standard deviation. This is

[1]Often, this estimator is rescaled by multiplying by the ad hoc bias adjustment $\frac{n}{n-k}$ in analogy to the bias-corrected error variance estimator.

so important when discussing the distribution of parameter estimates, we have a special name for estimates of their standard deviation.

Definition 4.10.1 *A **standard error** $s(\hat{\beta})$ for an real-valued estimator $\hat{\beta}$ is an estimate of the standard deviation of the distribution of $\hat{\beta}$.*

When $\boldsymbol{\beta}$ is a vector with estimate $\hat{\boldsymbol{\beta}}$ and covariance matrix estimate $n^{-1}\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$, standard errors for individual elements are the square roots of the diagonal elements of $n^{-1}\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$. That is,

$$s(\hat{\beta}_j) = \sqrt{n^{-1}\widehat{\boldsymbol{V}}_{\hat{\beta}_j}} = n^{-1/2}\sqrt{\left[\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\right]_{jj}}.$$

As we discussed in the previous section, there are multiple possible covariance matrix estimators, so standard errors are not unique. It is therefore important to understand what formula and method is used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions.

4.11 Multicollinearity

If $\operatorname{rank}(\boldsymbol{X}'\boldsymbol{X}) < k$, then $\hat{\boldsymbol{\beta}}$ is not defined[2]. This is called **strict multicollinearity**. This happens when the columns of $\boldsymbol{X}$ are linearly dependent, i.e., there is some $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if $\boldsymbol{X}$ includes both the logs of two prices and the log of the relative prices, $\log(p_1)$, $\log(p_2)$ and $\log(p_1/p_2)$. When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Since the error is discovered quickly, this is rarely a *problem* for applied econometric practice.

The more relevant situation is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $\boldsymbol{X}'\boldsymbol{X}$ matrix is *near* singular, when the columns of $\boldsymbol{X}$ are *close* to linearly dependent. This definition is not precise, because we have not said what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced. In extreme cases it is possible that the reported calculations will be in error.

A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise. We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

and

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

In this case

$$\operatorname{var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

[2]See Appendix A.5 for the defintion of the rank of a matrix.

The correlation ρ indexes collinearity, since as ρ approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the variance of a coefficient estimate $\sigma^2 \left[n\left(1-\rho^2\right)\right]^{-1}$ approaches infinity as ρ approaches 1. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of β_2. As a consequence, the precision of individual estimates are reduced. The imprecision, however, will be reflected by large standard errors, so there is no distortion in inference.

Some earlier textbooks overemphasized a concern about multicollinearity. A very amusing parody of these texts appeared in Chapter 23.3 of Goldberger's *A Course in Econometrics* (1991), which is reprinted below. To understand his basic point, you should notice how the estimation variance $\sigma^2 \left[n\left(1-\rho^2\right)\right]^{-1}$ depends equally and symmetrically on the the correlation ρ and the sample size n.

Arthur S. Goldberger

Art Goldberger (1930-2009) was one of the most distinguished members of the Department of Economics at the University of Wisconsin. His PhD thesis developed an early macroeconometric forecasting model (known as the Klein-Goldberger model) but most of his career focused on microeconometric issues. He was the leading pioneer of what has been called the Wisconsin Tradition of empirical work – a combination of formal econometric theory with a careful critical analysis of empirical work. Goldberger wrote a series of highly regarded and influential graduate econometric textbooks, including including *Econometric Theory* (1964), *Topics in Regression Analysis* (1968), and *A Course in Econometrics* (1991).

Micronumerosity

Arthur S. Goldberger

A Course in Econometrics (1991), Chapter 23.3

Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimation a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for "small sample size." If so, we can remove that impediment by introducing the term *micronumerosity*.

Suppose an econometrician set out to write a chapter about small sample size in sampling from a univariate population. Judging from what is now written about multicollinearity, the chapter might look like this:

1. *Micronumerosity*

   The extreme case, "exact micronumerosity," arises when $n = 0$, in which case the sample estimate of μ is not unique. (Technically, there is a violation of the rank condition $n > 0$: the matrix 0 is singular.) The extreme case is easy enough to recognize. "Near micronumerosity" is more subtle, and yet very serious. It arises when the rank condition $n > 0$ is barely satisfied. Near micronumerosity is very prevalent in empirical economics.

2. *Consequences of micronumerosity*

   The consequences of micronumerosity are serious. Precision of estimation is reduced. There are two aspects of this reduction: estimates of μ may have large errors, and not only that, but $V_{\bar{y}}$ will be large.

   Investigators will sometimes be led to accept the hypothesis $\mu = 0$ because $\bar{y}/\hat{\sigma}_{\bar{y}}$ is small, even though the true situation may be not that $\mu = 0$ but simply that the sample data have not enabled us to pick μ up.

   The estimate of μ will be very sensitive to sample data, and the addition of a few more observations can sometimes produce drastic shifts in the sample mean.

   The true μ may be sufficiently large for the null hypothesis $\mu = 0$ to be rejected, even though $V_{\bar{y}} = \sigma^2/n$ is large because of micronumerosity. But if the true μ is small (although nonzero) the hypothesis $\mu = 0$ may mistakenly be accepted.

3. *Testing for micronumerosity*

   Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule.

   A generally reliable guide may be obtained by counting the number of observations. Most of the time in econometric analysis, when n is close to zero, it is also far from infinity.

   Several test procedures develop critical values n^*, such that micronumerosity is a problem only if n is smaller than n^*. But those procedures are questionable.

4. *Remedies for micronumerosity*

   If micronumerosity proves serious in the sense that the estimate of μ has an unsatisfactorily low degree of precision, we are in the statistical position of not being able to make bricks without straw. The remedy lies essentially in the acquisition, if possible, of larger samples from the same population.

   But more data are no remedy for micronumerosity if the additional data are simply "more of the same." So obtaining lots of small samples from the same population will not help.

4.12 Omitted Variable Bias

Let the regressors be partitioned as

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix}.$$

We can write the regression of y_i on $\boldsymbol{x}_i$ as

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0}. \end{aligned} \tag{4.23}$$

Now suppose that instead of estimating equation (4.23) by least-squares, we regress y_i on $\boldsymbol{x}_{1i}$ only. Perhaps this is done because the variables $\boldsymbol{x}_{2i}$ are not in the data set, in order to reduce the number of estimated parameters. Effectively, we are estimating the equation

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\gamma}_1 + u_i \\ \mathbb{E}(\boldsymbol{x}_{1i} u_i) &= \mathbf{0} \end{aligned} \tag{4.24}$$

Notice that we have written the coefficient on $\boldsymbol{x}_{1i}$ as $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$ and the error as u_i rather than e_i. This is because the model being estimated is different than (4.23). Goldberger (1991) introduced the labels (4.23) the **long regression** and (4.24) the **short regression** to emphasize the distinction.

Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \left(\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_{1i} y_i\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_{1i}\left(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i\right)\right) \\ &= \boldsymbol{\beta}_1 + \left(\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{2i}'\right)\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2 \end{aligned}$$

where

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{2i}'\right)$$

is the coefficient from a regression of $\boldsymbol{x}_{2i}$ on $\boldsymbol{x}_{1i}$.

Observe that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\beta}_1$ unless $\boldsymbol{\Gamma} = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus the short and long regressions have the same coefficient on $\boldsymbol{x}_{1i}$ only under one of two conditions. First, the regression of $\boldsymbol{x}_{2i}$ on $\boldsymbol{x}_{1i}$ yields a set of zero coefficients (they are uncorrelated), or second, the coefficient on $\boldsymbol{x}_{2i}$ in (4.23) is zero. In general, least-squares estimation of (4.24) is an estimate of $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2$ rather than $\boldsymbol{\beta}_1$. The difference $\boldsymbol{\Gamma}\boldsymbol{\beta}_2$ is known as **omitted variable bias**. It is the consequence of omission of a relevant correlated variable.

To avoid omitted variables bias the standard advice is to include potentially relevant variables in the estimated model. By construction, the general model will be free of the omitted variables problem. Typically there are limits, as many desired variables are not available in a given dataset. In this case, the possibility of omitted variables bias should be acknowledged and discussed in the course of an empirical investigation.

4.13 Normal Regression Model

In the special case of the normal linear regression model introduced in Section 3.11, we can derive exact sampling distributions for the least-squares estimator, residuals, and variance estimator.

In particular, under the normality assumption $e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then we have the multivariate implication

$$\boldsymbol{e} \mid \boldsymbol{X} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{I}_n\sigma^2\right).$$

That is, the error vector $\boldsymbol{e}$ is independent of $\boldsymbol{X}$ and is normally distributed. Since linear functions of normals are also normal, this implies that conditional on $\boldsymbol{X}$

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{e}} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ \boldsymbol{M} \end{pmatrix} \boldsymbol{e} \sim \mathrm{N}\left(0, \begin{pmatrix} \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} & 0 \\ 0 & \sigma^2 \boldsymbol{M} \end{pmatrix}\right)$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Since uncorrelated normal variables are independent, it follows that $\hat{\boldsymbol{\beta}}$ is independent of any function of the OLS residuals including the estimated error variance s^2 or $\hat{\sigma}^2$ or prediction errors $\tilde{\boldsymbol{e}}$.

The spectral decomposition of $\boldsymbol{M}$ yields

$$\boldsymbol{M} = \boldsymbol{H} \begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{H}'$$

(see equation (A.4)) where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$. Let $\boldsymbol{u} = \sigma^{-1}\boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{H}'\boldsymbol{H}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then

$$\begin{aligned} \frac{n\hat{\sigma}^2}{\sigma^2} &= \frac{(n-k)\,s^2}{\sigma^2} \\ &= \frac{1}{\sigma^2}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e} \\ &= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{H} \begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{H}'\boldsymbol{e} \\ &= \boldsymbol{u}' \begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{u} \\ &\sim \chi^2_{n-k}, \end{aligned}$$

a chi-square distribution with $n-k$ degrees of freedom.

Furthermore, if standard errors are calculated using the homoskedastic formula (4.20)

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} \sim \frac{\mathrm{N}\left(0, \sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{n-k}\chi^2_{n-k}}\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} = \frac{\mathrm{N}(0,1)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim t_{n-k}$$

a t distribution with $n-k$ degrees of freedom.

Theorem 4.13.1 ***Normal Regression***
In the linear regression model (Assumption 4.1.1) if e_i is independent of $\boldsymbol{x}_i$ and distributed $\mathrm{N}(0, \sigma^2)$ then

- $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$
- $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$
- $\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$

These are the exact finite-sample distributions of the least-squares estimator and variance estimators, and are the basis for traditional inference in linear regression.

While elegant, the difficulty in applying Theorem 4.13.1 is that the normality assumption is too restrictive to be empirical plausible, and therefore inference based on Theorem 4.13.1 has no guarantee of accuracy. We develop a more broadly-applicable inference theory based on large sample (asymptotic) approximations in the following chapter.

Exercises

Exercise 4.1 Explain the difference between $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ and $\mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right)$.

Exercise 4.2 True or False. If $y_i = x_i\beta + e_i$, $x_i \in \mathbb{R}$, $\mathbb{E}(e_i \mid x_i) = 0$, and $\hat{e}_i$ is the OLS residual from the regression of y_i on x_i, then $\sum_{i=1}^{n} x_i^2 \hat{e}_i = 0$.

Exercise 4.3 Prove Theorem 4.5.1.2.

Exercise 4.4 In a linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = 0, \quad \operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{\Omega}$$

with $\boldsymbol{\Omega}$ known, the GLS estimator is

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}\right).$$

the residual vector is $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, and an estimate of σ^2 is

$$s^2 = \frac{1}{n-k}\hat{\boldsymbol{e}}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{e}}.$$

(a) Why is this a reasonable estimator for σ^2?

(b) Prove that $\hat{\boldsymbol{e}} = \boldsymbol{M}_1\boldsymbol{e}$, where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

(c) Prove that $\boldsymbol{M}_1'\boldsymbol{\Omega}^{-1}\boldsymbol{M}_1 = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

Exercise 4.5 Let $(y_i, \boldsymbol{x}_i)$ be a random sample with $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. Consider the **Weighted Least Squares** (WLS) estimator of $\boldsymbol{\beta}$

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\right)$$

where $\boldsymbol{W} = \operatorname{diag}(w_1, ..., w_n)$ and $w_i = x_{ji}^{-2}$, where x_{ji} is one of the $\boldsymbol{x}_i$.

(a) In which contexts would $\tilde{\boldsymbol{\beta}}$ be a good estimator?

(b) Using your intuition, in which situations would you expect that $\tilde{\boldsymbol{\beta}}$ would perform better than OLS?

Exercise 4.6 Show (4.19) in the homoskedastic regression model.

Exercise 4.7 Show (4.21) and (4.22) in the homoskedastic regression model.

Chapter 5

Asymptotic Theory

5.1 Introduction

As discussed in Section 4.2, the OLS estimator $\hat{\boldsymbol{\beta}}$ is has an unknown statistical distribution. Inference (confidence intervals and hypothesis testing) requires useful approximations to the sampling distribution. The most widely used and versatile method is asymptotic theory, which approximates sampling distributions by taking the limit of the finite sample distribution as the sample size n tends to infinity. The primary tools of asymptotic theory are the weak law of large numbers (WLLN), central limit theorem (CLT), and continuous mapping theorem (CMT). With these tools we can approximate the sampling distributions of most econometric estimators.

It turns out that most of this theory equally applies to the projection model and the linear conditional mean model, and therefore the results in this Chapter will be stated for the broader projection model unless otherwise stated. Throughout this chapter we maintain the following.

Assumption 5.1.1 ***Linear Projection Model***
The observations $(y_i, \boldsymbol{x}_i)$ *come from a random sample with finite second moments*

$$\mathbb{E}y_i^2 < \infty$$

and

$$\mathbb{E}x_{ji}^2 < \infty$$

for $j = 1, ..., k$, *and an invertible design matrix*

$$\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right) > 0.$$

From Theorems 2.9.1 and 2.9.2, under Assumtpion 5.1.1 the variables satisfy the linear projection equation

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \\ \boldsymbol{\beta} &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right). \end{aligned}$$

A review of the most important tools in asymptotic theory is contained in Appendix C.

5.2 Weak Law of Large Numbers

At the beginning of Chapter 4, we showed in Figure 4.1 how the sampling density of the least-square estimator varies with the sample size n. It is possible to see in the figure that the sampling

density concentrates about the true parameter value as the sample size increases. This is the property of estimator consistency – convergence in probability to the true parameter value. In this section we review the core theory explaining this phenomenon.

At its heart, estimator consistency is the effect of sample size on the variance of the sample mean. To review, suppose u_i is an iid random variable with finite mean $\mathbb{E}u_i = \mu$ and variance $\mathbb{E}(u_i - \mu)^2 = \sigma^2$, and consider the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n u_i$. The mean and variance of $\hat{\mu}$ are

$$\mathbb{E}\hat{\mu} = \mathbb{E}\frac{1}{n}\sum_{i=1}^n u_i = \frac{1}{n}\sum_{i=1}^n \mathbb{E}u_i = \mu$$

and

$$\operatorname{var}(\hat{\mu}) = \mathbb{E}(\hat{\mu} - \mu)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n (u_i - \mu)\right)^2 = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}(u_i - \mu)(u_j - \mu) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

where the second-to-last inequality is because $\mathbb{E}(u_i - \mu)(u_j - \mu) = \sigma^2$ for $i = j$ yet $\mathbb{E}(u_i - \mu)(u_j - \mu) = 0$ for $i \neq j$ due to independence.

We see that $\operatorname{var}(\hat{\mu}) = \sigma^2/n$ which is decreasing in n (as long as $\sigma^2 < \infty$). It follows that $\operatorname{var}(\hat{\mu}) = \sigma^2/n \to 0$ as $n \to \infty$. This means that the distribution of $\hat{\mu}$ is increasingly concentrated about its mean μ as n increases.

To be more precise, for any $\delta > 0$, an application of Chebyshev's inequality yields

$$\mathbb{P}(|\hat{\mu} - \mu| > \delta) \leq \frac{\operatorname{var}(\hat{\mu})}{\delta^2} = \frac{\sigma^2/n}{\delta^2} \to 0$$

as $n \to \infty$. This says that the probability that $\hat{\mu}$ differs from μ by more than δ declines to zero as $n \to \infty$. Equivalently, the distribution of $\hat{\mu}$ becomes concentrated within the region $[\mu - \delta, \mu + \delta]$ as n diverges. As this holds for *any* δ (even an extremely small value) it is reasonable to say that the distribution of $\hat{\mu}$ concentrates about μ as n increases.

We have described three distinct but intertwined concepts: convergence in probability (concentration of a sampling distribution), consistency (convergence in probability of an estimator to the parameter value), and the weak law of large numbers (convergence in probability of the sample mean). We now state these concepts formally.

Definition 5.2.1 *We say that a random variable $z_n \in \mathbb{R}$ **converges in probability** to z as $n \to \infty$, denoted $z_n \xrightarrow{p} z$, if for all $\delta > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|z_n - z| > \delta) = 0.$$

Definition 5.2.2 *An estimator $\hat{\boldsymbol{\theta}}$ of a parameter $\boldsymbol{\theta}$ is **consistent** if $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$*

Consistency is a good property for an estimator to possess. It means that for any given data distribution, there is a sample size n sufficiently large such that the estimator $\hat{\theta}$ will be arbitrarily close to the true value θ with high probability.

Above, we showed that the sample mean $\hat{\mu}$ converges in probability to the population mean μ as $n \to \infty$, and is thus consistent for μ. This result is known as the weak law of large numbers.

Theorem 5.2.1 ***Weak Law of Large Numbers (WLLN)***
If $u_i \in \mathbb{R}$ is iid and $\mathbb{E}|u_i| < \infty$, then

$$\overline{u}_n = \frac{1}{n}\sum_{i=1}^{n} u_i \xrightarrow{p} \mathbb{E}(u_i)$$

as $n \to \infty$.

Theorem 5.2.2 ***WLLN for Random Matrices***
If $\boldsymbol{U}_i \in \mathbb{R}^{k\times r}$ is iid and $\mathbb{E}|u_{jli}| < \infty$ for $1 \le j \le k$ and $1 \le l \le r$ then

$$\overline{\boldsymbol{U}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i \xrightarrow{p} \mathbb{E}(\boldsymbol{U}_i)$$

as $n \to \infty$.

In our derivation, we proved the WLLN under the assumption that u_i has a finite variance. Theorem 5.2.1 states that the WLLN holds under the weaker assumption of a finite mean. We provide a proof of this more general result for the technically-inclinded readers.

Proof of Theorem 5.2.1: Without loss of generality, we can assume $\mathbb{E}(u_i) = 0$ by recentering u_i on its expectation.

We need to show that for all $\delta > 0$ and $\eta > 0$ there is some $N < \infty$ so that for all $n \ge N$, $\mathbb{P}(|\overline{u}_n| > \delta) \le \eta$. Fix δ and η. Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$\mathbb{E}(|u_i| 1(|u_i| > C)) \le \varepsilon \tag{5.1}$$

(where $1(\cdot)$ is the indicator function) which is possible since $\mathbb{E}|u_i| < \infty$. Define the random variables

$$\begin{aligned} w_i &= u_i 1(|u_i| \le C) - \mathbb{E}(u_i 1(|u_i| \le C)) \\ z_i &= u_i 1(|u_i| > C) - \mathbb{E}(u_i 1(|u_i| > C)). \end{aligned}$$

By the Triangle Inequality (A.8), the Expectation Inequality (C.2), and (5.1),

$$\begin{aligned} \mathbb{E}|\overline{z}_n| &= \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|z_i| \\ &= \mathbb{E}|z_i| \\ &\le \mathbb{E}|u_i| 1(|u_i| > C) + |\mathbb{E}(u_i 1(|u_i| > C))| \\ &\le 2\mathbb{E}|u_i| 1(|u_i| > C) \\ &\le 2\varepsilon. \end{aligned} \tag{5.2}$$

By Jensen's Inequality (C.1), the fact that the w_i are iid and mean zero, and the bound $|w_i| \leq 2C$,

$$\begin{aligned}(\mathbb{E}\,|\overline{w}_n|)^2 &\leq \mathbb{E}\overline{w}_n^2 \\ &= \frac{\mathbb{E}w_i^2}{n} \\ &\leq \frac{4C^2}{n} \\ &\leq \varepsilon^2 \end{aligned} \tag{5.3}$$

the final inequality holding for $n \geq 4C^2/\varepsilon^2 = 36C^2/\delta^2\eta^2$.

Finally, by Markov's Inequality (C.6), the fact that $\overline{u}_n = \overline{w}_n + \overline{z}_n$, the triangle inequality, (5.2) and (5.3),

$$\mathbb{P}\left(|\overline{u}_n| > \delta\right) \leq \frac{\mathbb{E}\,|\overline{u}_n|}{\delta} \leq \frac{\mathbb{E}\,|\overline{w}_n| + \mathbb{E}\,|\overline{z}_n|}{\delta} \leq \frac{3\varepsilon}{\delta} = \eta,$$

the equality by the definition of ε. We have shown that for any $\delta > 0$ and $\eta > 0$ then for all $n \geq 36C^2/\delta^2\eta^2$, $\mathbb{P}\left(|\overline{u}_n| > \delta\right) \leq \eta$, as needed.

Proof of Theorem 5.2.2: A random vector or matrix converges in probability to its limit if (and only if) all elements in the vector or matrix converge in probability. Since each element of $\boldsymbol{U}_i$ has a finite mean by assumption, Theorem 5.2.1 applies to each element and therefore converges in probability, as needed.

Jacob Bernoulli

Jacob Bernoulli (1654 -1705) of Switzerland was one of many famous mathematicians in the Bernoulli family. One of Jacob Bernoulli's important contributions was the first proof of the weak law of large numbers, published in his posthumous masterpiece *Ars Conjectandi*.

5.3 Consistency of Least-Squares Estimation

In this section we use the WLLN and continuous mapping theorem (CMT, Theorem C.3.1) to show that the least-squares estimator $\hat{\boldsymbol{\beta}}$ is consistent for the projection coefficient $\boldsymbol{\beta}$.

This derivation is based on three key components. First, the OLS estimator can be written as a continuous function of a set of sample moments. Second, the weak law of large numbers (WLLN, Theorem 5.2.1) shows that sample moments converge in probability to population moments. And third, the continuous mapping theorem (CMT, Theorem C.3.1) states that continuous functions preserve convergence in probability. We now explain each step in brief and then in greater detail.

First, observe that the OLS estimator

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right)$$

is a function of the sample moments $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$ and $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i$.

Second, by an application of the WLLN these sample moments converge in probability to the population moments. Specifically, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) = \boldsymbol{Q} \tag{5.4}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i y_i\right). \tag{5.5}$$

Third, the CMT to allows us to combine these equations to show that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Specifically, as $n \to \infty$,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right) \\ &\xrightarrow{p} \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{x}_i y_i\right)\right) \\ &= \boldsymbol{\beta}. \end{aligned} \tag{5.6}$$

We have shown that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, as $n \to \infty$. In words, the OLS estimator converges in probability to the projection coefficient vector $\boldsymbol{\beta}$ as the sample size n gets large.

For a slightly different demonstration of this result, recall that (4.6) implies that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i e_i\right). \tag{5.7}$$

The WLLN and (2.14) imply

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i e_i \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i e_i\right) = 0. \tag{5.8}$$

Therefore

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i e_i\right) \\ &\xrightarrow{p} \boldsymbol{Q}^{-1}\mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

which is the same as $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Theorem 5.3.1 ***Consistency of Least-Squares***
Under Assumption 5.1.1, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

Theorem 5.3.1 states that the OLS estimator $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$ as n diverges to positive infinity, and thus $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

We now explain the application of the WLLN in (5.4) and (5.5) and the CMT in (5.6) in greater detail.

The weak law of large numbers (Theorem 5.2.1, Section 5.2) says that when when random variables are iid and have finite mean, then sample averages converge in probability to their population mean. Thus to apply the WLLN to (5.4) and (5.5) it is sufficient to verify that the elements of the random matrices $\boldsymbol{x}_i\boldsymbol{x}_i'$ and $\boldsymbol{x}_i y_i$ are iid and have finite mean. First, these random variables are iid because the observations $(y_i, \boldsymbol{x}_i)$ are mutually independent and identically distributed (Assumption 1.5.1), and so are any functions of the observations, including $\boldsymbol{x}_i\boldsymbol{x}_i'$ and $\boldsymbol{x}_i y_i$. Second, Assumption

5.1.1 is sufficient for $1 \leq j \leq k$ and $1 \leq l \leq k$, $\mathbb{E}\left|x_{ji}x_{li}\right| < \infty$ and $\mathbb{E}\left|x_{ji}y_i\right| < \infty$. Indeed, by an application of the Cauchy-Schwarz inequality and Assumption 5.1.1

$$\mathbb{E}\left|x_{ji}x_{li}\right| \leq \left(\mathbb{E}x_{ji}^2\mathbb{E}x_{li}^2\right)^{1/2} < \infty$$

and

$$\mathbb{E}\left|x_{ji}y_i\right| \leq \left(\mathbb{E}x_{ji}^2\mathbb{E}y_i^2\right)^{1/2} < \infty.$$

We have verified the conditions for the WLLN, and thus (5.4) and (5.5).

The final step of the proof is the application of the continuous mapping theorem to obtain (5.6). To fully understand its application we walk through it in detail. We can write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \\ &= \boldsymbol{g}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i', \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i\right)\end{aligned}$$

where $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{b}$ is a function of $\boldsymbol{A}$ and $\boldsymbol{b}$. The function $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is a continuous function of $\boldsymbol{A}$ and $\boldsymbol{b}$ at all values of the arguments such that $\boldsymbol{A}^{-1}$ exists. Assumption 5.1.1 implies that $\boldsymbol{Q}^{-1}$ exists and thus $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is continuous at $\boldsymbol{A} = \boldsymbol{Q}$. Hence by the continuous mapping theorem (Theorem C.3.1), as $n \to \infty$,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{g}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i', \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \\ &\xrightarrow{p} \boldsymbol{g}\left(\boldsymbol{Q}, \mathbb{E}\left(\boldsymbol{x}_i y_i\right)\right) \\ &= \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right) \\ &= \boldsymbol{\beta}.\end{aligned}$$

This completes the proof of Theorem 5.3.1.

5.4 Asymptotic Normality

We started this Chapter discussing the need for an approximation to the distribution of the OLS estimator $\hat{\boldsymbol{\beta}}$. In Section 5.3 we showed that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Consistency is a useful first step, but in itself does not provide a useful approximation to the distribution of the estimator. In this Section we derive an approximation typically called the **asymptotic distribution**.

The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied. The steps are as follows.

Take equation (5.7) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i e_i\right). \tag{5.9}$$

This shows that the normalized and centered estimator $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is a function of the sample average $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$ and the normalized sample average $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{x}_i e_i$. Furthermore, the latter has mean zero so the central limit theorem (CLT) applies.

Central Limit Theorem (Theorem C.2.1)
If $\boldsymbol{u}_i \in \mathbb{R}^k$ is iid, $\mathbb{E}\boldsymbol{u}_i = \mathbf{0}$ and $\mathbb{E}u_{ji}^2 < \infty$ for $j = 1, ..., k$, then as $n \to \infty$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{u}_i \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbb{E}\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right)\right).$$

For our application, $\boldsymbol{u}_i = \boldsymbol{x}_i e_i$ which is iid (since the observations are iid) and mean zero (since $\mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \mathbf{0}$). We calculate that $\mathbb{E}\left(\boldsymbol{u}_i \boldsymbol{u}_i'\right) = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right)$. By the CLT we conclude

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i e_i \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right) \tag{5.10}$$

as $n \to \infty$, where

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right). \tag{5.11}$$

Putting these steps together, using (5.4), (5.9), and (5.10),

$$\begin{aligned} \sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &\xrightarrow{d} \boldsymbol{Q}^{-1} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right) \\ &= \mathrm{N}\left(\mathbf{0}, \boldsymbol{Q}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}^{-1}\right) \end{aligned}$$

as $n \to \infty$, where the final equality follows from the property that linear combinations of normal vectors are also normal (Theorem B.9.1).

Formally, (5.10) requires that the elements of $\boldsymbol{u}_i = \boldsymbol{x}_i e_i$ have finite variances. Indeed, if this is not true then (5.11) is not well defined and (5.10) does not make sense. A sufficient condition can be found as follows. For any $j = 1, ..., k$, by the Cauchy-Schwarz Inequality (C.3), note that

$$\mathbb{E}\left|x_{ji} e_i\right|^2 = \mathbb{E}\left|x_{ji}^2 e_i^2\right| \leq \left(\mathbb{E}x_{ji}^4\right)^{1/2} \left(\mathbb{E}e_i^4\right)^{1/2} \tag{5.12}$$

which is finite if x_{ji} and e_i have finite fourth moments. As e_i is a linear combination of y_i and $\boldsymbol{x}_i$, it is sufficient that the observables have finite fourth moments.

Assumption 5.4.1 *In addition to Assumption 5.1.1,* $\mathbb{E}y_i^4 < \infty$ *and for* $j = 1, ..., k$, $\mathbb{E}x_{ji}^4 < \infty$.

We have derived the asymptotic normal approximation to the distribution of the least-squares estimator.

Theorem 5.4.1 ***Asymptotic Normality of Least-Squares Estimator***
Under Assumption 5.4.1, as $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}^{-1}.$$

As $\boldsymbol{V}_{\boldsymbol{\beta}}$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$, $\boldsymbol{V}_{\boldsymbol{\beta}}$ is often referred to as the **asymptotic covariance matrix** of $\hat{\boldsymbol{\beta}}$. The expression $\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$ is called a **sandwich** form.

Theorem 5.4.1 states that the sampling distribution of the least-squares estimator, after rescaling, is approximately normal when the sample size n is sufficiently large. This holds true for all joint distributions of $(y_i, \boldsymbol{x}_i)$ which satisfy the conditions of Assumption 5.4.1. However, for any fixed n the sampling distribution of $\hat{\boldsymbol{\beta}}$ can be arbitrarily far from the normal distribution. In Figure 4.1 we have already seen a simple example where the least-squares estimate is quite asymmetric and non-normal even for reasonably large sample sizes.

There is a special case where $\boldsymbol{\Omega}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$ simplify. We say that e_i is a **Homoskedastic Projection Error** when

$$\operatorname{cov}(\boldsymbol{x}_i\boldsymbol{x}_i', e_i^2)=\mathbf{0}. \tag{5.13}$$

Condition (5.13) holds in the homoskedastic linear regression model, but is somewhat broader. Under (5.13) the asymptotic variance formulas simplify as

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)E\left(e_i^2\right)=\boldsymbol{Q}\sigma^2 \tag{5.14}$$

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}=\boldsymbol{Q}^{-1}\sigma^2\equiv\boldsymbol{V}_{\boldsymbol{\beta}}^0 \tag{5.15}$$

In (5.15) we define $\boldsymbol{V}_{\boldsymbol{\beta}}^0=\boldsymbol{Q}^{-1}\sigma^2$ whether (5.13) is true or false. When (5.13) is true then $\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{V}_{\boldsymbol{\beta}}^0$, otherwise $\boldsymbol{V}_{\boldsymbol{\beta}}\neq\boldsymbol{V}_{\boldsymbol{\beta}}^0$. We call $\boldsymbol{V}^0$ the **homoskedastic covariance matrix**.

The asymptotic distribution of Theorem 5.4.1 is commonly used to approximate the finite sample distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. The approximation may be poor when n is small. How large should n be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. The trouble is that no matter how large is the sample size, the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions. We illustrate this problem using a simulation. Let $y_i=\beta_0+\beta_1x_i+e_i$ where x_i is $\mathrm{N}(0,1)$, and e_i is independent of x_i with the Double Pareto density $f(e)=\frac{\alpha}{2}|e|^{-\alpha-1}$, $|e|\geq 1$. If $\alpha>2$ the error e_i has zero mean and variance $\alpha/(\alpha-2)$. As α approaches 2, however, its variance diverges to infinity. In this context the normalized least-squares slope estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_2-\beta_2\right)$ has the $\mathrm{N}(0,1)$ asymptotic distibution for any $\alpha>2$. In Figure 5.1 we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_2-\beta_2\right)$, setting $n=100$ and varying the parameter α. For $\alpha=3.0$ the density is very close to the $\mathrm{N}(0,1)$ density. As α diminishes the density changes significantly, concentrating most of the probability mass around zero.

Vilfredo Pareto

Vilfredo Pareto (1848-1923) of Italy was a major economic theorist, introducing the economic concept of Pareto efficiency. His major econometric contribution was the Pareto (or power law) distribution which is commonly used to model the empirical distribution of wealth.

Another example is shown in Figure 5.2. Here the model is $y_i=\beta+e_i$ where (5.16)

$$e_i=\frac{u_i^k-\mathbb{E}\left(u_i^k\right)}{\left(\mathbb{E}\left(u_i^{2k}\right)-\left(\mathbb{E}\left(u_i^k\right)\right)^2\right)^{1/2}} \tag{5.16}$$

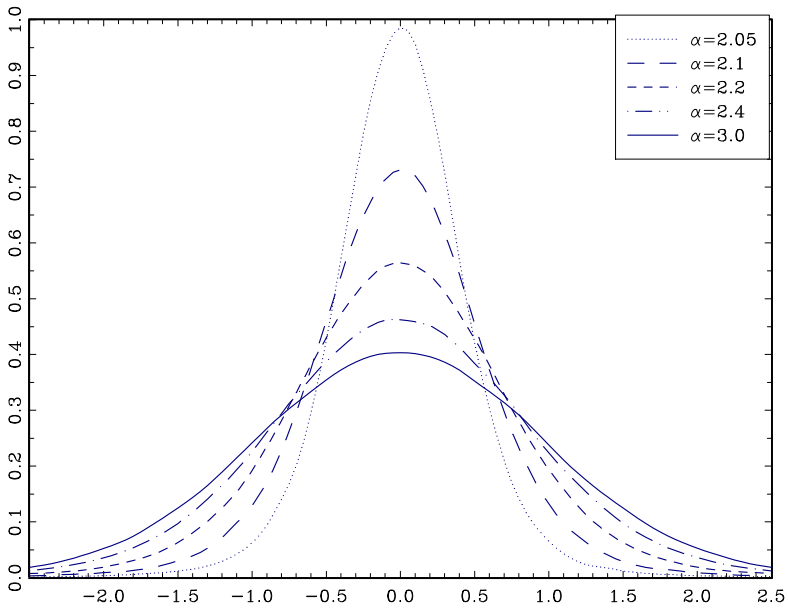

Figure 5.1: Density of Normalized OLS estimator with Double Pareto Error

and $u_i \sim \mathrm{N}(0,1)$. We show the sampling distribution of $\sqrt{n}\left(\hat{\beta}-\beta\right)$ setting $n = 100$, for $k = 1$, 4, 6 and 8. As k increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figures 5.1 and 5.2 is that the $\mathrm{N}(0,1)$ asymptotic approximation is never guaranteed to be accurate.

5.5 Consistency of Sample Variance Estimators

Using the methods of Section 5.3 we can show that the estimators $\hat{\sigma}^2$ and s^2 are consistent for σ^2.

Theorem 5.5.1 *Under Assumption 5.1.1,* $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ *and* $s^2 \xrightarrow{p} \sigma^2$ *as* $n \to \infty$.

One implication of this theorem is that multiple estimators can be consistent for the sample population parameter. While $\hat{\sigma}^2$ and s^2 are unequal in any given application, they are close in value when n is very large.

Proof of Theorem 5.5.1. Note that

$$\begin{aligned}\hat{e}_i &= y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{x}_i'\boldsymbol{\beta} - x_i'\hat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).\end{aligned}$$

Thus

$$\hat{e}_i^2 = e_i^2 - 2e_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \tag{5.17}$$

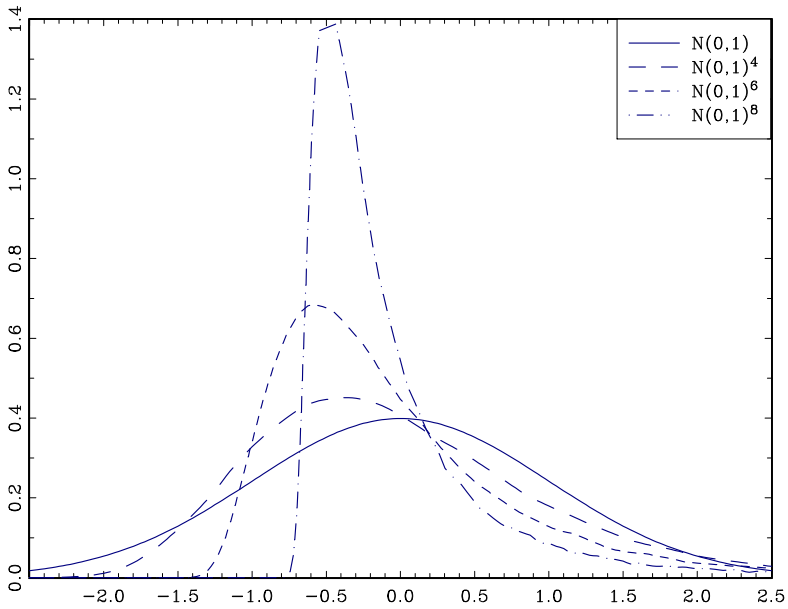

Figure 5.2: Density of Normalized OLS estimator with error process (5.16)

and

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^{n}e_i\boldsymbol{x}_i'\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) + \left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \\ &\xrightarrow{p} \sigma^2\end{aligned}$$

as $n \to \infty$, the last line using the WLLN, (5.4), (5.8) and Theorem 5.3.1. Thus $\hat{\sigma}^2$ is consistent for σ^2.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$, it follows that as $n \to \infty$,

$$s^2 = \left(\frac{n}{n-k}\right)\hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

5.6 Consistent Covariance Matrix Estimation

In Sections 4.8 and 4.9 we introduced estimators of the finite-sample covariance matrix of the least-squares estimator in the regression model. In this section we show that these estimators, when normalized, are consistent for the asymptotic covariance matrix.

First, consider $\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}}$, the covariance matrix estimate constructed under the assumption of homoskedasticity. Writing

$$\hat{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$$

as the moment estimator of $\boldsymbol{Q}$, we can write the covariance matrix estimator as

$$\widehat{\boldsymbol{V}}^0_{\hat{\boldsymbol{\beta}}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2 = \hat{\boldsymbol{Q}}^{-1}s^2.$$

Since $\hat{\boldsymbol{Q}} \xrightarrow{p} \boldsymbol{Q}$ and $s^2 \xrightarrow{p} \sigma^2$ (see (5.4) and Theorem 5.5.1), and the invertibility of $\boldsymbol{Q}$ (Assumption 5.1.1), it follows that

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0 = \hat{\boldsymbol{Q}}^{-1} s^2 \xrightarrow{p} \boldsymbol{Q}^{-1}\sigma^2 = \boldsymbol{V}_{\boldsymbol{\beta}}^0$$

so that $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$ is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}^0$, the homoskedastic covariance matrix.

Theorem 5.6.1 *Under Assumption 5.1.1,* $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0 \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}^0$ *as* $n \to \infty$.

Now consider $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0$, the White covariance matrix estimator. Writing

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 \tag{5.18}$$

as the moment estimator for $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right)$, then

$$\begin{aligned} \widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2\right) \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}^{-1}. \end{aligned}$$

With some work, we can show that $\hat{\boldsymbol{\Omega}}$ is consisent for $\boldsymbol{\Omega}$. Combined with the consistency of $\hat{\boldsymbol{Q}}$ for $\boldsymbol{Q}$ and the invertibility of $\boldsymbol{Q}$ we find that $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ converges in probability to $\boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1} = \boldsymbol{V}_{\boldsymbol{\beta}}$.

Theorem 5.6.2 *Under Assumption 5.4.1,* $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ *and* $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$ *as* $n \to \infty$.

To illustrate, we return to the log wage regression (3.9) of Section 3.3. We calculate that $s^2 = 0.20$ and

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} 0.199 & 2.80 \\ 2.80 & 40.6 \end{pmatrix}.$$

Therefore the two covariance matrix estimates are

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^0 = \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} 0.20 = \begin{pmatrix} 6.98 & -0.480 \\ -0.480 & .039 \end{pmatrix}$$

and

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} \begin{pmatrix} .199 & 2.80 \\ 2.80 & 40.6 \end{pmatrix} \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} = \begin{pmatrix} 7.20 & -0.493 \\ -0.493 & 0.035 \end{pmatrix}.$$

In this case the two estimates are quite similar. The (White) standard errors for $\hat{\beta}_0$ are $\sqrt{7.2/988} = .085$ and that for $\hat{\beta}_1$ is $\sqrt{.035/988} = .006$. We can write the estimated equation with standard errors using the format

$$\begin{array}{rccc} \widehat{\log(Wage)} = & 1.33 & + & 0.115 \quad Education. \\ & (.08) & & (.006) \end{array}$$

Proof of Theorem 5.6.2. We first show $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$. Using (5.17)

$$\begin{aligned}\hat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'e_i^2 - \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i e_i + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i\right)^2. \qquad (5.19)\end{aligned}$$

We now examine each $k \times k$ sum on the right-hand-side of (5.19) in turn.

Take the first term on the right-hand-side of (5.19). The jl'th element of $\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2$ is $x_{ji}x_{li}e_i^2$. Using the Cauchy-Schwarz Inequality (C.3) twice and Assumption 5.4.1,

$$\begin{aligned}\mathbb{E}\left|x_{ji}x_{li}e_i^2\right| &\leq \left(\mathbb{E}x_{ji}^2x_{li}^2\right)^{1/2}\left(\mathbb{E}e_i^4\right)^{1/2} \\ &\leq \left(\mathbb{E}x_{ji}^4\right)^{1/4}\left(\mathbb{E}x_{li}^4\right)^{1/4}\left(\mathbb{E}e_i^4\right)^{1/2}.\end{aligned}$$

Since this expectation is finite, we can apply the WLLN (Theorem 5.2.1) to find that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'e_i^2 \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right) = \boldsymbol{\Omega}.$$

Now take the second term on the right-hand-side of (5.19). Applying the Triangle Inequality (A.8) to the matrix Euclidean norm, the Matrix Schwarz Inequality (A.7), equation (A.5) and the Schwarz Inequality (A.6)

$$\begin{aligned}\left\|\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i e_i\right\| &\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i e_i\right\| \\ &\leq \frac{2}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\|\left|\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i\right||e_i| \\ &\leq \left(\frac{2}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^3|e_i|\right)\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|. \qquad (5.20)\end{aligned}$$

Using Holder's inequality (C.4) and Assumption 5.4.1,

$$\mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^3|e_i|\right) \leq \left(\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4\right)^{3/4}\left(\mathbb{E}\left|e_i^4\right|\right)^{1/4} < \infty.$$

By the WLLN

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^3|e_i| \xrightarrow{p} \mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^3|e_i|\right) < \infty.$$

Since $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta} \xrightarrow{p} 0$ it follows that (5.20) converges in probability to zero. This shows that the second term on the right-hand-side of (5.19) converges in probability to zero.

We now take the third term in (5.19). Again by the Triangle Inequality, the Matrix Schwarz Inequality, (A.5) and the Schwarz Inequality

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i\right)^2\right\| &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\|\left(\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{x}_i\right)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^4\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| \\ &\xrightarrow{p} 0\end{aligned}$$

the final convergence since $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} 0$ and $\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^4 \xrightarrow{p} E\|\boldsymbol{x}_i\|^4 < \infty$ under Assumption 5.4.1. This shows that the third term on the right-hand-side of (5.19) converges in probability to zero.

Considering the three terms on the right-hand-side of (5.19), we have shown that the first term converges in probability to $\boldsymbol{\Omega}$, and the second and third converge in probability to zero. We conclude that $\hat{\boldsymbol{\Omega}} \xrightarrow{\boldsymbol{p}} \boldsymbol{\Omega}$ as claimed.

Finally, combined with (5.4) and the invertibilility of $\boldsymbol{Q}$,

$$\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{Q}}^{-1} \xrightarrow{\boldsymbol{p}} \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1} = \boldsymbol{V}_{\boldsymbol{\beta}},$$

from which it follows that $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$ as $n \to \infty$.

5.7 Functions of Parameters

Sometimes we are interested in some lower-dimensional function of the parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$. For example, we may be interested in a single coefficient β_j or a ratio β_j/β_l. In these cases we can write the parameter of interest as a function of $\boldsymbol{\beta}$. Let $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^q$ denote this function and let

$$\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta})$$

denote the parameter of interest. The estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \boldsymbol{h}(\hat{\boldsymbol{\beta}}).$$

What is the asymptotic distribution of $\hat{\boldsymbol{\theta}}$? Assume that $\boldsymbol{h}(\boldsymbol{\beta})$ is differentiable at the true value of $\boldsymbol{\beta}$. By a first-order Taylor series approximation:

$$\boldsymbol{h}(\hat{\boldsymbol{\beta}}) \simeq \boldsymbol{h}(\boldsymbol{\beta}) + \boldsymbol{H}'_{\boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

where

$$\boldsymbol{H}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{h}(\boldsymbol{\beta}) \qquad k \times q.$$

Thus

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &= \sqrt{n}\left(\boldsymbol{h}(\hat{\boldsymbol{\beta}}) - \boldsymbol{h}(\boldsymbol{\beta})\right) \\
&\simeq \boldsymbol{H}'_{\boldsymbol{\beta}}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{d} \boldsymbol{H}'_{\boldsymbol{\beta}}\,\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) \\
&= \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right). \qquad (5.21)
\end{aligned}$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{H}'_{\boldsymbol{\beta}}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{H}_{\boldsymbol{\beta}}. \qquad (5.22)$$

The asymptotic approximation (5.21) is often called **the delta method** because it approximates the distribution of $\hat{\boldsymbol{\theta}}$ by a first-order expansion. It shows that (at least approximately), nonlinear functions of asymptotically normal estimators are themselves asymptotically normally distributed. It is a very powerful result, as most parameters of interest can be written in this form.

In many cases, the function $\boldsymbol{h}(\boldsymbol{\beta})$ is linear:

$$\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$$

for some $k \times q$ matrix $\boldsymbol{R}$. In this case, $\boldsymbol{H}_{\boldsymbol{\beta}} = \boldsymbol{R}$.

In particular, if $\boldsymbol{R}$ is a "selector matrix"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} \tag{5.23}$$

so that if $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, then $\boldsymbol{\theta} = \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\beta}_1$ and

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{V}_{11},$$

the upper-left block of $\boldsymbol{V}_{\boldsymbol{\beta}}$. In other words, (5.21)-(5.22) in this case is

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{11}\right)$$

where

$$\boldsymbol{V}_{11} = [\boldsymbol{V}]_{11}$$

How do we estimate the covariance matrix for $\hat{\boldsymbol{\theta}}$? From (5.22) we see we need an estimate of $\boldsymbol{H}_{\boldsymbol{\beta}}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$. We already have an estimate of the latter, $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$. To estimate $\boldsymbol{H}_{\boldsymbol{\beta}}$ we use

$$\widehat{\boldsymbol{H}}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{h}(\hat{\boldsymbol{\beta}}).$$

Putting the parts together we obtain

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{H}}_{\boldsymbol{\beta}}' \widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{H}}_{\boldsymbol{\beta}}$$

as the covariance matrix estimator for $\hat{\boldsymbol{\theta}}$. As the primary justification for $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$ is the asymptotic approximation (5.21), $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$ is often called an asymptotic covariance matrix estimator.

When $\boldsymbol{h}(\boldsymbol{\beta})$ is linear

$$\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$$

then $\boldsymbol{H}_{\boldsymbol{\beta}} = \boldsymbol{R}$ and

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{R}' \widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \boldsymbol{R}.$$

When $\boldsymbol{R}$ takes the form of a selector matrix as in (5.23) then

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{V}}_{11} = \left[\widehat{\boldsymbol{V}}\right]_{11},$$

the upper-left block of the covariance matrix estimate $\widehat{\boldsymbol{V}}$.

When $q = 1$ (so $\boldsymbol{h}(\boldsymbol{\beta})$ is real-valued), the standard error for $\hat{\theta}$ is the square root of $\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}}$, that is,

$$s(\hat{\theta}) = n^{-1/2}\sqrt{\widehat{\boldsymbol{V}}_{\hat{\theta}}} = n^{-1/2}\sqrt{\widehat{\boldsymbol{H}}_{\boldsymbol{\beta}}' \widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{H}}_{\boldsymbol{\beta}}}$$

.

Theorem 5.7.1 ***Asymptotic Distribution of Functions of Parameters***
Under Assumption 5.4.1,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{H}_{\boldsymbol{\beta}}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{H}_{\boldsymbol{\beta}}$$

and

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}$$

as $n \to \infty$.

Proof. We showed (5.21), we need only to show consistency of the covariance matrix estimator. First, from Theorem 5.6.2,

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}.$$

Second, since $\hat{\boldsymbol{\beta}} \xrightarrow{p} \hat{\boldsymbol{\beta}}$ and $\boldsymbol{h}(\boldsymbol{\beta})$ is continuously differentiable, by the continuous mapping theorem,

$$\widehat{\boldsymbol{H}}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{h}(\hat{\boldsymbol{\beta}}) \xrightarrow{p} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{H}_{\boldsymbol{\beta}}.$$

Putting these together

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{H}}_{\boldsymbol{\beta}}' n \widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \widehat{\boldsymbol{H}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{H}_{\boldsymbol{\beta}}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{H}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\theta}},$$

completing the proof .

5.8 t statistic

Let $\theta = h(\boldsymbol{\beta}) : \mathbb{R}^k \to \mathbb{R}$ be any parameter of interest (for example, θ could be a single element of $\boldsymbol{\beta}$), $\hat{\theta}$ its estimate and $s(\hat{\theta})$ its asymptotic standard error. Consider the statistic

$$t_n(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \tag{5.24}$$

which different writers alternatively call a **t-statistic**, a **z-statistic** or a **studentized statistic**. We won't be making such distinctions and will typically refer to $t_n(\theta)$ as a t-statistic. We also often suppress the parameter dependence, writing it as t_n. The t-statistic is a simple function of the estimate, its standard error, and the parameter.

Theorem 5.8.1 $t_n(\theta) \xrightarrow{d} \mathrm{N}(0, 1)$

Thus the asymptotic distribution of the t-ratio $t_n(\theta)$ is the standard normal. Since this distribution does not depend on the parameters, we say that $t_n(\theta)$ is **asymptotically pivotal**. In special cases (such as the normal regression model, see Section 3.11), the statistic t_n has an exact t distribution, and is therefore exactly free of unknowns. In this case, we say that t_n is **exactly pivotal**. In general, however, pivotal statistics are unavailable and we must rely on asymptotically pivotal statistics.

William Gosset

William S. Gosset (1876-1937) of England is most famous for his derivation of the student's t distribution, published in the paper "The probable error of a mean" in 1908. At the time, Gosset worked at Guiness brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the student's t rather than Gosset's t!

Proof of Theorem 5.8.1. By Theorem 5.7.1,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

and

$$\widehat{\boldsymbol{V}}_{\hat{\boldsymbol{\theta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}$$

Thus

$$\begin{aligned} t_n(\theta) &= \frac{\hat{\theta}-\theta}{s(\hat{\theta})} \\ &= \frac{\sqrt{n}\left(\hat{\theta}-\theta\right)}{\sqrt{\widehat{V}_{\hat{\theta}}}} \\ &\xrightarrow{d} \frac{\mathrm{N}\left(0, V_\theta\right)}{\sqrt{V_\theta}} \\ &= \mathrm{N}(0,1) \end{aligned}$$

The last equality is by the property that linear scales of normal distributions are normal.

5.9 Confidence Intervals

A confidence interval C_n is an interval estimate of $\theta \in \mathbb{R}$. It is a function of the data and hence is random. It is designed to cover θ with high probability. Either $\theta \in C_n$ or $\theta \notin C_n$. The coverage probability is $\mathbb{P}(\theta \in C_n)$. The convention is to design confidence intervals to have coverage probability approximately equal to a pre-specified target, typically 90% or 95%, or more generally written as $(1-\alpha)\%$ for some $\alpha \in (0,1)$. In this case, by reporting a $(1-\alpha)\%$ confidence interval C_n, we are stating that with $(1-\alpha)\%$ probability (in repeated samples) the true θ lies in C_n.

There is not a unique method to construct confidence intervals. For example, a simple (yet silly) interval is

$$C_n = \begin{cases} \mathbb{R} & \text{with probability } 1-\alpha \\ \hat{\theta} & \text{with probability } \alpha \end{cases}$$

By construction, if $\hat{\theta}$ has a continuous distribution, $\mathbb{P}(\theta \in C_n) = 1-\alpha$, so this confidence interval has perfect coverage, but C_n is uninformative about θ. This is not a useful confidence interval.

When we have an asymptotically normal parameter estimate $\hat{\theta}$ with standard error $s(\hat{\theta})$, it turns out that a generally reasonable confidence interval for θ takes the form

$$C_n = \left[\hat{\theta} - c \cdot s(\hat{\theta}), \quad \hat{\theta} + c \cdot s(\hat{\theta})\right] \tag{5.25}$$

where $c > 0$ is a pre-specified constant. This confidence interval is symmetric about the point estimate $\hat{\theta}$, and its length is proportional to the standard error $s(\hat{\theta})$.

Equivalently, C_n is the set of parameter values for θ such that the t-statistic $t_n(\theta)$ is smaller (in absolute value) than c, that is

$$C_n = \{\theta : |t_n(\theta)| \le c\} = \left\{\theta : -c \le \frac{\hat{\theta}-\theta}{s(\hat{\theta})} \le c\right\}.$$

The coverage probability of this confidence interval is

$$\mathbb{P}\left(\theta \in C_n\right) = \mathbb{P}\left(|t_n(\theta)| \le c\right)$$

which is generally unknown, but we can approximate the coverage probability by taking the asymptotic limit as $n \to \infty$. Since $t_n(\theta)$ is asymptotically standard normal (Theorem 5.8.1), it follows that as $n \to \infty$ that

$$\mathbb{P}(\theta \in C_n) \to \mathbb{P}(|Z| \leq c) = \Phi(c) - \Phi(-c)$$

where $Z \sim \mathrm{N}(0,1)$ and $\Phi(u) = \mathbb{P}(Z \leq u)$ is the standard normal distribution function. Thus the asymptotic coverage probability is a function only of c.

The convention is to design the confidence interval to have a pre-specified coverage probability $1-\alpha$, typically 90% or 95%. This means selecting the constant c so that

$$\Phi(c) - \Phi(-c) = 1 - \alpha.$$

Effectively, this makes c a function of α, and can be backed out of a normal distribution table. For example, $\alpha = 0.05$ (a 95% interval) implies $c = 1.96$ and $\alpha = 0.1$ (a 90% interval) implies $c = 1.645$. Rounding 1.96 to 2, this yields the most commonly implied confidence interval in applied econometric practice

$$C_n = \left[\hat{\theta} - 2s(\hat{\theta}), \quad \hat{\theta} + 2s(\hat{\theta})\right].$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval C_n is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is a 95.4% interval, due to the substitution of 2.0 for 1.96, but this distinction is meaningless.)

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results, look at the estimated coefficient estimates and the standard errors. For a parameter of interest, compute the confidence interval C_n and consider the meaning of the spread of the suggested values. If the rage of values in the confidence interval are too wide to learn about θ, then do not jump to a conclusion about θ based on the point estimate alone.

5.10 Semiparametric Efficiency

In Section 4.5 we presented the Gauss-Markov theorem as a limited efficiency justification for the least-squares estimator. A broader justification is provided in Chamberlain (1987), who established that in the projection model the OLS estimator has the smallest asymptotic mean-squared error among feasible estimators. This property is called **semiparametric efficiency**, and is a strong justification for the least-squares estimator. We discuss the intuition behind his result in this section.

Suppose that the joint distribution of $(y_i, \boldsymbol{x}_i)$ is discrete. That is, for finite r,

$$P\left(y_i = \tau_j, \quad \boldsymbol{x}_i = \boldsymbol{\xi}_j\right) = p_j, \qquad j = 1, ..., r \tag{5.26}$$

for some constants p_j, τ_j, and $\boldsymbol{\xi}_j$. Assume that the τ_j and $\boldsymbol{\xi}_j$ are known, but the p_j are unknown. (We know the values y_i and $\boldsymbol{x}_i$ can take, but we don't know the probabilities.)

In this discrete setting, the definition of linear the projection coefficient (2.10) can be rewritten as

$$\boldsymbol{\beta} = \left(\sum_{j=1}^{r} p_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'\right)^{-1} \left(\sum_{j=1}^{r} p_j \boldsymbol{\xi}_j \tau_j\right) \tag{5.27}$$

Thus $\boldsymbol{\beta}$ is a function of $(\pi_1, ..., \pi_r)$.

As the data are multinomial, the maximum likelihood estimator (MLE) is

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} 1(y_i = \tau_j)\, 1\left(\boldsymbol{x}_i = \boldsymbol{\xi}_j\right)$$

for $j = 1, ..., r$, where $1(\cdot)$ is the indicator function. That is, $\hat{p}_j$ is the percentage of the observations which fall in each category. The MLE $\widehat{\boldsymbol{\beta}}_{\text{mle}}$ for $\boldsymbol{\beta}$ is then the analog of (5.27) with the parameters p_j replaced by the estimates $\hat{p}_j$:

$$\widehat{\boldsymbol{\beta}}_{\text{mle}} = \left(\sum_{j=1}^{r} \hat{p}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j'\right)^{-1} \left(\sum_{j=1}^{r} \hat{p}_j \boldsymbol{\xi}_j \tau_j\right).$$

Substituting in the expressions for $\hat{p}_j$,

$$\begin{aligned}
\sum_{j=1}^{r} \hat{p}_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j' &= \sum_{j=1}^{r} \frac{1}{n} \sum_{i=1}^{n} 1(y_i = \tau_j) 1\left(\boldsymbol{x}_i = \boldsymbol{\xi}_j\right) \boldsymbol{\xi}_j \boldsymbol{\xi}_j' \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} 1(y_i = \tau_j) 1\left(\boldsymbol{x}_i = \boldsymbol{\xi}_j\right) \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'
\end{aligned}$$

and

$$\begin{aligned}
\sum_{j=1}^{r} \hat{p}_j \boldsymbol{\xi}_j \tau_j &= \sum_{j=1}^{r} \frac{1}{n} \sum_{i=1}^{n} 1(y_i = \tau_j) 1\left(\boldsymbol{x}_i = \boldsymbol{\xi}_j\right) \boldsymbol{\xi}_j \tau_j \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} 1(y_i = \tau_j) 1\left(\boldsymbol{x}_i = \boldsymbol{\xi}_j\right) \boldsymbol{x}_i y_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i y_i.
\end{aligned}$$

Thus

$$\widehat{\boldsymbol{\beta}}_{\text{mle}} = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) = \widehat{\boldsymbol{\beta}}_{\text{ols}}.$$

In other words, if the data have a discrete distribution, the maximum likelihood estimator is identical to the OLS estimator.

Since this is a regular parametric model the MLE is asymptotically efficient (see Appendix D). It follows that the OLS estimator is asymptotically efficient.

The hard part of the argument (which was rigorously developed in Chamberlain's paper, but we do not present it here) is the extension to the case of continuously-distributed data.The intuition is that all continuous distributions can be arbitrarily well approximated by some multinomial distribution, and for any multinomial distribution the moment estimator is asymptotically efficient. Formalizing this intuition using a rigorous mathematical argument, Chamberlain proved that the OLS estimator is asymptotically semiparametrically efficient for the projection coefficient $\boldsymbol{\beta}$ for the class of models satisfying Assumption 5.1.1.

5.11 Semiparametric Efficiency in the Projection Model

In this section we continue the investigation of semiparametric efficiency as raised in Section 5.10. There we presented the intuition behind Chamberlain's demonstration of the asymptotic efficiency of the least-squares estimator. In this section we provide an alternative demonstration based on the rich but technically challenging theory of semiparametric efficiency bounds. An excellent accessible review has been provided by Newey (1990).

Our treatment covers what is known as the smooth function model, which includes the projection model as a special case. Let $\boldsymbol{z} \in \mathbb{R}^m$ be a random vector with finite mean $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{z}$ and finite variance matrix $\boldsymbol{\Sigma} = \operatorname{var}(\boldsymbol{z})$, and let $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ be an iid sample from this distribution. The parameter of interest is $\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\mu})$ where $\mathbf{g}(\cdot)$ is a continuously differentiable function. The standard moment estimator for $\boldsymbol{\mu}$ is the sample mean $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i$ and that for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$. This setting includes the least-squares estimator for the projection model $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ by letting $\boldsymbol{z}$ be the vector with elements $x_j e$ and $x_j x_l$ for all $j \leq k$ and $l \leq k$.

The sample mean has the asymptotic distribution $\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{\Sigma})$. Applying the Delta Method (Theorem C.3.3), we see that the moment estimator $\hat{\boldsymbol{\beta}}$ has the asymptotic distribution $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{V})$ where $\boldsymbol{V} = \frac{\partial}{\partial \boldsymbol{\mu}'}\mathbf{g}(\boldsymbol{\mu})\,\boldsymbol{\Sigma}\,\frac{\partial}{\partial \boldsymbol{\mu}}\mathbf{g}(\boldsymbol{\mu})'$. We want to know if $\hat{\boldsymbol{\beta}}$ is the best feasible estimator. Is there another estimator with a smaller asymptotic variance? While it seems intuitively unlikely that another estimator could have a smaller asymptotic variance than $\hat{\boldsymbol{\beta}}$, how do we know that this is not the case?

To show that the answer is not immediately obvious, it might be helpful to review a setting where the sample mean is inefficient. Suppose that $z \in \mathbb{R}$ has the density $f(z \mid \mu) = 2^{-1/2}\exp\left(-|z - \mu|\sqrt{2}\right)$. Since $\operatorname{var}(z) = 1$ we see that the sample mean satisfies $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, 1)$. In this model the maximum likelihood estimator (MLE) $\tilde{\mu}$ for μ is different than the sample mean (and happens to be the sample median). Recall from the theory of maximum likelihood that the MLE satisfies $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, \mathcal{I}_0)$ where $\mathcal{I}_0 = \left(\mathbb{E}S_\mu^2\right)^{-1}$ and $S_\mu = \frac{\partial}{\partial \mu}\log f(z \mid \mu) = -\sqrt{2}\operatorname{sgn}(z - \mu)$ is the score. We can calculate that $\mathbb{E}S_\mu^2 = 2$ and thus conclude that $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, 1/2)$. The asymptotic variance of the MLE is one-half that of the sample mean. In this setting the sample mean is inefficient.

But the question at hand is whether or not the sample mean is efficient when the form of the distribution is unknown. We call this setting **semiparametric** as the parameter of interest (the mean) is finite dimensional while the remaining features of the distribution are unspecified. In the semiparametric context an estimator is called **semiparametrically efficient** if it has the smallest asymptotic variance among all semiparametric estimators.

The mathematical trick is to reduce the semiparametric model to a set of parametric "submodels". The classic Cramer-Rao variance bound can be found for each parametric submodel. The variance bound for the semiparametric model (the union of the submodels) is then defined as the supremum of the individual variance bounds.

Formally, suppose that the true density of $\boldsymbol{z}$ is the unknown function $f(\boldsymbol{z})$ with mean $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{z} = \int \boldsymbol{z} f(\boldsymbol{z}) d\boldsymbol{z}$ and the parameter of interest is $\boldsymbol{\beta} = \mathbf{g}(\boldsymbol{\mu})$. A parametric submodel η for $f(\boldsymbol{z})$ is a density $f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta})$ which is a smooth function of a parameter $\boldsymbol{\theta}$, and there is some $\boldsymbol{\theta}_0$ such that $f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{z})$. The index η indicates the submodels. The equality $f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{z})$ means that the submodel class passes through the true density, so the submodel is a true model. The class of submodels η and parameter $\boldsymbol{\theta}_0$ depend on the true density f. In the submodel $f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta})$, the mean is $\boldsymbol{\mu}_\eta(\boldsymbol{\theta}) = \int \boldsymbol{z} f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta})\, d\boldsymbol{z}$, and the parameter of interest is $\boldsymbol{\beta}_\eta(\boldsymbol{\theta}) = \mathbf{g}\left(\boldsymbol{\mu}_\eta(\boldsymbol{\theta})\right)$ which varies with the parameter $\boldsymbol{\theta}$. Let $\eta \in \aleph$ be the class of all submodels for f.

Since each submodel η is parametric we can calculate its Cramer-Rao bound for estimation of $\boldsymbol{\beta}$. Specifically, given the density $f_\eta(\boldsymbol{z} \mid \boldsymbol{\theta})$ we can construct the MLE $\hat{\boldsymbol{\theta}}_\eta$ for $\boldsymbol{\theta}$, the MLE $\hat{\boldsymbol{\mu}}_\eta = \int \boldsymbol{z} f_\eta\left(\boldsymbol{z} \mid \hat{\boldsymbol{\theta}}_\eta\right) d\boldsymbol{z}$ for $\boldsymbol{\mu}$, and the MLE $\hat{\boldsymbol{\beta}}_\eta = \mathbf{g}(\hat{\boldsymbol{\mu}}_\eta)$ for $\boldsymbol{\beta}$. The MLE satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\eta - \boldsymbol{\beta}_\eta(\boldsymbol{\theta})\right) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{V}_\eta)$$

where $\boldsymbol{V}_\eta$ is the smallest possible covariance matrix among regular estimators. By the Cramer-Rao theorem no estimator (and in particular no semiparametric estimator) has an asymptotic variance smaller than $\boldsymbol{V}_\eta$. This comparison is true for all submodels η, so the asymptotic variance of any semiparametric estimator cannot be smaller than the Cramer-Rao bound for any parametric submodel. The **semiparametric asymptotic variance bound** (which is sometimes called

the **semiparametric efficiency bound**) is the supremum of the Cramer-Rao bounds from all conceivable submodels.

$$\overline{\boldsymbol{V}} = \sup_{\eta \in \aleph} \boldsymbol{V}_\eta.$$

It is a lower bound for the asymptotic variance of any semiparametric estimator. If the asymptotic variance of a specific semiparametric estimator equals the bound $\overline{\boldsymbol{V}}$ we say that the estimator is **semiparametrically efficient**.

For many statistical problems it is quite challenging to calculate the semiparametric variance bound. However the solution is straightforward in the smooth function model. As the semiparametric variance bound cannot be smaller than the Cramer-Rao bound for any submodel, and cannot be larger than the asymptotic variance of any feasible semiparametric estimator, it follows that if the asymptotic variance of a feasible semiparametric estimator equals the Cramer-Rao bound for at least one submodel, then this is the semiparametric asymptotic variance bound, and the aforementioned feasible semiparametric estimator must be semiparametrically efficient. In these cases, it is sufficient to construct a parametric submodel for which the Cramer-Rao bound (equivalently, the asymptotic variance of the MLE) equals that of a known semiparametric estimator.

Formally, for any submodel η with Cramer-Rao variance $\boldsymbol{V}_\eta$ and any semiparametric estimator $\hat{\boldsymbol{\beta}}$ with asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$, then it is necessary that

$$\boldsymbol{V}_\eta \leq \overline{\boldsymbol{V}} \leq \boldsymbol{V}_{\boldsymbol{\beta}}.$$

The first inequality holds by the definition of $\overline{\boldsymbol{V}}$, and the second holds since no semiparametric estimator can be more efficient than the MLE in any parametric submodel. Thus if we find a submodel η and semiparametric estimator $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{V}_\eta = \boldsymbol{V}_{\boldsymbol{\beta}}$, then it must be the case that $\overline{\boldsymbol{V}} = \boldsymbol{V}_{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ is semiparametrically efficient.

We now show this for the moment estimator $\hat{\boldsymbol{\beta}} = \mathbf{g}(\hat{\boldsymbol{\mu}})$ discussed above. As $\hat{\boldsymbol{\beta}}$ has asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$, our goal is to find a parametric submodel whose Cramer-Rao bound for estimation of $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}}$. The solution involves creating a tilted version of the true density. Consider the parametric submodel

$$f(\boldsymbol{z} \mid \boldsymbol{\theta}) = f(\boldsymbol{z}) \left(1 + \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right) \tag{5.28}$$

where $f(\boldsymbol{z})$ is the true density and $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{z}$. Note that

$$\int f(\boldsymbol{z} \mid \boldsymbol{\theta})\, d\boldsymbol{z} = \int f(\boldsymbol{z}) d\boldsymbol{z} + \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \int f(\boldsymbol{z}) (\boldsymbol{z} - \boldsymbol{\mu})\, d\mathbf{z} = 1$$

and for all $\boldsymbol{\theta}$ close to zero $f(\boldsymbol{z} \mid \boldsymbol{\theta}) \geq 0$. Thus $f(\boldsymbol{z} \mid \boldsymbol{\theta})$ is a valid density function. It is a parametric submodel since $f(\boldsymbol{z} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{z})$ when $\boldsymbol{\theta}_0 = 0$. This parametric submodel has the mean

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \int \boldsymbol{z} f(\boldsymbol{z} \mid \boldsymbol{\theta})\, d\boldsymbol{z} \\ &= \int \boldsymbol{z} f(\boldsymbol{z}) d\boldsymbol{z} + \int f(\boldsymbol{z}) \boldsymbol{z} (\boldsymbol{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} d\boldsymbol{z} \\ &= \boldsymbol{\mu} + \boldsymbol{\theta} \end{aligned}$$

and parameter of interest $\boldsymbol{\beta}(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\mu} + \boldsymbol{\theta})$ both which are smooth functions of $\boldsymbol{\theta}$.

Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{z} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \left(1 + \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})\right) = \frac{\boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})}{1 + \boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})}$$

it follows that the score function for $\boldsymbol{\theta}$ is

$$\boldsymbol{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{z} \mid \boldsymbol{\theta}_0) = \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}). \tag{5.29}$$

By classic theory the asymptotic variance of the MLE $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is the Cramer-Rao bound $\left(\mathbb{E}\left(\boldsymbol{s}\boldsymbol{s}'\right)\right)^{-1} = \left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left(\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\left(\boldsymbol{z}-\boldsymbol{\mu}\right)'\right)\boldsymbol{\Sigma}^{-1}\right)^{-1} = \boldsymbol{\Sigma}$. The MLE for $\boldsymbol{\beta}$ is $\boldsymbol{\beta}(\hat{\boldsymbol{\theta}}) = \mathbf{g}\left(\boldsymbol{\mu}+\hat{\boldsymbol{\theta}}\right)$ which by the delta method has asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\mu}'}\mathbf{g}\left(\boldsymbol{\mu}\right)\boldsymbol{\Sigma}\frac{\partial}{\partial \boldsymbol{\mu}}\mathbf{g}\left(\boldsymbol{\mu}\right)'$, which is identical to the asymptotic variance of the moment estimator $\hat{\boldsymbol{\beta}}$. This shows that moment estimators are semiparametrically efficient, and this includes the OLS estimator in the projection model. We have established the following theorem.

Theorem 5.11.1 *Under Assumption 5.1.1, the semiparametric variance bound for estimation of $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$, and the OLS estimator is semiparametrically efficient.*

5.12 Semiparametric Efficiency in the Homoskedastic Regression Model

In Section 4.5 we presented the Gauss-Markov theorem, which stated that in the homoskedastic regression model, in the class of linear unbiased estimators the one with the smallest variance is least-squares. As we noted in that section, the restriction to linear unbiased estimators is unsatisfactory as it leaves open the possibility that an alternative (non-linear) estimator could have a smaller asymptotic variance. In Sections 5.10 and 5.11 we showed that the OLS estimator is efficient in the projection model, but this does not address the question of whether or not OLS is efficient in the homoskedastic regression model. In this section we return to the question of efficient estimation in this model using the theory of semiparametric variance bounds as presented in the previous section.

Recall that in the homoskedastic regression model the asymptotic variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}}^{0} = \boldsymbol{Q}^{-1}\sigma^2$. Therefore, as described in the previous section, it is sufficient to find a parametric submodel whose Cramer-Rao bound for estimation of $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}}^{0}$. This would establish that $\boldsymbol{V}_{\boldsymbol{\beta}}^{0}$ is the semiparametric variance bound and the OLS estimator $\hat{\boldsymbol{\beta}}$ is semiparametrically efficient for $\boldsymbol{\beta}$.

Let the joint density of y and $\boldsymbol{x}$ be written as $f\left(y,\boldsymbol{x}\right) = f_1\left(y \mid \boldsymbol{x}\right) f_2\left(\boldsymbol{x}\right)$, the product of the conditional density of y given $\boldsymbol{x}$, and the marginal density of $\boldsymbol{x}$. Now consider the parametric submodel

$$f\left(y,\boldsymbol{x} \mid \boldsymbol{\theta}\right) = f_1\left(y \mid \boldsymbol{x}\right)\left(1+\left(y-\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) f_2\left(\boldsymbol{x}\right). \tag{5.30}$$

You can check that in this submodel, the marginal density of $\boldsymbol{x}$ is $f_2\left(\boldsymbol{x}\right)$, and the conditional density of y given $\boldsymbol{x}$ is $f_1\left(y \mid \boldsymbol{x}\right)\left(1+\left(y-\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right)$. To see that the latter is a valid conditional density, observe that the regression assumption implies that $\int y f_1\left(y \mid \boldsymbol{x}\right) dy = \boldsymbol{x}'\boldsymbol{\beta}$ and therefore

$$\begin{aligned}\int f_1\left(y \mid \boldsymbol{x}\right)\left(1+\left(y-\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) dy &= \int f_1\left(y \mid \boldsymbol{x}\right) dy + \int f_1\left(y \mid \boldsymbol{x}\right)\left(y-\boldsymbol{x}'\boldsymbol{\beta}\right) dy\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 \\ &= 1.\end{aligned}$$

In this parametric submodel the conditional mean of y given $\boldsymbol{x}$ is

$$\begin{aligned}
\mathbb{E}_\theta(y \mid \boldsymbol{x}) &= \int y f_1(y \mid \boldsymbol{x}) \left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) dy \\
&= \int y f_1(y \mid \boldsymbol{x})\, dy + \int y f_1(y \mid \boldsymbol{x}) \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 dy \\
&= \int y f_1(y \mid \boldsymbol{x})\, dy + \int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_1(y \mid \boldsymbol{x}) \left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 dy \\
&\quad + \int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right) f_1(y \mid \boldsymbol{x})\, dy \left(\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 \\
&= \boldsymbol{x}'(\boldsymbol{\beta} + \boldsymbol{\theta}),
\end{aligned}$$

using the homoskedasticity assumption that $\int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_1(y \mid \boldsymbol{x})\, dy = \sigma^2$. This means that in this parametric submodel, the conditional mean is linear in $\boldsymbol{x}$ and the regression coefficient is $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\beta} + \boldsymbol{\theta}$.

We now calculate the score for estimation of $\boldsymbol{\theta}$. Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) = \frac{\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)/\sigma^2}{1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2}$$

the score is

$$\boldsymbol{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0) = \boldsymbol{x}e/\sigma^2.$$

The Cramer-Rao bound for estimation of $\boldsymbol{\theta}$ (and therefore $\boldsymbol{\beta}(\boldsymbol{\theta})$ as well) is

$$\left(\mathbb{E}\left(\boldsymbol{s}\boldsymbol{s}'\right)\right)^{-1} = \left(\sigma^{-4}\mathbb{E}\left(\left(\boldsymbol{x}e\right)\left(\boldsymbol{x}e\right)'\right)\right)^{-1} = \sigma^2\boldsymbol{Q}^{-1} = \boldsymbol{V}^0_{\boldsymbol{\beta}}.$$

We have shown that there is a parametric submodel (5.30) whose Cramer-Rao bound for estimation of $\boldsymbol{\beta}$ is identical to the asymptotic variance of the least-squares estimator, which therefore is the semiparametric variance bound.

Theorem 5.12.1 *In the homoskedastic regression model, the semiparametric variance bound for estimation of $\boldsymbol{\beta}$ is $\boldsymbol{V}^0 = \sigma^2\boldsymbol{Q}^{-1}$ and the OLS estimator is semiparametrically efficient.*

This result is similar to the Gauss-Markov theorem, in that it asserts the efficiency of the least-squares estimator in the context of the homoskedastic regression model. The difference is that the Gauss-Markov theorem states that OLS has the smallest variance among the set of unbiased linear estimators, while Theorem 5.12.1 states that OLS has the smallest asymptotic variance among regular estimators. This is a much more powerful statement.

Exercises

Exercise 5.1 You have two independent samples $(\boldsymbol{y}_1, \boldsymbol{X}_1)$ and $(\boldsymbol{y}_2, \boldsymbol{X}_2)$ which satisfy $\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1$ and $\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}_2$, where $\mathbb{E}(\boldsymbol{x}_{1i}e_{1i}) = 0$ and $\mathbb{E}(\boldsymbol{x}_{2i}e_{2i}) = 0$, and both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have k columns. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the OLS estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. For simplicity, you may assume that both samples have the same number of observations n.

(a) Find the asymptotic distribution of $\sqrt{n}\left(\left(\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1\right) - (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)\right)$ as $n \to \infty$.

(b) Find an appropriate test statistic for $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$.

(c) Find the asymptotic distribution of this statistic under $\mathbb{H}_0$.

Exercise 5.2 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \\ \boldsymbol{\Omega} &= \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' e_i^2\right). \end{aligned}$$

Find the method of moments estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}}\right)$ for $(\boldsymbol{\beta}, \boldsymbol{\Omega})$.

(a) In this model, are $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}}\right)$ efficient estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$?

(b) If so, in what sense are they efficient?

Exercise 5.3 Take the model $y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$ with $\mathbb{E}\boldsymbol{x}_i e_i = \boldsymbol{0}$. Suppose that $\boldsymbol{\beta}_1$ is estimated by regressing y_i on $\boldsymbol{x}_{1i}$ only. Find the probability limit of this estimator. In general, is it consistent for $\boldsymbol{\beta}_1$? If not, under what conditions is this estimator consistent for $\boldsymbol{\beta}_1$?

Exercise 5.4 Let $\boldsymbol{y}$ be $n \times 1$, $\boldsymbol{X}$ be $n \times k$ (rank k). $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Define the *ridge regression* estimator

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' + \lambda \boldsymbol{I}_k\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right)$$

where $\lambda > 0$ is a fixed constant. Find the probability limit of $\hat{\boldsymbol{\beta}}$ as $n \to \infty$. Is $\hat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$?

Exercise 5.5 Of the variables $(y_i^*, y_i, \boldsymbol{x}_i)$ only the pair $(y_i, \boldsymbol{x}_i)$ are observed. In this case, we say that y_i^* is a *latent* variable. Suppose

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \boldsymbol{0} \\ y_i &= y_i^* + u_i \end{aligned}$$

where u_i is a measurement error satisfying

$$\begin{aligned} \mathbb{E}(\boldsymbol{x}_i u_i) &= \boldsymbol{0} \\ \mathbb{E}(y_i^* u_i) &= 0 \end{aligned}$$

Let $\hat{\boldsymbol{\beta}}$ denote the OLS coefficient from the regression of y_i on $\boldsymbol{x}_i$.

(a) Is $\boldsymbol{\beta}$ the coefficient from the linear projection of y_i on $\boldsymbol{x}_i$?

(b) Is $\hat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$ as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$.

Exercise 5.6 The model is

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}\left(e_i \mid x_i\right) &= 0 \end{aligned}$$

where $x_i \in \mathbb{R}$. Consider the two estimators

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \tilde{\beta} &= \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}. \end{aligned}$$

(a) Under the stated assumptions, are both estimators consistent for β?

(b) Are there conditions under which either estimator is efficient?

Chapter 6

Testing

6.1 t tests

The t-test is routinely used to test hypotheses on θ. A simple null and composite hypothesis takes the form

$$\begin{aligned} \mathbb{H}_0 &: \theta = \theta_0 \\ \mathbb{H}_1 &: \theta \neq \theta_0 \end{aligned}$$

where θ_0 is some pre-specified value. A t-test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ when $|t_n(\theta_0)|$ is large. By "large" we mean that the observed value of the t-statistic would be unlikely if $\mathbb{H}_0$ were true.

Formally, we first pick an asymptotic significance level α. We then find $z_{\alpha/2}$, the upper $\alpha/2$ quantile of the standard normal distribution which has the property that if $Z \sim \mathrm{N}(0,1)$ then

$$\mathbb{P}\left(|Z| > z_{\alpha/2}\right) = \alpha.$$

For example, $z_{.025} = 1.96$ and $z_{.05} = 1.645$. A test of asymptotic significance α rejects $\mathbb{H}_0$ if $|t_n| > z_{\alpha/2}$. Otherwise the test does not reject, or "accepts" $\mathbb{H}_0$.

The asymptotic significance level is α because Theorem 5.8.1 implies that

$$\begin{aligned} \mathbb{P}\left(\text{reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}\right) &= \mathbb{P}\left(|t_n| > z_{\alpha/2} \mid \theta = \theta_0\right) \\ &\to \mathbb{P}\left(|Z| > z_{\alpha/2}\right) = \alpha. \end{aligned}$$

The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing.

While there is no objective scientific basis for choice of significance level α, the common practice is to set $\alpha = .05$ or 5%. This implies a critical value of $z_{.025} = 1.96 \approx 2$. When $|t_n| > 2$ it is common to say that the t-statistic is **statistically significant**. and if $|t_n| < 2$ it is common to say that the t-statistic is **statistically insignificant**. It is helpful to remember that this is simply a way of saying "Using a t-test, the hypothesis that $\theta = \theta_0$ can [cannot] be rejected at the asymptotic 5% level."

A related statistic is the asymptotic p-value, which can be interpreted as a measure of the evidence against the null hypothesis. The asymptotic p-value of the statistic t_n is

$$p_n = p(t_n)$$

where $p(t)$ is the tail probability function

$$p(t) = \mathbb{P}\left(|Z| > |t|\right) = 2\left(1 - \Phi(|t|)\right).$$

If the p-value p_n is small (close to zero) then the evidence against $\mathbb{H}_0$ is strong.

An equivalent statement of a Neyman-Pearson test is to reject at the $\alpha\%$ level if and only if $p_n < \alpha$. Significance tests can be deduced directly from the p-value since for any α, $p_n < \alpha$ if and only if $|t_n| > z_{\alpha/2}$. The p-value is more general, however, in that the reader is allowed to pick the level of significance α, in contrast to Neyman-Pearson rejection/acceptance reporting where the researcher picks the significance level.

Another helpful observation is that the p-value function is a unit-free transformation of the t statistic. That is, under $\mathbb{H}_0$, $p_n \xrightarrow{d} \mathrm{U}[0,1]$, so the "unusualness" of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic. To see this fact, note that the asymptotic distribution of $|t_n|$ is $F(x) = 1 - p(x)$. Thus

$$\begin{aligned}
\mathbb{P}\left(1 - p_n \leq u\right) &= \mathbb{P}\left(1 - p(t_n) \leq u\right) \\
&= \mathbb{P}\left(F(t_n) \leq u\right) \\
&= \mathbb{P}\left(|t_n| \leq F^{-1}(u)\right) \\
&\to F\left(F^{-1}(u)\right) = u,
\end{aligned}$$

establishing that $1 - p_n \xrightarrow{d} \mathrm{U}[0,1]$, from which it follows that $p_n \xrightarrow{d} \mathrm{U}[0,1]$.

6.2 t-ratios

Some applied papers (especially older ones) report "t-ratios" for each estimated coefficient. For a coefficient θ these are

$$t_n = t_n(0) = \frac{\hat{\theta}}{s(\hat{\theta})},$$

the ratio of the coefficient estimate to its standard error, and equal the t-statistic for the test of the hypothesis $\mathbb{H}_0 : \theta = 0$. Such papers often discuss the "significance" of certain variables or coefficients, or describe "which regressors have a significant effect on y" by noting which t-ratios exceed 2 in absolute value.

This is very poor econometric practice, and should be studiously avoided. It is a receipe for banishment of your work to lower tier economics journals.

Fundamentally, the common t-ratio is a test for the hypothesis that a coefficient equals zero. This should be reported and discussed when this is an interesting economic hypothesis of interest. But if this is not the case, it is distracting.

Instead, when a coefficient θ is of interest, it is constructive to focus on the point estimate, its standard error, and its confidence interval. The point estimate gives our "best guess" for the value. The standard error is a measure of precision. The confidence interval gives us the range of values consistent with the data. If the standard error is large then the point estimate is not a good summary about θ. The endpoints of the confidence interval describe the bounds on the likely possibilities. If the confidence interval embraces too broad a set of values for θ, then the dataset is not sufficiently informative to render inferences about θ. On the other hand if the confidence interval is tight, then the data have produced an accurate estimate, and the focus should be on the value and interpretation of this estimate. In contrast, the widely-seen statement "the t-ratio is highly significant" has little interpretive value.

The above discussion requires that the researcher knows what the coefficient θ means (in terms of the economic problem) and can interpret values and magnitudes, not just signs. This is critical for good applied econometric practice.

6.3 Wald Tests

Sometimes $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta})$ is a $q \times 1$ vector, and it is desired to test the joint restrictions simultaneously. In this case the t-statistic approach does not work. We have the null and alternative

$$\begin{aligned} \mathbb{H}_0 &: \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ \mathbb{H}_1 &: \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \end{aligned}$$

The natural estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \boldsymbol{h}(\hat{\boldsymbol{\beta}})$ and has asymptotic covariance matrix estimate

$$\hat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \hat{\boldsymbol{H}}_{\boldsymbol{\beta}}' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \hat{\boldsymbol{H}}_{\boldsymbol{\beta}}$$

where

$$\hat{\boldsymbol{H}}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{h}(\hat{\boldsymbol{\beta}}).$$

The Wald statistic for $\mathbb{H}_0$ against $\mathbb{H}_1$ is

$$\begin{aligned} W_n &= n \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &= n \left(\boldsymbol{h}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\hat{\boldsymbol{H}}_{\boldsymbol{\beta}}' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \hat{\boldsymbol{H}}_{\boldsymbol{\beta}}\right)^{-1} \left(\boldsymbol{h}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right). \end{aligned} \tag{6.1}$$

When $\boldsymbol{h}$ is a linear function of $\boldsymbol{\beta}$, $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$, then the Wald statistic takes the form

$$W_n = n \left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

The delta method (5.21) showed that $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_\theta\right)$, and Theorem 5.6.2 showed that $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$. Furthermore, $\boldsymbol{H}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$, so by the continuous mapping theorem, $\boldsymbol{H}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) \xrightarrow{p} \boldsymbol{H}_{\boldsymbol{\beta}}$. Thus $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \hat{\boldsymbol{H}}_{\boldsymbol{\beta}}' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}} \hat{\boldsymbol{H}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{H}_{\boldsymbol{\beta}}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{H}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\theta}} > 0$ if $\boldsymbol{H}_{\boldsymbol{\beta}}$ has full rank q. Hence

$$W_n = n \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \boldsymbol{Z}' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \boldsymbol{Z} = \chi_q^2,$$

by Theorem B.9.3. We have established:

Theorem 6.3.1 *Under $\mathbb{H}_0$ and Assumption 5.4.1, if $\operatorname{rank}(\boldsymbol{H}_{\boldsymbol{\beta}}) = q$, then $W_n \xrightarrow{d} \chi_q^2$, a chi-square random variable with q degrees of freedom.*

An asymptotic Wald test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if W_n exceeds $\chi_q^2(\alpha)$, the upper-α quantile of the χ_q^2 distribution. For example, $\chi_1^2(.05) = 3.84 = z_{.025}^2$. The Wald test fails to reject if W_n is less than $\chi_q^2(\alpha)$. As with t-tests, it is conventional to describe a Wald test as "significant" if W_n exceeds the 5% critical value.

Notice that the asymptotic distribution in Theorem 6.3.1 depends solely on q – the number of restrictions being tested. It does not depend on k – the number of parameters estimated.

The asymptotic p-value for W_n is $p_n = p(W_n)$, where $p(x) = \mathbb{P}\left(\chi_q^2 \geq x\right)$ is the tail probability function of the χ_q^2 distribution. The Wald test rejects at the α% level if and only if $p_n < \alpha$, and p_n is asymptotically U$[0, 1]$ under $\mathbb{H}_0$. In applied work it is good practice to report the p-value of a Wald statistic, as it helps readers intrepret the magnitude of the statistic.

6.4 F Tests

Take the linear model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$$

where $\boldsymbol{X}_1$ is $n \times k_1$, $\boldsymbol{X}_2$ is $n \times k_2$, $k = k_1 + k_2$, and the null hypothesis is

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}.$$

In this case, $\boldsymbol{\theta} = \boldsymbol{\beta}_2$, and there are $q = k_2$ restrictions. Also $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$ is linear with $\boldsymbol{R} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix}$ a selector matrix. We know that the Wald statistic takes the form

$$\begin{aligned} W_n &= n\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{\theta}} \\ &= n\hat{\boldsymbol{\beta}}_2'\left(\boldsymbol{R}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\hat{\boldsymbol{\beta}}_2. \end{aligned}$$

Now suppose that covariance matrix is computed under the assumption of homoskedasticity, so that $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is replaced with $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}} = s^2\left(n^{-1}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. We define the "homoskedastic" Wald statistic

$$\begin{aligned} W_n^0 &= n\hat{\boldsymbol{\theta}}'\left(\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}^0\right)^{-1}\hat{\boldsymbol{\theta}} \\ &= n\hat{\boldsymbol{\beta}}_2'\left(\boldsymbol{R}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}}\boldsymbol{R}\right)^{-1}\hat{\boldsymbol{\beta}}_2. \end{aligned}$$

What we show in this section is that this Wald statistic can be written very simply using the formula

$$W_n^0 = (n-k)\left(\frac{\tilde{\boldsymbol{e}}'\tilde{\boldsymbol{e}} - \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}\right) \tag{6.2}$$

where

$$\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1, \qquad \tilde{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$$

are from OLS of $\boldsymbol{y}$ on $\boldsymbol{X}_1$, and

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

are from OLS of $\boldsymbol{y}$ on $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$.

The elegant feature about (6.2) is that it is directly computable from the standard output from two simple OLS regressions, as the sum of squared errors is a typical output from statistical packages. This statistic is typically reported as an "F-statistic" which is defined as

$$F_n = \frac{W_n^0}{k_2} = \frac{\left(\tilde{\boldsymbol{e}}'\tilde{\boldsymbol{e}} - \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}\right)/k_2}{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n-k)}.$$

While it should be emphasized that equality (6.2) only holds if $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}} = s^2\left(n^{-1}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$, still this formula often finds good use in reading applied papers. Because of this connection we call (6.2) the F form of the Wald statistic. (We can also call W_n^0 a homoskedastic form of the Wald statistic.)

We now derive expression (6.2). First, note that by partitioned matrix inversion (A.3)

$$\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R} = \boldsymbol{R}'\begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}^{-1}\boldsymbol{R} = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}$$

where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. Thus

$$\left(\boldsymbol{R}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}}\boldsymbol{R}\right)^{-1} = s^{-2}n^{-1}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1} = s^{-2}n^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)$$

and

$$\begin{aligned} W_n^0 &= n\hat{\boldsymbol{\beta}}_2' \left(\boldsymbol{R}' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}} \boldsymbol{R}\right)^{-1} \hat{\boldsymbol{\beta}}_2 \\ &= \frac{\hat{\boldsymbol{\beta}}_2' \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2\right) \hat{\boldsymbol{\beta}}_2}{s^2}. \end{aligned}$$

To simplify this expression further, note that if we regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ alone, the residual is $\tilde{\boldsymbol{e}} = \boldsymbol{M}_1 \boldsymbol{y}$. Now consider the residual regression of $\tilde{\boldsymbol{e}}$ on $\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1 \boldsymbol{X}_2$. By the FWL theorem, $\tilde{\boldsymbol{e}} = \tilde{\boldsymbol{X}}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}$ and $\tilde{\boldsymbol{X}}_2' \hat{\boldsymbol{e}} = \boldsymbol{0}$. Thus

$$\begin{aligned} \tilde{\boldsymbol{e}}' \tilde{\boldsymbol{e}} &= \left(\tilde{\boldsymbol{X}}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}\right)' \left(\tilde{\boldsymbol{X}}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}\right) \\ &= \hat{\boldsymbol{\beta}}_2' \tilde{\boldsymbol{X}}_2' \tilde{\boldsymbol{X}}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}} \\ &= \hat{\boldsymbol{\beta}}_2' \boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}}, \end{aligned}$$

or alternatively,

$$\hat{\boldsymbol{\beta}}_2' \boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{e}}' \tilde{\boldsymbol{e}} - \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}}.$$

Also, since

$$s^2 = (n-k)^{-1} \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}}$$

we conclude that

$$W_n^0 = (n-k) \left(\frac{\tilde{\boldsymbol{e}}' \tilde{\boldsymbol{e}} - \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}}}{\hat{\boldsymbol{e}}' \hat{\boldsymbol{e}}}\right)$$

as claimed.

In many statistical packages, when an OLS regression is estimated, an "F-statistic" is reported. This is F_n when $\boldsymbol{X}_1$ is a vector is ones, so $\mathbb{H}_0$ is an intercept-only model. This special F statistic is testing the hypothesis that *all* slope coefficients (all coefficients other than the intercept) are zero. This was a popular statistic in the early days of econometric reporting, when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is nearly always highly significant. While there are special cases where this F statistic is useful, these cases are atypical. As a general rule, there is no reason to report this F statistic.

6.5 Normal Regression Model

Now let us partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and consider tests of the linear restriction

$$\begin{aligned} \mathbb{H}_0 &: \boldsymbol{\beta}_2 = \boldsymbol{0} \\ \mathbb{H}_1 &: \boldsymbol{\beta}_2 \neq \boldsymbol{0} \end{aligned}$$

in the normal regression model. In parametric models, a good test statistic is the likelihood ratio, which is twice the difference in the log-likelihood function evaluated under the null and alternative hypotheses. The estimator under the alternative is the unrestricted estimator $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\sigma}^2)$ discussed above. The Gaussian log-likelihood at these estimates is

$$\begin{aligned} \log L(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\sigma}^2) &= -\frac{n}{2} \log\left(2\pi\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2} \hat{\boldsymbol{e}}' \hat{\boldsymbol{e}} \\ &= -\frac{n}{2} \log\left(\hat{\sigma}^2\right) - \frac{n}{2} \log\left(2\pi\right) - \frac{n}{2}. \end{aligned}$$

The MLE under the null hypothesis is the restricted estimates $(\tilde{\boldsymbol{\beta}}_1, \boldsymbol{0}, \tilde{\sigma}^2)$ where $\tilde{\boldsymbol{\beta}}_1$ is the OLS estimate from a regression of y_i on $\boldsymbol{x}_{1i}$ only, with residual variance $\tilde{\sigma}^2$. The log-likelihood of this model is

$$\log L(\tilde{\boldsymbol{\beta}}_1, \boldsymbol{0}, \tilde{\sigma}^2) = -\frac{n}{2} \log\left(\tilde{\sigma}^2\right) - \frac{n}{2} \log\left(2\pi\right) - \frac{n}{2}.$$

The LR statistic for $\mathbb{H}_0$ against $\mathbb{H}_1$ is

$$\begin{aligned} LR_n &= 2\left(\log L(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\sigma}^2) - \log L(\tilde{\boldsymbol{\beta}}_1, \mathbf{0}, \tilde{\sigma}^2)\right) \\ &= n\left(\log\left(\tilde{\sigma}^2\right) - \log\left(\hat{\sigma}^2\right)\right) \\ &= n\log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right). \end{aligned}$$

By a first-order Taylor series approximation

$$LR_n = n\log\left(1 + \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) \simeq n\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) = W_n^0.$$

the homoskedastic Wald statistic. This shows that the two statistics (LR_n and W_n^0) can be numerically close. It also shows that the homoskedastic Wald statistic for linear hypotheses can also be interpreted as an appropriate likelihood ratio statistic under normality.

6.6 Problems with Tests of NonLinear Hypotheses

While the t and Wald tests work well when the hypothesis is a linear restriction on $\boldsymbol{\beta}$, they can work quite poorly when the restrictions are nonlinear. This can be seen by a simple example introduced by Lafontaine and White (1986). Take the model

$$\begin{aligned} y_i &= \beta + e_i \\ e_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

and consider the hypothesis

$$\mathbb{H}_0 : \beta = 1.$$

Let $\hat{\beta}$ and $\hat{\sigma}^2$ be the sample mean and variance of y_i. The standard Wald test for $\mathbb{H}_0$ is

$$W_n = n\frac{\left(\hat{\beta} - 1\right)^2}{\hat{\sigma}^2}.$$

Now notice that $\mathbb{H}_0$ is equivalent to the hypothesis

$$\mathbb{H}_0(r) : \beta^r = 1$$

for any positive integer r. Letting $h(\beta) = \beta^r$, and noting $H_\beta = r\beta^{r-1}$, we find that the standard Wald test for $\mathbb{H}_0(r)$ is

$$W_n(r) = n\frac{\left(\hat{\beta}^r - 1\right)^2}{\hat{\sigma}^2 r^2 \hat{\beta}^{2r-2}}.$$

While the hypothesis $\beta^r = 1$ is unaffected by the choice of r, the statistic $W_n(r)$ varies with r. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 6.1 the Wald statistic $W_n(r)$ as a function of r, setting $n/\sigma^2 = 10$. The increasing solid line is for the case $\hat{\beta} = 0.8$. The decreasing dashed line is for the case $\hat{\beta} = 1.6$. It is easy to see that in each case there are values of r for which the test statistic is significant relative to asymptotic critical values, while there are other values of r for which the test statistic is insignificant. This is distressing since the choice of r is arbitrary and irrelevant to the actual hypothesis.

Our first-order asymptotic theory is not useful to help pick r, as $W_n(r) \xrightarrow{d} \chi_1^2$ under $\mathbb{H}_0$ for any r. This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and

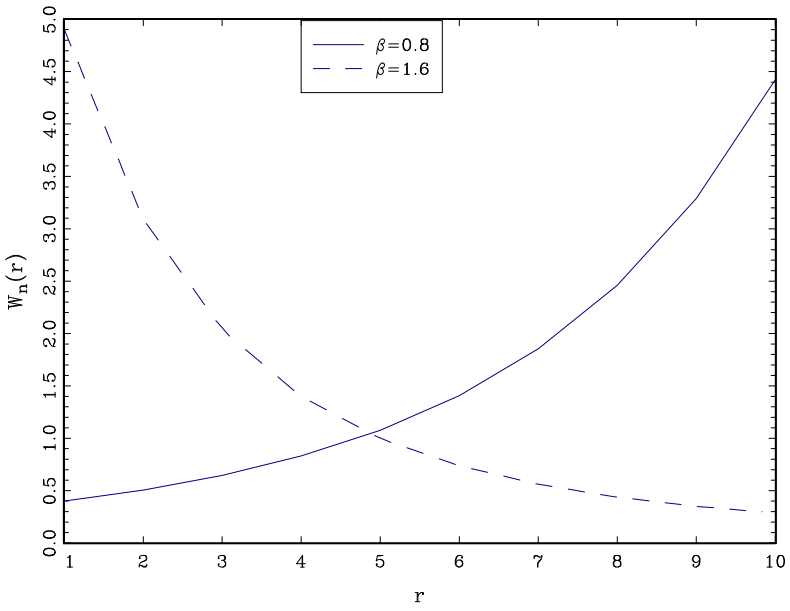

Figure 6.1: Wald Statistic as a function of s

compare the exact distributions of statistical procedures in finite samples. The method uses random simulation to create artificial datasets, to which we apply the statistical tools of interest. This produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated.

In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 – the probability of a false rejection, $\mathbb{P}\left(W_n(r) > 3.84 \mid \beta = 1\right)$. Given the simplicity of the model, this probability depends only on r, n, and σ^2. In Table 2.1 we report the results of a Monte Carlo simulation where we vary these three parameters. The value of r is varied from 1 to 10, n is varied among 20, 100 and 500, and σ is varied among 1 and 3. Table 4.1 reports the simulation estimate of the Type I error probability from 50,000 random samples. Each row of the table corresponds to a different value of r – and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of n and σ. These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics $W_n(r)$ which are larger than 3.84. The null hypothesis $\beta^r = 1$ is true, so these probabilities are Type I error.

To interpret the table, remember that the ideal Type I error probability is 5% (.05) with deviations indicating distortion. Type I error rates between 3% and 8% are considered reasonable. Error rates above 10% are considered excessive. Rates above 20% are unacceptable. When comparing statistical procedures, we compare the rates row by row, looking for tests for which rejection rates are close to 5% and rarely fall outside of the 3%-8% range. For this particular example the only test which meets this criterion is the conventional $W_n = W_n(1)$ test. Any other choice of r leads to a test with unacceptable Type I error probabilities.

In Table 4.1 you can also see the impact of variation in sample size. In each case, the Type I error probability improves towards 5% as the sample size n increases. There is, however, no magic choice of n for which all tests perform uniformly well. Test performance deteriorates as r increases, which is not surprising given the dependence of $W_n(r)$ on r as shown in Figure 6.1.

Table 4.1
Type I error Probability of Asymptotic 5% $W_n(r)$ Test

	$\sigma = 1$			$\sigma = 3$		
r	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	.06	.05	.05	.07	.05	.05
2	.08	.06	.05	.15	.08	.06
3	.10	.06	.05	.21	.12	.07
4	.13	.07	.06	.25	.15	.08
5	.15	.08	.06	.28	.18	.10
6	.17	.09	.06	.30	.20	.11
7	.19	.10	.06	.31	.22	.13
8	.20	.12	.07	.33	.24	.14
9	.22	.13	.07	.34	.25	.15
10	.23	.14	.08	.35	.26	.16

Note: Rejection frequencies from 50,000 simulated random samples

In this example it is not surprising that the choice $r = 1$ yields the best test statistic. Other choices are arbitrary and would not be used in practice. While this is clear in this particular example, in other examples natural choices are not always obvious and the best choices may in fact appear counter-intuitive at first.

This point can be illustrated through another example which is similar to one developed in Gregory and Veall (1985). Take the model

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned} \tag{6.3}$$

and the hypothesis

$$\mathbb{H}_0 : \frac{\beta_1}{\beta_2} = r$$

where r is a known constant. Equivalently, define $\theta = \beta_1/\beta_2$, so the hypothesis can be stated as $\mathbb{H}_0 : \theta = r$.

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ be the least-squares estimates of (6.3), let $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}$ be an estimate of the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ and set $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$. Define

$$\hat{\boldsymbol{H}}_1 = \begin{pmatrix} 0 \\ \dfrac{1}{\hat{\beta}_2} \\ -\dfrac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{pmatrix}$$

so that the standard error for $\hat{\theta}$ is $s(\hat{\theta}) = \left(n^{-1}\hat{\boldsymbol{H}}_1'\hat{\boldsymbol{V}}\hat{\boldsymbol{H}}_1\right)^{1/2}$. In this case a t-statistic for $\mathbb{H}_0$ is

$$t_{1n} = \frac{\left(\frac{\hat{\beta}_1}{\hat{\beta}_2} - r\right)}{s(\hat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$\mathbb{H}_0 : \beta_1 - r\beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$t_{2n} = \frac{\hat{\beta}_1 - r\hat{\beta}_2}{\left(n^{-1}\boldsymbol{H}_2'\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{H}_2\right)^{1/2}}.$$

where

$$\boldsymbol{H}_2 = \begin{pmatrix} 0 \\ 1 \\ -r \end{pmatrix}.$$

To compare t_{1n} and t_{2n} we perform another simple Monte Carlo simulation. We let x_{1i} and x_{2i} be mutually independent $\mathrm{N}(0,1)$ variables, e_i be an independent $\mathrm{N}(0,\sigma^2)$ draw with $\sigma = 3$, and normalize $\beta_0 = 0$ and $\beta_1 = 1$. This leaves β_2 as a free parameter, along with sample size n. We vary β_2 among .1, .25, .50, .75, and 1.0 and n among 100 and 500.

Table 4.2
Type I error Probability of Asymptotic 5% t-tests

	$n = 100$				$n = 500$			
	$\mathbb{P}(t_n < -1.645)$		$\mathbb{P}(t_n > 1.645)$		$\mathbb{P}(t_n < -1.645)$		$\mathbb{P}(t_n > 1.645)$	
β_2	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}
.10	.47	.06	.00	.06	.28	.05	.00	.05
.25	.26	.06	.00	.06	.15	.05	.00	.05
.50	.15	.06	.00	.06	.10	.05	.00	.05
.75	.12	.06	.00	.06	.09	.05	.00	.05
1.00	.10	.06	.00	.06	.07	.05	.02	.05

The one-sided Type I error probabilities $\mathbb{P}(t_n < -1.645)$ and $\mathbb{P}(t_n > 1.645)$ are calculated from 50,000 simulated samples. The results are presented in Table 4.2. Ideally, the entries in the table should be 0.05. However, the rejection rates for the t_{1n} statistic diverge greatly from this value, especially for small values of β_2. The left tail probabilities $\mathbb{P}(t_{1n} < -1.645)$ greatly exceed 5%, while the right tail probabilities $\mathbb{P}(t_{1n} > 1.645)$ are close to zero in most cases. In contrast, the rejection rates for the linear t_{2n} statistic are invariant to the value of β_2, and are close to the ideal 5% rate for both sample sizes. The implication of Table 4.2 is that the two t-ratios have dramatically different sampling behavior.

The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis. In all cases, if the hypothesis can be expressed as a linear restriction on the model parameters, this formulation should be used. If no linear formulation is feasible, then the "most linear" formulation should be selected (as suggested by the theory of Park and Phillips (1988)), and alternatives to asymptotic critical values should be considered. It is also prudent to consider alternative tests to the Wald statistic, such as the GMM distance statistic which will be presented in Section 9.7 (as advocated by Hansen (2006)).

6.7 Monte Carlo Simulation

In the previous section we introduced the method of Monte Carlo simulation to illustrate the small sample problems with tests of nonlinear hypotheses. In this section we describe the method in more detail.

Recall, our data consist of observations $(y_i, \boldsymbol{x}_i)$ which are random draws from a population distribution F. Let $\boldsymbol{\theta}$ be a parameter and let $T_n = T_n((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), \boldsymbol{\theta})$ be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $(\hat{\theta} - \theta)/s(\hat{\theta})$. The exact distribution of T_n is

$$G_n(u, F) = \mathbb{P}(T_n \leq u \mid F).$$

While the asymptotic distribution of T_n might be known, the exact (finite sample) distribution G_n is generally unknown.

Monte Carlo simulation uses numerical simulation to compute $G_n(u, F)$ for selected choices of F. This is useful to investigate the performance of the statistic T_n in reasonable situations and sample sizes. The basic idea is that for any given F, the distribution function $G_n(u, F)$ can be calculated

numerically through simulation. The name Monte Carlo derives from the famous Mediterranean gambling resort where games of chance are played.

The method of Monte Carlo is quite simple to describe. The researcher chooses F (the distribution of the data) and the sample size n. A "true" value of $\boldsymbol{\theta}$ is implied by this choice, or equivalently the value $\boldsymbol{\theta}$ is selected directly by the researcher which implies restrictions on F.

Then the following experiment is conducted

- n independent random pairs $\left(y_i^*, \boldsymbol{x}_i^*\right)$, $i = 1, ..., n$, are drawn from the distribution F using the computer's random number generator.
- The statistic $T_n = T_n\left(\left(y_1^*, \boldsymbol{x}_1^*\right), ..., \left(y_n^*, \boldsymbol{x}_n^*\right), \boldsymbol{\theta}\right)$ is calculated on this pseudo data.

For step 1, most computer packages have built-in procedures for generating U[0, 1] and N(0, 1) random numbers, and from these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the "true" value of $\boldsymbol{\theta}$ corresponding to the choice of F.

The above experiment creates one random draw from the distribution $G_n(u, F)$. This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment B times, where B is a large number. Typically, we set $B = 1000$ or $B = 5000$. We will discuss this choice later.

Notationally, let the b'th experiment result in the draw T_{nb}, $b = 1, ..., B$. These results are stored. They constitute a random sample of size B from the distribution of $G_n(u, F) = \mathbb{P}\left(T_{nb} \leq u\right) = \mathbb{P}\left(T_n \leq u \mid F\right)$.

From a random sample, we can estimate any feature of interest using (typically) a method of moments estimator. For example:

Suppose we are interested in the bias, mean-squared error (MSE), or variance of the distribution of $\hat{\theta} - \theta$. We then set $T_n = \hat{\theta} - \theta$, run the above experiment, and calculate

$$\begin{aligned}
\widehat{Bias(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} T_{nb} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b - \theta \\
\widehat{MSE(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} \left(T_{nb}\right)^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\theta}_b - \theta\right)^2 \\
\widehat{\text{var}(\hat{\theta})} &= \widehat{MSE(\hat{\theta})} - \left(\widehat{Bias(\hat{\theta})}\right)^2
\end{aligned}$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T_n = \left|\hat{\theta} - \theta\right| / s(\hat{\theta})$ and calculate

$$\hat{\mathbb{P}} = \frac{1}{B}\sum_{b=1}^{B} 1\left(T_{nb} \geq 1.96\right), \tag{6.4}$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of $T_n = \hat{\theta}$. We then compute the 5% and 95% sample quantiles of the sample $\{T_{nb}\}$. The α% sample quantile is a number q_α such that α% of the sample are less than q_α. A simple way to compute sample quantiles is to sort the sample $\{T_{nb}\}$ from low to high. Then q_α is the N'th number in this ordered sequence, where $N = (B + 1)\alpha$. It is therefore convenient to pick B so that N is an integer. For example, if we set $B = 999$, then the 5% sample quantile is 50'th sorted value and the 95% sample quantile is the 950'th sorted value.

The typical purpose of a Monte Carlo simulation is to investigate the performance of a statistical procedure (estimator or test) in realistic settings. Generally, the performance will depend on n and

F. In many cases, an estimator or test may perform wonderfully for some values, and poorly for others. It is therefore useful to conduct a variety of experiments, for a selection of choices of n and F.

As discussed above, the researcher must select the number of experiments, B. Often this is called the number of **replications**. Quite simply, a larger B results in more precise estimates of the features of interest of G_n, but requires more computational time. In practice, therefore, the choice of B is often guided by the computational demands of the statistical procedure. Since the results of a Monte Carlo experiment are estimates computed from a random sample of size B, it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference, then B will have to be increased.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (6.4). The random variable $1\left(T_{nb} \geq 1.96\right)$ is iid Bernoulli, equalling 1 with probability $p = \mathbb{E}1\left(T_{nb} \geq 1.96\right)$. The average (6.4) is therefore an unbiased estimator of p with standard error $s\left(\hat{p}\right) = \sqrt{p\left(1-p\right)/B}$. As p is unknown, this may be approximated by replacing p with $\hat{p}$ or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s\left(\hat{p}\right) = \sqrt{(.05)\left(.95\right)/B} \simeq .22/\sqrt{B}$. Hence, standard errors for $B = 100$, 1000, and 5000, are, respectively, $s\left(\hat{p}\right) = .022$, .007, and .003.

6.8 Estimating a Wage Equation

We again return to our wage equation. We use the sample of wage earners from the March 2004 Current Population Survey, excluding military. For the dependent variable we use the natural log of wages so that coefficients may be interpreted as semi-elasticities. For regressors we include years of education, potential work experience, experience squared, and dummy variable indicators for the following: married, female, union member, immigrant, hispanic, and non-white. Furthermore, we included a dummy variable for state of residence (including the District of Columbia, this adds 50 regressors). The available sample is 18,808 so the parameter estimates are quite precise and reported in Table 4.1, excluding the coefficients on the state dummy variables.

Table 4.1 displays the parameter estimates in a standard format. The Table clearly states the estimation method (OLS), the dependent variable (log(Wage)), and the regressors are clearly labeled. Parameter estimates are both reported for the coefficients of interest (the coefficients on the state dummy variables are omitted) and standard errors are reported for all reported coefficient estimates. In addition to the coefficient estimates, the table also reports the estimated error standard deviation, and the sample size. These are useful summary measures of fit which aid readers.

Table 4.1
OLS Estimates of Linear Equation for Log(Wage)

	$\hat{\beta}$	$s(\hat{\beta})$
Intercept	1.027	.032
Education	.101	.002
Experience	.033	.001
Experience^2	−.00057	.00002
Married	.102	.008
Female	−.232	.007
Union Member	.097	.010
Immigrant	−.121	.013
Hispanic	−.102	.014
Non-White	−.070	.010
$\hat{\sigma}$	.4877	
Sample Size	18,808	

Note: Equation also includes state dummy variables.

As a general rule, it is best to always report standard errors along with parameter estimates (as done in Table 4.1). This allows readers to assess the precision of the parameter estimates, and form confidence intervals and t-tests on individual coefficients if desired. For example, if you are interested in the difference in mean wages between men and women, you can read from the table that the estimated coefficient on the Female dummy variable is -0.232, implying a mean wage difference of 23%. To assess the precision, you can see that the standard error for this coefficient estimate is 0.007. This implies a 95% asymptotic confidence interval for the coefficient estimate of $[-.246, -.218]$. This means that we have estimated the difference in mean wages between men and women to lie between 22% and 25%. I interpret this as a precise estimate because there is not an important difference between the lower and upper bound.

Instead of reporting standard errors, some empirical researchers report t-ratios for each parameter estimate. "t-ratios" are t-statistics which test the hypothesis that the coefficient equals zero. An example is reported in Table 4.2. In this example, all the t-ratios are highly significant, ranging in magnitude from 9.3 to 50. What we learn from these statistics is that these coefficients are non-zero, but not much more. In a sample of this size this finding is rather uninteresting; consequently the reporting of t-ratios is a waste of space. Again consider the male-female wage difference. Table 4.2 reports that the t-ratio is 33, enabling us to reject the hypothesis that the coefficient is zero. But how precise is the reported estimate of a wage gap of 23%? It is hard to assess from a quick reading of Table 4.2 Standard errors are much more useful, for they enable for quick and easy assessment of the degree of estimation uncertainty.

Table 4.2
OLS Estimates of Linear Equation for Log(Wage)
Improper Reporting: t-ratios replacing standard errors

	$\hat{\beta}$	t
Intercept	1.027	32
Education	.101	50
Experience	.033	33
Experience^2	$-.00057$	28
Married	.102	12.8
Female	$-.232$	33
Union Member	.097	9.7
Immigrant	$-.121$	9.3
Hispanic	$-.102$	7.3
Non-White	$-.070$	7

Returning to the estimated wage equation, one might question whether or not the state dummy variables are relevant. Computing the Wald statistic (6.1) that the state coefficients are jointly zero, we find $W_n = 550$. Alternatively, re-estimating the model with the 50 state dummies excluded, the restricted standard deviation estimate is $\tilde{\sigma} = .4945$. The F form of the Wald statistic (6.2) is

$$W_n = n\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) = 18,808\left(\frac{.4945^2}{.4877^2} - 1\right) = 528.$$

Notice that the two statistics are close, but not equal. Using either statistic the hypothesis is easily rejected, as the 1% critical value for the χ^2_{50} distribution is 76.

Another interesting question which can be addressed from these estimates is the maximal impact of experience on mean wages. Ignoring the other coefficients, we can write this effect as

$$\log(Wage) = \beta_2 Experience + \beta_3 Experience^2 + \cdots$$

Our question is: At which level of experience θ do workers achieve the highest wage? In this quadratic model, if $\beta_2 > 0$ and $\beta_3 < 0$ the solution is

$$\theta = -\frac{\beta_2}{2\beta_3}.$$

From Table 4.1 we find the point estimate

$$\hat{\theta} = -\frac{\hat{\beta}_2}{2\hat{\beta}_3} = 28.69.$$

Using the Delta Method, we can calculate a standard error of $s(\hat{\theta}) = .40$, implying a 95% confidence interval of [27.9, 29.5].

However, this is a poor choice, as the coverage probability of this confidence interval is one minus the Type I error of the hypothesis test based on the t-test. In Section 6.6 we discovered that such t-tests have very poor Type I error rates. Instead, we found better Type I error rates by reformulating the hypothesis as a linear restriction. These t-statistics take the form

$$t_n(\theta) = \frac{\hat{\beta}_2 + 2\hat{\beta}_3\theta}{\left(\boldsymbol{h}_\theta' \hat{\boldsymbol{V}} \boldsymbol{h}_\theta\right)^{1/2}}$$

where

$$\boldsymbol{h}_\theta = \begin{pmatrix} 1 \\ 2\theta \end{pmatrix}$$

and $\hat{\boldsymbol{V}}$ is the covariance matrix for $(\hat{\beta}_2\ \hat{\beta}_3)$.

In the present context we are interested in forming a confidence interval, not testing a hypothesis, so we have to go one step further. Our desired confidence interval will be the set of parameter values θ which are not rejected by the hypothesis test. This is the set of θ such that $|t_n(\theta)| \leq 1.96$. Since $t_n(\theta)$ is a non-linear function of θ, there is not a simple expression for this set, but it can be found numerically quite easily. This set is [27.0, 29.5]. Notice that the upper end of the confidence interval is the same as that from the delta method, but the lower end is substantially lower.

Exercises

For exercises 1-4, the following definition is used. In the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the least-squares estimate of $\boldsymbol{\beta}$ subject to the restriction $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{0}$ is

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}} S_n(\boldsymbol{\beta}) \\ S_n(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) . \end{aligned}$$

That is, $\tilde{\boldsymbol{\beta}}$ minimizes the sum of squared errors $S_n(\boldsymbol{\beta})$ over all $\boldsymbol{\beta}$ such that the restriction holds.

Exercise 6.1 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show that the least-squares estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ subject to the constraint that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is the OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

Exercise 6.2 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show that the least-squares estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, subject to the constraint that $\boldsymbol{\beta}_1 = \boldsymbol{c}$ (where $\boldsymbol{c}$ is some given vector) is simply the OLS regression of $\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{c}$ on $\boldsymbol{X}_2$.

Exercise 6.3 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ each $n \times k$, find the least-squares estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, subject to the constraint that $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$.

Exercise 6.4 Take the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with the restriction $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{r}$ where $\boldsymbol{R}$ is a known $k \times s$ matrix, $\boldsymbol{r}$ is a known $s \times 1$ vector, $0 < s < k$, and $\operatorname{rank}(\boldsymbol{R}) = s$. Explain why $\tilde{\boldsymbol{\beta}}$ solves the minimization of the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} S_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}' \left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{r}\right)$$

where $\boldsymbol{\lambda}$ is $s \times 1$.

(a) Show that the solution is

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R} \left[\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right]^{-1} \left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \\ \hat{\boldsymbol{\lambda}} &= \left[\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right]^{-1} \left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \end{aligned}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

is the unconstrained OLS estimator.

(b) Verify that $\boldsymbol{R}'\tilde{\boldsymbol{\beta}} = \boldsymbol{r}$.

(c) Show that if $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{r}$ is true, then

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{I}_k - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R} \left[\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right]^{-1} \boldsymbol{R}'\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e}.$$

(d) Under the standard assumptions plus $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{r}$, find the asymptotic distribution of $\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$.

(e) Find an appropriate formula to calculate standard errors for the elements of $\tilde{\boldsymbol{\beta}}$.

Exercise 6.5 Prove that if an additional regressor $\boldsymbol{X}_{k+1}$ is added to $\boldsymbol{X}$, Theil's adjusted $\overline{R}^2$ increases if and only if $|t_{k+1}| > 1$, where $t_{k+1} = \hat{\beta}_{k+1}/s(\hat{\beta}_{k+1})$ is the t-ratio for $\hat{\beta}_{k+1}$ and

$$s(\hat{\beta}_{k+1}) = \left(s^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{k+1,k+1}\right)^{1/2}$$

is the homoskedasticity-formula standard error.

Exercise 6.6 The data set `invest.dat` contains data on 565 U.S. firms extracted from Compustat for the year 1987. The variables, in order, are

- I_i Investment to Capital Ratio (multiplied by 100).
- Q_i Total Market Value to Asset Ratio (Tobin's Q).
- C_i Cash Flow to Asset Ratio.
- D_i Long Term Debt to Asset Ratio.

The flow variables are annual sums for 1987. The stock variables are beginning of year.

(a) Estimate a linear regression of I_i on the other variables. Calculate appropriate standard errors.

(b) Calculate asymptotic confidence intervals for the coefficients.

(c) This regression is related to Tobin's q theory of investment, which suggests that investment should be predicted solely by Q_i. Thus the coefficient on Q_i should be positive and the others should be zero. Test the joint hypothesis that the coefficients on C_i and D_i are zero. Test the hypothesis that the coefficient on Q_i is zero. Are the results consistent with the predictions of the theory?

(d) Now try a non-linear (quadratic) specification. Regress I_i on Q_i, C_i, D_i, Q_i^2, C_i^2, D_i^2, Q_iC_i, Q_iD_i, C_iD_i. Test the joint hypothesis that the six interaction and quadratic coefficients are zero.

Exercise 6.7 In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. (The problem is discussed in Example 8.3 of Greene, section 1.7 of Hayashi, and the empirical exercise in Chapter 1 of Hayashi). The data file `nerlov.dat` contains his data. The variables are described on page 77 of Hayashi. Nerlov was interested in estimating a *cost function*: $TC = f(Q, PL, PF, PK)$.

(a) First estimate an unrestricted Cobb-Douglass specification

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{6.5}$$

Report parameter estimates and standard errors. You should obtain the same OLS estimates as in Hayashi's equation (1.7.7), but your standard errors may differ.

(b) Using a Wald statistic, test the hypothesis $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$.

(c) Estimate (6.5) by least-squares imposing this restriction by substitution. Report your parameter estimates and standard errors.

(d) Estimate (6.5) subject to $\beta_3 + \beta_4 + \beta_5 = 1$ using the restricted least-squares estimator from problem 4. Do you obtain the same estimates as in part (c)?

Chapter 7

Additional Regression Topics

7.1 Generalized Least Squares

In the projection model, we know that the least-squares estimator is semi-parametrically efficient for the projection coefficient. However, in the linear regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0, \end{aligned}$$

the least-squares estimator is inefficient. The theory of Chamberlain (1987) can be used to show that in this model the semiparametric efficiency bound is obtained by the **Generalized Least Squares** (GLS) estimator (4.11) introduced in Section 4.5.1. The GLS estimator is sometimes called the Aitken estimator. The GLS estimator (7.1) is infeasible since the matrix $\boldsymbol{D}$ is unknown. A feasible GLS (FGLS) estimator replaces the unknown $\boldsymbol{D}$ with an estimate $\hat{\boldsymbol{D}} = \text{diag}\{\hat{\sigma}_1^2, ..., \hat{\sigma}_n^2\}$. We now discuss this estimation problem.

Suppose that we model the conditional variance using the parametric form

$$\begin{aligned} \sigma_i^2 &= \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1 \\ &= \boldsymbol{\alpha}'\boldsymbol{z}_i, \end{aligned}$$

where $\boldsymbol{z}_{1i}$ is some $q \times 1$ function of $\boldsymbol{x}_i$. Typically, $\boldsymbol{z}_{1i}$ are squares (and perhaps levels) of some (or all) elements of $\boldsymbol{x}_i$. Often the functional form is kept simple for parsimony.

Let $\eta_i = e_i^2$. Then

$$\mathbb{E}\left(\eta_i \mid \boldsymbol{x}_i\right) = \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1$$

and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1 + \xi_i \\ \mathbb{E}\left(\xi_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned} \tag{7.1}$$

This regression error ξ_i is generally heteroskedastic and has the conditional variance

$$\begin{aligned} \text{var}\left(\xi_i \mid \boldsymbol{x}_i\right) &= \text{var}\left(e_i^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(\left(e_i^2 - \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(e_i^4 \mid \boldsymbol{x}_i\right) - \left(\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2. \end{aligned}$$

Suppose e_i (and thus η_i) were observed. Then we could estimate $\boldsymbol{\alpha}$ by OLS:

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{\eta} \xrightarrow{p} \boldsymbol{\alpha}$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\alpha}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}}=\left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\xi_i^2\right)\left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}. \tag{7.2}$$

While e_i is not observed, we have the OLS residual $\hat{e}_i=y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}=e_i-\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$. Thus

$$\begin{aligned}\phi_i &\equiv \hat{\eta}-\eta_i\\ &= \hat{e}_i^2-e_i^2\\ &= -2e_i\boldsymbol{x}_i'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}).\end{aligned}$$

And then

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{z}_i\phi_i &= \frac{-2}{n}\sum_{i=1}^n\boldsymbol{z}_ie_i\boldsymbol{x}_i'\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}_i(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\sqrt{n}\\ &\xrightarrow{p}\mathbf{0}\end{aligned}$$

Let

$$\tilde{\boldsymbol{\alpha}}=\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\hat{\boldsymbol{\eta}} \tag{7.3}$$

be from OLS regression of $\hat{\eta}_i$ on $\boldsymbol{z}_i$. Then

$$\begin{aligned}\sqrt{n}\left(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right) &= \sqrt{n}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right)+\left(n^{-1}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}n^{-1/2}\boldsymbol{Z}'\boldsymbol{\phi}\\ &\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\alpha}}\right)\end{aligned} \tag{7.4}$$

Thus the fact that η_i is replaced with $\hat{\eta}_i$ is asymptotically irrelevant. We call (7.3) the *skedastic* regression, as it is estimating the conditional variance of the regression of y_i on $\boldsymbol{x}_i$. We have shown that $\boldsymbol{\alpha}$ is consistently estimated by a simple procedure, and hence we can estimate $\sigma_i^2=\boldsymbol{z}_i'\boldsymbol{\alpha}$ by

$$\tilde{\sigma}_i^2=\tilde{\boldsymbol{\alpha}}'\boldsymbol{z}_i. \tag{7.5}$$

Suppose that $\tilde{\sigma}_i^2>0$ for all i. Then set

$$\tilde{\boldsymbol{D}}=\mathrm{diag}\{\tilde{\sigma}_1^2,...,\tilde{\sigma}_n^2\}$$

and

$$\tilde{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{y}.$$

This is the feasible GLS, or FGLS, estimator of $\boldsymbol{\beta}$. Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression.

One typical problem with implementation of FGLS estimation is that in the linear specification (7.1), there is no guarantee that $\tilde{\sigma}_i^2>0$ for all i. If $\tilde{\sigma}_i^2<0$ for some i, then the FGLS estimator is not well defined. Furthermore, if $\tilde{\sigma}_i^2\approx 0$ for some i then the FGLS estimator will force the regression equation through the point $(y_i,\boldsymbol{x}_i)$, which is undesirable. This suggests that there is a need to bound the estimated variances away from zero. A trimming rule takes the form

$$\overline{\sigma}_i^2=\max[\tilde{\sigma}_i^2,c\hat{\sigma}^2]$$

for some $c>0$. For example, setting $c=1/4$ means that the conditional variance function is constrained to exceed one-fourth of the unconditional variance. As there is no clear method to select c, this introduces a degree of arbitrariness. In this context it is useful to re-estimate the model with several choices for the trimming parameter. If the estimates turn out to be sensitive to its choice, the estimation method should probably be reconsidered.

It is possible to show that if the skedastic regression is correctly specified, then FGLS is asymptotically equivalent to GLS. As the proof is tricky, we just state the result without proof.

Theorem 7.1.1 *If the skedastic regression is correctly specified,*

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{GLS} - \tilde{\boldsymbol{\beta}}_{FGLS}\right) \xrightarrow{p} \mathbf{0},$$

and thus

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\mathbb{E}\left(\sigma_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

Examining the asymptotic distribution of Theorem 7.1.1, the natural estimator of the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0} = \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\frac{1}{n}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}.$$

which is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$ as $n \to \infty$. This estimator $\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$ is appropriate when the skedastic regression (7.1) is correctly specified.

It may be the case that $\boldsymbol{\alpha}'\boldsymbol{z}_i$ is only an approximation to the true conditional variance $\sigma_i^2 = \mathbb{E}(e_i^2 \mid \boldsymbol{x}_i)$. In this case we interpret $\boldsymbol{\alpha}'\boldsymbol{z}_i$ as a linear projection of e_i^2 on $\boldsymbol{z}_i$. $\tilde{\boldsymbol{\beta}}$ should perhaps be called a quasi-FGLS estimator of $\boldsymbol{\beta}$. Its asymptotic variance is not that given in Theorem 7.1.1. Instead,

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-2}\sigma_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

$\boldsymbol{V}_{\boldsymbol{\beta}}$ takes a sandwich form similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2 = \boldsymbol{\alpha}'\boldsymbol{z}_i$, $\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$ is inconsistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$.

An appropriate solution is to use a White-type estimator in place of $\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$. This may be written as

$$\begin{aligned}
\tilde{\boldsymbol{V}}_{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\sigma}_i^{-4}\hat{e}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \\
&= \left(\frac{1}{n}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\hat{\boldsymbol{D}}\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}
\end{aligned}$$

where $\hat{\boldsymbol{D}} = \mathrm{diag}\{\hat{e}_1^2, ..., \hat{e}_n^2\}$. This is estimator is robust to misspecification of the conditional variance, and was proposed by Cragg (1992).

In the linear regression model, FGLS is asymptotically superior to OLS. Why then do we not exclusively estimate regression models by FGLS? This is a good question. There are three reasons.

First, FGLS estimation depends on specification and estimation of the skedastic regression. Since the form of the skedastic regression is unknown, and it may be estimated with considerable error, the estimated conditional variances may contain more noise than information about the true conditional variances. In this case, FGLS can do worse than OLS in practice.

Second, individual estimated conditional variances may be negative, and this requires trimming to solve. This introduces an element of arbitrariness which is unsettling to empirical researchers.

Third, and probably most importantly, OLS is a robust estimator of the parameter vector. It is consistent not only in the regression model, but also under the assumptions of linear projection. The GLS and FGLS estimators, on the other hand, require the assumption of a correct conditional mean. If the equation of interest is a linear projection and not a conditional mean, then the OLS

and FGLS estimators will converge in probability to different limits as they will be estimating two different projections. The FGLS probability limit will depend on the particular function selected for the skedastic regression. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct conditional mean, and the cost is a loss of robustness to misspecification.

7.2 Testing for Heteroskedasticity

The hypothesis of homoskedasticity is that $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, or equivalently that

$$\mathbb{H}_0 : \boldsymbol{\alpha}_1 = 0$$

in the regression (7.1). We may therefore test this hypothesis by the estimation (7.3) and constructing a Wald statistic. In the classic literature it is typical to impose the stronger assumption that e_i is independent of $\boldsymbol{x}_i$, in which case ξ_i is independent of $\boldsymbol{x}_i$ and the asymptotic variance (7.2) for $\tilde{\boldsymbol{\alpha}}$ simplifies to

$$V_{\boldsymbol{\alpha}} = \left(\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)\right)^{-1} \mathbb{E}\left(\xi_i^2\right). \tag{7.6}$$

Hence the standard test of $\mathbb{H}_0$ is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (7.3). The asymptotic distribution (7.4) and the asymptotic variance (7.6) under independence show that this test has an asymptotic chi-square distribution.

Theorem 7.2.1 *Under $\mathbb{H}_0$ and e_i independent of $\boldsymbol{x}_i$, the Wald test of $\mathbb{H}_0$ is asymptotically χ_q^2.*

Most tests for heteroskedasticity take this basic form. The main differences between popular tests are which transformations of $\boldsymbol{x}_i$ enter $\boldsymbol{z}_i$. Motivated by the form of the asymptotic variance of the OLS estimator $\hat{\boldsymbol{\beta}}$, White (1980) proposed that the test for heteroskedasticity be based on setting $\boldsymbol{z}_i$ to equal all non-redundant elements of $\boldsymbol{x}_i$, its squares, and all cross-products. Breusch-Pagan (1979) proposed what might appear to be a distinct test, but the only difference is that they allowed for general choice of $\boldsymbol{z}_i$, and replaced $\mathbb{E}\left(\xi_i^2\right)$ with $2\sigma^4$ which holds when e_i is $\mathrm{N}\left(\boldsymbol{0}, \sigma^2\right)$. If this simplification is replaced by the standard formula (under independence of the error), the two tests coincide.

It is important not to misuse tests for heteroskedasticity. It should not be used to determine whether to estimate a regression equation by OLS or FGLS, nor to determine whether classic or White standard errors should be reported. Hypothesis tests are not designed for these purposes. Rather, tests for heteroskedasticity should be used to answer the scientific question of whether or not the conditional variance is a function of the regressors. If this question is not of economic interest, then there is no value in conducting a test for heteorskedasticity.

7.3 Forecast Intervals

In the linear regression model the conditional mean of y_i given $\boldsymbol{x}_i = \boldsymbol{x}$ is

$$m(\boldsymbol{x}) = \mathbb{E}\left(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\beta}.$$

In some cases, we want to estimate $m(\boldsymbol{x})$ at a particular point $\boldsymbol{x}$. Notice that this is a (linear) function of $\boldsymbol{\beta}$. Letting $h(\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ and $\theta = h(\boldsymbol{\beta})$, we see that $\hat{m}(\boldsymbol{x}) = \hat{\theta} = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$ and $\boldsymbol{H}_{\boldsymbol{\beta}} = \boldsymbol{x}$, so $s(\hat{\theta}) = \sqrt{n^{-1}\boldsymbol{x}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{x}}$. Thus an asymptotic 95% confidence interval for $m(\boldsymbol{x})$ is

$$\left[\boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm 2\sqrt{n^{-1}\boldsymbol{x}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{x}}\right].$$

It is interesting to observe that if this is viewed as a function of $\boldsymbol{x}$, the width of the confidence set is dependent on $\boldsymbol{x}$.

For a given value of $\boldsymbol{x}_i = \boldsymbol{x}$, we may want to forecast (guess) y_i out-of-sample. A reasonable rule is the conditional mean $m(\boldsymbol{x})$ as it is the mean-square-minimizing forecast. A point forecast is the estimated conditional mean $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}'\hat{\boldsymbol{\beta}}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\hat{e}_i = y_i - \hat{m}(\boldsymbol{x}) = e_i - \boldsymbol{x}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. As the out-of-sample error e_i is independent of the in-sample estimate $\hat{\boldsymbol{\beta}}$, this has variance

$$\begin{aligned}\mathbb{E}\hat{e}_i^2 &= \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i = \boldsymbol{x}\right) + \boldsymbol{x}'\mathbb{E}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x} \\ &= \sigma^2(\boldsymbol{x}) + n^{-1}\boldsymbol{x}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{x}.\end{aligned}$$

Assuming $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, the natural estimate of this variance is $\hat{\sigma}^2 + n^{-1}\boldsymbol{x}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{x}$, so a standard error for the forecast is $\hat{s}(\boldsymbol{x}) = \sqrt{\hat{\sigma}^2 + n^{-1}\boldsymbol{x}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{x}}$. Notice that this is different from the standard error for the conditional mean. If we have an estimate of the conditional variance function, e.g. $\tilde{\sigma}^2(\boldsymbol{x}) = \tilde{\boldsymbol{\alpha}}'\boldsymbol{z}$ from (7.5), then the forecast standard error is $\hat{s}(\boldsymbol{x}) = \sqrt{\tilde{\sigma}^2(\boldsymbol{x}) + n^{-1}\boldsymbol{x}'\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{x}}$

It would appear natural to conclude that an asymptotic 95% forecast interval for y_i is

$$\left[\boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm 2\hat{s}(\boldsymbol{x})\right],$$

but this turns out to be incorrect. In general, the validity of an asymptotic confidence interval is based on the asymptotic normality of the studentized ratio. In the present case, this would require the asymptotic normality of the ratio

$$\frac{e_i - \boldsymbol{x}'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\hat{s}(\boldsymbol{x})}.$$

But no such asymptotic approximation can be made. The only special exception is the case where e_i has the exact distribution $\mathrm{N}(0, \sigma^2)$, which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of e_i given $\boldsymbol{x}_i = \boldsymbol{x}$, which is a much more difficult task. Perhaps due to this difficulty, many applied forecasters use the simple approximate interval $\left[\boldsymbol{x}'\hat{\boldsymbol{\beta}} \pm 2\hat{s}(\boldsymbol{x})\right]$ despite the lack of a convincing justification.

7.4 NonLinear Least Squares

In some cases we might use a parametric regression function $m(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{E}(y_i \mid \boldsymbol{x}_i = \boldsymbol{x})$ which is a non-linear function of the parameters $\boldsymbol{\theta}$. We describe this setting as **non-linear regression**. Examples of nonlinear regression functions include

$$\begin{aligned}m(x, \boldsymbol{\theta}) &= \theta_1 + \theta_2 \frac{x}{1 + \theta_3 x} \\ m(x, \boldsymbol{\theta}) &= \theta_1 + \theta_2 x^{\theta_3} \\ m(x, \boldsymbol{\theta}) &= \theta_1 + \theta_2 \exp(\theta_3 x) \\ m(\boldsymbol{x}, \boldsymbol{\theta}) &= G(\boldsymbol{x}'\theta),\ G \text{ known} \\ m(\boldsymbol{x}, \boldsymbol{\theta}) &= \theta_1'\boldsymbol{x}_1 + \left(\theta_2'\boldsymbol{x}_1\right)\Phi\left(\frac{x_2 - \theta_3}{\theta_4}\right) \\ m(x, \boldsymbol{\theta}) &= \theta_1 + \theta_2 x + \theta_3 (x - \theta_4)\, 1(x > \theta_4) \\ m(\boldsymbol{x}, \boldsymbol{\theta}) &= \left(\theta_1'\boldsymbol{x}_1\right) 1(x_2 < \theta_3) + \left(\theta_2'\boldsymbol{x}_1\right) 1(x_2 > \theta_3)\end{aligned}$$

In the first five examples, $m(\boldsymbol{x}, \boldsymbol{\theta})$ is (generically) differentiable in the parameters $\boldsymbol{\theta}$. In the final two examples, m is not differentiable with respect to θ_4 and θ_3 which alters some of the analysis. When it exists, let

$$\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} m(\boldsymbol{x}, \boldsymbol{\theta}).$$

Nonlinear regression is sometimes adopted because the functional form $m(\boldsymbol{x}, \boldsymbol{\theta})$ is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

The least squares estimator $\hat{\boldsymbol{\theta}}$ minimizes the normalized sum-of-squared-errors

$$S_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2.$$

When the regression function is nonlinear, we call this the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are $\hat{e}_i = y_i - m\left(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)$.

One motivation for the choice of NLLS as the estimation method is that the parameter $\boldsymbol{\theta}$ is the solution to the population problem $\min_{\boldsymbol{\theta}} \mathbb{E}\left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2$

Since sum-of-squared-errors function $S_n(\boldsymbol{\theta})$ is not quadratic, $\hat{\boldsymbol{\theta}}$ must be found by numerical methods. See Appendix E. When $m(\boldsymbol{x}, \boldsymbol{\theta})$ is differentiable, then the FOC for minimization are

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{m}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right) \hat{e}_i. \tag{7.7}$$

Theorem 7.4.1 ***Asymptotic Distribution of NLLS Estimator***
If the model is identified and $m\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$ is differentiable with respect to $\boldsymbol{\theta}$,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'e_i^2\right)\right)\left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1}$$

where $\boldsymbol{m}_{\boldsymbol{\theta}i} = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$.

Based on Theorem 7.4.1, an estimate of the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\hat{e}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1}$$

where $\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i} = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$ and $\hat{e}_i = y_i - m(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$.

Identification is often tricky in nonlinear regression models. Suppose that

$$m(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{\beta}_1'\boldsymbol{z}_i + \boldsymbol{\beta}_2'\boldsymbol{x}_i(\gamma)$$

where $\boldsymbol{x}_i\left(\gamma\right)$ is a function of $\boldsymbol{x}_i$ and the unknown parameter $\boldsymbol{\gamma}$. Examples include $x_i\left(\gamma\right) = x_i^{\gamma}$, $x_i\left(\gamma\right) = \exp\left(\gamma x_i\right)$, and $x_i\left(\boldsymbol{\gamma}\right) = x_i 1\left(g\left(x_i\right) > \gamma\right)$. The model is linear when $\boldsymbol{\beta}_2 = \mathbf{0}$, and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

However, under $\mathbb{H}_0$, the model is

$$y_i = \boldsymbol{\beta}_1'\boldsymbol{z}_i + e_i$$

and both $\boldsymbol{\beta}_2$ and γ have dropped out. This means that under $\mathbb{H}_0$, γ is not identified. This renders the distribution theory presented in the previous section invalid. Thus when the truth is that

$\boldsymbol{\beta}_2 = 0$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of $\mathbb{H}_0$ do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests have been worked out by Andrews and Ploberger (1994) and B. Hansen (1996). In particular, Hansen shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p-values) in a given application.

Proof of Theorem 7.4.1 (Sketch). NLLS estimation falls in the class of optimization estimators. For this theory, it is useful to denote the true value of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0$.

The first step is to show that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Proving that nonlinear estimators are consistent is more challenging than for linear estimators. We sketch the main argument. The idea is that $\hat{\boldsymbol{\theta}}$ minimizes the sample criterion function $S_n(\boldsymbol{\theta})$, which (for any $\boldsymbol{\theta}$) converges in probability to the mean-squared error function $\mathbb{E}\left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2$. Thus it seems reasonable that the minimizer $\hat{\boldsymbol{\theta}}$ will converge in probability to $\boldsymbol{\theta}_0$, the minimizer of $\mathbb{E}\left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2$. It turns out that to show this rigorously, we need to show that $S_n(\boldsymbol{\theta})$ converges *uniformly* to its expectation $\mathbb{E}\left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2$, which means that the maximum discrepancy must converge in probability to zero, to exclude the possibility that $S_n(\boldsymbol{\theta})$ is excessively wiggly in $\boldsymbol{\theta}$. Proving uniform convergence is technically challenging, but it can be shown to hold broadly for relevant nonlinear regression models, especially if the regression function $m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)$ is differentiabel in $\boldsymbol{\theta}$. For a complete treatment of the theory of optimization estimators see Newey and McFadden (1994).

Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$ for n large, so the minimization of $S_n(\boldsymbol{\theta})$ only needs to be examined for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$. Let

$$y_i^0 = e_i + \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}_0.$$

For $\boldsymbol{\theta}$ close to the true value $\boldsymbol{\theta}_0$, by a first-order Taylor series approximation,

$$m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right) \simeq m\left(\boldsymbol{x}_i, \boldsymbol{\theta}_0\right) + \boldsymbol{m}_{\boldsymbol{\theta}i}'\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right).$$

Thus

$$\begin{aligned} y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right) &\simeq \left(e_i + m\left(\boldsymbol{x}_i, \boldsymbol{\theta}_0\right)\right) - \left(m\left(\boldsymbol{x}_i, \boldsymbol{\theta}_0\right) + \boldsymbol{m}_{\boldsymbol{\theta}i}'\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)\right) \\ &= e_i - \boldsymbol{m}_{\boldsymbol{\theta}i}'\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) \\ &= y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}. \end{aligned}$$

Hence the sum of squared errors function is

$$S_n(\boldsymbol{\theta}) = \sum_{i=1}^n \left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2 \simeq \sum_{i=1}^n \left(y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}\right)^2$$

and the right-hand-side is the SSE function for a linear regression of y_i^0 on $\boldsymbol{m}_{\boldsymbol{\theta}i}$. Thus the NLLS estimator $\hat{\boldsymbol{\theta}}$ has the same asymptotic distribution as the (infeasible) OLS regression of y_i^0 on $\boldsymbol{m}_{\boldsymbol{\theta}i}$, which is that stated in the theorem.

7.5 Least Absolute Deviations

We stated that a conventional goal in econometrics is estimation of impact of variation in $\boldsymbol{x}_i$ on the central tendency of y_i. We have discussed projections and conditional means, but these are not the only measures of central tendency. An alternative good measure is the conditional median.

To recall the definition and properties of the median, let y be a continuous random variable. The median $\theta = \text{med}(y)$ is the value such that $\mathbb{P}(y \leq \theta) = \mathbb{P}\left(y \geq \theta_0\right) = .5$. Two useful facts about the median are that

$$\theta = \underset{\theta}{\text{argmin}}\, \mathbb{E}\left|y - \theta\right| \tag{7.8}$$

and

$$\mathbb{E}\operatorname{sgn}(y-\theta)=0$$

where

$$\operatorname{sgn}(u)=\begin{cases}1 & \text{if } u\geq 0\\ -1 & \text{if } u<0\end{cases}$$

is the sign function.

These facts and definitions motivate three estimators of θ. The first definition is the $50th$ empirical quantile. The second is the value which minimizes $\frac{1}{n}\sum_{i=1}^{n}|y_i-\theta|$, and the third definition is the solution to the moment equation $\frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(y_i-\theta)$. These distinctions are illusory, however, as these estimators are indeed identical.

Now let's consider the conditional median of y given a random vector $\boldsymbol{x}$. Let $m(\boldsymbol{x})=\operatorname{med}(y\mid\boldsymbol{x})$ denote the conditional median of y given $\boldsymbol{x}$. The linear median regression model takes the form

$$\begin{aligned}y_i &= \boldsymbol{x}_i'\boldsymbol{\beta}+e_i\\ \operatorname{med}(e_i\mid\boldsymbol{x}_i) &= 0\end{aligned}$$

In this model, the linear function $\operatorname{med}(y_i\mid\boldsymbol{x}_i=\boldsymbol{x})=\boldsymbol{x}'\boldsymbol{\beta}$ is the conditional median function, and the substantive assumption is that the median function is linear in $\boldsymbol{x}$.

Conditional analogs of the facts about the median are

- $\mathbb{P}(y_i\leq\boldsymbol{x}'\boldsymbol{\beta}_0\mid\boldsymbol{x}_i=\boldsymbol{x})=\mathbb{P}(y_i>\boldsymbol{x}'\boldsymbol{\beta}\mid\boldsymbol{x}_i=\boldsymbol{x})=.5$
- $\mathbb{E}(\operatorname{sgn}(e_i)\mid\boldsymbol{x}_i)=0$
- $\mathbb{E}(\boldsymbol{x}_i\operatorname{sgn}(e_i))=0$
- $\boldsymbol{\beta}=\min_{\boldsymbol{\beta}}\mathbb{E}|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|$

These facts motivate the following estimator. Let

$$LAD_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}\right|$$

be the average of absolute deviations. The **least absolute deviations** (LAD) estimator of $\boldsymbol{\beta}$ minimizes this function

$$\hat{\boldsymbol{\beta}}=\operatorname*{argmin}_{\boldsymbol{\beta}}LAD_n(\boldsymbol{\beta})$$

Equivalently, it is a solution to the moment condition

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\operatorname{sgn}\left(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)=0. \tag{7.9}$$

The LAD estimator has an asymptotic normal distribution.

Theorem 7.5.1 ***Asymptotic Distribution of LAD Estimator***
When the conditional median is linear in $\boldsymbol{x}$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}(\mathbf{0},\boldsymbol{V})$$

where

$$V=\frac{1}{4}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f(0\mid\boldsymbol{x}_i)\right)\right)^{-1}\left(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f(0\mid\boldsymbol{x}_i)\right)\right)^{-1}$$

and $f(e\mid\boldsymbol{x})$ *is the conditional density of* e_i *given* $\boldsymbol{x}_i=\boldsymbol{x}$.

The variance of the asymptotic distribution inversely depends on $f(0 \mid \boldsymbol{x})$, the conditional density of the error at its median. When $f(0 \mid \boldsymbol{x})$ is large, then there are many innovations near to the median, and this improves estimation of the median. In the special case where the error is independent of $\boldsymbol{x}_i$, then $f(0 \mid \boldsymbol{x}) = f(0)$ and the asymptotic variance simplifies

$$\boldsymbol{V} = \frac{\left(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}}{4f(0)^2} \tag{7.10}$$

This simplification is similar to the simplification of the asymptotic covariance of the OLS estimator under homoskedasticity.

Computation of standard error for LAD estimates typically is based on equation (7.10). The main difficulty is the estimation of $f(0)$, the height of the error density at its median. This can be done with kernel estimation techniques. See Chapter 16. While a complete proof of Theorem 7.5.1 is advanced, we provide a sketch here for completeness.

Proof of Theorem 7.5.1: Similar to NLLS, LAD is an optimization estimator. Let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}_0$.

The first step is to show that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. The general nature of the proof is similar to that for the NLLS estimator, and is sketched here. For any fixed $\boldsymbol{\beta}$, by the WLLN, $LAD_n(\boldsymbol{\beta}) \xrightarrow{p} \mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$. Furthermore, it can be shown that this convergence is uniform in $\boldsymbol{\beta}$. (Proving uniform convergence is more challenging than for the NLLS criterion since the LAD criterion is not differentiable in $\boldsymbol{\beta}$.) It follows that $\hat{\boldsymbol{\beta}}$, the minimizer of $LAD_n(\boldsymbol{\beta})$, converges in probability to $\boldsymbol{\beta}_0$, the minimizer of $\mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$.

Since $\operatorname{sgn}(a) = 1 - 2 \cdot 1(a \leq 0)$, (7.9) is equivalent to $\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = 0$, where $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta})$ and $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\left(1 - 2 \cdot 1\left(y_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)$. Let $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta})$. We need three preliminary results. First, by the central limit theorem (Theorem C.2.1)

$$\sqrt{n}\left(\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}(\boldsymbol{\beta}_0)\right) = -n^{-1/2}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)$$

since $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)' = \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'$. Second using the law of iterated expectations and the chain rule of differentiation,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\boldsymbol{x}_i\left(1 - 2 \cdot 1\left(y_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) \\
&= -2\frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\left[\boldsymbol{x}_i\mathbb{E}\left(1\left(e_i \leq \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0\right) \mid \boldsymbol{x}_i\right)\right] \\
&= -2\frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\left[\boldsymbol{x}_i\int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0} f(e \mid \boldsymbol{x}_i)\, de\right] \\
&= -2\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i' f\left(\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0 \mid \boldsymbol{x}_i\right)\right]
\end{aligned}$$

so

$$\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}) = -2\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i' f(0 \mid \boldsymbol{x}_i)\right].$$

Third, by a Taylor series expansion and the fact $\boldsymbol{g}(\boldsymbol{\beta}) = 0$

$$\boldsymbol{g}(\hat{\boldsymbol{\beta}}) \simeq \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta})\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

Together

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) &\simeq \left(\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}_0)\right)^{-1}\sqrt{n}\boldsymbol{g}(\hat{\boldsymbol{\beta}}) \\
&= \left(-2\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\sqrt{n}\left(\boldsymbol{g}(\hat{\boldsymbol{\beta}})-\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right) \\
&\simeq \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\sqrt{n}\left(\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)-\boldsymbol{g}(\boldsymbol{\beta}_0)\right) \\
&\xrightarrow{d} \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\mathrm{N}\left(\mathbf{0}, \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\
&= \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right).
\end{aligned}$$

The third line follows from an asymptotic empirical process argument and the fact that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

7.6 Quantile Regression

Quantile regression has become quite popular in recent econometric practice. For $\tau \in [0,1]$ the τ'th quantile Q_τ of a random variable with distribution function $F(u)$ is defined as

$$Q_\tau = \inf\left\{u : F(u) \geq \tau\right\}$$

When $F(u)$ is continuous and strictly monotonic, then $F\left(Q_\tau\right) = \tau$, so you can think of the quantile as the inverse of the distribution function. The quantile Q_τ is the value such that τ (percent) of the mass of the distribution is less than Q_τ. The median is the special case $\tau = .5$.

The following alternative representation is useful. If the random variable U has τ'th quantile Q_τ, then

$$Q_\tau = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}\rho_\tau\left(U-\theta\right). \tag{7.11}$$

where $\rho_\tau\left(q\right)$ is the piecewise linear function

$$\begin{aligned}
\rho_\tau\left(q\right) &= \begin{cases} -q\left(1-\tau\right) & q < 0 \\ q\tau & q \geq 0 \end{cases} \\
&= q\left(\tau - 1\left(q < 0\right)\right).
\end{aligned} \tag{7.12}$$

This generalizes representation (7.8) for the median to all quantiles.

For the random variables $(y_i, \boldsymbol{x}_i)$ with conditional distribution function $F\left(y \mid \boldsymbol{x}\right)$ the conditional quantile function $q_\tau(\boldsymbol{x})$ is

$$Q_\tau(\boldsymbol{x}) = \inf\left\{y : F\left(y \mid \boldsymbol{x}\right) \geq \tau\right\}.$$

Again, when $F\left(y \mid \boldsymbol{x}\right)$ is continuous and strictly monotonic in y, then $F\left(Q_\tau(\boldsymbol{x}) \mid \boldsymbol{x}\right) = \tau$. For fixed τ, the quantile regression function $q_\tau(\boldsymbol{x})$ describes how the τ'th quantile of the conditional distribution varies with the regressors.

As functions of $\boldsymbol{x}$, the quantile regression functions can take any shape. However for computational convenience it is typical to assume that they are (approximately) linear in $\boldsymbol{x}$ (after suitable transformations). This linear specification assumes that $Q_\tau(\boldsymbol{x}) = \boldsymbol{\beta}_\tau'\boldsymbol{x}$ where the coefficients $\boldsymbol{\beta}_\tau$ vary across the quantiles τ. We then have the linear quantile regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + e_i$$

where e_i is the error defined to be the difference between y_i and its τ'th conditional quantile $\boldsymbol{x}_i'\boldsymbol{\beta}_\tau$. By construction, the τ'th conditional quantile of e_i is zero, otherwise its properties are unspecified without further restrictions.

Given the representation (7.11), the quantile regression estimator $\hat{\boldsymbol{\beta}}_\tau$ for $\boldsymbol{\beta}_\tau$ solves the minimization problem

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} S_n^\tau(\boldsymbol{\beta})$$

where

$$S_n^\tau(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

and $\rho_\tau(q)$ is defined in (7.12).

Since the quanitle regression criterion function $S_n^\tau(\boldsymbol{\beta})$ does not have an algebraic solution, numerical methods are necessary for its minimization. Furthermore, since it has discontinuous derivatives, conventional Newton-type optimization methods are inappropriate. Fortunately, fast linear programming methods have been developed for this problem, and are widely available.

An asymptotic distribution theory for the quantile regression estimator can be derived using similar arguments as those for the LAD estimator in Theorem 7.5.1.

Theorem 7.6.1 ***Asymptotic Distribution of the Quantile Regression Estimator***
When the τ'th conditional quantile is linear in $\boldsymbol{x}$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_\tau\right),$$

where

$$\boldsymbol{V}_\tau = \tau(1-\tau)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\left(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}$$

and $f(e \mid \boldsymbol{x})$ is the conditional density of e_i given $\boldsymbol{x}_i = \boldsymbol{x}$.

In general, the asymptotic variance depends on the conditional density of the quantile regression error. When the error e_i is independent of $\boldsymbol{x}_i$, then $f(0 \mid \boldsymbol{x}_i) = f(0)$, the unconditional density of e_i at 0, and we have the simplification

$$\boldsymbol{V}_\tau = \frac{\tau(1-\tau)}{f(0)^2}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

A recent monograph on the details of quantile regression is Koenker (2005).

7.7 Testing for Omitted NonLinearity

If the goal is to estimate the conditional expectation $\mathbb{E}(y_i \mid \boldsymbol{x}_i)$, it is useful to have a general test of the adequacy of the specification.

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. Thus, if the model $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$ has been fit by OLS, let $\boldsymbol{z}_i = \boldsymbol{h}(\boldsymbol{x}_i)$ denote functions of $\boldsymbol{x}_i$ which are not linear functions of $\boldsymbol{x}_i$ (perhaps squares of non-binary regressors) and then fit $y_i = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\tilde{\boldsymbol{\gamma}} + \tilde{e}_i$ by OLS, and form a Wald statistic for $\boldsymbol{\gamma} = 0$.

Another popular approach is the RESET test proposed by Ramsey (1969). The null model is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$

which is estimated by OLS, yielding predicted values $\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$. Now let

$$\boldsymbol{z}_i = \begin{pmatrix} \hat{y}_i^2 \\ \vdots \\ \hat{y}_i^m \end{pmatrix}$$

be an $(m-1)$-vector of powers of $\hat{y}_i$. Then run the auxiliary regression

$$y_i = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\tilde{\boldsymbol{\gamma}} + \tilde{e}_i \tag{7.13}$$

by OLS, and form the Wald statistic W_n for $\boldsymbol{\gamma} = \boldsymbol{0}$. It is easy (although somewhat tedious) to show that under the null hypothesis, $W_n \xrightarrow{d} \chi^2_{m-1}$. Thus the null is rejected at the $\alpha\%$ level if W_n exceeds the upper $\alpha\%$ tail critical value of the χ^2_{m-1} distribution.

To implement the test, m must be selected in advance. Typically, small values such as $m = 2$, 3, or 4 seem to work best.

The RESET test appears to work well as a test of functional form against a wide range of smooth alternatives. It is particularly powerful at detecting *single-index* models of the form

$$y_i = G(\boldsymbol{x}_i'\boldsymbol{\beta}) + e_i$$

where $G(\cdot)$ is a smooth "link" function. To see why this is the case, note that (7.13) may be written as

$$y_i = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^2\tilde{\gamma}_1 + \left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^3\tilde{\gamma}_2 + \cdots \left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^m\tilde{\gamma}_{m-1} + \tilde{e}_i$$

which has essentially approximated $G(\cdot)$ by a m'th order polynomial.

7.8 Irrelevant Variables

In the model

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0, \end{aligned}$$

$\boldsymbol{x}_{2i}$ is "irrelevant" if $\boldsymbol{\beta}_1$ is the parameter of interest and $\boldsymbol{\beta}_2 = 0$. One estimator of $\boldsymbol{\beta}_1$ is to regress y_i on $\boldsymbol{x}_{1i}$ alone, $\tilde{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{y}\right)$. Another is to regress y_i on $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ jointly, yielding $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$. Under which conditions is $\hat{\boldsymbol{\beta}}_1$ or $\tilde{\boldsymbol{\beta}}_1$ superior?

It is easy to see that both estimators are consistent for $\boldsymbol{\beta}_1$. However, they will (typically) have different asymptotic variances.

The comparison between the two estimators is straightforward when the error is conditionally homoskedastic $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$. In this case

$$\lim_{n\to\infty} n \operatorname{var}(\tilde{\boldsymbol{\beta}}_1) = \left(\mathbb{E}\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)^{-1}\sigma^2 = \boldsymbol{Q}_{11}^{-1}\sigma^2,$$

say, and

$$\lim_{n\to\infty} n \operatorname{var}(\hat{\boldsymbol{\beta}}_1) = \left(\mathbb{E}\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}' - \mathbb{E}\boldsymbol{x}_{1i}\boldsymbol{x}_{2i}'\left(\mathbb{E}\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}'\right)^{-1}\mathbb{E}\boldsymbol{x}_{2i}\boldsymbol{x}_{1i}'\right)^{-1}\sigma^2 = \left(\boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)^{-1}\sigma^2,$$

say. If $\boldsymbol{Q}_{12} = 0$ (so the variables are orthogonal) then these two variance matrices equal, and the two estimators have equal asymptotic efficiency. Otherwise, since $\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} > 0$, then $\boldsymbol{Q}_{11} > \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$, and consequently

$$\boldsymbol{Q}_{11}^{-1}\sigma^2 < \left(\boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)^{-1}\sigma^2.$$

This means that $\tilde{\boldsymbol{\beta}}_1$ has a lower asymptotic variance matrix than $\hat{\boldsymbol{\beta}}_1$. We conclude that the inclusion of irrelevant variable reduces estimation efficiency if these variables are correlated with the relevant variables.

For example, take the model $y_i = \beta_0 + \beta_1 x_i + e_i$ and suppose that $\beta_0 = 0$. Let $\hat{\beta}_1$ be the estimate of β_1 from the unconstrained model, and $\tilde{\beta}_1$ be the estimate under the constraint $\beta_0 = 0$. (The least-squares estimate with the intercept omitted.). Let $\mathbb{E}x_i = \mu$, and $\mathbb{E}(x_i - \mu)^2 = \sigma_x^2$. Then under (5.13),

$$\lim_{n\to\infty} n \operatorname{var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sigma_x^2 + \mu^2}$$

while

$$\lim_{n\to\infty} n \operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sigma_x^2}.$$

When $\mu \neq 0$, we see that $\tilde{\beta}_1$ has a lower asymptotic variance.

However, this result can be reversed when the error is conditionally heteroskedastic. In the absence of the homoskedasticity assumption, there is no clear ranking of the efficiency of the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ versus the unrestricted estimator.

7.9 Model Selection

In earlier sections we discussed the costs and benefits of inclusion/exclusion of variables. How does a researcher go about selecting an econometric specification, when economic theory does not provide complete guidance? This is the question of model selection. It is important that the model selection question be well-posed. For example, the question: "What is the right model for y?" is not well-posed, because it does not make clear the conditioning set. In contrast, the question, "Which subset of $(x_1, ..., x_K)$ enters the regression function $\mathbb{E}(y_i \mid x_{1i} = x_1, ..., x_{Ki} = x_K)$?" is well posed.

In many cases the problem of model selection can be reduced to the comparison of two nested models, as the larger problem can be written as a sequence of such comparisons. We thus consider the question of the inclusion of $\boldsymbol{X}_2$ in the linear regression

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e},$$

where $\boldsymbol{X}_1$ is $n \times k_1$ and $\boldsymbol{X}_2$ is $n \times k_2$. This is equivalent to the comparison of the two models

$$\begin{aligned} \mathcal{M}_1 &: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}, \qquad \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}_1, \boldsymbol{X}_2) = 0 \\ \mathcal{M}_2 &: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}, \qquad \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}_1, \boldsymbol{X}_2) = 0. \end{aligned}$$

Note that $\mathcal{M}_1 \subset \mathcal{M}_2$. To be concrete, we say that $\mathcal{M}_2$ is true if $\boldsymbol{\beta}_2 \neq 0$.

To fix notation, models 1 and 2 are estimated by OLS, with residual vectors $\hat{\boldsymbol{e}}_1$ and $\hat{\boldsymbol{e}}_2$, estimated variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, etc., respectively. To simplify some of the statistical discussion, we will on occasion use the homoskedasticity assumption $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}\right) = \sigma^2$.

A model selection procedure is a data-dependent rule which selects one of the two models. We can write this as $\widehat{\mathcal{M}}$. There are many possible desirable properties for a model selection procedure. One useful property is consistency, that it selects the true model with probability one if the sample is sufficiently large. A model selection procedure is consistent if

$$\begin{aligned} \mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &\to 1 \\ \mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right) &\to 1 \end{aligned}$$

However, this rule only makes sense when the true model is finite dimensional. If the truth is infinite dimensional, it is more appropriate to view model selection as determining the best finite sample approximation.

A common approach to model selection is to base the decision on a statistical test such as the Wald W_n. The model selection rule is as follows. For some critical level α, let c_α satisfy $\mathbb{P}\left(\chi^2_{k_2} > c_\alpha\right) = \alpha$. Then select $\mathcal{M}_1$ if $W_n \leq c_\alpha$, else select $\mathcal{M}_2$.

A major problem with this approach is that the critical level α is indeterminate. The reasoning which helps guide the choice of α in hypothesis testing (controlling Type I error) is not relevant for model selection. That is, if α is set to be a small number, then $\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) \approx 1 - \alpha$ but $\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right)$ could vary dramatically, depending on the sample size, etc. Another problem is that if α is held fixed, then this model selection procedure is inconsistent, as $\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) \to 1 - \alpha < 1$.

Another common approach to model selection is to use a selection criterion. One popular choice is the Akaike Information Criterion (AIC). The AIC under normality for model m is

$$AIC_m = \log\left(\hat{\sigma}^2_m\right) + 2\frac{k_m}{n}. \tag{7.14}$$

where $\hat{\sigma}^2_m$ is the variance estimate for model m, and k_m is the number of coefficients in the model. The AIC can be derived as an estimate of the KullbackLeibler information distance $K(\mathcal{M}) = \mathbb{E}\left(\log f(\boldsymbol{y} \mid \boldsymbol{X}) - \log f(\boldsymbol{y} \mid \boldsymbol{X}, \mathcal{M})\right)$ between the true density and the model density. The rule is to select $\mathcal{M}_1$ if $AIC_1 < AIC_2$, else select $\mathcal{M}_2$. AIC selection is inconsistent, as the rule tends to overfit. Indeed, since under $\mathcal{M}_1$,

$$LR_n = n\left(\log\hat{\sigma}^2_1 - \log\hat{\sigma}^2_2\right) \simeq W_n \xrightarrow{d} \chi^2_{k_2}, \tag{7.15}$$

then

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &= \mathbb{P}\left(AIC_1 < AIC_2 \mid \mathcal{M}_1\right) \\
&= \mathbb{P}\left(\log(\hat{\sigma}^2_1) + 2\frac{k_1}{n} < \log(\hat{\sigma}^2_2) + 2\frac{k_1 + k_2}{n} \mid \mathcal{M}_1\right) \\
&= \mathbb{P}\left(LR_n < 2k_2 \mid \mathcal{M}_1\right) \\
&\to \mathbb{P}\left(\chi^2_{k_2} < 2k_2\right) < 1.
\end{aligned}$$

While many criterions similar to the AIC have been proposed, the most popular is one proposed by Schwarz based on Bayesian arguments. His criterion, known as the BIC, is

$$BIC_m = \log\left(\hat{\sigma}^2_m\right) + \log(n)\frac{k_m}{n}. \tag{7.16}$$

Since $\log(n) > 2$ (if $n > 8$), the BIC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious.

In contrast to the AIC, BIC model selection is consistent. Indeed, since (7.15) holds under $\mathcal{M}_1$,

$$\frac{LR_n}{\log(n)} \xrightarrow{p} 0,$$

so

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &= \mathbb{P}\left(BIC_1 < BIC_2 \mid \mathcal{M}_1\right) \\
&= \mathbb{P}\left(LR_n < \log(n)k_2 \mid \mathcal{M}_1\right) \\
&= \mathbb{P}\left(\frac{LR_n}{\log(n)} < k_2 \mid \mathcal{M}_1\right) \\
&\to \mathbb{P}\left(0 < k_2\right) = 1.
\end{aligned}$$

Also under $\mathcal{M}_2$, one can show that

$$\frac{LR_n}{\log(n)} \xrightarrow{p} \infty,$$

thus

$$\begin{aligned}\mathbb{P}\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right) &= \mathbb{P}\left(\frac{LR_n}{\log(n)} > k_2 \mid \mathcal{M}_2\right) \\ &\to 1.\end{aligned}$$

We have discussed model selection between two models. The methods extend readily to the issue of selection among multiple regressors. The general problem is the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + e_i, \qquad \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0$$

and the question is which subset of the coefficients are non-zero (equivalently, which regressors enter the regression).

There are two leading cases: ordered regressors and unordered.

In the ordered case, the models are

$$\begin{aligned}\mathcal{M}_1 &: \beta_1 \neq 0, \beta_2 = \beta_3 = \cdots = \beta_K = 0 \\ \mathcal{M}_2 &: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = \cdots = \beta_K = 0 \\ &\vdots \\ \mathcal{M}_K &: \beta_1 \neq 0, \beta_2 \neq 0, \ldots, \beta_K \neq 0.\end{aligned}$$

which are nested. The AIC selection criteria estimates the K models by OLS, stores the residual variance $\hat{\sigma}^2$ for each model, and then selects the model with the lowest AIC (7.14). Similarly for the BIC, selecting based on (7.16).

In the unordered case, a model consists of any possible subset of the regressors $\{x_{1i}, ..., x_{Ki}\}$, and the AIC or BIC in principle can be implemented by estimating all possible subset models. However, there are 2^K such models, which can be a very large number. For example, $2^{10} = 1024$, and $2^{20} = 1,048,576$. In the latter case, a full-blown implementation of the BIC selection criterion would seem computationally prohibitive.

Exercises

Exercise 7.1 The data file `cps78.dat` contains 550 observations on 20 variables taken from the May 1978 current population survey. Variables are listed in the file `cps78.pdf`. The goal of the exercise is to estimate a model for the log of earnings (variable LNWAGE) as a function of the conditioning variables.

(a) Start by an OLS regression of LNWAGE on the other variables. Report coefficient estimates and standard errors.

(b) Consider augmenting the model by squares and/or cross-products of the conditioning variables. Estimate your selected model and report the results.

(c) Are there any variables which seem to be unimportant as a determinant of wages? You may re-estimate the model without these variables, if desired.

(d) Test whether the error variance is different for men and women. Interpret.

(e) Test whether the error variance is different for whites and nonwhites. Interpret.

(f) Construct a model for the conditional variance. Estimate such a model, test for general heteroskedasticity and report the results.

(g) Using this model for the conditional variance, re-estimate the model from part (c) using FGLS. Report the results.

(h) Do the OLS and FGLS estimates differ greatly? Note any interesting differences.

(i) Compare the estimated standard errors. Note any interesting differences.

Exercise 7.2 In the homoskedastic regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$ and $\mathbb{E}(e_i^2 \mid \boldsymbol{x}_i) = \sigma^2$, suppose $\hat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$ with covariance matrix $\hat{\boldsymbol{V}}$, based on a sample of size n. Let $\hat{\sigma}^2$ be the estimate of σ^2. You wish to forecast an out-of-sample value of y_{n+1} given that $\boldsymbol{x}_{n+1} = \boldsymbol{x}$. Thus the available information is the sample $(\boldsymbol{y}, \boldsymbol{X})$, the estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{V}}, \hat{\sigma}^2)$, the residuals $\hat{\boldsymbol{e}}$, and the out-of-sample value of the regressors, $\boldsymbol{x}_{n+1}$.

(a) Find a point forecast of y_{n+1}.

(b) Find an estimate of the variance of this forecast.

Exercise 7.3 Suppose that $y_i = g(\boldsymbol{x}_i, \boldsymbol{\theta}) + e_i$ with $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$, $\hat{\boldsymbol{\theta}}$ is the NLLS estimator, and $\hat{\boldsymbol{V}}$ is the estimate of $\operatorname{var}\left(\hat{\boldsymbol{\theta}}\right)$. You are interested in the conditional mean function $\mathbb{E}(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}) = g(\boldsymbol{x})$ at some $\boldsymbol{x}$. Find an asymptotic 95% confidence interval for $g(\boldsymbol{x})$.

Exercise 7.4 For any predictor $g(\boldsymbol{x}_i)$ for y_i, the mean absolute error (MAE) is

$$\mathbb{E}\left|y_i - g(\boldsymbol{x}_i)\right|.$$

Show that the function $g(\boldsymbol{x})$ which minimizes the MAE is the conditional median $m(\boldsymbol{x}) = \operatorname{med}(y_i \mid \boldsymbol{x}_i)$.

Exercise 7.5 Define

$$g(u) = \tau - 1(u < 0)$$

where $1(\cdot)$ is the indicator function (takes the value 1 if the argument is true, else equals zero). Let θ satisfy $\mathbb{E}g(y_i - \theta) = 0$. Is θ a quantile of the distribution of y_i?

Exercise 7.6 Verify equation (7.11).

Exercise 7.7 In Exercise 6.7, you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{7.17}$$

(a) Following Nerlove, add the variable $(\log Q_i)^2$ to the regression. Do so. Assess the merits of this new specification using (i) a hypothesis test; (ii) AIC criterion; (iii) BIC criterion. Do you agree with this modification?

(b) Now try a non-linear specification. Consider model (7.17) plus the extra term $\beta_6 \boldsymbol{z}_i$, where

$$\boldsymbol{z}_i = \log Q_i \left(1 + \exp\left(-\left(\log Q_i - \beta_7\right)\right)\right)^{-1}.$$

In addition, impose the restriction $\beta_3 + \beta_4 + \beta_5 = 1$. This model is called a smooth threshold model. For values of $\log Q_i$ much below β_7, the variable $\log Q_i$ has a regression slope of β_2. For values much above β_7, the regression slope is $\beta_2 + \beta_6$, and the model imposes a smooth transition between these regimes. The model is non-linear because of the parameter β_7.

The model works best when β_7 is selected so that several values (in this example, at least 10 to 15) of $\log Q_i$ are both below and above β_7. Examine the data and pick an appropriate range for β_7.

(c) Estimate the model by non-linear least squares. I recommend the concentration method: Pick 10 (or more or you like) values of β_7 in this range. For each value of β_7, calculate $\boldsymbol{z}_i$ and estimate the model by OLS. Record the sum of squared errors, and find the value of β_7 for which the sum of squared errors is minimized.

(d) Calculate standard errors for all the parameters $(\beta_1, ..., \beta_7)$.

Chapter 8

The Bootstrap

8.1 Definition of the Bootstrap

Let F denote a distribution function for the population of observations $(y_i, \boldsymbol{x}_i)$. Let

$$T_n = T_n\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), F\right)$$

be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Note that we write T_n as possibly a function of F. For example, the t-statistic is a function of the parameter θ which itself is a function of F.

The exact CDF of T_n when the data are sampled from the distribution F is

$$G_n(u, F) = \mathbb{P}(T_n \leq u \mid F)$$

In general, $G_n(u, F)$ depends on F, meaning that G changes as F changes.

Ideally, inference would be based on $G_n(u, F)$. This is generally impossible since F is unknown.

Asymptotic inference is based on approximating $G_n(u, F)$ with $G(u, F) = \lim_{n\to\infty} G_n(u, F)$. When $G(u, F) = G(u)$ does not depend on F, we say that T_n is asymptotically pivotal and use the distribution function $G(u)$ for inferential purposes.

In a seminal contribution, Efron (1979) proposed the bootstrap, which makes a different approximation. The unknown F is replaced by a consistent estimate F_n (one choice is discussed in the next section). Plugged into $G_n(u, F)$ we obtain

$$G_n^*(u) = G_n(u, F_n). \tag{8.1}$$

We call G_n^* the bootstrap distribution. Bootstrap inference is based on $G_n^*(u)$.

Let $(y_i^*, \boldsymbol{x}_i^*)$ denote random variables with the distribution F_n. A random sample from this distribution is called the bootstrap data. The statistic $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ constructed on this sample is a random variable with distribution G_n^*. That is, $\mathbb{P}(T_n^* \leq u) = G_n^*(u)$. We call T_n^* the bootstrap statistic. The distribution of T_n^* is identical to that of T_n when the true CDF of F_n rather than F.

The bootstrap distribution is itself random, as it depends on the sample through the estimator F_n.

In the next sections we describe computation of the bootstrap distribution.

8.2 The Empirical Distribution Function

Recall that $F(y, \boldsymbol{x}) = \mathbb{P}\left(y_i \leq y, \boldsymbol{x}_i \leq \boldsymbol{x}\right) = \mathbb{E}\left(1\left(y_i \leq y\right) 1\left(\boldsymbol{x}_i \leq \boldsymbol{x}\right)\right)$, where $1(\cdot)$ is the indicator function. This is a population moment. The method of moments estimator is the corresponding

sample moment:

$$F_n(y, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x}). \tag{8.2}$$

$F_n(y, \boldsymbol{x})$ is called the empirical distribution function (EDF). F_n is a nonparametric estimate of F. Note that while F may be either discrete or continuous, F_n is by construction a step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any $(y, \boldsymbol{x})$, $1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x})$ is an iid random variable with expectation $F(y, \boldsymbol{x})$. Thus by the WLLN (Theorem 5.2.1), $F_n(y, \boldsymbol{x}) \xrightarrow{p} F(y, \boldsymbol{x})$. Furthermore, by the CLT (Theorem C.2.1),

$$\sqrt{n}\left(F_n(y, \boldsymbol{x}) - F(y, \boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F(y, \boldsymbol{x})\left(1 - F(y, \boldsymbol{x})\right)\right).$$

To see the effect of sample size on the EDF, in the Figure below, I have plotted the EDF and true CDF for three random samples of size $n = 25$, 50, 100, and 500. The random draws are from the $\mathrm{N}(0, 1)$ distribution. For $n = 25$, the EDF is only a crude approximation to the CDF, but the approximation appears to improve for the large n. In general, as the sample size gets larger, the EDF step function gets uniformly close to the true CDF.

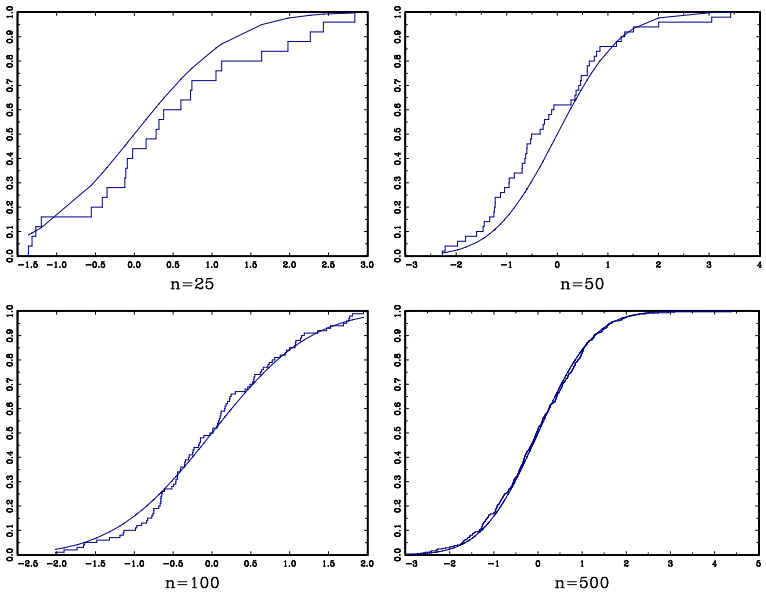

Figure 8.1: Empirical Distribution Functions

The EDF is a valid discrete probability distribution which puts probability mass $1/n$ at each pair $(y_i, \boldsymbol{x}_i)$, $i = 1, ..., n$. Notationally, it is helpful to think of a random pair $(y_i^*, \boldsymbol{x}_i^*)$ with the distribution F_n. That is,

$$\mathbb{P}(y_i^* \leq y, \boldsymbol{x}_i^* \leq \boldsymbol{x}) = F_n(y, \boldsymbol{x}).$$

We can easily calculate the moments of functions of $(y_i^*, \boldsymbol{x}_i^*)$:

$$\begin{aligned}
\mathbb{E}h(y_i^*, \boldsymbol{x}_i^*) &= \int h(y, \boldsymbol{x}) dF_n(y, \boldsymbol{x}) \\
&= \sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i)\, \mathbb{P}(y_i^* = y_i, \boldsymbol{x}_i^* = \boldsymbol{x}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i),
\end{aligned}$$

the empirical sample average.

8.3 Nonparametric Bootstrap

The **nonparametric bootstrap** is obtained when the bootstrap distribution (8.1) is defined using the EDF (8.2) as the estimate F_n of F.

Since the EDF F_n is a multinomial (with n support points), in principle the distribution G_n^* could be calculated by direct methods. However, as there are 2^n possible samples $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$, such a calculation is computationally infeasible. The popular alternative is to use simulation to approximate the distribution. The algorithm is identical to our discussion of Monte Carlo simulation, with the following points of clarification:

- The sample size n used for the simulation is the same as the sample size.
- The random vectors $(y_i^*, \boldsymbol{x}_i^*)$ are drawn randomly from the empirical distribution. This is equivalent to sampling a pair $(y_i, \boldsymbol{x}_i)$ randomly from the sample.

The bootstrap statistic $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ is calculated for each bootstrap sample. This is repeated B times. B is known as the number of bootstrap replications. A theory for the determination of the number of bootstrap replications B has been developed by Andrews and Buchinsky (2000). It is desirable for B to be large, so long as the computational costs are reasonable. $B = 1000$ typically suffices.

When the statistic T_n is a function of F, it is typically through dependence on a parameter. For example, the t-ratio $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ depends on θ. As the bootstrap statistic replaces F with F_n, it similarly replaces θ with θ_n, the value of θ implied by F_n. Typically $\theta_n = \hat{\theta}$, the parameter estimate. (When in doubt use $\hat{\theta}$.)

Sampling from the EDF is particularly easy. Since F_n is a discrete probability distribution putting probability mass $1/n$ at each sample point, sampling from the EDF is equivalent to random sampling a pair $(y_i, \boldsymbol{x}_i)$ from the observed data **with replacement**. In consequence, a bootstrap sample $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$ will necessarily have some ties and multiple values, which is generally not a problem.

8.4 Bootstrap Estimation of Bias and Variance

The bias of $\hat{\theta}$ is $\tau_n = \mathbb{E}(\hat{\theta} - \theta_0)$. Let $T_n(\theta) = \hat{\theta} - \theta$. Then $\tau_n = \mathbb{E}(T_n(\theta_0))$. The bootstrap counterparts are $\hat{\theta}^* = \hat{\theta}((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*))$ and $T_n^* = \hat{\theta}^* - \theta_n = \hat{\theta}^* - \hat{\theta}$. The bootstrap estimate of τ_n is

$$\tau_n^* = \mathbb{E}(T_n^*).$$

If this is calculated by the simulation described in the previous section, the estimate of τ_n^* is

$$\begin{aligned} \hat{\tau}_n^* &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}. \end{aligned}$$

If $\hat{\boldsymbol{\theta}}$ is biased, it might be desirable to construct a biased-corrected estimator (one with reduced bias). Ideally, this would be

$$\tilde{\theta} = \hat{\theta} - \tau_n,$$

but τ_n is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned}\tilde{\theta}^* &= \hat{\theta} - \hat{\tau}_n^* \\ &= \hat{\theta} - (\overline{\hat{\theta}^*} - \hat{\theta}) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}.\end{aligned}$$

Note, in particular, that the biased-corrected estimator is *not* $\overline{\hat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\hat{\theta}$ is the truth. Then what is the average value of $\hat{\theta}$ calculated from such samples? The answer is $\overline{\hat{\theta}^*}$. If this is lower than $\hat{\theta}$, this suggests that the estimator is *downward-biased*, so a biased-corrected estimator of θ should be *larger* than $\hat{\theta}$, and the best guess is the difference between $\hat{\theta}$ and $\overline{\hat{\theta}^*}$. Similarly if $\overline{\hat{\theta}^*}$ is higher than $\hat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\hat{\theta}$.

Let $T_n = \hat{\theta}$. The variance of $\hat{\theta}$ is

$$V_n = \mathbb{E}(T_n - \mathbb{E}T_n)^2.$$

Let $T_n^* = \hat{\theta}^*$. It has variance

$$V_n^* = \mathbb{E}(T_n^* - \mathbb{E}T_n^*)^2.$$

The simulation estimate is

$$\hat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2.$$

A bootstrap standard error for $\hat{\theta}$ is the square root of the bootstrap estimate of variance, $s^*(\hat{\theta}) = \sqrt{\hat{V}_n^*}$.

While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct asymptotic confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspected. It appears superior to calculate bootstrap confidence intervals, and we turn to this next.

8.5 Percentile Intervals

For a distribution function $G_n(u, F)$, let $q_n(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q_n(\alpha, F), F) = \alpha.$$

[When $G_n(u, F)$ is discrete, $q_n(\alpha, F)$ may be non-unique, but we will ignore such complications.] Let $q_n(\alpha)$ denote the quantile function of the true sampling distribution, and $q_n^*(\alpha) = q_n(\alpha, F_n)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T_n whose distribution is G_n.

Let $T_n = \hat{\theta}$, an estimate of a parameter of interest. In $(1-\alpha)\%$ of samples, $\hat{\theta}$ lies in the region $[q_n(\alpha/2), q_n(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:

$$C_1 = [q_n^*(\alpha/2), \quad q_n^*(1-\alpha/2)].$$

This is often called the *percentile confidence interval*.

Computationally, the quantile $q_n^*(\alpha)$ is estimated by $\hat{q}_n^*(\alpha)$, the α'th sample quantile of the simulated statistics $\{T_{n1}^*, ..., T_{nB}^*\}$, as discussed in the section on Monte Carlo simulation. The $(1-\alpha)\%$ Efron percentile interval is then $[\hat{q}_n^*(\alpha/2), \quad \hat{q}_n^*(1-\alpha/2)]$.

The interval C_1 is a popular bootstrap confidence interval often used in empirical practice. This is because it is easy to compute, simple to motivate, was popularized by Efron early in the history of the bootstrap, and also has the feature that it is translation invariant. That is, if we define $\phi = f(\theta)$ as the parameter of interest for a monotonically increasing function f, then percentile method applied to this problem will produce the confidence interval $[f(q_n^*(\alpha/2)), \quad f(q_n^*(1-\alpha/2))]$, which is a naturally good property.

However, as we show now, C_1 is in a deep sense very poorly motivated.

It will be useful if we introduce an alternative definition C_1. Let $T_n(\theta) = \hat{\theta} - \theta$ and let $q_n(\alpha)$ be the quantile function of its distribution. (These are the original quantiles, with θ subtracted.) Then C_1 can alternatively be written as

$$C_1 = [\hat{\theta} + q_n^*(\alpha/2), \quad \hat{\theta} + q_n^*(1-\alpha/2)].$$

This is a bootstrap estimate of the "ideal" confidence interval

$$C_1^0 = [\hat{\theta} + q_n(\alpha/2), \quad \hat{\theta} + q_n(1-\alpha/2)].$$

The latter has coverage probability

$$\begin{aligned}
\mathbb{P}\left(\theta_0 \in C_1^0\right) &= \mathbb{P}\left(\hat{\theta} + q_n(\alpha/2) \leq \theta_0 \leq \hat{\theta} + q_n(1-\alpha/2)\right) \\
&= \mathbb{P}\left(-q_n(1-\alpha/2) \leq \hat{\theta} - \theta_0 \leq -q_n(\alpha/2)\right) \\
&= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0)
\end{aligned}$$

which generally is not $1-\alpha$! There is one important exception. If $\hat{\theta}-\theta_0$ has a symmetric distribution, then $G_n(-u, F_0) = 1 - G_n(u, F_0)$, so

$$\begin{aligned}
\mathbb{P}\left(\theta_0 \in C_1^0\right) &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \\
&= (1 - G_n(q_n(\alpha/2), F_0)) - (1 - G_n(q_n(1-\alpha/2), F_0)) \\
&= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha
\end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\hat{\theta}$ has a symmetric distribution about θ_0.

When $\hat{\theta}$ does not have a symmetric distribution, C_1 may perform quite poorly.

However, by the translation invariance argument presented above, it also follows that if there exists some monotonically increasing transformation $f(\cdot)$ such that $f(\hat{\theta})$ is symmetrically distributed about $f(\theta_0)$, then the idealized percentile bootstrap method will be accurate.

Based on these arguments, many argue that the percentile interval should not be used unless the sampling distribution is close to unbiased and symmetric.

The problems with the percentile method can be circumvented, at least in principle, by an alternative method.

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left(q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1-\alpha/2)\right) \\
&= \mathbb{P}\hat{\theta} - q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - q_n(\alpha/2),
\end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_2^0 = [\hat{\theta} - q_n(1-\alpha/2), \quad \hat{\theta} - q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_2 = [\hat{\theta} - q_n^*(1-\alpha/2), \quad \hat{\theta} - q_n^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval C_1! They coincide in the special case that $G_n^*(u)$ is symmetric about $\hat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)$, which are centered at the sample estimate $\hat{\theta}$. These are sorted to yield the quantile estimates $\hat{q}_n^*(.025)$ and $\hat{q}_n^*(.975)$. The 95% confidence interval is then $[\hat{\theta} - \hat{q}_n^*(.975), \quad \hat{\theta} - \hat{q}_n^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

8.6 Percentile-t Equal-Tailed Interval

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta < \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $T_n(\theta_0) < c$, where c would be selected so that

$$\mathbb{P}\left(T_n(\theta_0) < c\right) = \alpha.$$

Thus $c = q_n(\alpha)$. Since this is unknown, a bootstrap test replaces $q_n(\alpha)$ with the bootstrap estimate $q_n^*(\alpha)$, and the test rejects if $T_n(\theta_0) < q_n^*(\alpha)$.

Similarly, if the alternative is $\mathbb{H}_1 : \theta > \theta_0$, the bootstrap test rejects if $T_n(\theta_0) > q_n^*(1-\alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)/s(\hat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\hat{\theta}$, and the standard error $s(\hat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\hat{q}_n^*(\alpha)$ and/or $\hat{q}_n^*(1-\alpha)$.

Let $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Then taking the intersection of two one-sided intervals,

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(q_n(\alpha/2) \le T_n(\theta_0) \le q_n(1-\alpha/2)\right) \\ &= \mathbb{P}\left(q_n(\alpha/2) \le \left(\hat{\theta} - \theta_0\right)/s(\hat{\theta}) \le q_n(1-\alpha/2)\right) \\ &= \mathbb{P}\left(\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2) \le \theta_0 \le \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1-\alpha)$% confidence interval for θ_0 would be

$$C_3^0 = [\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_3 = [\hat{\theta} - s(\hat{\theta})q_n^*(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n^*(\alpha/2)].$$

This is often called a *percentile-t confidence interval.* It is *equal-tailed* or *central* since the probability that θ_0 is below the left endpoint approximately equals the probability that θ_0 is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

8.7 Symmetric Percentile-t Intervals

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T_n(\theta_0)| > c$, where c would be selected so that

$$\mathbb{P}\left(|T_n(\theta_0)| > c\right) = \alpha.$$

Note that

$$\begin{aligned}\mathbb{P}\left(|T_n(\theta_0)| < c\right) &= \mathbb{P}\left(-c < T_n(\theta_0) < c\right) \\ &= G_n(c) - G_n(-c) \\ &\equiv \overline{G}_n(c),\end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q_n(\alpha)$ solves the equation

$$\overline{G}_n(q_n(\alpha)) = 1 - \alpha.$$

Equivalently, $q_n(\alpha)$ is the $1-\alpha$ quantile of the distribution of $|T_n(\theta_0)|$.

The bootstrap estimate is $q_n^*(\alpha)$, the $1-\alpha$ quantile of the distribution of $|T_n^*|$, or the number which solves the equation

$$\overline{G}_n^*(q_n^*(\alpha)) = G_n^*(q_n^*(\alpha)) - G_n^*(-q_n^*(\alpha)) = 1 - \alpha.$$

Computationally, $q_n^*(\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T_n^*| = \left|\hat{\theta}^* - \hat{\theta}\right| / s(\hat{\theta}^*)$, and taking the upper $\alpha\%$ quantile. The bootstrap test rejects if $|T_n(\theta_0)| > q_n^*(\alpha)$.

Let

$$C_4 = [\hat{\theta} - s(\hat{\theta})q_n^*(\alpha), \quad \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)],$$

where $q_n^*(\alpha)$ is the bootstrap critical value for a two-sided hypothesis test. C_4 is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned}\mathbb{P}\left(\theta_0 \in C_4\right) &= \mathbb{P}\left(\hat{\theta} - s(\hat{\theta})q_n^*(\alpha) \le \theta_0 \le \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)\right) \\ &= \mathbb{P}\left(|T_n(\theta_0)| < q_n^*(\alpha)\right) \\ &\simeq \mathbb{P}\left(|T_n(\theta_0)| < q_n(\alpha)\right) \\ &= 1 - \alpha.\end{aligned}$$

If $\boldsymbol{\theta}$ is a vector, then to test $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ at size α, we would use a Wald statistic

$$W_n(\boldsymbol{\theta}) = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \hat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

or some other asymptotically chi-square statistic. Thus here $T_n(\boldsymbol{\theta}) = W_n(\boldsymbol{\theta})$. The ideal test rejects if $W_n \ge q_n(\alpha)$, where $q_n(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of W_n. The bootstrap test rejects if $W_n \ge q_n^*(\alpha)$, where $q_n^*(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of

$$W_n^* = n\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)' \hat{\mathbf{V}}_{\theta}^{*-1}\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right).$$

Computationally, the critical value $q_n^*(\alpha)$ is found as the quantile from simulated values of W_n^*. Note in the simulation that the Wald statistic is a quadratic form in $\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)$, not $\left(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\right)$. [This is a typical mistake made by practitioners.]

8.8 Asymptotic Expansions

Let $T_n \in \mathbb{R}$ be a statistic such that

$$T_n \xrightarrow{d} \mathrm{N}(0, \sigma^2). \tag{8.3}$$

In some cases, such as when T_n is a t-ratio, then $\sigma^2 = 1$. In other cases σ^2 is unknown. Equivalently, writing $T_n \sim G_n(u, F)$ then

$$\lim_{n\to\infty} G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right),$$

or

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + o(1). \tag{8.4}$$

While (8.4) says that G_n converges to $\Phi\left(\frac{u}{\sigma}\right)$ as $n \to \infty$, it says nothing, however, about the rate of convergence, or the size of the divergence for any particular sample size n. A better asymptotic approximation may be obtained through an *asymptotic expansion.*

The following notation will be helpful. Let a_n be a sequence.

Definition 8.8.1 $a_n = o(1)$ *if* $a_n \to 0$ *as* $n \to \infty$

Definition 8.8.2 $a_n = O(1)$ *if* $|a_n|$ *is uniformly bounded.*

Definition 8.8.3 $a_n = o(n^{-r})$ *if* $n^r |a_n| \to 0$ *as* $n \to \infty$.

Basically, $a_n = O(n^{-r})$ if it declines to zero like n^{-r}.

We say that a function $g(u)$ is *even* if $g(-u) = g(u)$, and a function $h(u)$ is *odd* if $h(-u) = -h(u)$. The derivative of an even function is odd, and vice-versa.

Theorem 8.8.1 *Under regularity conditions and (8.3),*

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F) + O(n^{-3/2})$$

uniformly over u, where g_1 is an even function of u, and g_2 is an odd function of u. Moreover, g_1 and g_2 are differentiable functions of u and continuous in F relative to the supremum norm on the space of distribution functions.

The expansion in Theorem 8.8.1 is often called an **Edgeworth expansion**.

We can interpret Theorem 8.8.1 as follows. First, $G_n(u, F)$ converges to the normal limit at rate $n^{1/2}$. To a second order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F).$$

Since the derivative of g_1 is odd, the density function is skewed. To a third order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F) + n^{-1} g_2(u, F)$$

which adds a symmetric non-normal component to the approximate density (for example, adding leptokurtosis).

[Side Note: When $T_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, a standardized sample mean, then

$$\begin{aligned} g_1(u) &= -\frac{1}{6}\kappa_3\left(u^2 - 1\right)\phi(u) \\ g_2(u) &= -\left(\frac{1}{24}\kappa_4\left(u^3 - 3u\right) + \frac{1}{72}\kappa_3^2\left(u^5 - 10u^3 + 15u\right)\right)\phi(u) \end{aligned}$$

where $\phi(u)$ is the standard normal pdf, and

$$\begin{aligned}\kappa_3 &= \mathbb{E}\left(X-\mu\right)^3/\sigma^3 \\ \kappa_4 &= \mathbb{E}\left(X-\mu\right)^4/\sigma^4 - 3\end{aligned}$$

the standardized skewness and excess kurtosis of the distribution of X. Note that when $\kappa_3 = 0$ and $\kappa_4 = 0$, then $g_1 = 0$ and $g_2 = 0$, so the second-order Edgeworth expansion corresponds to the normal distribution.]

Francis Edgeworth

Francis Ysidro Edgeworth (1845-1926) of Ireland, founding editor of the *Economic Journal*, was a profound economic and statistical theorist, developing the theories of indifference curves and asymptotic expansions. He also could be viewed as the first econometrician due to his early use of mathematical statistics in the study of economic data.

8.9 One-Sided Tests

Using the expansion of Theorem 8.8.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on an asymptotically normal t-ratio T_n. An asymptotic test is based on $\Phi(u)$.

To the second order, the exact distribution is

$$\mathbb{P}\left(T_n < u\right) = G_n(u, F_0) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, F_0) + O(n^{-1})$$

since $\sigma = 1$. The difference is

$$\begin{aligned}\Phi(u) - G_n(u, F_0) &= \frac{1}{n^{1/2}} g_1(u, F_0) + O(n^{-1}) \\ &= O(n^{-1/2}),\end{aligned}$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(u)$, which from Theorem 8.8.1 has the expansion

$$G_n^*(u) = G_n(u, F_n) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, F_n) + O(n^{-1}).$$

Because $\Phi(u)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(u) - G_n(u, F_0) = \frac{1}{n^{1/2}} \left(g_1(u, F_n) - g_1(u, F_0)\right) + O(n^{-1}).$$

Since F_n converges to F at rate $\sqrt{n}$, and g_1 is continuous with respect to F, the difference $(g_1(u, F_n) - g_1(u, F_0))$ converges to 0 at rate $\sqrt{n}$. Heuristically,

$$\begin{aligned}g_1(u, F_n) - g_1(u, F_0) &\approx \frac{\partial}{\partial F} g_1(u, F_0) \left(F_n - F_0\right) \\ &= O(n^{-1/2}),\end{aligned}$$

The "derivative" $\frac{\partial}{\partial F} g_1(u, F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(u) - G_n(u, F_0) = O(n^{-1}),$$

or

$$\mathbb{P}\left(T_n^* \leq u\right) = \mathbb{P}\left(T_n \leq u\right) + O(n^{-1}),$$

which is an improved rate of convergence over the asymptotic test (which converged at rate $O(n^{-1/2})$). This rate can be used to show that one-tailed bootstrap inference based on the t-ratio achieves a so-called *asymptotic refinement* – the Type I error of the test converges at a faster rate than an analogous asymptotic test.

8.10 Symmetric Two-Sided Tests

If a random variable y has distribution function $H(u) = \mathbb{P}(y \leq u)$, then the random variable $|y|$ has distribution function

$$\overline{H}(u) = H(u) - H(-u)$$

since

$$\begin{aligned} \mathbb{P}\left(|y| \leq u\right) &= \mathbb{P}\left(-u \leq y \leq u\right) \\ &= \mathbb{P}\left(y \leq u\right) - \mathbb{P}\left(y \leq -u\right) \\ &= H(u) - H(-u). \end{aligned}$$

For example, if $Z \sim \mathrm{N}(0,1)$, then $|Z|$ has distribution function

$$\overline{\Phi}(u) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1.$$

Similarly, if T_n has exact distribution $G_n(u, F)$, then $|T_n|$ has the distribution function

$$\overline{G}_n(u, F) = G_n(u, F) - G_n(-u, F).$$

A two-sided hypothesis test rejects $\mathbb{H}_0$ for large values of $|T_n|$. Since $T_n \xrightarrow{d} Z$, then $|T_n| \xrightarrow{d} |Z| \sim \overline{\Phi}$. Thus asymptotic critical values are taken from the $\overline{\Phi}$ distribution, and exact critical values are taken from the $\overline{G}_n(u, F_0)$ distribution. From Theorem 8.8.1, we can calculate that

$$\begin{aligned} \overline{G}_n(u, F) &= G_n(u, F) - G_n(-u, F) \\ &= \left(\Phi(u) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F)\right) \\ &\quad - \left(\Phi(-u) + \frac{1}{n^{1/2}} g_1(-u, F) + \frac{1}{n} g_2(-u, F)\right) + O(n^{-3/2}) \\ &= \overline{\Phi}(u) + \frac{2}{n} g_2(u, F) + O(n^{-3/2}), \end{aligned} \tag{8.5}$$

where the simplifications are because g_1 is even and g_2 is odd. Hence the difference between the asymptotic distribution and the exact distribution is

$$\overline{\Phi}(u) - \overline{G}_n(u, F_0) = \frac{2}{n} g_2(u, F_0) + O(n^{-3/2}) = O(n^{-1}).$$

The order of the error is $O(n^{-1})$.

Interestingly, the asymptotic two-sided test has a better coverage rate than the asymptotic one-sided test. This is because the first term in the asymptotic expansion, g_1, is an even function, meaning that the errors in the two directions exactly cancel out.

Applying (8.5) to the bootstrap distribution, we find

$$\overline{G}_n^*(u) = \overline{G}_n(u, F_n) = \overline{\Phi}(u) + \frac{2}{n} g_2(u, F_n) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned}\overline{G}_n^*(u) - \overline{G}_n(u, F_0) &= \frac{2}{n}\left(g_2(u, F_n) - g_2(u, F_0)\right) + O(n^{-3/2}) \\ &= O(n^{-3/2}),\end{aligned}$$

the last equality because F_n converges to F_0 at rate $\sqrt{n}$, and g_2 is continuous in F. Another way of writing this is

$$\mathbb{P}\left(|T_n^*| < u\right) = \mathbb{P}\left(|T_n| < u\right) + O(n^{-3/2})$$

so the error from using the bootstrap distribution (relative to the true unknown distribution) is $O(n^{-3/2})$. This is in contrast to the use of the asymptotic distribution, whose error is $O(n^{-1})$. Thus a two-sided bootstrap test also achieves an asymptotic refinement, similar to a one-sided test.

A reader might get confused between the two simultaneous effects. Two-sided tests have better rates of convergence than the one-sided tests, and bootstrap tests have better rates of convergence than asymptotic tests.

The analysis shows that there may be a trade-off between one-sided and two-sided tests. Two-sided tests will have more accurate size (Reported Type I error), but one-sided tests might have more power against alternatives of interest. Confidence intervals based on the bootstrap can be asymmetric if based on one-sided tests (equal-tailed intervals) and can therefore be more informative and have smaller length than symmetric intervals. Therefore, the choice between symmetric and equal-tailed confidence intervals is unclear, and needs to be determined on a case-by-case basis.

8.11 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T_n = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$. We know that $T_n \xrightarrow{d} \mathrm{N}(0, V)$, which is not pivotal, as it depends on the unknown V. Theorem 8.8.1 shows that a first-order approximation

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}),$$

where $\sigma = \sqrt{V}$, and for the bootstrap

$$G_n^*(u) = G_n(u, F_n) = \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}),$$

where $\hat{\sigma} = V(F_n)$ is the bootstrap estimate of σ. The difference is

$$\begin{aligned}G_n^*(u) - G_n(u, F_0) &= \Phi\left(\frac{u}{\sigma}\right) - \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}) \\ &= -\phi\left(\frac{u}{\sigma}\right)\frac{u}{\sigma}\left(\hat{\sigma} - \sigma\right) + O(n^{-1/2}) \\ &= O(n^{-1/2})\end{aligned}$$

Hence the order of the error is $O(n^{-1/2})$.

The good news is that the percentile-type methods (if appropriately used) can yield $\sqrt{n}$-convergent asymptotic inference. Yet these methods do not require the calculation of standard errors! This means that in contexts where standard errors are not available or are difficult to calculate, the percentile bootstrap methods provide an attractive inference method.

The bad news is that the rate of convergence is disappointing. It is no better than the rate obtained from an asymptotic one-sided confidence region. Therefore if standard errors are available, it is unclear if there are any benefits from using the percentile bootstrap over simple asymptotic methods.

Based on these arguments, the theoretical literature (e.g. Hall, 1992, Horowitz, 2001) tends to advocate the use of the percentile-t bootstrap methods rather than percentile methods.

8.12 Bootstrap Methods for Regression Models

The bootstrap methods we have discussed have set $G_n^*(u) = G_n(u, F_n)$, where F_n is the EDF. Any other consistent estimate of F may be used to define a feasible bootstrap estimator. The advantage of the EDF is that it is fully nonparametric, it imposes no conditions, and works in nearly any context. But since it is fully nonparametric, it may be inefficient in contexts where more is known about F. We discuss bootstrap methods appropriate for the linear regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

The non-parametric bootstrap resamples the observations $(y_i^*, \boldsymbol{x}_i^*)$ from the EDF, which implies

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^* \\ \mathbb{E}\left(\boldsymbol{x}_i^* e_i^*\right) &= 0 \end{aligned}$$

but generally

$$\mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i^*\right) \neq 0.$$

The the bootstrap distribution does not impose the regression assumption, and is thus an inefficient estimator of the true distribution (when in fact the regression assumption is true.)

One approach to this problem is to impose the very strong assumption that the error ε_i is independent of the regressor $\boldsymbol{x}_i$. The advantage is that in this case it is straightforward to construct bootstrap distributions. The disadvantage is that the bootstrap distribution may be a poor approximation when the error is not independent of the regressors.

To impose independence, it is sufficient to sample the $\boldsymbol{x}_i^*$ and e_i^* independently, and then create $y_i^* = \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^*$. There are different ways to impose independence. A non-parametric method is to sample the bootstrap errors e_i^* randomly from the OLS residuals $\{\hat{e}_1, ..., \hat{e}_n\}$. A parametric method is to generate the bootstrap errors e_i^* from a parametric distribution, such as the normal $e_i^* \sim \mathrm{N}(0, \hat{\sigma}^2)$.

For the regressors $\boldsymbol{x}_i^*$, a nonparametric method is to sample the $\boldsymbol{x}_i^*$ randomly from the EDF or sample values $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. A parametric method is to sample $\boldsymbol{x}_i^*$ from an estimated parametric distribution. A third approach sets $\boldsymbol{x}_i^* = \boldsymbol{x}_i$. This is equivalent to treating the regressors as *fixed in repeated samples*. If this is done, then all inferential statements are made conditionally on the observed values of the regressors, which is a valid statistical approach. It does not really matter, however, whether or not the $\boldsymbol{x}_i$ are really "fixed" or random.

The methods discussed above are unattractive for most applications in econometrics because they impose the stringent assumption that $\boldsymbol{x}_i$ and e_i are independent. Typically what is desirable is to impose only the regression condition $\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0$. Unfortunately this is a harder problem.

One proposal which imposes the regression condition without independence is the *Wild Bootstrap*. The idea is to construct a conditional distribution for e_i^* so that

$$\begin{aligned} \mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^{*2} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^2 \\ \mathbb{E}\left(e_i^{*3} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^3. \end{aligned}$$

A conditional distribution with these features will preserve the main important features of the data. This can be achieved using a two-point distribution of the form

$$\begin{aligned} \mathbb{P}\left(e_i^* = \left(\frac{1+\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \mathbb{P}\left(e_i^* = \left(\frac{1-\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}} \end{aligned}$$

For each $\boldsymbol{x}_i$, you sample e_i^* using this two-point distribution.

Exercises

Exercise 8.1 Let $F_n(\boldsymbol{x})$ denote the EDF of a random sample. Show that

$$\sqrt{n}\left(F_n(\boldsymbol{x}) - F_0(\boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F_0(\boldsymbol{x})\left(1 - F_0(\boldsymbol{x})\right)\right).$$

Exercise 8.2 Take a random sample $\{y_1, ..., y_n\}$ with $\mu = \mathbb{E}y_i$ and $\sigma^2 = \mathrm{var}\left(y_i\right)$. Let the statistic of interest be the sample mean $T_n = \overline{y}_n$. Find the population moments $\mathbb{E}T_n$ and $\mathrm{var}\left(T_n\right)$. Let $\{y_1^*, ..., y_n^*\}$ be a random sample from the empirical distribution function and let $T_n^* = \overline{y}_n^*$ be its sample mean. Find the bootstrap moments $\mathbb{E}T_n^*$ and $\mathrm{var}\left(T_n^*\right)$.

Exercise 8.3 Consider the following bootstrap procedure for a regression of y_i on $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\beta}}$ denote the OLS estimator from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the OLS residuals.

(a) Draw a random vector $(\boldsymbol{x}^*, e^*)$ from the pair $\{(\boldsymbol{x}_i, \hat{e}_i) : i = 1, ..., n\}$. That is, draw a random integer i' from $[1, 2, ..., n]$, and set $\boldsymbol{x}^* = \boldsymbol{x}_{i'}$ and $e^* = \hat{e}_{i'}$. Set $y^* = \boldsymbol{x}^{*\prime}\hat{\boldsymbol{\beta}} + e^*$. Draw (with replacement) n such vectors, creating a random bootstrap data set $(\boldsymbol{y}^*, \boldsymbol{X}^*)$.

(b) Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$, yielding OLS estimates $\hat{\boldsymbol{\beta}}^*$ and any other statistic of interest.

Show that this bootstrap procedure is (numerically) *identical* to the non-parametric bootstrap.

Exercise 8.4 Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = (\hat{\theta}^* - \hat{\theta})/s(\hat{\theta}),$$

where $s(\hat{\theta})$ is the standard error in the original data. Let $q_n^*(.05)$ and $q_n^*(.95)$ denote the 5% and 95% quantiles of T_n^*, and define the bootstrap confidence interval

$$C = \left[\hat{\theta} - s(\hat{\theta})q_n^*(.95), \quad \hat{\theta} - s(\hat{\theta})q_n^*(.05)\right].$$

Show that C exactly equals the Alternative percentile interval (not the percentile-t interval).

Exercise 8.5 You want to test $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. The test for $\mathbb{H}_0$ is to reject if $T_n = \hat{\theta}/s(\hat{\theta}) > c$ where c is picked so that Type I error is α. You do this as follows. Using the non-parametric bootstrap, you generate bootstrap samples, calculate the estimates $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = \hat{\theta}^*/s(\hat{\theta}^*).$$

Let $q_n^*(.95)$ denote the 95% quantile of T_n^*. You replace c with $q_n^*(.95)$, and thus reject $\mathbb{H}_0$ if $T_n = \hat{\theta}/s(\hat{\theta}) > q_n^*(.95)$. What is wrong with this procedure?

Exercise 8.6 Suppose that in an application, $\hat{\theta} = 1.2$ and $s(\hat{\theta}) = .2$. Using the non-parametric bootstrap, 1000 samples are generated from the bootstrap distribution, and $\hat{\theta}^*$ is calculated on each sample. The $\hat{\theta}^*$ are sorted, and the 2.5% and 97.5% quantiles of the $\hat{\theta}^*$ are .75 and 1.3, respectively.

(a) Report the 95% Efron Percentile interval for θ.

(b) Report the 95% Alternative Percentile interval for θ.

(c) With the given information, can you report the 95% Percentile-t interval for θ?

Exercise 8.7 The datafile `hprice1.dat` contains data on house prices (sales), with variables listed in the file `hprice1.pdf`. Estimate a linear regression of price on the number of bedrooms, lot size, size of house, and the colonial dummy. Calculate 95% confidence intervals for the regression coefficients using both the asymptotic normal approximation and the percentile-t bootstrap.

Chapter 9

Generalized Method of Moments

9.1 Overidentified Linear Model

Consider the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

where $\boldsymbol{x}_{1i}$ is $k \times 1$ and $\boldsymbol{x}_2$ is $r \times 1$ with $\ell = k + r$. We know that without further restrictions, an asymptotically efficient estimator of $\boldsymbol{\beta}$ is the OLS estimator. Now suppose that we are given the information that $\boldsymbol{\beta}_2 = \mathbf{0}$. Now we can write the model as

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

In this case, how should $\boldsymbol{\beta}_1$ be estimated? One method is OLS regression of y_i on $\boldsymbol{x}_{1i}$ alone. This method, however, is not necessarily efficient, as there are ℓ restrictions in $\mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \mathbf{0}$, while $\boldsymbol{\beta}_1$ is of dimension $k < \ell$. This situation is called **overidentified**. There are $\ell - k = r$ more moment restrictions than free parameters. We call r the **number of overidentifying restrictions**.

This is a special case of a more general class of moment condition models. Let $\boldsymbol{g}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta})$ be an $\ell \times 1$ function of a $k \times 1$ parameter $\boldsymbol{\beta}$ with $\ell \geq k$ such that

$$\mathbb{E}\boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\beta}_0) = \mathbf{0} \tag{9.1}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. In our previous example, $\boldsymbol{g}(y, \boldsymbol{z}, \boldsymbol{\beta}) = \boldsymbol{z}\cdot(y - \boldsymbol{x}_1'\boldsymbol{\beta})$. In econometrics, this class of models are called **moment condition models**. In the statistics literature, these are known as **estimating equations**.

As an important special case we will devote special attention to linear moment condition models, which can be written as

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

where the dimensions of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are $k \times 1$ and $\ell \times 1$, with $\ell \geq k$. If $k = \ell$ the model is **just identified**, otherwise it is **overidentified**. The variables $\boldsymbol{x}_i$ may be components and functions of $\boldsymbol{z}_i$, but this is not required. This model falls in the class (9.1) by setting

$$\boldsymbol{g}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}_0) = \boldsymbol{z}\cdot(y - \boldsymbol{x}'\boldsymbol{\beta}) \tag{9.2}$$

9.2 GMM Estimator

Define the sample analog of (9.2)

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) = \frac{1}{n}\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right). \tag{9.3}$$

The method of moments estimator for $\boldsymbol{\beta}$ is defined as the parameter value which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \mathbf{0}$. This is generally not possible when $\ell > k$, as there are more equations than free parameters. The idea of the generalized method of moments (GMM) is to define an estimator which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ "close" to zero.

For some $\ell \times \ell$ weight matrix $\boldsymbol{W}_n > 0$, let

$$J_n(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{W}_n \overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

This is a non-negative measure of the "length" of the vector $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$. For example, if $\boldsymbol{W}_n = \boldsymbol{I}$, then, $J_n(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n \cdot \left\|\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})\right\|^2$, the square of the Euclidean length. The GMM estimator minimizes $J_n(\boldsymbol{\beta})$.

Definition 9.2.1 $\boldsymbol{\beta}_{GMM} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} J_n(\boldsymbol{\beta})$.

Note that if $k = \ell$, then $\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, and the GMM estimator is the method of moments estimator. The first order conditions for the GMM estimator are

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} J_n(\hat{\boldsymbol{\beta}}) \\
&= 2\frac{\partial}{\partial \boldsymbol{\beta}} \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})' \boldsymbol{W}_n \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) \\
&= -2\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\frac{1}{n}\boldsymbol{Z}'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right)
\end{aligned}$$

so

$$2\left(\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\boldsymbol{Z}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} = 2\left(\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\boldsymbol{Z}'\boldsymbol{y}\right)$$

which establishes the following.

Proposition 9.2.1

$$\hat{\boldsymbol{\beta}}_{GMM} = \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\boldsymbol{Z}'\boldsymbol{y}\right).$$

While the estimator depends on $\boldsymbol{W}_n$, the dependence is only up to scale, for if $\boldsymbol{W}_n$ is replaced by $c\boldsymbol{W}_n$ for some $c > 0$, $\hat{\boldsymbol{\beta}}_{GMM}$ does not change.

9.3 Distribution of GMM Estimator

Assume that $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{W} > 0$. Let

$$\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)$$

and

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i' e_i^2\right) = \mathbb{E}\left(\boldsymbol{g}_i \boldsymbol{g}_i'\right),$$

where $\boldsymbol{g}_i = \boldsymbol{z}_i e_i$. Then

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right) \xrightarrow{p} \boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}$$

and

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right) \xrightarrow{d} \boldsymbol{Q}'\boldsymbol{W}\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right).$$

We conclude:

Theorem 9.3.1 ***Asymptotic Distribution of GMM Estimator***

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1} \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\right) \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}.$$

In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances.

The optimal weight matrix $\boldsymbol{W}_0$ is one which minimizes $\boldsymbol{V}_{\boldsymbol{\beta}}$. This turns out to be $\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$. The proof is left as an exercise. This yields the *efficient GMM* estimator:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Thus we have

Theorem 9.3.2 ***Asymptotic Distribution of Efficient GMM Estimator***

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}\right).$$

$\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$ is not known in practice, but it can be estimated consistently. For any $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{W}_0$, we still call $\hat{\boldsymbol{\beta}}$ the efficient GMM estimator, as it has the same asymptotic distribution.

By "efficient", we mean that this estimator has the smallest asymptotic variance in the class of GMM estimators with this set of moment conditions. This is a weak concept of optimality, as we are only considering alternative weight matrices $\boldsymbol{W}_n$. However, it turns out that the GMM estimator is semiparametrically efficient, as shown by Gary Chamberlain (1987).

If it is known that $\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = 0$, and this is all that is known, this is a semi-parametric problem, as the distribution of the data is unknown. Chamberlain showed that in this context, no semiparametric estimator (one which is consistent globally for the class of models considered) can have a smaller asymptotic variance than $\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}$ where $\boldsymbol{G} = \mathbb{E}\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta})$. Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

This result shows that in the linear model, no estimator has greater asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

9.4 Estimation of the Efficient Weight Matrix

Given any weight matrix $\boldsymbol{W}_n > 0$, the GMM estimator $\hat{\boldsymbol{\beta}}$ is consistent yet inefficient. For example, we can set $\boldsymbol{W}_n = \boldsymbol{I}_\ell$. In the linear model, a better choice is $\boldsymbol{W}_n = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Given any such first-step estimator, we can define the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ and moment equations $\hat{\boldsymbol{g}}_i = \boldsymbol{z}_i\hat{e}_i = \boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \hat{\boldsymbol{\beta}})$. Construct

$$\overline{\boldsymbol{g}}_n = \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i,$$

$$\hat{\boldsymbol{g}}_i^* = \hat{\boldsymbol{g}}_i - \overline{\boldsymbol{g}}_n,$$

and define

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i^*\hat{\boldsymbol{g}}_i^{*\prime}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i' - \overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}. \tag{9.4}$$

Then $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{\Omega}^{-1} = \boldsymbol{W}_0$, and GMM using $\boldsymbol{W}_n$ as the weight matrix is asymptotically efficient.

A common alternative choice is to set

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i'\right)^{-1}$$

which uses the uncentered moment conditions. Since $\mathbb{E}\boldsymbol{g}_i = 0$, these two estimators are asymptotically equivalent under the hypothesis of correct specification. However, Alastair Hall (2000) has shown that the uncentered estimator is a poor choice. When constructing hypothesis tests, under the alternative hypothesis the moment conditions are violated, i.e. $\mathbb{E}\boldsymbol{g}_i \neq 0$, so the uncentered estimator will contain an undesirable bias term and the power of the test will be adversely affected. A simple solution is to use the centered moment conditions to construct the weight matrix, as in (9.4) above.

Here is a simple way to compute the efficient GMM estimator for the linear model. First, set $\boldsymbol{W}_n = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, estimate $\hat{\boldsymbol{\beta}}$ using this weight matrix, and construct the residual $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$. Then set $\hat{\boldsymbol{g}}_i = \boldsymbol{z}_i\hat{e}_i$, and let $\hat{\boldsymbol{g}}$ be the associated $n \times \ell$ matrix. Then the efficient GMM estimator is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

In most cases, when we say "GMM", we actually mean "efficient GMM". There is little point in using an inefficient GMM estimator when the efficient estimator is easy to compute.

An estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ can be seen from the above formula. Set

$$\hat{\boldsymbol{V}} = n\left(\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}.$$

Asymptotic standard errors are given by the square roots of the diagonal elements of $\hat{\boldsymbol{V}}$.

There is an important alternative to the two-step GMM estimator just described. Instead, we can let the weight matrix be considered as a function of $\boldsymbol{\beta}$. The criterion function is then

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i^*(\boldsymbol{\beta})\boldsymbol{g}_i^*(\boldsymbol{\beta})'\right)^{-1}\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

where

$$\boldsymbol{g}_i^*(\boldsymbol{\beta}) = \boldsymbol{g}_i(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

The $\hat{\boldsymbol{\beta}}$ which minimizes this function is called the **continuously-updated GMM estimator**, and was introduced by L. Hansen, Heaton and Yaron (1996).

The estimator appears to have some better properties than traditional GMM, but can be numerically tricky to obtain in some cases. This is a current area of research in econometrics.

9.5 GMM: The General Case

In its most general form, GMM applies whenever an economic or statistical model implies the $\ell \times 1$ moment condition

$$\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \mathbf{0}.$$

Often, this is *all* that is known. Identification requires $l \geq k = \dim(\boldsymbol{\beta})$. The GMM estimator minimizes

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{W}_n \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

where

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta})$$

and

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{g}}_i \hat{\boldsymbol{g}}_i' - \overline{\boldsymbol{g}}_n \overline{\boldsymbol{g}}_n'\right)^{-1},$$

with $\hat{\boldsymbol{g}}_i = \boldsymbol{g}_i(\tilde{\boldsymbol{\beta}})$ constructed using a preliminary consistent estimator $\tilde{\boldsymbol{\beta}}$, perhaps obtained by first setting $\boldsymbol{W}_n = \boldsymbol{I}$. Since the GMM estimator depends upon the first-stage estimator, often the weight matrix $\boldsymbol{W}_n$ is updated, and then $\hat{\boldsymbol{\beta}}$ recomputed. This estimator can be iterated if needed.

Theorem 9.5.1 ***Distribution of Nonlinear GMM Estimator***
Under general regularity conditions,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\right),$$

where

$$\boldsymbol{\Omega} = \left(\mathbb{E}\left(\boldsymbol{g}_i \boldsymbol{g}_i'\right)\right)^{-1}$$

and

$$\boldsymbol{G} = \mathbb{E}\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta}).$$

The variance of $\hat{\boldsymbol{\beta}}$ may be estimated by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{G}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}\right)^{-1}$$

where

$$\hat{\boldsymbol{\Omega}} = n^{-1}\sum_{i} \hat{\boldsymbol{g}}_i^* \hat{\boldsymbol{g}}_i^{*\prime}$$

and

$$\hat{\boldsymbol{G}} = n^{-1}\sum_{i} \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}).$$

The general theory of GMM estimation and testing was exposited by L. Hansen (1982).

9.6 Over-Identification Test

Overidentified models ($\ell > k$) are special in the sense that there may not be a parameter value $\boldsymbol{\beta}$ such that the moment condition

$$\mathbb{E}\boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\beta}) = \mathbf{0}$$

holds. Thus the model – the overidentifying restrictions – are testable.

For example, take the linear model $y_i = \boldsymbol{\beta}_1'\boldsymbol{x}_{1i} + \boldsymbol{\beta}_2'\boldsymbol{x}_{2i} + e_i$ with $\mathbb{E}(\boldsymbol{x}_{1i}e_i) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{x}_{\boldsymbol{2i}}e_i) = \mathbf{0}$. It is possible that $\boldsymbol{\beta}_2 = \mathbf{0}$, so that the linear equation may be written as $y_i = \boldsymbol{\beta}_1'\boldsymbol{x}_{1i} + e_i$. However, it is possible that $\boldsymbol{\beta}_2 \neq \mathbf{0}$, and in this case it would be impossible to find a value of $\boldsymbol{\beta}_1$ so that both $\mathbb{E}(\boldsymbol{x}_{1i}(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1)) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{x}_{2i}(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1)) = \mathbf{0}$ hold simultaneously. In this sense an exclusion restriction can be seen as an overidentifying restriction.

Note that $\overline{\boldsymbol{g}}_n \xrightarrow{p} \mathbb{E}\boldsymbol{g}_i$, and thus $\overline{\boldsymbol{g}}_n$ can be used to assess whether or not the hypothesis that $\mathbb{E}\boldsymbol{g}_i = \mathbf{0}$ is true or not. The criterion function at the parameter estimates is

$$\begin{aligned} J &= n\,\overline{\boldsymbol{g}}_n'\boldsymbol{W}_n\overline{\boldsymbol{g}}_n \\ &= n^2\overline{\boldsymbol{g}}_n'\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\overline{\boldsymbol{g}}_n. \end{aligned}$$

is a quadratic form in $\overline{\boldsymbol{g}}_n$, and is thus a natural test statistic for $\mathbb{H}_0 : \mathbb{E}\boldsymbol{g}_i = \mathbf{0}$.

Theorem 9.6.1 *(Sargan-Hansen). Under the hypothesis of correct specification, and if the weight matrix is asymptotically efficient,*

$$J = J(\hat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2_{\ell-k}.$$

The proof of the theorem is left as an exercise. This result was established by Sargan (1958) for a specialized case, and by L. Hansen (1982) for the general case.

The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. If the statistic J exceeds the chi-square critical value, we can reject the model. Based on this information alone, it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J whenever GMM is the estimation method.

When over-identified models are estimated by GMM, it is customary to report the J statistic as a general test of model adequacy.

9.7 Hypothesis Testing: The Distance Statistic

We described before how to construct estimates of the asymptotic covariance matrix of the GMM estimates. These may be used to construct Wald tests of statistical hypotheses.

If the hypothesis is non-linear, a better approach is to directly use the GMM criterion function. This is sometimes called the GMM Distance statistic, and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987).

For a given weight matrix $\boldsymbol{W}_n$, the GMM criterion function is

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\boldsymbol{W}_n\,\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

For $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^r$, the hypothesis is

$$\mathbb{H}_0 : \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}.$$

The estimates under $\mathbb{H}_1$ are

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, J(\boldsymbol{\beta})$$

and those under $\mathbb{H}_0$ are

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} J(\boldsymbol{\beta}).$$

The two minimizing criterion functions are $J(\hat{\boldsymbol{\beta}})$ and $J(\tilde{\boldsymbol{\beta}})$. The GMM distance statistic is the difference

$$D = J(\tilde{\boldsymbol{\beta}}) - J(\hat{\boldsymbol{\beta}}).$$

Proposition 9.7.1 *If the same weight matrix $\boldsymbol{W}_n$ is used for both null and alternative,*

1. $D \geq 0$
2. $D \xrightarrow{d} \chi^2_r$
3. *If $\boldsymbol{h}$ is linear in $\boldsymbol{\beta}$, then D equals the Wald statistic.*

If $\boldsymbol{h}$ is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D statistic appears to have quite good sampling properties, and is the preferred test statistic.

Newey and West (1987) suggested to use the same weight matrix $\boldsymbol{W}_n$ for both null and alternative, as this ensures that $D \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test.

This test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

9.8 Conditional Moment Restrictions

In many contexts, the model implies more than an unconditional moment restriction of the form $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}$. It implies a conditional moment restriction of the form

$$\mathbb{E}\left(\boldsymbol{e}_i(\boldsymbol{\beta}) \mid \boldsymbol{z}_i\right) = \mathbf{0}$$

where $\boldsymbol{e}_i(\boldsymbol{\beta})$ is some $s \times 1$ function of the observation and the parameters. In many cases, $s = 1$.

It turns out that this conditional moment restriction is much more powerful, and restrictive, than the unconditional moment restriction discussed above.

Our linear model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with instruments $\boldsymbol{z}_i$ falls into this class under the stronger assumption $\mathbb{E}\left(e_i \mid \boldsymbol{z}_i\right) = 0$. Then $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$.

It is also helpful to realize that conventional regression models also fall into this class, except that in this case $\boldsymbol{x}_i = \boldsymbol{z}_i$. For example, in linear regression, $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, while in a nonlinear regression model $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$. In a joint model of the conditional mean and variance

$$\boldsymbol{e}_i\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \begin{cases} y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \\ \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 - f\left(\boldsymbol{x}_i\right)'\boldsymbol{\gamma} \end{cases}.$$

Here $s = 2$.

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any $\ell \times 1$ function $\boldsymbol{\phi}\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)$, we can set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{\phi}\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) e_i(\boldsymbol{\beta})$ which satisfies $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}$ and hence defines a GMM estimator. The obvious problem is that the class of functions $\boldsymbol{\phi}$ is infinite. Which should be selected?

This is equivalent to the problem of selection of the best instruments. If $x_i \in \mathbb{R}$ is a valid instrument satisfying $\mathbb{E}(e_i \mid x_i) = 0$, then x_i, x_i^2, x_i^3, ..., etc., are all valid instruments. Which should be used?

One solution is to construct an infinite list of potent instruments, and then use the first k instruments. How is k to be determined? This is an area of theory still under development. A recent study of this problem is Donald and Newey (2001).

Another approach is to construct the *optimal instrument*. The form was uncovered by Chamberlain (1987). Take the case $s = 1$. Let

$$\boldsymbol{R}_i = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} e_i(\boldsymbol{\beta}) \mid \boldsymbol{z}_i\right)$$

and

$$\sigma_i^2 = \mathbb{E}\left(e_i(\boldsymbol{\beta})^2 \mid \boldsymbol{z}_i\right).$$

Then the "optimal instrument" is

$$\boldsymbol{A}_i = -\sigma_i^{-2}\boldsymbol{R}_i$$

so the optimal moment is

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i e_i(\boldsymbol{\beta}).$$

Setting $\boldsymbol{g}_i(\boldsymbol{\beta})$ to be this choice (which is $k \times 1$, so is just-identified) yields the best GMM estimator possible.

In practice, $\boldsymbol{A}_i$ is unknown, but its form does help us think about construction of optimal instruments.

In the linear model $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, note that

$$\boldsymbol{R}_i = -\mathbb{E}(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$$

and

$$\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right),$$

so

$$\boldsymbol{A}_i = \sigma_i^{-2}\mathbb{E}(\boldsymbol{x}_i \mid \boldsymbol{z}_i).$$

In the case of linear regression, $\boldsymbol{x}_i = \boldsymbol{z}_i$, so $\boldsymbol{A}_i = \sigma_i^{-2}\boldsymbol{z}_i$. Hence efficient GMM is GLS, as we discussed earlier in the course.

In the case of endogenous variables, note that the efficient instrument $\boldsymbol{A}_i$ involves the estimation of the conditional mean of $\boldsymbol{x}_i$ given $\boldsymbol{z}_i$. In other words, to get the best instrument for $\boldsymbol{x}_i$, we need the best conditional mean model for $\boldsymbol{x}_i$ given $\boldsymbol{z}_i$, not just an arbitrary linear projection. The efficient instrument is also inversely proportional to the conditional variance of e_i. This is the same as the GLS estimator; namely that improved efficiency can be obtained if the observations are weighted inversely to the conditional variance of the errors.

9.9 Bootstrap GMM Inference

Let $\hat{\boldsymbol{\beta}}$ be the 2SLS or GMM estimator of $\boldsymbol{\beta}$. Using the EDF of $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, we can apply the bootstrap methods discussed in Chapter 8 to compute estimates of the bias and variance of $\hat{\boldsymbol{\beta}}$, and construct confidence intervals for $\boldsymbol{\beta}$, identically as in the regression model. However, caution should be applied when interpreting such results.

A straightforward application of the nonparametric bootstrap works in the sense of consistently achieving the first-order asymptotic distribution. This has been shown by Hahn (1996). However, it fails to achieve an asymptotic refinement when the model is over-identified, jeopardizing the theoretical justification for percentile-t methods. Furthermore, the bootstrap applied J test will yield the wrong answer.

The problem is that in the sample, $\hat{\boldsymbol{\beta}}$ is the "true" value and yet $\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) \neq 0$. Thus according to random variables $(y_i^*, \boldsymbol{z}_i^*, \boldsymbol{x}_i^*)$ drawn from the EDF F_n,

$$\mathbb{E}\left(\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right) = \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) \neq \mathbf{0}.$$

This means that $(y_i^*, \boldsymbol{z}_i^*, \boldsymbol{x}_i^*)$ do not satisfy the same moment conditions as the population distribution.

A correction suggested by Hall and Horowitz (1996) can solve the problem. Given the bootstrap sample $(\boldsymbol{y}^*, \boldsymbol{Z}^*, \boldsymbol{X}^*)$, define the bootstrap GMM criterion

$$J^*(\boldsymbol{\beta}) = n \cdot \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right)' \boldsymbol{W}_n^* \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right)$$

where $\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})$ is from the in-sample data, not from the bootstrap data.

Let $\hat{\boldsymbol{\beta}}^*$ minimize $J^*(\boldsymbol{\beta})$, and define all statistics and tests accordingly. In the linear model, this implies that the bootstrap estimator is

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\boldsymbol{W}_n^*\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\boldsymbol{W}_n^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{y}^* - \boldsymbol{Z}'\hat{\boldsymbol{e}}\right)\right).$$

where $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are the in-sample residuals. The bootstrap J statistic is $J^*(\hat{\boldsymbol{\beta}}^*)$.

Brown and Newey (2002) have an alternative solution. They note that we can sample from the observations with the empirical likelihood probabilities $\hat{p}_i$ described in Chapter 10. Since $\sum_{i=1}^n \hat{p}_i \boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$, this sampling scheme preserves the moment conditions of the model, so no recentering or adjustments is needed. Brown and Newey argue that this bootstrap procedure will be more efficient than the Hall-Horowitz GMM bootstrap.

Exercises

Exercise 9.1 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \\ e_i^2 &= \boldsymbol{z}_i'\boldsymbol{\gamma} + \eta_i \\ \mathbb{E}\left(\boldsymbol{z}_i \eta_i\right) &= \mathbf{0}. \end{aligned}$$

Find the method of moments estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Exercise 9.2 Take the single equation

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{Z}\right) &= \mathbf{0} \end{aligned}$$

Assume $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$. Show that if $\hat{\boldsymbol{\beta}}$ is estimated by GMM with weight matrix $\boldsymbol{W}_n = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \sigma^2\left(\boldsymbol{Q}'\boldsymbol{M}^{-1}\boldsymbol{Q}\right)^{-1}\right)$$

where $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$ and $\boldsymbol{M} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$.

Exercise 9.3 Take the model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with $\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = 0$. Let $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (e.g. a GMM estimator with arbitrary weight matrix). Define the estimate of the optimal GMM weight matrix

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'\hat{e}_i^2\right)^{-1}.$$

Show that $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{\Omega}^{-1}$ where $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Exercise 9.4 In the linear model estimated by GMM with general weight matrix $\boldsymbol{W}$, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{GMM}$ is

$$\boldsymbol{V} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}$$

(a) Let $\boldsymbol{V}_0$ be this matrix when $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$. Show that $\boldsymbol{V}_0 = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}$.

(b) We want to show that for any $\boldsymbol{W}$, $\boldsymbol{V} - \boldsymbol{V}_0$ is positive semi-definite (for then $\boldsymbol{V}_0$ is the smaller possible covariance matrix and $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$ is the efficient weight matrix). To do this, start by finding matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{\Omega}\boldsymbol{A}$ and $\boldsymbol{V}_0 = \boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{B}$.

(c) Show that $\boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{A} = \boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{B}$ and therefore that $\boldsymbol{B}'\boldsymbol{\Omega}\left(\boldsymbol{A} - \boldsymbol{B}\right) = \mathbf{0}$.

(d) Use the expressions $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{\Omega}\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{B} + \left(\boldsymbol{A} - \boldsymbol{B}\right)$, and $\boldsymbol{B}'\boldsymbol{\Omega}\left(\boldsymbol{A} - \boldsymbol{B}\right) = \mathbf{0}$ to show that $\boldsymbol{V} \geq \boldsymbol{V}_0$.

Exercise 9.5 The equation of interest is

$$\begin{aligned} y_i &= \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

The observed data is $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$. $\boldsymbol{z}_i$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$, $\ell \geq k$. Show how to construct an efficient GMM estimator for $\boldsymbol{\beta}$.

Exercise 9.6 In the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$, a Generalized Method of Moments (GMM) criterion function for $\boldsymbol{\beta}$ is defined as

$$J_n(\boldsymbol{\beta}) = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)' \boldsymbol{X}\hat{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{9.5}$$

where $\hat{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2$, $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ are the OLS residuals, and $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is LS. The GMM estimator of $\boldsymbol{\beta}$, subject to the restriction $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{0}$, is defined as

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}} J_n(\boldsymbol{\beta}).$$

The GMM test statistic (the distance statistic) of the hypothesis $\boldsymbol{h}(\boldsymbol{\beta}) = 0$ is

$$D = J_n(\tilde{\boldsymbol{\beta}}) = \min_{\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{0}} J_n(\boldsymbol{\beta}). \tag{9.6}$$

(a) Show that you can rewrite $J_n(\boldsymbol{\beta})$ in (9.5) as

$$J_n(\boldsymbol{\beta}) = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \hat{\boldsymbol{V}}_n^{-1}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$$

where

$$\hat{\boldsymbol{V}}_n = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

(b) Now focus on linear restrictions: $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{r}$. Thus

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}-\boldsymbol{r}=\boldsymbol{0}}{\operatorname{argmin}} J_n(\boldsymbol{\beta})$$

and hence $\boldsymbol{R}'\tilde{\boldsymbol{\beta}} = \boldsymbol{r}$. Define the Lagrangian $(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}J_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{r}\right)$ where $\boldsymbol{\lambda}$ is $s \times 1$. Show that the minimizer is

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{V}}_n\boldsymbol{R}\left(\boldsymbol{R}_n'\hat{\boldsymbol{V}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right) \\
\hat{\boldsymbol{\lambda}} &= \left(\boldsymbol{R}_n'\hat{\boldsymbol{V}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\hat{\boldsymbol{\beta}} - \boldsymbol{r}\right).
\end{aligned}$$

(c) Show that if $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{r}$ then $\sqrt{n}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{R}}\right)$ where

$$\boldsymbol{V}_{\boldsymbol{R}} = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}.$$

(d) Show that in this setting, the distance statistic D in (9.6) equals the Wald statistic.

Exercise 9.7 Take the linear model

$$\begin{aligned}
y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\
\mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \boldsymbol{0}.
\end{aligned}$$

and consider the GMM estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Let

$$J_n = n\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Omega}}^{-1}\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})$$

denote the test of overidentifying restrictions. Show that $J_n \xrightarrow{d} \chi^2_{\ell-k}$ as $n \to \infty$ by demonstrating each of the following:

(a) Since $\boldsymbol{\Omega} > 0$, we can write $\boldsymbol{\Omega}^{-1} = \boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{\Omega} = \boldsymbol{C}'^{-1}\boldsymbol{C}^{-1}$

(b) $J_n = n\left(\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right)'\left(\boldsymbol{C}'\hat{\boldsymbol{\Omega}}\boldsymbol{C}\right)^{-1}\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})$

(c) $\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{D}_n\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)$ where

$$\begin{aligned}\boldsymbol{D}_n &= \boldsymbol{I}_\ell - \boldsymbol{C}'\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\boldsymbol{\Omega}}^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{C}'^{-1} \\ \overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) &= \frac{1}{n}\boldsymbol{Z}'e.\end{aligned}$$

(d) $\boldsymbol{D}_n \xrightarrow{p} \boldsymbol{I}_\ell - \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'$ where $\boldsymbol{R} = \boldsymbol{C}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$

(e) $n^{1/2}\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) \xrightarrow{d} \boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_\ell\right)$

(f) $J_n \xrightarrow{d} \boldsymbol{X}'\left(\boldsymbol{I}_\ell - \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\boldsymbol{X}$

(g) $\boldsymbol{X}'\left(\boldsymbol{I}_\ell - \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\boldsymbol{X} \sim \chi^2_{\ell-k}$.

Hint: $\boldsymbol{I}_\ell - \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'$ is a projection matrix.

Chapter 10

Empirical Likelihood

10.1 Non-Parametric Likelihood

An alternative to GMM is **empirical likelihood**. The idea is due to Art Owen (1988, 2001) and has been extended to moment condition models by Qin and Lawless (1994). It is a non-parametric analog of likelihood estimation.

The idea is to construct a multinomial distribution $F(p_1, ..., p_n)$ which places probability p_i at each observation. To be a valid multinomial distribution, these probabilities must satisfy the requirements that $p_i \geq 0$ and

$$\sum_{i=1}^{n} p_i = 1. \tag{10.1}$$

Since each observation is observed once in the sample, the log-likelihood function for this multinomial distribution is

$$\log L(p_1, ..., p_n) = \sum_{i=1}^{n} \log(p_i). \tag{10.2}$$

First let us consider a just-identified model. In this case the moment condition places no additional restrictions on the multinomial distribution. The maximum likelihood estimators of the probabilities $(p_1, ..., p_n)$ are those which maximize the log-likelihood subject to the constraint (10.1). This is equivalent to maximizing

$$\sum_{i=1}^{n} \log(p_i) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right)$$

where μ is a Lagrange multiplier. The n first order conditions are $0 = p_i^{-1} - \mu$. Combined with the constraint (10.1) we find that the MLE is $p_i = n^{-1}$ yielding the log-likelihood $-n \log(n)$.

Now consider the case of an overidentified model with moment condition

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0) = \mathbf{0}$$

where $\boldsymbol{g}$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$ and for simplicity we write $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{g}(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{\beta})$. The multinomial distribution which places probability p_i at each observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ will satisfy this condition if and only if

$$\sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0} \tag{10.3}$$

The **empirical likelihood estimator** is the value of $\boldsymbol{\beta}$ which maximizes the multinomial log-likelihood (10.2) subject to the restrictions (10.1) and (10.3).

The Lagrangian for this maximization problem is

$$\mathcal{L}\left(\boldsymbol{\beta}, p_1, ..., p_n, \boldsymbol{\lambda}, \mu\right) = \sum_{i=1}^{n} \log(p_i) - \mu\left(\sum_{i=1}^{n} p_i - 1\right) - n\boldsymbol{\lambda}' \sum_{i=1}^{n} p_i \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)$$

where $\boldsymbol{\lambda}$ and μ are Lagrange multipliers. The first-order-conditions of $\mathcal{L}$ with respect to p_i, μ, and $\boldsymbol{\lambda}$ are

$$\begin{aligned} \frac{1}{p_i} &= \mu + n\boldsymbol{\lambda}' \boldsymbol{g}_i\left(\boldsymbol{\beta}\right) \\ \sum_{i=1}^{n} p_i &= 1 \\ \sum_{i=1}^{n} p_i \boldsymbol{g}_i\left(\boldsymbol{\beta}\right) &= \mathbf{0}. \end{aligned}$$

Multiplying the first equation by p_i, summing over i, and using the second and third equations, we find $\mu = n$ and

$$p_i = \frac{1}{n\left(1 + \boldsymbol{\lambda}' \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right)}.$$

Substituting into $\mathcal{L}$ we find

$$R\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = -n\log\left(n\right) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}' \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right). \tag{10.4}$$

For given $\boldsymbol{\beta}$, the Lagrange multiplier $\boldsymbol{\lambda}(\boldsymbol{\beta})$ minimizes $R\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right)$:

$$\boldsymbol{\lambda}(\boldsymbol{\beta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, R(\boldsymbol{\beta}, \boldsymbol{\lambda}). \tag{10.5}$$

This minimization problem is the dual of the constrained maximization problem. The solution (when it exists) is well defined since $R(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is a convex function of $\boldsymbol{\lambda}$. The solution cannot be obtained explicitly, but must be obtained numerically (see section 6.5). This yields the (profile) empirical log-likelihood function for $\boldsymbol{\beta}$.

$$\begin{aligned} R(\boldsymbol{\beta}) &= R(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) \\ &= -n\log\left(n\right) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}(\boldsymbol{\beta})' \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right) \end{aligned}$$

The EL estimate $\hat{\boldsymbol{\beta}}$ is the value which maximizes $R(\boldsymbol{\beta})$, or equivalently minimizes its negative

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[-R(\boldsymbol{\beta})\right] \tag{10.6}$$

Numerical methods are required for calculation of $\hat{\boldsymbol{\beta}}$ (see Section 10.5).

As a by-product of estimation, we also obtain the Lagrange multiplier $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\hat{\boldsymbol{\beta}})$, probabilities

$$\hat{p}_i = \frac{1}{n\left(1 + \hat{\boldsymbol{\lambda}}' \boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right)}.$$

and maximized empirical likelihood

$$R(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \log\left(\hat{p}_i\right). \tag{10.7}$$

10.2 Asymptotic Distribution of EL Estimator

Define

$$\boldsymbol{G}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta}) \tag{10.8}$$

$$\boldsymbol{G} = \mathbb{E}\boldsymbol{G}_i(\boldsymbol{\beta}_0)$$

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\,\boldsymbol{g}_i(\boldsymbol{\beta}_0)'\right)$$

and

$$\boldsymbol{V} = \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1} \tag{10.9}$$

$$\boldsymbol{V}_{\boldsymbol{\lambda}} = \boldsymbol{\Omega} - \boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}' \tag{10.10}$$

For example, in the linear model, $\boldsymbol{G}_i(\boldsymbol{\beta}) = -\boldsymbol{z}_i\boldsymbol{x}_i'$, $\boldsymbol{G} = -\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$, and $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Theorem 10.2.1 *Under regularity conditions,*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}})$$

$$\sqrt{n}\hat{\boldsymbol{\lambda}} \xrightarrow{d} \boldsymbol{\Omega}^{-1}\mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\lambda}})$$

where $\boldsymbol{V}$ and $\boldsymbol{V}_{\boldsymbol{\lambda}}$ are defined in (10.9) and (10.10), and $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ and $\sqrt{n}\hat{\boldsymbol{\lambda}}$ are asymptotically independent.

The theorem shows that asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}$ is the same as for efficient GMM. Thus the EL estimator is asymptotically efficient.

Chamberlain (1987) showed that $\boldsymbol{V}_{\boldsymbol{\beta}}$ is the semiparametric efficiency bound for $\boldsymbol{\beta}$ in the overidentified moment condition model. This means that no consistent estimator for this class of models can have a lower asymptotic variance than $\boldsymbol{V}_{\boldsymbol{\beta}}$. Since the EL estimator achieves this bound, it is an asymptotically efficient estimator for $\boldsymbol{\beta}$.

Proof of Theorem 10.2.1. $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ jointly solve

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\lambda}}R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -\sum_{i=1}^{n}\frac{\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)}{\left(1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right)} \tag{10.11}$$

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\beta}}R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -\sum_{i=1}^{n}\frac{\boldsymbol{G}_i\left(\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\lambda}}{1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)}. \tag{10.12}$$

Let $\boldsymbol{G}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}_i(\boldsymbol{\beta}_0)$, $\overline{\boldsymbol{g}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}_0)$ and $\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\,\boldsymbol{g}_i(\boldsymbol{\beta}_0)'$.

Expanding (10.12) around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq \boldsymbol{G}_n'\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right). \tag{10.13}$$

Expanding (10.11) around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq -\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + \boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}} \tag{10.14}$$

Premultiplying by $\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}$ and using (10.13) yields

$$\begin{aligned}\mathbf{0} &\simeq -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) + \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}} \\ &= -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$ and using the WLLN and CLT yields

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) &\simeq -\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d} \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right) \\ &= \mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right)\end{aligned} \tag{10.15}$$

Solving (10.14) for $\hat{\boldsymbol{\lambda}}$ and using (10.15) yields

$$\begin{aligned}\sqrt{n}\hat{\boldsymbol{\lambda}} &\simeq \boldsymbol{\Omega}_n^{-1}\left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d} \boldsymbol{\Omega}^{-1}\left(\boldsymbol{I}-\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\right)\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right) \\ &= \boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\lambda}}\right)\end{aligned} \tag{10.16}$$

Furthermore, since

$$\boldsymbol{G}'\left(\boldsymbol{I}-\boldsymbol{\Omega}^{-1}\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\right)=\mathbf{0}$$

$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$ and $\sqrt{n}\hat{\boldsymbol{\lambda}}$ are asymptotically uncorrelated and hence independent.

10.3 Overidentifying Restrictions

In a parametric likelihood context, tests are based on the difference in the log likelihood functions. The same statistic can be constructed for empirical likelihood. Twice the difference between the unrestricted empirical log-likelihood $-n\log(n)$ and the maximized empirical log-likelihood for the model (10.7) is

$$LR_n = \sum_{i=1}^{n} 2\log\left(1+\hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right). \tag{10.17}$$

Theorem 10.3.1 *If* $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)=\mathbf{0}$ *then* $LR_n \xrightarrow{d} \chi^2_{\ell-k}$.

The EL overidentification test is similar to the GMM overidentification test. They are asymptotically first-order equivalent, and have the same interpretation. The overidentification test is a very useful by-product of EL estimation, and it is advisable to report the statistic LR_n whenever EL is the estimation method.

Proof of Theorem 10.3.1. First, by a Taylor expansion, (10.15), and (10.16),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right) &\simeq \sqrt{n}\left(\overline{\boldsymbol{g}}_n+\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right) \\ &\simeq \left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\simeq \boldsymbol{\Omega}_n\sqrt{n}\hat{\boldsymbol{\lambda}}.\end{aligned}$$

Second, since $\log(1+u) \simeq u - u^2/2$ for u small,

$$\begin{aligned} LR_n &= \sum_{i=1}^{n} 2\log\left(1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right) \\ &\simeq 2\hat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right) - \hat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\lambda}} \\ &\simeq n\hat{\boldsymbol{\lambda}}'\boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}} \\ &\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\lambda}}\right)'\boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\lambda}}\right) \\ &= \chi^2_{\ell-k} \end{aligned}$$

where the proof of the final equality is left as an exercise.

10.4 Testing

Let the maintained model be

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0} \tag{10.18}$$

where $\boldsymbol{g}$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$. By "maintained" we mean that the overidentfying restrictions contained in (10.18) are assumed to hold and are not being challenged (at least for the test discussed in this section). The hypothesis of interest is

$$\boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}.$$

where $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^a$. The restricted EL estimator and likelihood are the values which solve

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmax}} R(\boldsymbol{\beta}) \\ R(\tilde{\boldsymbol{\beta}}) &= \max_{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}} R(\boldsymbol{\beta}). \end{aligned}$$

Fundamentally, the restricted EL estimator $\tilde{\boldsymbol{\beta}}$ is simply an EL estimator with $\ell - k + a$ overidentifying restrictions, so there is no fundamental change in the distribution theory for $\tilde{\boldsymbol{\beta}}$ relative to $\hat{\boldsymbol{\beta}}$. To test the hypothesis $\boldsymbol{h}(\boldsymbol{\beta})$ while maintaining (10.18), the simple overidentifying restrictions test (10.17) is not appropriate. Instead we use the difference in log-likelihoods:

$$LR_n = 2\left(R(\hat{\boldsymbol{\beta}}) - R(\tilde{\boldsymbol{\beta}})\right).$$

This test statistic is a natural analog of the GMM distance statistic.

Theorem 10.4.1 *Under (10.18) and* $\mathbb{H}_0 : \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}$, $LR_n \xrightarrow{d} \chi^2_a$.

The proof of this result is more challenging and is omitted.

10.5 Numerical Computation

Gauss code which implements the methods discussed below can be found at

`http://www.ssc.wisc.edu/~bhansen/progs/elike.prc`

Derivatives

The numerical calculations depend on derivatives of the dual likelihood function (10.4). Define

$$\begin{aligned} \boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{\boldsymbol{g}_i(\boldsymbol{\beta})}{\left(1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})\right)} \\ \boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}) &= \frac{\boldsymbol{G}_i(\boldsymbol{\beta})'\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})} \end{aligned}$$

The first derivatives of (10.4) are

$$\begin{aligned} \boldsymbol{R}_{\boldsymbol{\lambda}} &= \frac{\partial}{\partial\boldsymbol{\lambda}}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = -\sum_{i=1}^{n}\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}) \\ \boldsymbol{R}_{\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = -\sum_{i=1}^{n}\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}). \end{aligned}$$

The second derivatives are

$$\begin{aligned} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} &= \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' \\ \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} &= \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\boldsymbol{G}_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right) \\ \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\left(\boldsymbol{g}_i(\boldsymbol{\beta})'\boldsymbol{\lambda}\right)}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right) \end{aligned}$$

Inner Loop

The so-called "inner loop" solves (10.5) for given $\boldsymbol{\beta}$. The modified Newton method takes a quadratic approximation to $R_n(\boldsymbol{\beta},\boldsymbol{\lambda})$ yielding the iteration rule

$$\boldsymbol{\lambda}_{j+1} = \boldsymbol{\lambda}_j - \delta\left(\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j). \tag{10.19}$$

where $\delta > 0$ is a scalar steplength (to be discussed next). The starting value $\boldsymbol{\lambda}_1$ can be set to the zero vector. The iteration (10.19) is continued until the gradient $R_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)$ is smaller than some prespecified tolerance.

Efficient convergence requires a good choice of steplength δ. One method uses the following quadratic approximation. Set $\delta_0 = 0$, $\delta_1 = \frac{1}{2}$ and $\delta_2 = 1$. For $p = 0, 1, 2$, set

$$\begin{aligned} \boldsymbol{\lambda}_p &= \boldsymbol{\lambda}_j - \delta_p\left(\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)) \\ R_p &= R(\boldsymbol{\beta},\boldsymbol{\lambda}_p) \end{aligned}$$

A quadratic function can be fit exactly through these three points. The value of δ which minimizes this quadratic is

$$\hat{\delta} = \frac{R_2 + 3R_0 - 4R_1}{4R_2 + 4R_0 - 8R_1}.$$

yielding the steplength to be plugged into (10.19).

A complication is that $\boldsymbol{\lambda}$ must be constrained so that $0 \leq p_i \leq 1$ which holds if

$$n\left(1 + \boldsymbol{\lambda}'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right) \geq 1 \tag{10.20}$$

for all i. If (10.20) fails, the stepsize δ needs to be decreased.

Outer Loop

The outer loop is the minimization (10.6). This can be done by the modified Newton method described in the previous section. The gradient for (10.6) is

$$\boldsymbol{R}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} R(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{R}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}} = \boldsymbol{R}_{\boldsymbol{\beta}}$$

since $\boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = 0$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\beta})$, where

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}'} \boldsymbol{\lambda}(\boldsymbol{\beta}) = -\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}},$$

the second equality following from the implicit function theorem applied to $\boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) = 0$.

The Hessian for (10.6) is

$$\begin{aligned}
\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\frac{\partial}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} R(\boldsymbol{\beta}) \\
&= -\frac{\partial}{\partial \boldsymbol{\beta}'} \left[\boldsymbol{R}_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right)\right] \\
&= -\left(\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) + \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{\lambda}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \boldsymbol{\lambda}_{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} - \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}.
\end{aligned}$$

It is not guaranteed that $\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} > 0$. If not, the eigenvalues of $\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ should be adjusted so that all are positive. The Newton iteration rule is

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \delta \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \boldsymbol{R}_{\boldsymbol{\beta}}$$

where δ is a scalar stepsize, and the rule is iterated until convergence.

Chapter 11

Endogeneity

We say that there is endogeneity in the linear model $y = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ if $\boldsymbol{\beta}$ is the parameter of interest and $\mathbb{E}(\boldsymbol{x}_i e_i) \neq 0$. This cannot happen if $\boldsymbol{\beta}$ is defined by linear projection, so requires a structural interpretation. The coefficient $\boldsymbol{\beta}$ must have meaning separately from the definition of a conditional mean or linear projection.

Example: Measurement error in the regressor. Suppose that $(y_i, \boldsymbol{x}_i^*)$ are joint random variables, $\mathbb{E}(y_i \mid \boldsymbol{x}_i^*) = \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta}$ is linear, $\boldsymbol{\beta}$ is the parameter of interest, and $\boldsymbol{x}_i^*$ is not observed. Instead we observe $\boldsymbol{x}_i = \boldsymbol{x}_i^* + \boldsymbol{u}_i$ where $\boldsymbol{u}_i$ is an $k \times 1$ measurement error, independent of y_i and $\boldsymbol{x}_i^*$. Then

$$\begin{aligned} y_i &= \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta} + e_i \\ &= (\boldsymbol{x}_i - \boldsymbol{u}_i)'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + v_i \end{aligned}$$

where

$$v_i = e_i - \boldsymbol{u}_i'\boldsymbol{\beta}.$$

The problem is that

$$\mathbb{E}\left(\boldsymbol{x}_i v_i\right) = \mathbb{E}\left[\left(\boldsymbol{x}_i^* + \boldsymbol{u}_i\right)\left(e_i - \boldsymbol{u}_i'\boldsymbol{\beta}\right)\right] = -\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} \neq 0$$

if $\boldsymbol{\beta} \neq 0$ and $\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right) \neq 0$. It follows that if $\hat{\boldsymbol{\beta}}$ is the OLS estimator, then

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^* = \boldsymbol{\beta} - \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

This is called **measurement error bias**.

Example: Supply and Demand. The variables q_i and p_i (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that $\boldsymbol{e}_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$ is iid, $\mathbb{E}\boldsymbol{e}_i = 0$, $\beta_1 + \beta_2 = 1$ and $\mathbb{E}\boldsymbol{e}_i\boldsymbol{e}_i' = \boldsymbol{I}_2$ (the latter for simplicity). The question is, if we regress q_i on p_i, what happens?

It is helpful to solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}\begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

so

$$\begin{aligned}\begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{pmatrix} \beta_2 e_{1i} + \beta_1 e_{2i} \\ (e_{1i} - e_{2i}) \end{pmatrix}.\end{aligned}$$

The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + \varepsilon_i \\ \mathbb{E}(p_i \varepsilon_i) &= 0 \end{aligned}$$

where

$$\beta^* = \frac{\mathbb{E}(p_i q_i)}{\mathbb{E}(p_i^2)} = \frac{\beta_2 - \beta_1}{2}$$

Hence if it is estimated by OLS, $\hat{\beta} \xrightarrow{p} \beta^*$, which does not equal either β_1 or β_2. This is called **simultaneous equations bias**.

11.1 Instrumental Variables

Let the equation of interest be

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + e_i \tag{11.1}$$

where $\boldsymbol{x}_i$ is $k \times 1$, and assume that $\mathbb{E}(\boldsymbol{x}_i e_i) \neq 0$ so there is **endogeneity**. We call (11.1) the structural equation. In matrix notation, this can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{11.2}$$

Any solution to the problem of endogeneity requires additional information which we call **instruments**.

Definition 11.1.1 *The $\ell \times 1$ random vector $\boldsymbol{z}_i$ is an **instrumental variable** for (11.1) if $\mathbb{E}(\boldsymbol{z}_i e_i) = \boldsymbol{0}$.*

In a typical set-up, some regressors in $\boldsymbol{x}_i$ will be uncorrelated with e_i (for example, at least the intercept). Thus we make the partition

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix} \tag{11.3}$$

where $\mathbb{E}(\boldsymbol{x}_{1i} e_i) = 0$ yet $\mathbb{E}(\boldsymbol{x}_{2i} e_i) \neq 0$. We call $\boldsymbol{x}_{1i}$ exogenous and $\boldsymbol{x}_{2i}$ endogenous. By the above definition, $\boldsymbol{x}_{1i}$ is an instrumental variable for (11.1), so should be included in $\boldsymbol{z}_i$. So we have the partition

$$\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ \ell_2 \end{matrix} \tag{11.4}$$

where $\boldsymbol{x}_{1i} = \boldsymbol{z}_{1i}$ are the **included exogenous variables**, and $\boldsymbol{z}_{2i}$ are the **excluded exogenous variables**. That is $\boldsymbol{z}_{2i}$ are variables which could be included in the equation for y_i (in the sense that they are uncorrelated with e_i) yet can be *excluded*, as they would have true zero coefficients in the equation.

The model is **just-identified** if $\ell = k$ (i.e., if $\ell_2 = k_2$) and **over-identified** if $\ell > k$ (i.e., if $\ell_2 > k_2$).

We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

11.2 Reduced Form

The reduced form relationship between the variables or "regressors" $\boldsymbol{x}_i$ and the instruments $\boldsymbol{z}_i$ is found by linear projection. Let

$$\boldsymbol{\Gamma} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)$$

be the $\ell \times k$ matrix of coefficients from a projection of $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$, and define

$$\boldsymbol{u}_i = \boldsymbol{x}_i - \boldsymbol{\Gamma}' \boldsymbol{z}_i$$

as the projection error. Then the reduced form linear relationship between $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is

$$\boldsymbol{x}_i = \boldsymbol{\Gamma}' \boldsymbol{z}_i + \boldsymbol{u}_i. \tag{11.5}$$

In matrix notation, we can write (11.5) as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \tag{11.6}$$

where $\boldsymbol{U}$ is $n \times k$.

By construction,

$$\mathbb{E}(\boldsymbol{z}_i \boldsymbol{u}_i') = \boldsymbol{0},$$

so (11.5) is a projection and can be estimated by OLS:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{z}\hat{\Gamma} + \hat{\boldsymbol{u}} \\ \hat{\boldsymbol{\Gamma}} &= \left(\boldsymbol{z}'\boldsymbol{z}\right)^{-1}\left(\boldsymbol{z}'\boldsymbol{x}\right). \end{aligned}$$

Substituting (11.6) into (11.2), we find

$$\begin{aligned} \boldsymbol{y} &= \left(\boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U}\right)\boldsymbol{\beta} + \boldsymbol{e} \\ &= \boldsymbol{Z}\boldsymbol{\lambda} + \boldsymbol{v}, \end{aligned} \tag{11.7}$$

where

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta} \tag{11.8}$$

and

$$\boldsymbol{v} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{e}.$$

Observe that

$$\mathbb{E}\left(\boldsymbol{z}_i v_i\right) = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{u}_i'\right)\boldsymbol{\beta} + \mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \boldsymbol{0}.$$

Thus (11.7) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\hat{\boldsymbol{\lambda}} + \hat{\boldsymbol{v}}, \\ \hat{\boldsymbol{\lambda}} &= \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right) \end{aligned}$$

The equation (11.7) is the reduced form for $\boldsymbol{y}$. (11.6) and (11.7) together are the **reduced form equations** for the system

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{\lambda} + \boldsymbol{v} \\ \boldsymbol{x} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U}. \end{aligned}$$

As we showed above, OLS yields the reduced-form estimates $\left(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\Gamma}}\right)$

11.3 Identification

The structural parameter $\boldsymbol{\beta}$ relates to $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ through (11.8). The parameter $\boldsymbol{\beta}$ is **identified**, meaning that it can be recovered from the reduced form, if

$$\operatorname{rank}(\boldsymbol{\Gamma}) = k. \tag{11.9}$$

Assume that (11.9) holds. If $\ell = k$, then $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$. If $\ell > k$, then for any $\boldsymbol{W} > 0$, $\boldsymbol{\beta} = (\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\lambda}$.

If (11.9) is not satisfied, then $\boldsymbol{\beta}$ cannot be recovered from $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$. Note that a necessary (although not sufficient) condition for (11.9) is $\ell \geq k$.

Since $\boldsymbol{Z}$ and $\boldsymbol{X}$ have the common variables $\boldsymbol{X}_1$, we can rewrite some of the expressions. Using (11.3) and (11.4) to make the matrix partitions $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ and $\boldsymbol{X} = [\boldsymbol{Z}_1, \boldsymbol{X}_2]$, we can partition $\boldsymbol{\Gamma}$ as

$$\begin{aligned} \boldsymbol{\Gamma} &= \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \end{aligned}$$

(11.6) can be rewritten as

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{Z}_1 \\ \boldsymbol{X}_2 &= \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2. \end{aligned} \tag{11.10}$$

$\boldsymbol{\beta}$ is identified if $\operatorname{rank}(\boldsymbol{\Gamma}) = k$, which is true if and only if $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$ (by the upper-diagonal structure of $\boldsymbol{\Gamma}$). Thus the key to identification of the model rests on the $\ell_2 \times k_2$ matrix $\boldsymbol{\Gamma}_{22}$ in (11.10).

11.4 Estimation

The model can be written as

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= \boldsymbol{0} \end{aligned}$$

or

$$\begin{aligned} \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) &= \boldsymbol{0} \\ \boldsymbol{g}_i(\boldsymbol{\beta}) &= \boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right). \end{aligned}$$

This is a moment condition model. Appropriate estimators include GMM and EL. The estimators and distribution theory developed in those Chapter 8 and 9 directly apply. Recall that the GMM estimator, for given weight matrix $\boldsymbol{W}_n$, is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y}.$$

11.5 Special Cases: IV and 2SLS

If the model is just-identified, so that $k = \ell$, then the formula for GMM simplifies. We find that

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{W}_n^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \end{aligned}$$

This estimator is often called the **instrumental variables estimator** (IV) of $\boldsymbol{\beta}$, where $\boldsymbol{Z}$ is used as an instrument for $\boldsymbol{X}$. Observe that the weight matrix $\boldsymbol{W}_n$ has disappeared. In the just-identified case, the weight matrix places no role. This is also the MME estimator of $\boldsymbol{\beta}$, and the EL estimator. Another interpretation stems from the fact that since $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$, we can construct the **Indirect Least Squares** (ILS) estimator:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\lambda}} \\
&= \left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right)\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right).
\end{aligned}$$

which again is the IV estimator.

Recall that the optimal weight matrix is an estimate of the inverse of $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$. In the special case that $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$ (homoskedasticity), then $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\sigma^2 \propto \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$ suggesting the weight matrix $\boldsymbol{W}_n = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$. Using this choice, the GMM estimator equals

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

This is called the **two-stage-least squares** (2SLS) estimator. It was originally proposed by Theil (1953) and Basmann (1957), and is the classic estimator for linear equations with instruments. Under the homoskedasticity assumption, the 2SLS estimator is efficient GMM, but otherwise it is inefficient.

It is useful to observe that writing

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \\
\hat{\boldsymbol{X}} &= \boldsymbol{P}\boldsymbol{X} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}}
\end{aligned}$$

then

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y} \\
&= \left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y}.
\end{aligned}$$

The source of the "two-stage" name is since it can be computed as follows

- First regress $\boldsymbol{X}$ on $\boldsymbol{Z}$, vis., $\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$ and $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}} = \boldsymbol{P}\boldsymbol{X}$.
- Second, regress $\boldsymbol{y}$ on $\hat{\boldsymbol{Z}}$, vis., $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{Z}}'\hat{\boldsymbol{Z}}\right)^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{y}$.

It is useful to scrutinize the projection $\hat{\boldsymbol{Z}}$. Recall, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{Z}_2]$. Then

$$\begin{aligned}
\hat{\boldsymbol{X}} &= \left[\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2\right] \\
&= [\boldsymbol{P}\boldsymbol{X}_1, \boldsymbol{P}\boldsymbol{X}_2] \\
&= [\boldsymbol{X}_1, \boldsymbol{P}\boldsymbol{X}_2] \\
&= \left[\boldsymbol{X}_1, \hat{\boldsymbol{X}}_2\right],
\end{aligned}$$

since $\boldsymbol{X}_1$ lies in the span of $\boldsymbol{X}$. Thus in the second stage, we regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and $\hat{\boldsymbol{X}}_2$. So only the endogenous variables $\boldsymbol{X}_2$ are replaced by their fitted values:

$$\hat{\boldsymbol{X}}_2 = \boldsymbol{Z}_1\hat{\boldsymbol{\Gamma}}_{12} + \boldsymbol{Z}_2\hat{\boldsymbol{\Gamma}}_{22}.$$

11.6 Bekker Asymptotics

Bekker (1994) used an alternative asymptotic framework to analyze the finite-sample bias in the 2SLS estimator. Here we present a simplified version of one of his results. In our notation, the model is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \qquad (11.11) \\ \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \qquad (11.12) \\ \boldsymbol{\xi} &= (\boldsymbol{e}, \boldsymbol{U}) \\ \mathbb{E}(\boldsymbol{\xi} \mid \boldsymbol{Z}) &= \boldsymbol{0} \\ \mathbb{E}(\boldsymbol{\xi}'\boldsymbol{\xi} \mid \boldsymbol{Z}) &= \boldsymbol{S} \end{aligned}$$

As before, $\boldsymbol{Z}$ is $n \times l$ so there are l instruments.

First, let's analyze the approximate bias of OLS applied to (11.11). Using (11.12),

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e}\right) = \mathbb{E}(\boldsymbol{x}_i e_i) = \boldsymbol{\Gamma}'\mathbb{E}(\boldsymbol{z}_i e_i) + \mathbb{E}(\boldsymbol{u}_i e_i) = \boldsymbol{s}_{21}$$

and

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) &= \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\ &= \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right) \\ &= \boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \boldsymbol{S}_{22} \end{aligned}$$

where $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$. Hence by a first-order approximation

$$\begin{aligned} \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right) &\approx \left(\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e}\right) \\ &= \left(\boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \boldsymbol{S}_{22}\right)^{-1}\boldsymbol{s}_{21} \qquad (11.13) \end{aligned}$$

which is zero only when $\boldsymbol{s}_{21} = \boldsymbol{0}$ (when $\boldsymbol{X}$ is exogenous).

We now derive a similar result for the 2SLS estimator.

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y}\right).$$

Let $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$. By the spectral decomposition of an idempotent matrix, $\boldsymbol{P} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{I}_l, \boldsymbol{0})$. Let $\boldsymbol{Q} = \boldsymbol{H}'\boldsymbol{\xi}\boldsymbol{S}^{-1/2}$ which satisfies $\mathbb{E}\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_n$ and partition $\boldsymbol{Q} = (\boldsymbol{q}_1' \ \boldsymbol{Q}_2')$ where $\boldsymbol{q}_1$ is $l \times 1$. Hence

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n}\boldsymbol{\xi}'\boldsymbol{P}\boldsymbol{\xi} \mid \boldsymbol{Z}\right) &= \frac{1}{n}\boldsymbol{S}^{1/2\prime}\mathbb{E}\left(\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q} \mid \boldsymbol{Z}\right)\boldsymbol{S}^{1/2} \\ &= \frac{1}{n}\boldsymbol{S}^{1/2\prime}\mathbb{E}\left(\frac{1}{n}\boldsymbol{q}_1'\boldsymbol{q}_1\right)\boldsymbol{S}^{1/2} \\ &= \frac{l}{n}\boldsymbol{S}^{1/2\prime}\boldsymbol{S}^{1/2} \\ &= \alpha\boldsymbol{S} \end{aligned}$$

where

$$\alpha = \frac{l}{n}.$$

Using (11.12) and this result,

$$\frac{1}{n}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{e}\right) = \frac{1}{n}\mathbb{E}\left(\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{e}\right) + \frac{1}{n}\mathbb{E}\left(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{e}\right) = \alpha\boldsymbol{s}_{21},$$

and

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right) &= \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i\right) + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \frac{1}{n}\mathbb{E}\left(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\right) \\ &= \boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \alpha\boldsymbol{S}_{22}.\end{aligned}$$

Together

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}\right) &\approx \left(\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)\right)^{-1}\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{e}\right) \\ &= \alpha\left(\boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \alpha\boldsymbol{S}_{22}\right)^{-1}\boldsymbol{s}_{21}. \qquad (11.14)\end{aligned}$$

In general this is non-zero, except when $\boldsymbol{s}_{21} = 0$ (when $\boldsymbol{X}$ is exogenous). It is also close to zero when $\alpha = 0$. Bekker (1994) pointed out that it also has the reverse implication – that when $\alpha = l/n$ is large, the bias in the 2SLS estimator will be large. Indeed as $\alpha \to 1$, the expression in (11.14) approaches that in (11.13), indicating that the bias in 2SLS approaches that of OLS as the number of instruments increases.

Bekker (1994) showed further that under the alternative asymptotic approximation that α is fixed as $n \to \infty$ (so that the number of instruments goes to infinity proportionately with sample size) then the expression in (11.14) is the probability limit of $\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}$

11.7 Identification Failure

Recall the reduced form equation

$$\boldsymbol{X}_2 = \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2.$$

The parameter $\boldsymbol{\beta}$ fails to be identified if $\boldsymbol{\Gamma}_{22}$ has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where $k = l = 1$ (so there is no $\boldsymbol{Z}_1$). Then the model may be written as

$$\begin{aligned}y_i &= x_i\beta + e_i \\ x_i &= z_i\gamma + u_i\end{aligned}$$

and $\Gamma_{22} = \gamma = \mathbb{E}\left(z_i x_i\right)/\mathbb{E}z_i^2$. We see that β is identified if and only if $\gamma \neq 0$, which occurs when $\mathbb{E}\left(x_i z_i\right) \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails, so $\mathbb{E}\left(x_i z_i\right) = 0$. Then by the CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i \xrightarrow{d} N_1 \sim \mathrm{N}\left(\mathbf{0}, \mathbb{E}\left(z_i^2 e_i^2\right)\right) \qquad (11.15)$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \xrightarrow{d} N_2 \sim \mathrm{N}\left(\mathbf{0}, \mathbb{E}\left(z_i^2 u_i^2\right)\right) \qquad (11.16)$$

therefore

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \boldsymbol{x}_i} \xrightarrow{d} \frac{N_1}{N_2} \sim \text{Cauchy},$$

since the ratio of two normals is Cauchy. This is particularly nasty, as the Cauchy distribution does not have a finite mean. This result carries over to more general settings, and was examined by Phillips (1989) and Choi and Phillips (1992).

Suppose that identification does not completely fail, but is *weak*. This occurs when Γ_{22} is full rank, but *small*. This can be handled in an asymptotic analysis by modeling it as local-to-zero, viz

$$\boldsymbol{\Gamma}_{22} = n^{-1/2}\boldsymbol{C},$$

where $\boldsymbol{C}$ is a full rank matrix. The $n^{-1/2}$ is picked because it provides just the right balancing to allow a rich distribution theory.

To see the consequences, once again take the simple case $k = l = 1$. Here, the instrument x_i is weak for z_i if

$$\gamma = n^{-1/2}c.$$

Then (11.15) is unaffected, but (11.16) instead takes the form

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i^2\gamma + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \\
&= \frac{1}{n}\sum_{i=1}^{n} z_i^2 c + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \\
&\xrightarrow{d} Qc + N_2
\end{aligned}$$

therefore

$$\hat{\beta} - \beta \xrightarrow{d} \frac{N_1}{Qc + N_2}.$$

As in the case of complete identification failure, we find that $\hat{\beta}$ is inconsistent for β and the asymptotic distribution of $\hat{\beta}$ is non-normal. In addition, standard test statistics have non-standard distributions, meaning that inferences about parameters of interest can be misleading.

The distribution theory for this model was developed by Staiger and Stock (1997) and extended to nonlinear GMM estimation by Stock and Wright (2000). Further results on testing were obtained by Wang and Zivot (1998).

The bottom line is that it is highly desirable to avoid identification failure. Once again, the equation to focus on is the reduced form

$$\boldsymbol{X}_2 = \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2$$

and identification requires $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$. If $k_2 = 1$, this requires $\Gamma_{22} \neq \mathbf{0}$, which is straightforward to assess using a hypothesis test on the reduced form. Therefore in the case of $k_2 = 1$ (one RHS endogenous variable), one constructive recommendation is to explicitly estimate the reduced form equation for $\boldsymbol{X}_2$, construct the test of $\boldsymbol{\Gamma}_{22} = \mathbf{0}$, and at a minimum check that the test rejects $\mathbb{H}_0 : \Gamma_{22} = \mathbf{0}$.

When $k_2 > 1$, $\boldsymbol{\Gamma}_{22} \neq 0$ is not sufficient for identification. It is not even sufficient that each column of $\boldsymbol{\Gamma}_{22}$ is non-zero (each column corresponds to a distinct endogenous variable in $\boldsymbol{Z}_2$). So while a minimal check is to test that each columns of $\boldsymbol{\Gamma}_{22}$ is non-zero, this cannot be interpreted as definitive proof that $\boldsymbol{\Gamma}_{22}$ has full rank. Unfortunately, tests of deficient rank are difficult to implement. In any event, it appears reasonable to explicitly estimate and report the reduced form equations for $\boldsymbol{Z}_2$, and attempt to assess the likelihood that $\boldsymbol{\Gamma}_{22}$ has deficient rank.

Exercises

1. Consider the single equation model

$$y_i = x_i\beta + e_i,$$

where y_i and z_i are both real-valued (1×1). Let $\hat{\beta}$ denote the IV estimator of β using as an instrument a dummy variable d_i (takes only the values 0 and 1). Find a simple expression for the IV estimator in this context.

2. In the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

suppose $\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid \boldsymbol{X}_i\right)$ is known. Show that the GLS estimator of $\boldsymbol{\beta}$ can be written as an IV estimator using some instrument $\boldsymbol{z}_i$. (Find an expression for $\boldsymbol{z}_i$.)

3. Take the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Let the OLS estimator for $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}$ and the OLS residual be $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Let the IV estimator for $\boldsymbol{\beta}$ using some instrument $\boldsymbol{Z}$ be $\tilde{\boldsymbol{\beta}}$ and the IV residual be $\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$. If $\boldsymbol{Z}$ is indeed endogeneous, will IV "fit" better than OLS, in the sense that $\tilde{\boldsymbol{e}}'\tilde{\boldsymbol{e}} < \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, at least in large samples?

4. The reduced form between the regressors $\boldsymbol{x}_i$ and instruments $\boldsymbol{z}_i$ takes the form

$$\boldsymbol{x}_i = \boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i$$

or

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U}$$

where $\boldsymbol{X}_i$ is $k \times 1$, $\boldsymbol{z}_i$ is $l \times 1$, $\boldsymbol{X}$ is $n \times k$, $\boldsymbol{Z}$ is $n \times l$, $\boldsymbol{U}$ is $n \times k$, and $\boldsymbol{\Gamma}$ is $l \times k$. The parameter $\boldsymbol{\Gamma}$ is defined by the population moment condition

$$\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) = \boldsymbol{0}$$

Show that the method of moments estimator for $\boldsymbol{\Gamma}$ is $\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$.

5. In the structural model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \end{aligned}$$

with $\boldsymbol{\Gamma}$ $l \times k$, $l \geq k$, we claim that $\boldsymbol{\beta}$ is identified (can be recovered from the reduced form) if $\operatorname{rank}(\boldsymbol{\Gamma}) = k$. Explain why this is true. That is, show that if $\operatorname{rank}(\boldsymbol{\Gamma}) < k$ then $\boldsymbol{\beta}$ cannot be identified.

6. Take the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

where x_i and $\boldsymbol{\beta}$ are 1×1.

(a) Show that $\mathbb{E}(x_i e_i) = 0$ and $\mathbb{E}\left(x_i^2 e_i\right) = 0$. Is $\boldsymbol{z}_i = (x_i \quad x_i^2)'$ a valid instrumental variable for estimation of $\boldsymbol{\beta}$?

(b) Define the 2SLS estimator of $\boldsymbol{\beta}$, using $\boldsymbol{z}_i$ as an instrument for x_i. How does this differ from OLS?

(c) Find the efficient GMM estimator of $\boldsymbol{\beta}$ based on the moment condition

$$\mathbb{E}\left(\boldsymbol{z}_i\left(y_i - x_i\boldsymbol{\beta}\right)\right) = \mathbf{0}.$$

Does this differ from 2SLS and/or OLS?

7. Suppose that price and quantity are determined by the intersection of the linear demand and supply curves

$$\begin{aligned} \text{Demand} \quad &: \quad Q = a_0 + a_1 P + a_2 Y + \boldsymbol{e}_1 \\ \text{Supply} \quad &: \quad Q = b_0 + b_1 P + b_2 W + \boldsymbol{e}_2 \end{aligned}$$

where income (Y) and wage (W) are determined outside the market. In this model, are the parameters identified?

8. The data file `card.dat` is taken from David Card "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in *Aspects of Labour Market Behavior* (1995). There are 2215 observations with 29 variables, listed in card.pdf. We want to estimate a wage equation

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 South + \beta_5 Black + \boldsymbol{e}$$

where $Educ = Eduation$ (Years) $Exper = Experience$ (Years), and $South$ and $Black$ are regional and racial dummy variables.

(a) Estimate the model by OLS. Report estimates and standard errors.

(b) Now treat *Education* as endogenous, and the remaining variables as exogenous. Estimate the model by 2SLS, using the instrument $near4$, a dummy indicating that the observation lives near a 4-year college. Report estimates and standard errors.

(c) Re-estimate by 2SLS (report estimates and standard errors) adding three additional instruments: *near2* (a dummy indicating that the observation lives near a 2-year college), $fatheduc$ (the education, in years, of the father) and *motheduc* (the education, in years, of the mother).

(d) Re-estimate the model by efficient GMM. I suggest that you use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix without further iteration. Report the estimates and standard errors.

(e) Calculate and report the J statistic for overidentification.

(f) Discuss your findings.

Chapter 12

Univariate Time Series

A time series y_t is a process observed in sequence over time, $t = 1, ..., T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations.

Because of the sequential nature of time series, we expect that y_t and y_{t-1} are *not* independent, so classical assumptions are not valid.

We can separate time series into two categories: univariate ($y_t \in \mathbb{R}$ is scalar); and multivariate ($y_t \in \mathbb{R}^m$ is vector-valued). The primary model for univariate time series is autoregressions (ARs). The primary model for multivariate time series is vector autoregressions (VARs).

12.1 Stationarity and Ergodicity

Definition 12.1.1 *$\{y_t\}$ is **covariance (weakly) stationary** if*

$$\mathbb{E}(y_t) = \mu$$

is independent of t, and

$$\operatorname{cov}(y_t, y_{t-k}) = \gamma(k)$$

*is independent of t for all k.$\gamma(k)$ is called the **autocovariance function**.*

$$\rho(k) = \gamma(k)/\gamma(0) = \operatorname{corr}(y_t, y_{t-k})$$

*is the **autocorrelation function**.*

Definition 12.1.2 *$\{y_t\}$ is **strictly stationary** if the joint distribution of $(y_t, ..., y_{t-k})$ is independent of t for all k.*

Definition 12.1.3 *A stationary time series is **ergodic** if $\gamma(k) \to 0$ as $k \to \infty$.*

The following two theorems are essential to the analysis of stationary time series. There proofs are rather difficult, however.

Theorem 12.1.1 *If y_t is strictly stationary and ergodic and $x_t = f(y_t, y_{t-1}, ...)$ is a random variable, then x_t is strictly stationary and ergodic.*

Theorem 12.1.2 *(Ergodic Theorem). If y_t is strictly stationary and ergodic and $\mathbb{E}|y_t| < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} \mathbb{E}(y_t).$$

This allows us to consistently estimate parameters using time-series moments:
The sample mean:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The sample autocovariance

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu}).$$

The sample autocorrelation

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}.$$

Theorem 12.1.3 *If y_t is strictly stationary and ergodic and $\mathbb{E}y_t^2 < \infty$, then as $T \to \infty$,*

1. $\hat{\mu} \xrightarrow{p} \mathbb{E}(y_t)$;
2. $\hat{\gamma}(k) \xrightarrow{p} \gamma(k)$;
3. $\hat{\rho}(k) \xrightarrow{p} \rho(k)$.

Proof of Theorem 12.1.3. Part (1) is a direct consequence of the Ergodic theorem. For Part (2), note that

$$\begin{aligned}
\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{\mu})(y_{t-k} - \hat{\mu}) \\
&= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} - \frac{1}{T}\sum_{t=1}^{T} y_t\hat{\mu} - \frac{1}{T}\sum_{t=1}^{T} y_{t-k}\hat{\mu} + \hat{\mu}^2.
\end{aligned}$$

By Theorem 12.1.1 above, the sequence $y_t y_{t-k}$ is strictly stationary and ergodic, and it has a finite mean by the assumption that $\mathbb{E}y_t^2 < \infty$. Thus an application of the Ergodic Theorem yields

$$\frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} \xrightarrow{p} \mathbb{E}(y_t y_{t-k}).$$

Thus

$$\hat{\gamma}(k) \xrightarrow{p} \mathbb{E}(y_t y_{t-k}) - \mu^2 - \mu^2 + \mu^2 = \mathbb{E}(y_t y_{t-k}) - \mu^2 = \gamma(k).$$

Part (3) follows by the continuous mapping theorem: $\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0) \xrightarrow{p} \gamma(k)/\gamma(0) = \rho(k)$.

12.2 Autoregressions

In time-series, the series $\{..., y_1, y_2, ..., y_T, ...\}$ are jointly random. We consider the conditional expectation

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right)$$

where $\mathcal{F}_{t-1} = \{y_{t-1}, y_{t-2}, ...\}$ is the past history of the series.

An autoregressive (AR) model specifies that only a finite number of past lags matter:

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left(y_t \mid y_{t-1}, ..., y_{t-k}\right).$$

A linear AR model (the most common type used in practice) specifies linearity:

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right) = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-1} + \cdots + \rho_k y_{t-k}.$$

Letting

$$e_t = y_t - \mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right),$$

then we have the autoregressive model

$$\begin{aligned} y_t &= \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-1} + \cdots + \rho_k y_{t-k} + e_t \\ \mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) &= 0. \end{aligned}$$

The last property defines a special time-series process.

Definition 12.2.1 *e_t is a **martingale difference sequence** (MDS) if* $\mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) = 0$.

Regression errors are naturally a MDS. Some time-series processes may be a MDS as a consequence of optimizing behavior. For example, some versions of the life-cycle hypothesis imply that either changes in consumption, or consumption growth rates, should be a MDS. Most asset pricing models imply that asset returns should be the sum of a constant plus a MDS.

The MDS property for the regression error plays the same role in a time-series regression as does the conditional mean-zero property for the regression error in a cross-section regression. In fact, it is even more important in the time-series context, as it is difficult to derive distribution theories without this property.

A useful property of a MDS is that e_t is uncorrelated with any function of the lagged information $\mathcal{F}_{t-1}$. Thus for $k > 0$, $\mathbb{E}\left(y_{t-k} e_t\right) = 0$.

12.3 Stationarity of AR(1) Process

A mean-zero AR(1) is

$$y_t = \rho y_{t-1} + e_t.$$

Assume that e_t is iid, $\mathbb{E}(e_t) = 0$ and $\mathbb{E}e_t^2 = \sigma^2 < \infty$.

By back-substitution, we find

$$\begin{aligned} y_t &= e_t + \rho e_{t-1} + \rho^2 e_{t-2} + ... \\ &= \sum_{k=0}^{\infty} \rho^k e_{t-k}. \end{aligned}$$

Loosely speaking, this series converges if the sequence $\rho^k e_{t-k}$ gets small as $k \to \infty$. This occurs when $|\rho| < 1$.

Theorem 12.3.1 *If $|\rho| < 1$ then y_t is strictly stationary and ergodic.*

We can compute the moments of y_t using the infinite sum:

$$\mathbb{E}y_t = \sum_{k=0}^{\infty} \rho^k \mathbb{E}\left(e_{t-k}\right) = 0$$

$$\operatorname{var}(y_t) = \sum_{k=0}^{\infty} \rho^{2k} \operatorname{var}\left(e_{t-k}\right) = \frac{\sigma^2}{1-\rho^2}.$$

If the equation for y_t has an intercept, the above results are unchanged, except that the mean of y_t can be computed from the relationship

$$\mathbb{E}y_t = \alpha + \rho \mathbb{E}y_{t-1},$$

and solving for $\mathbb{E}y_t = \mathbb{E}y_{t-1}$ we find $\mathbb{E}y_t = \alpha/(1-\rho)$.

12.4 Lag Operator

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 12.4.1 *The **lag operator** L satisfies $\mathrm{L}y_t = y_{t-1}$.*

Defining $\mathrm{L}^2 = \mathrm{LL}$, we see that $\mathrm{L}^2 y_t = \mathrm{L}y_{t-1} = y_{t-2}$. In general, $\mathrm{L}^k y_t = y_{t-k}$.

The AR(1) model can be written in the format

$$y_t - \rho y_{t-1} + e_t$$

or

$$\left(1 - \rho \mathrm{L}\right) y_{t-1} = e_t.$$

The operator $\rho(\mathrm{L}) = (1 - \rho \mathrm{L})$ is a polynomial in the operator L. We say that the *root* of the polynomial is $1/\rho$, since $\rho(z) = 0$ when $z = 1/\rho$. We call $\rho(\mathrm{L})$ the autoregressive polynomial of y_t.

From Theorem 12.3.1, an AR(1) is stationary iff $|\rho| < 1$. Note that an equivalent way to say this is that an AR(1) is stationary iff the root of the autoregressive polynomial is larger than one (in absolute value).

12.5 Stationarity of AR(k)

The AR(k) model is

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_k y_{t-k} + e_t.$$

Using the lag operator,

$$y_t - \rho_1 \mathrm{L} y_t - \rho_2 \mathrm{L}^2 y_t - \cdots - \rho_k \mathrm{L}^k y_t = e_t,$$

or

$$\rho(\mathrm{L}) y_t = e_t$$

where

$$\rho(\mathrm{L}) = 1 - \rho_1 \mathrm{L} - \rho_2 \mathrm{L}^2 - \cdots - \rho_k \mathrm{L}^k.$$

We call $\rho(\mathrm{L})$ the autoregressive polynomial of y_t.

The *Fundamental Theorem of Algebra* says that any polynomial can be factored as

$$\rho(z) = \left(1 - \lambda_1^{-1} z\right)\left(1 - \lambda_2^{-1} z\right) \cdots \left(1 - \lambda_k^{-1} z\right)$$

where the $\lambda_1, ..., \lambda_k$ are the complex *roots* of $\rho(z)$, which satisfy $\rho(\lambda_j) = 0$.

We know that an AR(1) is stationary iff the absolute value of the root of its autoregressive polynomial is larger than one. For an AR(k), the requirement is that all roots are larger than one. Let $|\lambda|$ denote the modulus of a complex number λ.

Theorem 12.5.1 *The AR(k) is strictly stationary and ergodic if and only if $|\lambda_j| > 1$ for all j.*

One way of stating this is that "All roots lie outside the unit circle."

If one of the roots equals 1, we say that $\rho(\mathrm{L})$, and hence y_t, "has a unit root". This is a special case of non-stationarity, and is of great interest in applied time series.

12.6 Estimation

Let

$$\begin{aligned} \boldsymbol{x}_t &= \begin{pmatrix} 1 & y_{t-1} & y_{t-2} & \cdots & y_{t-k} \end{pmatrix}' \\ \boldsymbol{\beta} &= \begin{pmatrix} \alpha & \rho_1 & \rho_2 & \cdots & \rho_k \end{pmatrix}'. \end{aligned}$$

Then the model can be written as

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + e_t.$$

The OLS estimator is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

To study $\hat{\boldsymbol{\beta}}$, it is helpful to define the process $u_t = \boldsymbol{x}_t e_t$. Note that u_t is a MDS, since

$$\mathbb{E}\left(u_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left(\boldsymbol{x}_t e_t \mid \mathcal{F}_{t-1}\right) = \boldsymbol{x}_t \mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) = 0.$$

By Theorem 12.1.1, it is also strictly stationary and ergodic. Thus

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t = \frac{1}{T}\sum_{t=1}^{T} u_t \xrightarrow{p} \mathbb{E}\left(u_t\right) = 0. \tag{12.1}$$

The vector $\boldsymbol{x}_t$ is strictly stationary and ergodic, and by Theorem 12.1.1, so is $\boldsymbol{x}_t\boldsymbol{x}_t'$. Thus by the Ergodic Theorem,

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t' \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_t\boldsymbol{x}_t'\right) = \boldsymbol{Q}.$$

Combined with (12.1) and the continuous mapping theorem, we see that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right) \xrightarrow{p} \boldsymbol{Q}^{-1}\mathbf{0} = \mathbf{0}.$$

We have shown the following:

Theorem 12.6.1 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}y_t^2 < \infty$, then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ as $T \to \infty$.*

12.7 Asymptotic Distribution

Theorem 12.7.1 *MDS CLT. If $\boldsymbol{u}_t$ is a strictly stationary and ergodic MDS and $\mathbb{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right) = \boldsymbol{\Omega} < \infty$, then as $T \to \infty$,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right).$$

Since $\boldsymbol{x}_t e_t$ is a MDS, we can apply Theorem 12.7.1 to see that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right),$$

where

$$\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t' e_t^2).$$

Theorem 12.7.2 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}y_t^4 < \infty$, then as $T \to \infty$,*

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}\right).$$

This is identical in form to the asymptotic distribution of OLS in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

12.8 Bootstrap for Autoregressions

In the non-parametric bootstrap, we constructed the bootstrap sample by randomly resampling from the data values $\{y_t, \boldsymbol{x}_t\}$. This creates an iid bootstrap sample. Clearly, this cannot work in a time-series application, as this imposes inappropriate independence.

Briefly, there are two popular methods to implement bootstrap resampling for time-series data.

Method 1: Model-Based (Parametric) Bootstrap.

1. Estimate $\hat{\boldsymbol{\beta}}$ and residuals $\hat{e}_t$.
2. Fix an initial condition $(y_{-k+1}, y_{-k+2}, ..., y_0)$.
3. Simulate iid draws e_i^* from the empirical distribution of the residuals $\{\hat{e}_1, ..., \hat{e}_T\}$.
4. Create the bootstrap series y_t^* by the recursive formula

$$y_t^* = \hat{\alpha} + \hat{\rho}_1 y_{t-1}^* + \hat{\rho}_2 y_{t-2}^* + \cdots + \hat{\rho}_k y_{t-k}^* + e_t^*.$$

This construction imposes homoskedasticity on the errors e_i^*, which may be different than the properties of the actual e_i. It also presumes that the AR(k) structure is the truth.

Method 2: Block Resampling

1. Divide the sample into T/m blocks of length m.
2. Resample complete blocks. For each simulated sample, draw T/m blocks.
3. Paste the blocks together to create the bootstrap time-series y_t^*.
4. This allows for arbitrary stationary serial correlation, heteroskedasticity, and for model-misspecification.
5. The results may be sensitive to the block length, and the way that the data are partitioned into blocks.
6. May not work well in small samples.

12.9 Trend Stationarity

$$\begin{aligned} y_t &= \mu_0 + \mu_1 t + S_t \qquad (12.2) \\ S_t &= \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-l} + e_t, \qquad (12.3) \end{aligned}$$

or

$$y_t = \alpha_0 + \alpha_1 t + \rho_1 y_{t-1} + \rho_2 y_{t-1} + \cdots + \rho_k y_{t-k} + e_t. \qquad (12.4)$$

There are two essentially equivalent ways to estimate the autoregressive parameters $(\rho_1, ..., \rho_k)$.

- You can estimate (12.4) by OLS.
- You can estimate (12.2)-(12.3) sequentially by OLS. That is, first estimate (12.2), get the residual $\hat{S}_t$, and then perform regression (12.3) replacing S_t with $\hat{S}_t$. This procedure is sometimes called *Detrending*.

The reason why these two procedures are (essentially) the same is the Frisch-Waugh-Lovell theorem.

Seasonal Effects

There are three popular methods to deal with seasonal data.

- Include dummy variables for each season. This presumes that "seasonality" does not change over the sample.
- Use "seasonally adjusted" data. The seasonal factor is typically estimated by a two-sided weighted average of the data for that season in neighboring years. Thus the seasonally adjusted data is a "filtered" series. This is a flexible approach which can extract a wide range of seasonal factors. The seasonal adjustment, however, also alters the time-series correlations of the data.
- First apply a seasonal differencing operator. If s is the number of seasons (typically $s = 4$ or $s = 12$),
$$\Delta_s y_t = y_t - y_{t-s},$$
or the season-to-season change. The series $\Delta_s y_t$ is clearly free of seasonality. But the long-run trend is also eliminated, and perhaps this was of relevance.

12.10 Testing for Omitted Serial Correlation

For simplicity, let the null hypothesis be an AR(1):

$$y_t = \alpha + \rho y_{t-1} + u_t. \tag{12.5}$$

We are interested in the question if the error u_t is serially correlated. We model this as an AR(1):

$$u_t = \theta u_{t-1} + e_t \tag{12.6}$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$\begin{aligned} \mathbb{H}_0 &: \theta = 0 \\ \mathbb{H}_1 &: \theta \neq 0. \end{aligned}$$

We want to test $\mathbb{H}_0$ against $\mathbb{H}_1$.

To combine (12.5) and (12.6), we take (12.5) and lag the equation once:

$$y_{t-1} = \alpha + \rho y_{t-2} + u_{t-1}.$$

We then multiply this by θ and subtract from (12.5), to find

$$y_t - \theta y_{t-1} = \alpha - \theta\alpha + \rho y_{t-1} - \theta\rho y_{t-1} + u_t - \theta u_{t-1},$$

or

$$y_t = \alpha(1-\theta) + (\rho + \theta)\, y_{t-1} - \theta\rho y_{t-2} + e_t = AR(2).$$

Thus under $\mathbb{H}_0$, y_t is an AR(1), and under $\mathbb{H}_1$ it is an AR(2). $\mathbb{H}_0$ may be expressed as the restriction that the coefficient on y_{t-2} is zero.

An appropriate test of $\mathbb{H}_0$ against $\mathbb{H}_1$ is therefore a Wald test that the coefficient on y_{t-2} is zero. (A simple exclusion test).

In general, if the null hypothesis is that y_t is an AR(k), and the alternative is that the error is an AR(m), this is the same as saying that under the alternative y_t is an AR(k+m), and this is equivalent to the restriction that the coefficients on $y_{t-k-1}, ..., y_{t-k-m}$ are jointly zero. An appropriate test is the Wald test of this restriction.

12.11 Model Selection

What is the appropriate choice of k in practice? This is a problem of model selection.

One approach to model selection is to choose k based on a Wald tests.

Another is to minimize the AIC or BIC information criterion, e.g.

$$AIC(k) = \log \hat{\sigma}^2(k) + \frac{2k}{T},$$

where $\hat{\sigma}^2(k)$ is the estimated residual variance from an AR(k)

One ambiguity in defining the AIC criterion is that the sample available for estimation changes as k changes. (If you increase k, you need more initial conditions.) This can induce strange behavior in the AIC. The best remedy is to fix a upper value $\overline{k}$, and then reserve the first $\overline{k}$ as initial conditions, and then estimate the models AR(1), AR(2), ..., AR($\overline{k}$) on this (unified) sample.

12.12 Autoregressive Unit Roots

The AR(k) model is

$$\begin{aligned} \rho(\mathrm{L})y_t &= \mu + e_t \\ \rho(\mathrm{L}) &= 1 - \rho_1 \mathrm{L} - \cdots - \rho_k \mathrm{L}^k. \end{aligned}$$

As we discussed before, y_t has a unit root when $\rho(1) = 0$, or

$$\rho_1 + \rho_2 + \cdots + \rho_k = 1.$$

In this case, y_t is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the equation is the so-called Dickey-Fuller reparameterization:

$$\Delta y_t = \mu + \alpha_0 y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{12.7}$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter α_0 summarizes the information about the unit root, since $\rho(1) = -\alpha_0$. To see this, observe that the lag polynomial for the y_t computed from (12.7) is

$$(1 - \mathrm{L}) - \alpha_0 \mathrm{L} - \alpha_1 (\mathrm{L} - \mathrm{L}^2) - \cdots - \alpha_{k-1} (\mathrm{L}^{k-1} - \mathrm{L}^k)$$

But this must equal $\rho(\mathrm{L})$, as the models are equivalent. Thus

$$\rho(1) = (1 - 1) - \alpha_0 - (1 - 1) - \cdots - (1 - 1) = -\alpha_0.$$

Hence, the hypothesis of a unit root in y_t can be stated as

$$\mathbb{H}_0 : \alpha_0 = 0.$$

Note that the model is stationary if $\alpha_0 < 0$. So the natural alternative is

$$\mathbb{H}_1 : \alpha_0 < 0.$$

Under $\mathbb{H}_0$, the model for y_t is

$$\Delta y_t = \mu + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_{k-1} \Delta y_{t-(k-1)} + e_t,$$

which is an AR(k-1) in the first-difference Δy_t. Thus if y_t has a (single) unit root, then Δy_t is a stationary AR process. Because of this property, we say that if y_t is non-stationary but $\Delta^d y_t$ is stationary, then y_t is "integrated of order d", or $I(d)$. Thus a time series with unit root is $I(1)$.

Since α_0 is the parameter of a linear regression, the natural test statistic is the t-statistic for $\mathbb{H}_0$ from OLS estimation of (12.7). Indeed, this is the most popular unit root test, and is called the Augmented Dickey-Fuller (ADF) test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under $\mathbb{H}_0$, y_t is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

Theorem 12.12.1 ***Dickey-Fuller Theorem.***
Assume $\alpha_0 = 0$. *As* $T \to \infty$,

$$T\hat{\alpha}_0 \xrightarrow{d} (1 - \alpha_1 - \alpha_2 - \cdots - \alpha_{k-1})\, DF_\alpha$$

$$ADF = \frac{\hat{\alpha}_0}{s(\hat{\alpha}_0)} \to DF_t.$$

The limit distributions DF_α and DF_t are non-normal. They are skewed to the left, and have negative means.

The first result states that $\hat{\alpha}_0$ converges to its true value (of zero) at rate T, rather than the conventional rate of $T^{1/2}$. This is called a "super-consistent" rate of convergence.

The second result states that the t-statistic for $\hat{\alpha}_0$ converges to a limit distribution which is non-normal, but does not depend on the parameters α. This distribution has been extensively tabulated, and may be used for testing the hypothesis $\mathbb{H}_0$. Note: The standard error $s(\hat{\alpha}_0)$ is the conventional ("homoskedastic") standard error. But the theorem does not require an assumption of homoskedasticity. Thus the Dickey-Fuller test is robust to heteroskedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ when $ADF < c$, where c is the critical value from the ADF table. If the test rejects $\mathbb{H}_0$, this means that the evidence points to y_t being stationary. If the test does not reject $\mathbb{H}_0$, a common conclusion is that the data suggests that y_t is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data are stationary or not.

We have described the test for the setting of with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta y_t = \mu_1 + \mu_2 t + \alpha_0 y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{12.8}$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for α_0 from OLS estimation of (12.8).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

Chapter 13

Multivariate Time Series

A multivariate time series $\boldsymbol{y}_t$ is a vector process $m \times 1$. Let $\mathcal{F}_{t-1} = (\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...)$ be all lagged information at time t. The typical goal is to find the conditional expectation $\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})$. Note that since $\boldsymbol{y}_t$ is a vector, this conditional expectation is also a vector.

13.1 Vector Autoregressions (VARs)

A VAR model specifies that the conditional mean is a function of only a finite number of lags:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \mathbb{E}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}\right).$$

A linear VAR specifies that this conditional mean is linear in the arguments:

$$\mathbb{E}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}\right) = \boldsymbol{a}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots \boldsymbol{A}_k\boldsymbol{y}_{t-k}.$$

Observe that $\boldsymbol{a}_0$ is $m \times 1$,and each of $\boldsymbol{A}_1$ through $\boldsymbol{A}_k$ are $m \times m$ matrices.

Defining the $m \times 1$ regression error

$$e_t = \boldsymbol{y}_t - \mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}),$$

we have the VAR model

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{a}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots \boldsymbol{A}_k\boldsymbol{y}_{t-k} + \boldsymbol{e}_t \\ \mathbb{E}(\boldsymbol{e}_t \mid \mathcal{F}_{t-1}) &= \boldsymbol{0}. \end{aligned}$$

Alternatively, defining the $mk + 1$ vector

$$\boldsymbol{x}_t = \begin{pmatrix} 1 \\ \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \vdots \\ \boldsymbol{y}_{t-k} \end{pmatrix}$$

and the $m \times (mk + 1)$ matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_k \end{pmatrix},$$

then

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{e}_t.$$

The VAR model is a system of m equations. One way to write this is to let a_j' be the jth row of $\boldsymbol{A}$. Then the VAR system can be written as the equations

$$Y_{jt} = a_j'\boldsymbol{x}_t + e_{jt}.$$

Unrestricted VARs were introduced to econometrics by Sims (1980).

13.2 Estimation

Consider the moment conditions

$$\mathbb{E}\left(\boldsymbol{x}_t e_{jt}\right) = \mathbf{0},$$

$j = 1, ..., m$. These are implied by the VAR model, either as a regression, or as a linear projection.

The GMM estimator corresponding to these moment conditions is equation-by-equation OLS

$$\hat{\boldsymbol{a}}_j = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_j.$$

An alternative way to compute this is as follows. Note that

$$\hat{\boldsymbol{a}}_j' = \boldsymbol{y}_j'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

And if we stack these to create the estimate $\hat{A}$, we find

$$\begin{aligned}\hat{\boldsymbol{A}} &= \begin{pmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_{m+1}' \end{pmatrix} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},\end{aligned}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_m \end{pmatrix}$$

the $T \times m$ matrix of the stacked $\boldsymbol{y}_t'$.

This (system) estimator is known as the SUR (Seemingly Unrelated Regressions) estimator, and was originally derived by Zellner (1962)

13.3 Restricted VARs

The unrestricted VAR is a system of m equations, each with the same set of regressors. A restricted VAR imposes restrictions on the system. For example, some regressors may be excluded from some of the equations. Restrictions may be imposed on individual equations, or across equations. The GMM framework gives a convenient method to impose such restrictions on estimation.

13.4 Single Equation from a VAR

Often, we are only interested in a single equation out of a VAR system. This takes the form

$$y_{jt} = \boldsymbol{a}_j'\boldsymbol{x}_t + e_t,$$

and $\boldsymbol{x}_t$ consists of lagged values of y_{jt} and the other $y_{lt}'s$. In this case, it is convenient to re-define the variables. Let $y_t = y_{jt}$, and $\boldsymbol{z}_t$ be the other variables. Let $e_t = e_{jt}$ and $\beta = a_j$. Then the single equation takes the form

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \tag{13.1}$$

and

$$\boldsymbol{x}_t = \left[\begin{pmatrix} 1 & \boldsymbol{y}_{t-1} & \cdots & \boldsymbol{y}_{t-k} & \boldsymbol{z}_{t-1}' & \cdots & \boldsymbol{z}_{t-k}' \end{pmatrix}'\right].$$

This is just a conventional regression with time series data.

13.5 Testing for Omitted Serial Correlation

Consider the problem of testing for omitted serial correlation in equation (13.1). Suppose that e_t is an AR(1). Then

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + e_t \\ e_t &= \theta e_{t-1} + u_t \\ \mathbb{E}\left(u_t \mid \mathcal{F}_{t-1}\right) &= 0. \end{aligned} \tag{13.2}$$

Then the null and alternative are

$$\mathbb{H}_0 : \theta = 0 \qquad \mathbb{H}_1 : \theta \neq 0.$$

Take the equation $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$, and subtract off the equation once lagged multiplied by θ, to get

$$\begin{aligned} y_t - \theta y_{t-1} &= \left(\boldsymbol{x}_t'\boldsymbol{\beta} + e_t\right) - \theta\left(\boldsymbol{x}_{t-1}'\boldsymbol{\beta} + e_{t-1}\right) \\ &= \boldsymbol{x}_t'\boldsymbol{\beta} - \theta\boldsymbol{x}_{t-1}\boldsymbol{\beta} + e_t - \theta e_{t-1}, \end{aligned}$$

or

$$y_t = \theta y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{x}_{t-1}'\boldsymbol{\gamma} + u_t, \tag{13.3}$$

which is a valid regression model.

So testing $\mathbb{H}_0$ versus $\mathbb{H}_1$ is equivalent to testing for the significance of adding $(y_{t-1}, \boldsymbol{x}_{t-1})$ to the regression. This can be done by a Wald test. We see that an appropriate, general, and simple way to test for omitted serial correlation is to test the significance of extra lagged values of the dependent variable and regressors.

You may have heard of the Durbin-Watson test for omitted serial correlation, which once was very popular, and is still routinely reported by conventional regression packages. The DW test is appropriate only when regression $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$ is not dynamic (has no lagged values on the RHS), and e_t is iid $\mathrm{N}(0, \sigma^2)$. Otherwise it is invalid.

Another interesting fact is that (13.2) is a special case of (13.3), under the restriction $\gamma = -\boldsymbol{\beta}\theta$. This restriction, which is called a common factor restriction, may be tested if desired. If valid, the model (13.2) may be estimated by iterated GLS. (A simple version of this estimator is called Cochrane-Orcutt.) Since the common factor restriction appears arbitrary, and is typically rejected empirically, direct estimation of (13.2) is uncommon in recent applications.

13.6 Selection of Lag Length in an VAR

If you want a data-dependent rule to pick the lag length k in a VAR, you may either use a testing-based approach (using, for example, the Wald statistic), or an information criterion approach. The formula for the AIC and BIC are

$$\begin{aligned} AIC(k) &= \log\det\left(\hat{\boldsymbol{\Omega}}(k)\right) + 2\frac{p}{T} \\ BIC(k) &= \log\det\left(\hat{\boldsymbol{\Omega}}(k)\right) + \frac{p\log(T)}{T} \\ \hat{\boldsymbol{\Omega}}(k) &= \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t(k)\hat{\boldsymbol{e}}_t(k)' \\ p &= m(km+1) \end{aligned}$$

where p is the number of parameters in the model, and $\hat{\boldsymbol{e}}_t(k)$ is the OLS residual vector from the model with k lags. The log determinant is the criterion from the multivariate normal likelihood.

13.7 Granger Causality

Partition the data vector into $(\boldsymbol{y}_t, \boldsymbol{z}_t)$. Define the two information sets

$$\begin{aligned}\mathcal{F}_{1t} &= \left(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...\right) \\ \mathcal{F}_{2t} &= \left(\boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{y}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{y}_{t-2}, \boldsymbol{z}_{t-2}, , ...\right)\end{aligned}$$

The information set $\mathcal{F}_{1t}$ is generated only by the history of $\boldsymbol{y}_t$, and the information set $\mathcal{F}_{2t}$ is generated by both $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$. The latter has more information.

We say that $\boldsymbol{z}_t$ does not *Granger-cause* $\boldsymbol{y}_t$ if

$$\mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{1,t-1}\right) = \mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{2,t-1}\right).$$

That is, conditional on information in lagged $\boldsymbol{y}_t$, lagged $\boldsymbol{z}_t$ does not help to forecast $\boldsymbol{y}_t$. If this condition does not hold, then we say that $\boldsymbol{z}_t$ Granger-causes $\boldsymbol{y}_t$.

The reason why we call this "Granger Causality" rather than "causality" is because this is not a physical or structure definition of causality. If $\boldsymbol{z}_t$ is some sort of forecast of the future, such as a futures price, then $\boldsymbol{z}_t$ may help to forecast $\boldsymbol{y}_t$ even though it does not "cause" $\boldsymbol{y}_t$. This definition of causality was developed by Granger (1969) and Sims (1972).

In a linear VAR, the equation for $\boldsymbol{y}_t$ is

$$\boldsymbol{y}_t = \alpha + \rho_1 \boldsymbol{y}_{t-1} + \cdots + \rho_k \boldsymbol{y}_{t-k} + \boldsymbol{z}_{t-1}'\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{z}_{t-k}'\boldsymbol{\gamma}_k + e_t.$$

In this equation, $\boldsymbol{z}_t$ does not Granger-cause $\boldsymbol{y}_t$ if and only if

$$\mathbb{H}_0 : \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \cdots = \boldsymbol{\gamma}_k = 0.$$

This may be tested using an exclusion (Wald) test.

This idea can be applied to blocks of variables. That is, $\boldsymbol{y}_t$ and/or $\boldsymbol{z}_t$ can be vectors. The hypothesis can be tested by using the appropriate multivariate Wald test.

If it is found that $\boldsymbol{z}_t$ does not Granger-cause $\boldsymbol{y}_t$, then we deduce that our time-series model of $\mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}\right)$ does not require the use of $\boldsymbol{z}_t$. Note, however, that $\boldsymbol{z}_t$ may still be useful to explain other features of $\boldsymbol{y}_t$, such as the conditional variance.

Clive W. J. Granger

Clive Granger (1934-2009) of England was one of the leading figures in time-series econometrics, and co-winner in 2003 of the Nobel Memorial Prize in Economic Sciences (along with Robert Engle). In addition to formalizing the definition of causality known as Granger causality, he invented the concept of cointegration, introduced spectral methods into econometrics, and formalized methods for the combination of forecasts.

13.8 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

Definition 13.8.1 *The $m \times 1$ series $\boldsymbol{y}_t$ is **cointegrated** if $\boldsymbol{y}_t$ is $I(1)$ yet there exists $\boldsymbol{\beta}$, $m \times r$, of rank r, such that $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{y}_t$ is $I(0)$. The r vectors in $\boldsymbol{\beta}$ are called the **cointegrating vectors**.*

If the series $\boldsymbol{y}_t$ is not cointegrated, then $r = 0$. If $r = m$, then $\boldsymbol{y}_t$ is $I(0)$. For $0 < r < m$, $\boldsymbol{y}_t$ is $I(1)$ and cointegrated.

In some cases, it may be believed that $\boldsymbol{\beta}$ is known a priori. Often, $\boldsymbol{\beta} = (1 \quad -1)'$. For example, if $\boldsymbol{y}_t$ is a pair of interest rates, then $\boldsymbol{\beta} = (1 \quad -1)'$ specifies that the spread (the difference in returns) is stationary. If $\boldsymbol{y} = (\log(Consumption) \quad \log(Income))'$, then $\boldsymbol{\beta} = (1 \quad -1)'$ specifies that $\log(Consumption/Income)$ is stationary.

In other cases, $\boldsymbol{\beta}$ may not be known.

If $\boldsymbol{y}_t$ is cointegrated with a single cointegrating vector ($r = 1$), then it turns out that $\boldsymbol{\beta}$ can be consistently estimated by an OLS regression of one component of $\boldsymbol{y}_t$ on the others. Thus $\boldsymbol{y}_t = (Y_{1t}, Y_{2t})$ and $\boldsymbol{\beta} = (\beta_1 \ \beta_2)$ and normalize $\beta_1 = 1$. Then $\hat{\beta}_2 = (\boldsymbol{y}_2'\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{y}_1 \xrightarrow{p} \beta_2$. Furthermore this estimation is super-consistent: $T(\hat{\beta}_2 - \beta_2) \xrightarrow{d} Limit$, as first shown by Stock (1987). This is not, in general, a good method to estimate $\boldsymbol{\beta}$, but it is useful in the construction of alternative estimators and tests.

We are often interested in testing the hypothesis of no cointegration:

$$\begin{aligned} \mathbb{H}_0 &: r = 0 \\ \mathbb{H}_1 &: r > 0. \end{aligned}$$

Suppose that $\boldsymbol{\beta}$ is known, so $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{y}_t$ is known. Then under $\mathbb{H}_0$ $\boldsymbol{z}_t$ is $I(1)$, yet under $\mathbb{H}_1$ $\boldsymbol{z}_t$ is $I(0)$. Thus $\mathbb{H}_0$ can be tested using a univariate ADF test on $\boldsymbol{z}_t$.

When $\boldsymbol{\beta}$ is unknown, Engle and Granger (1987) suggested using an ADF test on the estimated residual $\hat{z}_t = \hat{\boldsymbol{\beta}}'\boldsymbol{y}_t$, from OLS of y_{1t} on y_{2t}. Their justification was Stock's result that $\hat{\boldsymbol{\beta}}$ is super-consistent under $\mathbb{H}_1$. Under $\mathbb{H}_0$, however, $\hat{\boldsymbol{\beta}}$ is not consistent, so the ADF critical values are not appropriate. The asymptotic distribution was worked out by Phillips and Ouliaris (1990).

When the data have time trends, it may be necessary to include a time trend in the estimated cointegrating regression. Whether or not the time trend is included, the asymptotic distribution of the test is affected by the presence of the time trend. The asymptotic distribution was worked out in B. Hansen (1992).

13.9 Cointegrated VARs

We can write a VAR as

$$\begin{aligned} \boldsymbol{A}(\mathrm{L})\boldsymbol{y}_t &= \boldsymbol{e}_t \\ \boldsymbol{A}(\mathrm{L}) &= \boldsymbol{I} - \boldsymbol{A}_1\mathrm{L} - \boldsymbol{A}_2\mathrm{L}^2 - \cdots - \boldsymbol{A}_k\mathrm{L}^k \end{aligned}$$

or alternatively as

$$\Delta\boldsymbol{y}_t = \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta\boldsymbol{y}_{t-1} + \boldsymbol{e}_t$$

where

$$\begin{aligned} \boldsymbol{\Pi} &= -\boldsymbol{A}(1) \\ &= -\boldsymbol{I} + \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k. \end{aligned}$$

Theorem 13.9.1 ***Granger Representation Theorem***
$\boldsymbol{y}_t$ is cointegrated with $m \times r$ $\boldsymbol{\beta}$ if and only if $\operatorname{rank}(\boldsymbol{\Pi}) = r$ *and* $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ *where α is $m \times r$,* $\operatorname{rank}(\boldsymbol{\alpha}) = r$.

Thus cointegration imposes a restriction upon the parameters of a VAR. The restricted model can be written as

$$\begin{aligned}\Delta \boldsymbol{y}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta \boldsymbol{y}_{t-1} + \boldsymbol{e}_t \\ \Delta \boldsymbol{y}_t &= \boldsymbol{\alpha}\boldsymbol{z}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta \boldsymbol{y}_{t-1} + \boldsymbol{e}_t.\end{aligned}$$

If $\boldsymbol{\beta}$ is known, this can be estimated by OLS of $\Delta \boldsymbol{y}_t$ on $\boldsymbol{z}_{t-1}$ and the lags of $\Delta \boldsymbol{y}_t$.

If $\boldsymbol{\beta}$ is unknown, then estimation is done by "reduced rank regression", which is least-squares subject to the stated restriction. Equivalently, this is the MLE of the restricted parameters under the assumption that $\boldsymbol{e}_t$ is iid $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$.

One difficulty is that $\boldsymbol{\beta}$ is not identified without normalization. When $r = 1$, we typically just normalize one element to equal unity. When $r > 1$, this does not work, and different authors have adopted different identification schemes.

In the context of a cointegrated VAR estimated by reduced rank regression, it is simple to test for cointegration by testing the rank of $\boldsymbol{\Pi}$. These tests are constructed as likelihood ratio (LR) tests. As they were discovered by Johansen (1988, 1991, 1995), they are typically called the "Johansen Max and Trace" tests. Their asymptotic distributions are non-standard, and are similar to the Dickey-Fuller distributions.

Chapter 14

Limited Dependent Variables

A "limited dependent variable" y is one which takes a "limited" set of values. The most common cases are

- Binary: $y \in \{0, 1\}$
- Multinomial: $y \in \{0, 1, 2, ..., k\}$
- Integer: $y \in \{0, 1, 2, ...\}$
- Censored: $y \in \mathbb{R}^+$

The traditional approach to the estimation of limited dependent variable (LDV) models is parametric maximum likelihood. A parametric model is constructed, allowing the construction of the likelihood function. A more modern approach is semi-parametric, eliminating the dependence on a parametric distributional assumption. We will discuss only the first (parametric) approach, due to time constraints. They still constitute the majority of LDV applications. If, however, you were to write a thesis involving LDV estimation, you would be advised to consider employing a semi-parametric estimation approach.

For the parametric approach, estimation is by MLE. A major practical issue is construction of the likelihood function.

14.1 Binary Choice

The dependent variable $y_i \in \{0, 1\}$. This represents a Yes/No outcome. Given some regressors $\boldsymbol{x}_i$, the goal is to describe $\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i)$, as this is the full conditional distribution.

The linear probability model specifies that

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

As $\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = \mathbb{E}(y_i \mid \boldsymbol{x}_i)$, this yields the regression: $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ which can be estimated by OLS. However, the linear probability model does not impose the restriction that $0 \leq \mathbb{P}(y_i \mid \boldsymbol{x}_i) \leq 1$. Even so estimation of a linear probability model is a useful starting point for subsequent analysis.

The standard alternative is to use a function of the form

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

where $F(\cdot)$ is a known CDF, typically assumed to be symmetric about zero, so that $F(u) = 1 - F(-u)$. The two standard choices for F are

- Logistic: $F(u) = (1 + e^{-u})^{-1}$.

- Normal: $F(u) = \Phi(u)$.

If F is logistic, we call this the *logit* model, and if F is normal, we call this the *probit* model. This model is identical to the latent variable model

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\sim F(\cdot) \\ y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

For then

$$\begin{aligned} \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) &= \mathbb{P}(y_i^* > 0 \mid \boldsymbol{x}_i) \\ &= \mathbb{P}\left(\boldsymbol{x}_i'\boldsymbol{\beta} + e_i > 0 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{P}\left(e_i > -\boldsymbol{x}_i'\boldsymbol{\beta} \mid \boldsymbol{x}_i\right) \\ &= 1 - F\left(-\boldsymbol{x}_i'\boldsymbol{\beta}\right) \\ &= F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right). \end{aligned}$$

Estimation is by maximum likelihood. To construct the likelihood, we need the conditional distribution of an individual observation. Recall that if y is Bernoulli, such that $\mathbb{P}(y = 1) = p$ and $\mathbb{P}(y = 0) = 1 - p$, then we can write the density of y as

$$f(y) = p^y(1-p)^{1-y}, \qquad y = 0, 1.$$

In the Binary choice model, y_i is conditionally Bernoulli with $\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = p_i = F(\boldsymbol{x}_i'\boldsymbol{\beta})$. Thus the conditional density is

$$\begin{aligned} f(y_i \mid \boldsymbol{x}_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{y_i}\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-y_i}. \end{aligned}$$

Hence the log-likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} \log\left(F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{y_i}\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \left[y_i \log F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + (1-y_i)\log\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)\right] \\ &= \sum_{y_i=1} \log F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + \sum_{y_i=0} \log\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right). \end{aligned}$$

The MLE $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which maximizes $\log L(\boldsymbol{\beta})$. Standard errors and test statistics are computed by asymptotic approximations. Details of such calculations are left to more advanced courses.

14.2 Count Data

If $y \in \{0, 1, 2, ...\}$, a typical approach is to employ *Poisson regression.* This model specifies that

$$\begin{aligned} \mathbb{P}(y_i = k \mid \boldsymbol{x}_i) &= \frac{\exp(-\lambda_i)\lambda_i^k}{k!}, \qquad k = 0, 1, 2, ... \\ \lambda_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}). \end{aligned}$$

The conditional density is the Poisson with parameter λ_i. The functional form for λ_i has been picked to ensure that $\lambda_i > 0$.

The log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) = \sum_{i=1}^{n} \left(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) + y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!) \right).$$

The MLE is the value $\hat{\boldsymbol{\beta}}$ which maximizes $\log L(\boldsymbol{\beta})$.

Since

$$\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

is the conditional mean, this motivates the label Poisson "regression."

Also observe that the model implies that

$$\operatorname{var}\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

so the model imposes the restriction that the conditional mean and variance of y_i are the same. This may be considered restrictive. A generalization is the negative binomial.

14.3 Censored Data

The idea of "censoring" is that some data above or below a threshold are mis-reported at the threshold. Thus the model is that there is some latent process y_i^* with unbounded support, but we observe only

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \tag{14.1}$$

(This is written for the case of the threshold being zero, any known value can substitute.) The observed data y_i therefore come from a mixed continuous/discrete distribution.

Censored models are typically applied when the data set has a meaningful proportion (say 5% or higher) of data at the boundary of the sample support. The censoring process may be explicit in data collection, or it may be a by-product of economic constraints.

An example of a data collection censoring is top-coding of income. In surveys, incomes above a threshold are typically reported at the threshold.

The first censored regression model was developed by Tobin (1958) to explain consumption of durable goods. Tobin observed that for many households, the consumption level (purchases) in a particular period was zero. He proposed the latent variable model

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\sim iid\ \mathrm{N}(0, \sigma^2) \end{aligned}$$

with the observed variable y_i generated by the censoring equation (14.1). This model (now called the Tobit) specifies that the latent (or ideal) value of consumption may be negative (the household would prefer to sell than buy). All that is reported is that the household purchased zero units of the good.

The naive approach to estimate $\boldsymbol{\beta}$ is to regress y_i on $\boldsymbol{x}_i$. This does not work because regression estimates $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$, not $\mathbb{E}\left(y_i^* \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$, and the latter is of interest. Thus OLS will be biased for the parameter of interest $\boldsymbol{\beta}$.

[Note: it is still possible to estimate $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$ by LS techniques. The Tobit framework postulates that this is not inherently interesting, that the parameter of $\boldsymbol{\beta}$ is defined by an alternative statistical structure.]

Consistent estimation will be achieved by the MLE. To construct the likelihood, observe that the probability of being censored is

$$\begin{aligned}\mathbb{P}\left(y_i = 0 \mid \boldsymbol{x}_i\right) &= \mathbb{P}\left(y_i^* < 0 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{P}\left(\boldsymbol{x}_i'\boldsymbol{\beta} + e_i < 0 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{P}\left(\frac{e_i}{\sigma} < -\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \mid \boldsymbol{x}_i\right) \\ &= \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right).\end{aligned}$$

The conditional distribution function above zero is Gaussian:

$$\mathbb{P}\left(y_i = y \mid \boldsymbol{x}_i\right) = \int_0^y \sigma^{-1}\phi\left(\frac{z - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) dz, \qquad y > 0.$$

Therefore, the density function can be written as

$$f\left(y \mid \boldsymbol{x}_i\right) = \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)^{1(y=0)} \left[\sigma^{-1}\phi\left(\frac{z - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]^{1(y>0)},$$

where $1\left(\cdot\right)$ is the indicator function.

Hence the log-likelihood is a mixture of the probit and the normal:

$$\begin{aligned}\log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) \\ &= \sum_{y_i=0} \log \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) + \sum_{y_i>0} \log\left[\sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right].\end{aligned}$$

The MLE is the value $\hat{\boldsymbol{\beta}}$ which maximizes $\log L(\boldsymbol{\beta})$.

14.4 Sample Selection

The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is systematically different from the population of interest. For example, if you ask for volunteers for an experiment, and they wish to extrapolate the effects of the experiment on a general population, you should worry that the people who volunteer may be systematically different from the general population. This has great relevance for the evaluation of anti-poverty and job-training programs, where the goal is to assess the effect of "training" on the general population, not just on the volunteers.

A simple sample selection model can be written as the latent model

$$\begin{aligned}y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_{1i} \\ T_i &= 1\left(\boldsymbol{z}_i'\boldsymbol{\gamma} + e_{0i} > 0\right)\end{aligned}$$

where $1\left(\cdot\right)$ is the indicator function. The dependent variable y_i is observed if (and only if) $T_i = 1$. Else it is unobserved.

For example, y_i could be a wage, which can be observed only if a person is employed. The equation for T_i is an equation specifying the probability that the person is employed.

The model is often completed by specifying that the errors are jointly normal

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

It is presumed that we observe $\{\boldsymbol{x}_i, \boldsymbol{z}_i, T_i\}$ for all observations.

Under the normality assumption,

$$e_{1i} = \rho e_{0i} + v_i,$$

where v_i is independent of $e_{0i} \sim \mathrm{N}(0,1)$. A useful fact about the standard normal distribution is that

$$\mathbb{E}\left(e_{0i} \mid e_{0i} > -x\right) = \lambda(x) = \frac{\phi(x)}{\Phi(x)},$$

and the function $\lambda(x)$ is called the inverse Mills ratio.

The naive estimator of $\boldsymbol{\beta}$ is OLS regression of y_i on $\boldsymbol{x}_i$ for those observations for which y_i is available. The problem is that this is equivalent to conditioning on the event $\{T_i = 1\}$. However,

$$\begin{aligned}
\mathbb{E}\left(e_{1i} \mid T_i = 1, \boldsymbol{z}_i\right) &= \mathbb{E}\left(e_{1i} \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) \\
&= \rho\mathbb{E}\left(e_{0i} \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) + \mathbb{E}\left(v_i \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) \\
&= \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right),
\end{aligned}$$

which is non-zero. Thus

$$e_{1i} = \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right) + u_i,$$

where

$$\mathbb{E}\left(u_i \mid T_i = 1, \boldsymbol{z}_i\right) = 0.$$

Hence

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right) + u_i \tag{14.2}$$

is a valid regression equation for the observations for which $T_i = 1$.

Heckman (1979) observed that we could consistently estimate $\boldsymbol{\beta}$ and ρ from this equation, if $\boldsymbol{\gamma}$ were known. It is unknown, but also can be consistently estimated by a Probit model for selection. The "Heckit" estimator is thus calculated as follows

- Estimate $\hat{\boldsymbol{\gamma}}$ from a Probit, using regressors $\boldsymbol{z}_i$. The binary dependent variable is T_i.
- Estimate $\left(\hat{\boldsymbol{\beta}}, \hat{\rho}\right)$ from OLS of y_i on $\boldsymbol{x}_i$ and $\lambda(\boldsymbol{z}_i'\hat{\boldsymbol{\gamma}})$.
- The OLS standard errors will be incorrect, as this is a two-step estimator. They can be corrected using a more complicated formula. Or, alternatively, by viewing the Probit/OLS estimation equations as a large joint GMM problem.

The Heckit estimator is frequently used to deal with problems of sample selection. However, the estimator is built on the assumption of normality, and the estimator can be quite sensitive to this assumption. Some modern econometric research is exploring how to relax the normality assumption.

The estimator can also work quite poorly if $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ does not have much in-sample variation. This can happen if the Probit equation does not "explain" much about the selection choice. Another potential problem is that if $\boldsymbol{z}_i = \boldsymbol{x}_i$, then $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ can be highly collinear with $\boldsymbol{x}_i$, so the second step OLS estimator will not be able to precisely estimate $\boldsymbol{\beta}$. Based this observation, it is typically recommended to find a valid exclusion restriction: a variable should be in $\boldsymbol{z}_i$ which is not in $\boldsymbol{x}_i$. If this is valid, it will ensure that $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ is not collinear with $\boldsymbol{x}_i$, and hence improve the second stage estimator's precision.

Chapter 15

Panel Data

A panel is a set of observations on individuals, collected over time. An observation is the pair $\{y_{it}, \boldsymbol{x}_{it}\}$, where the i subscript denotes the individual, and the t subscript denotes time. A panel may be *balanced*:

$$\{y_{it}, \boldsymbol{x}_{it}\} : t = 1, ..., T; \quad i = 1, ..., n,$$

or *unbalanced*:

$$\{y_{it}, \boldsymbol{x}_{it}\} : \text{For } i = 1, ..., n, \quad t = \underline{t}_i, ..., \bar{t}_i.$$

15.1 Individual-Effects Model

The standard panel data specification is that there is an individual-specific effect which enters linearly in the regression

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it}.$$

The typical maintained assumptions are that the individuals i are mutually independent, that u_i and e_{it} are independent, that e_{it} is iid across individuals and time, and that e_{it} is uncorrelated with $\boldsymbol{x}_{it}$.

OLS of y_{it} on $\boldsymbol{x}_{it}$ is called pooled estimation. It is consistent if

$$\mathbb{E}\left(\boldsymbol{x}_{it} u_i\right) = 0 \tag{15.1}$$

If this condition fails, then OLS is inconsistent. (15.1) fails if the individual-specific unobserved effect u_i is correlated with the observed explanatory variables $\boldsymbol{x}_{it}$. This is often believed to be plausible if u_i is an omitted variable.

If (15.1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice.

Condition (15.1) is called the *random effects hypothesis.* It is a strong assumption, and most applied researchers try to avoid its use.

15.2 Fixed Effects

This is the most common technique for estimation of non-dynamic linear panel regressions.

The motivation is to allow u_i to be arbitrary, and have arbitrary correlated with $\boldsymbol{x}_i$. The goal is to eliminate u_i from the estimator, and thus achieve invariance.

There are several derivations of the estimator.

First, let

$$d_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases},$$

and

$$\boldsymbol{d}_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix},$$

an $n \times 1$ dummy vector with a "1" in the $i'th$ place. Let

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

Then note that

$$u_i = \boldsymbol{d}_i' \boldsymbol{u},$$

and

$$y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta} + \boldsymbol{d}_i' \boldsymbol{u} + e_{it}. \tag{15.2}$$

Observe that

$$\mathbb{E}\left(e_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{d}_i\right) = 0,$$

so (15.2) is a valid regression, with $\boldsymbol{d}_i$ as a regressor along with $\boldsymbol{x}_i$.

OLS on (15.2) yields estimator $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}\right)$. Conventional inference applies.

Observe that

- This is generally consistent.
- If $\boldsymbol{x}_{it}$ contains an intercept, it will be collinear with $\boldsymbol{d}_i$, so the intercept is typically omitted from $\boldsymbol{x}_{it}$.
- Any regressor in $\boldsymbol{x}_{it}$ which is constant over time for all individuals (e.g., their gender) will be collinear with $\boldsymbol{d}_i$, so will have to be omitted.
- There are $n + k$ regression parameters, which is quite large as typically n is very large.

Computationally, you do not want to actually implement conventional OLS estimation, as the parameter space is too large. OLS estimation of $\boldsymbol{\beta}$ proceeds by the FWL theorem. Stacking the observations together:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{u} + e,$$

then by the FWL theorem,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{y}^{*}\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{y}^{*} &= \boldsymbol{y} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{y} \\ \boldsymbol{X}^{*} &= \boldsymbol{X} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{X}. \end{aligned}$$

Since the regression of y_{it} on $\boldsymbol{d}_i$ is a regression onto individual-specific dummies, the predicted value from these regressions is the individual specific mean $\overline{y}_i$, and the residual is the demean value

$$y_{it}^{*} = y_{it} - \overline{y}_i.$$

The fixed effects estimator $\hat{\boldsymbol{\beta}}$ is OLS of y_{it}^{*} on $\boldsymbol{x}_{it}^{*}$, the dependent variable and regressors in deviation-from-mean form.

Another derivation of the estimator is to take the equation

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it},$$

and then take individual-specific means by taking the average for the $i'th$ individual:

$$\frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} y_{it} = \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\bar{t}_i} e_{it}$$

or

$$\overline{y}_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + u_i + \overline{e}_i.$$

Subtracting, we find

$$y_{it}^* = \boldsymbol{x}_{it}^{*\prime}\boldsymbol{\beta} + e_{it}^*,$$

which is free of the individual-effect u_i.

15.3 Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \alpha y_{it-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it}. \tag{15.3}$$

This is a model suitable for studying dynamic behavior of individual agents.

Unfortunately, the fixed effects estimator is inconsistent, at least if T is held finite as $n \to \infty$. This is because the sample mean of y_{it-1} is correlated with that of e_{it}.

The standard approach to estimate a dynamic panel is to combine first-differencing with IV or GMM. Taking first-differences of (15.3) eliminates the individual-specific effect:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \Delta e_{it}. \tag{15.4}$$

However, if e_{it} is iid, then it will be correlated with Δy_{it-1} :

$$\mathbb{E}\left(\Delta y_{it-1}\Delta e_{it}\right) = \mathbb{E}\left(\left(y_{it-1} - y_{it-2}\right)\left(e_{it} - e_{it-1}\right)\right) = -\mathbb{E}\left(y_{it-1}e_{it-1}\right) = -\sigma_e^2.$$

So OLS on (15.4) will be inconsistent.

But if there are valid instruments, then IV or GMM can be used to estimate the equation. Typically, we use lags of the dependent variable, two periods back, as y_{t-2} is uncorrelated with Δe_{it}. Thus values of y_{it-k}, $k \geq 2$, are valid instruments.

Hence a valid estimator of α and $\boldsymbol{\beta}$ is to estimate (15.4) by IV using y_{t-2} as an instrument for Δy_{t-1} (which is just identified). Alternatively, GMM using y_{t-2} and y_{t-3} as instruments (which is overidentified, but loses a time-series observation).

A more sophisticated GMM estimator recognizes that for time-periods later in the sample, there are more instruments available, so the instrument list should be different for each equation. This is conveniently organized by the GMM principle, as this enables the moments from the different time-periods to be stacked together to create a list of all the moment conditions. A simple application of GMM yields the parameter estimates and standard errors.

Chapter 16

Nonparametrics

16.1 Kernel Density Estimation

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $(X_1, ..., X_n\}$ While $F(x)$ can be estimated by the EDF $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$, we cannot define $\frac{d}{dx}\hat{F}(x)$ since $\hat{F}(x)$ is a step function. The standard **nonparametric** method to estimate $f(x)$ is based on **smoothing** using a kernel.

While we are typically interested in estimating the entire function $f(x)$, we can simply focus on the problem where x is a specific fixed number, and then see how the method generalizes to estimating the entire function.

Definition 16.1.1 *$K(u)$ is a **second-order kernel function** if it is a symmetric zero-mean density function.*

Three common choices for kernels include the **Normal**

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

the **Epanechnikov**

$$K(u) = \begin{cases} \frac{3}{4}\left(1-u^2\right), & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

and the **Biweight** or **Quartic**

$$K(u) = \begin{cases} \frac{15}{16}\left(1-u^2\right)^2, & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

In practice, the choice between these three rarely makes a meaningful difference in the estimates.

The kernel functions are used to smooth the data. The amount of smoothing is controlled by the **bandwidth** $h > 0$. Let

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

be the kernel K rescaled by the bandwidth h. The kernel density estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x).$$

This estimator is the average of a set of weights. If a large number of the observations X_i are near x, then the weights are relatively large and $\hat{f}(x)$ is larger. Conversely, if only a few X_i are near x, then the weights are small and $\hat{f}(x)$ is small. The bandwidth h controls the meaning of "near".

Interestingly, $\hat{f}(x)$ is a valid density. That is, $\hat{f}(x) \geq 0$ for all x, and

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_h\left(X_i - x\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(u\right) du = 1$$

where the second-to-last equality makes the change-of-variables $u = (X_i - x)/h$.

We can also calculate the moments of the density $\hat{f}(x)$. The mean is

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} xK_h\left(X_i - x\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(X_i + uh\right) K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} K\left(u\right) du + \frac{1}{n}\sum_{i=1}^{n} h \int_{-\infty}^{\infty} uK\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

the sample mean of the X_i, where the second-to-last equality used the change-of-variables $u = (X_i - x)/h$ which has Jacobian h.

The second moment of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2K_h\left(X_i - x\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(X_i + uh\right)^2 K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n} X_i h \int_{-\infty}^{\infty} K(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2 \int_{-\infty}^{\infty} u^2K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2
\end{aligned}$$

where

$$\sigma_K^2 = \int_{-\infty}^{\infty} u^2K\left(u\right) du$$

is the variance of the kernel. It follows that the variance of the density $\hat{f}(x)$ is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\hat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \hat{\sigma}^2 + h^2\sigma_K^2
\end{aligned}$$

Thus the variance of the estimated density is inflated by the factor $h^2\sigma_K^2$ relative to the sample moment.

16.2 Asymptotic MSE for Kernel Estimates

For fixed x and bandwidth h observe that

$$\mathbb{E}K_h(X-x) = \int_{-\infty}^{\infty} K_h(z-x) f(z)dz = \int_{-\infty}^{\infty} K_h(uh) f(x+hu)hdu = \int_{-\infty}^{\infty} K(u) f(x+hu)du$$

The second equality uses the change-of variables $u = (z-x)/h$. The last expression shows that the expected value is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a second order Taylor expansion of $f(x+hu)$ in the argument hu about $hu = 0$, which is valid as $h \to 0$. Thus

$$f(x+hu) \simeq f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$$

and therefore

$$\begin{aligned}
\mathbb{E}K_h(X-x) &\simeq \int_{-\infty}^{\infty} K(u)\left(f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2\right)du \\
&= f(x)\int_{-\infty}^{\infty} K(u)\,du + f'(x)h\int_{-\infty}^{\infty} K(u)\,udu + \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} K(u)\,u^2du \\
&= f(x) + \frac{1}{2}f''(x)h^2\sigma_K^2.
\end{aligned}$$

The bias of $\hat{f}(x)$ is then

$$Bias(x) = \mathbb{E}\hat{f}(x) - f(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}K_h(X_i-x) - f(x) = \frac{1}{2}f''(x)h^2\sigma_K^2.$$

We see that the bias of $\hat{f}(x)$ at x depends on the second derivative $f''(x)$. The sharper the derivative, the greater the bias. Intuitively, the estimator $\hat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

We now examine the variance of $\hat{f}(x)$. Since it is an average of iid random variables, using first-order Taylor approximations and the fact that n^{-1} is of smaller order than $(nh)^{-1}$

$$\begin{aligned}
\operatorname{var}(x) &= \frac{1}{n}\operatorname{var}(K_h(X_i-x)) \\
&= \frac{1}{n}\mathbb{E}K_h(X_i-x)^2 - \frac{1}{n}(\mathbb{E}K_h(X_i-x))^2 \\
&\simeq \frac{1}{nh^2}\int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)^2 f(z)dz - \frac{1}{n}f(x)^2 \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K(u)^2 f(x+hu)\,du \\
&\simeq \frac{f(x)}{nh}\int_{-\infty}^{\infty} K(u)^2\,du \\
&= \frac{f(x)R(K)}{nh}.
\end{aligned}$$

where $R(K) = \int_{-\infty}^{\infty} K(u)^2\,du$ is called the **roughness** of K.

Together, the asymptotic mean-squared error (AMSE) for fixed x is the sum of the approximate squared bias and approximate variance

$$AMSE_h(x) = \frac{1}{4}f''(x)^2h^4\sigma_K^4 + \frac{f(x)R(K)}{nh}.$$

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$AMISE_h = \int AMSE_h(x)dx = \frac{h^4\sigma_K^4 R(f'')}{4} + \frac{R(K)}{nh}. \tag{16.1}$$

where $R(f'') = \int (f''(x))^2 dx$ is the roughness of f''. Notice that the first term (the squared bias) is increasing in h and the second term (the variance) is decreasing in nh. Thus for the AMISE to decline with n, we need $h \to 0$ but $nh \to \infty$. That is, h must tend to zero, but at a slower rate than n^{-1}.

Equation (16.1) is an asymptotic approximation to the MSE. We define the asymptotically optimal bandwidth h_0 as the value which minimizes this approximate MSE. That is,

$$h_0 = \underset{h}{\operatorname{argmin}}\, AMISE_h$$

It can be found by solving the first order condition

$$\frac{d}{dh} AMISE_h = h^3\sigma_K^4 R(f'') - \frac{R(K)}{nh^2} = 0$$

yielding

$$h_0 = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{1/5} n^{-1/2}. \tag{16.2}$$

This solution takes the form $h_0 = cn^{-1/5}$ where c is a function of K and f, but not of n. We thus say that the optimal bandwidth is of order $O(n^{-1/5})$. Note that this h declines to zero, but at a very slow rate.

In practice, how should the bandwidth be selected? This is a difficult problem, and there is a large and continuing literature on the subject. The asymptotically optimal choice given in (16.2) depends on $R(K)$, σ_K^2, and $R(f'')$. The first two are determined by the kernel function. Their values for the three functions introduced in the previous section are given here.

K	$\sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u)\, du$	$R(K) = \int_{-\infty}^{\infty} K(u)^2\, du$
Gaussian	1	$1/(2\sqrt{\pi})$
Epanechnikov	1/5	1/5
Biweight	1/7	5/7

An obvious difficulty is that $R(f'')$ is unknown. A classic simple solution proposed by Silverman (1986)has come to be known as the **reference bandwidth** or **Silverman's Rule-of-Thumb**. It uses formula (16.2) but replaces $R(f'')$ with $\hat{\sigma}^{-5}R(\phi'')$, where ϕ is the N(0, 1) distribution and $\hat{\sigma}^2$ is an estimate of $\sigma^2 = \text{var}(X)$. This choice for h gives an optimal rule when $f(x)$ is normal, and gives a nearly optimal rule when $f(x)$ is close to normal. The downside is that if the density is very far from normal, the rule-of-thumb h can be quite inefficient. We can calculate that $R(\phi'') = 3/(8\sqrt{\pi})$. Together with the above table, we find the reference rules for the three kernel functions introduced earlier.

Gaussian Kernel: $h_{rule} = 1.06\hat{\sigma}n^{-1/5}$

Epanechnikov Kernel: $h_{rule} = 2.34\hat{\sigma}n^{-1/5}$

Biweight (Quartic) Kernel: $h_{rule} = 2.78\hat{\sigma}n^{-1/5}$

Unless you delve more deeply into kernel estimation methods the rule-of-thumb bandwidth is a good practical bandwidth choice, perhaps adjusted by visual inspection of the resulting estimate $\hat{f}(x)$. There are other approaches, but implementation can be delicate. I now discuss some of these choices. The **plug-in** approach is to estimate $R(f'')$ in a first step, and then plug this estimate into the formula (16.2). This is more treacherous than may first appear, as the optimal h for estimation of the roughness $R(f'')$ is quite different than the optimal h for estimation of $f(x)$. However, there

are modern versions of this estimator work well, in particular the iterative method of Sheather and Jones (1991). Another popular choice for selection of h is **cross-validation**. This works by constructing an estimate of the MISE using leave-one-out estimators. There are some desirable properties of cross-validation bandwidths, but they are also known to converge very slowly to the optimal values. They are also quite ill-behaved when the data has some discretization (as is common in economics), in which case the cross-validation rule can sometimes select very small bandwidths leading to dramatically undersmoothed estimates. Fortunately there are remedies, which are known as **smoothed cross-validation** which is a close cousin of the **bootstrap**.

Appendix A

Matrix Algebra

A.1 Notation

A **scalar** a is a single number.

A **vector** $\boldsymbol{a}$ is a $k \times 1$ list of numbers, typically arranged in a column. We write this as

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector $\boldsymbol{a}$ is an element of Euclidean k space, written as $\boldsymbol{a} \in \mathbb{R}^k$. If $k = 1$ then $\boldsymbol{a}$ is a scalar.

A **matrix** $\boldsymbol{A}$ is a $k \times r$ rectangular array of numbers, written as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

By convention a_{ij} refers to the element in the $i'th$ row and $j'th$ column of $\boldsymbol{A}$. If $r = 1$ then $\boldsymbol{A}$ is a column vector. If $k = 1$ then $\boldsymbol{A}$ is a row vector. If $r = k = 1$, then $\boldsymbol{A}$ is a scalar.

A standard convention (which we will follow in this text whenever possible) is to denote scalars by lower-case italics (a), vectors by lower-case bold italics ($\boldsymbol{a}$), and matrices by upper-case bold italics ($\boldsymbol{A}$). Sometimes a matrix $\boldsymbol{A}$ is denoted by the symbol (a_{ij}).

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix}$$

where

$$\boldsymbol{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\boldsymbol{\alpha}_j = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jr} \end{bmatrix}$$

are row vectors.

The **transpose** of a matrix, denoted $\boldsymbol{A}'$, is obtained by flipping the matrix on its diagonal. Thus

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Alternatively, letting $\boldsymbol{B} = \boldsymbol{A}'$, then $b_{ij} = a_{ji}$. Note that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{A}'$ is $r \times k$. If $\boldsymbol{a}$ is a $k \times 1$ vector, then $\boldsymbol{a}'$ is a $1 \times k$ row vector. An alternative notation for the transpose of $\boldsymbol{A}$ is $\boldsymbol{A}^\top$.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $\boldsymbol{A} = \boldsymbol{A}'$, which requires $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the off-diagonal elements are all zero, so that $a_{ij} = 0$ if $i \neq j$. A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. The $k \times k$ identity matrix is denoted as

$$\boldsymbol{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A **partitioned matrix** takes the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kr} \end{bmatrix}$$

where the A_{ij} denote matrices, vectors and/or scalars.

A.2 Matrix Addition

If the matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are of the same order, we define the sum

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij} + b_{ij}).$$

Matrix addition follows the communtative and associative laws:

$$\begin{aligned} \boldsymbol{A} + \boldsymbol{B} &= \boldsymbol{B} + \boldsymbol{A} \\ \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) &= (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}. \end{aligned}$$

A.3 Matrix Multiplication

If $\boldsymbol{A}$ is $k \times r$ and c is real, we define their product as

$$\boldsymbol{A}c = c\boldsymbol{A} = (a_{ij}c).$$

If $\boldsymbol{a}$ and $\boldsymbol{b}$ are both $k \times 1$, then their inner product is

$$\boldsymbol{a}'\boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = \sum_{j=1}^{k} a_j b_j.$$

Note that $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a}$. We say that two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.

If $\boldsymbol{A}$ is $k \times r$ and $\boldsymbol{B}$ is $r \times s$, so that the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$, we say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are **conformable**. In this event the matrix product $\boldsymbol{AB}$ is defined. Writing $\boldsymbol{A}$ as a set of row vectors and $\boldsymbol{B}$ as a set of column vectors (each of length r), then the matrix product is defined as

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_k' \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_s \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_s \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_s \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_s \end{bmatrix}.\end{aligned}$$

Matrix multiplication is not communicative: in general $\boldsymbol{AB} \neq \boldsymbol{BA}$**.** However, it is associative and distributive:

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{BC}) &= (\boldsymbol{AB})\boldsymbol{C} \\ \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{AB}+\boldsymbol{AC}\end{aligned}$$

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11}+\boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12}+\boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11}+\boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12}+\boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix}.\end{aligned}$$

As another example,

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_r \end{bmatrix} \\ &= \boldsymbol{A}_1\boldsymbol{B}_1+\boldsymbol{A}_2\boldsymbol{B}_2+\cdots+\boldsymbol{A}_r\boldsymbol{B}_r \\ &= \sum_{j=1}^{r} \boldsymbol{A}_j\boldsymbol{B}_j\end{aligned}$$

An important property of the identity matrix is that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{AI}_r = \boldsymbol{A}$ and $\boldsymbol{I}_k\boldsymbol{A} = \boldsymbol{A}$. The $k \times r$ matrix $\boldsymbol{A}$, $r \leq k$, is called orthogonal if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_r$.

A.4 Trace

The **trace** of a $k \times k$ square matrix $\boldsymbol{A}$ is the sum of its diagonal elements

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} a_{ii}.$$

Some straightforward properties for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and real c are

$$\begin{aligned}\operatorname{tr}(c\boldsymbol{A}) &= c\operatorname{tr}(\boldsymbol{A}) \\ \operatorname{tr}\left(\boldsymbol{A}'\right) &= \operatorname{tr}(\boldsymbol{A}) \\ \operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) &= \operatorname{tr}(\boldsymbol{A})+\operatorname{tr}(\boldsymbol{B}) \\ \operatorname{tr}(\boldsymbol{I}_k) &= k.\end{aligned}$$

Also, for $k \times r$ $\boldsymbol{A}$ and $r \times k$ $\boldsymbol{B}$ we have

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}).$$

Indeed,

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{AB}) &= \operatorname{tr}\begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_k \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_k \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_k \end{bmatrix} \\
&= \sum_{i=1}^{k} \boldsymbol{a}_i'\boldsymbol{b}_i \\
&= \sum_{i=1}^{k} \boldsymbol{b}_i'\boldsymbol{a}_i \\
&= \operatorname{tr}(\boldsymbol{BA}).
\end{aligned}$$

A.5 Rank and Inverse

The rank of the $k \times r$ matrix ($r \leq k$)

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{bmatrix}$$

is the number of linearly independent columns $\boldsymbol{a}_j$, and is written as $\operatorname{rank}(\boldsymbol{A})$. We say that $\boldsymbol{A}$ has full rank if $\operatorname{rank}(\boldsymbol{A}) = r$.

A square $k \times k$ matrix $\boldsymbol{A}$ is said to be **nonsingular** if it is has full rank, e.g. $\operatorname{rank}(\boldsymbol{A}) = k$. This means that there is no $k \times 1$ $\boldsymbol{c} \neq \boldsymbol{0}$ such that $\boldsymbol{Ac} = \boldsymbol{0}$.

If a square $k \times k$ matrix $\boldsymbol{A}$ is nonsingular then there exists a unique matrix $k \times k$ matrix $\boldsymbol{A}^{-1}$ called the **inverse** of $\boldsymbol{A}$ which satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k.$$

For non-singular $\boldsymbol{A}$ and $\boldsymbol{C}$, some important properties include

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^{-1} &= \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k \\
\left(\boldsymbol{A}^{-1}\right)' &= \left(\boldsymbol{A}'\right)^{-1} \\
(\boldsymbol{AC})^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{A}^{-1} \\
(\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{C}^{-1} \\
\boldsymbol{A}^{-1} - (\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)\boldsymbol{A}^{-1}
\end{aligned}$$

Also, if $\boldsymbol{A}$ is an orthogonal matrix, then $\boldsymbol{A}^{-1} = \boldsymbol{A}$.

Another useful result for non-singular $\boldsymbol{A}$ is known as the **Woodbury matrix identity**

$$(\boldsymbol{A}+\boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{BC}\left(\boldsymbol{C}+\boldsymbol{CDA}^{-1}\boldsymbol{BC}\right)^{-1}\boldsymbol{CDA}^{-1}. \tag{A.1}$$

In particular, for $\boldsymbol{C} = -1$, $\boldsymbol{B} = \boldsymbol{b}$ and $\boldsymbol{D} = \boldsymbol{b}'$ for vector $\boldsymbol{b}$ we find what is known as the **Sherman–Morrison formula**

$$\left(\boldsymbol{A}-\boldsymbol{bb}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1-\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{bb}'\boldsymbol{A}^{-1}. \tag{A.2}$$

The following fact about inverting partitioned matrices is quite useful. If $\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ and $\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$ are non-singular, then

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \left(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -\left(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\left(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & \left(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{bmatrix}. \tag{A.3}$$

Even if a matrix $\boldsymbol{A}$ does not possess an inverse, we can still define the **Moore-Penrose generalized inverse** $\boldsymbol{A}^{-}$ as the matrix which satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}$$
$$\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{-} = \boldsymbol{A}^{-}$$
$$\boldsymbol{A}\boldsymbol{A}^{-} \text{ is symmetric}$$
$$\boldsymbol{A}^{-}\boldsymbol{A} \text{ is symmetric}$$

For any matrix $\boldsymbol{A}$, the Moore-Penrose generalized inverse $\boldsymbol{A}^{-}$ exists and is unique.

For example, if

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

then

$$\boldsymbol{A}^{-} = \begin{bmatrix} \boldsymbol{A}_{11}^{-} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

A.6 Determinant

The **determinant** is a measure of the volume of a square matrix.

While the determinant is widely used, its precise definition is rarely needed. However, we present the definition here for completeness. Let $\boldsymbol{A} = (a_{ij})$ be a general $k \times k$ matrix . Let $\pi = (j_1, ..., j_k)$ denote a permutation of $(1, ..., k)$. There are $k!$ such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order $(1, ..., k)$, and let $\varepsilon_\pi = +1$ if this count is even and $\varepsilon_\pi = -1$ if the count is odd. Then the determinant of $\boldsymbol{A}$ is defined as

$$\det \boldsymbol{A} = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}.$$

For example, if $\boldsymbol{A}$ is 2×2, then the two permutations of $(1, 2)$ are $(1, 2)$ and $(2, 1)$, for which $\varepsilon_{(1,2)} = 1$ and $\varepsilon_{(2,1)} = -1$. Thus

$$\begin{aligned} \det \boldsymbol{A} &= \varepsilon_{(1,2)} a_{11} a_{22} + \varepsilon_{(2,1)} a_{21} a_{12} \\ &= a_{11} a_{22} - a_{12} a_{21}. \end{aligned}$$

Some properties include

- $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
- $\det(c\boldsymbol{A}) = c^k \det \boldsymbol{A}$
- $\det(\boldsymbol{A}\boldsymbol{B}) = (\det \boldsymbol{A})(\det \boldsymbol{B})$
- $\det\left(\boldsymbol{A}^{-1}\right) = (\det \boldsymbol{A})^{-1}$
- $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = (\det \boldsymbol{D}) \det\left(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)$ if $\det \boldsymbol{D} \neq 0$
- $\det \boldsymbol{A} \neq 0$ if and only if $\boldsymbol{A}$ is nonsingular.
- If $\boldsymbol{A}$ is triangular (upper or lower), then $\det \boldsymbol{A} = \prod_{i=1}^{k} a_{ii}$
- If $\boldsymbol{A}$ is orthogonal, then $\det \boldsymbol{A} = \pm 1$

A.7 Eigenvalues

The characteristic equation of a square matrix $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}_k) = 0.$$

The left side is a polynomial of degree k in λ so it has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots** or **characteristic roots** or **eigenvalues** of $\boldsymbol{A}$. If λ_i is an eigenvalue of $\boldsymbol{A}$, then $\boldsymbol{A} - \lambda_i \boldsymbol{I}_k$ is singular so there exists a non-zero vector $\boldsymbol{h}_i$ such that

$$(\boldsymbol{A} - \lambda_i \boldsymbol{I}_k)\, \boldsymbol{h}_i = \mathbf{0}.$$

The vector $\boldsymbol{h}_i$ is called a **latent vector** or **characteristic vector** or **eigenvector** of $\boldsymbol{A}$ corresponding to λ_i.

We now state some useful properties. Let λ_i and $\boldsymbol{h}_i$, $i = 1, ..., k$ denote the k eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$. Let $\boldsymbol{\Lambda}$ be a diagonal matrix with the characteristic roots in the diagonal, and let $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_k]$.

- $\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i$
- $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i$
- $\boldsymbol{A}$ is non-singular if and only if all its characteristic roots are non-zero.
- If $\boldsymbol{A}$ has distinct characteristic roots, there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1} = \boldsymbol{\Lambda}$.
- If $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ and $\boldsymbol{H}'\boldsymbol{A}\boldsymbol{H} = \boldsymbol{\Lambda}$, and the characteristic roots are all real. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ is called the **spectral decomposition** of a matrix.
- The characteristic roots of $\boldsymbol{A}^{-1}$ are λ_1^{-1}, λ_2^{-1}, ..., λ_k^{-1}.
- The matrix $\boldsymbol{H}$ has the **orthonormal** properties $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}$.
- $\boldsymbol{H}^{-1} = \boldsymbol{H}'$ and $(\boldsymbol{H}')^{-1} = \boldsymbol{H}$

A.8 Positive Definiteness

We say that a $k \times k$ symmetric square matrix $\boldsymbol{A}$ is **positive semi-definite** if for all $\boldsymbol{c} \neq \mathbf{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$. This is written as $\boldsymbol{A} \geq 0$. We say that $\boldsymbol{A}$ is **positive definite** if for all $\boldsymbol{c} \neq \mathbf{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$. This is written as $\boldsymbol{A} > 0$.

Some properties include:

- If $\boldsymbol{A} = \boldsymbol{G}'\boldsymbol{G}$ for some matrix $\boldsymbol{G}$, then $\boldsymbol{A}$ is positive semi-definite. (For any $\boldsymbol{c} \neq \mathbf{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \boldsymbol{\alpha}'\boldsymbol{\alpha} \geq 0$ where $\boldsymbol{\alpha} = \boldsymbol{G}\boldsymbol{c}$.) If $\boldsymbol{G}$ has full rank, then $\boldsymbol{A}$ is positive definite.
- If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}$ is non-singular and $\boldsymbol{A}^{-1}$ exists. Furthermore, $\boldsymbol{A}^{-1} > 0$.
- $\boldsymbol{A} > 0$ if and only if it is symmetric and all its characteristic roots are positive.
- By the spectral decomposition, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is diagonal with non-negative diagonal elements. All diagonal elements of $\boldsymbol{\Lambda}$ are strictly positive if (and only if) $\boldsymbol{A} > 0$.
- If $A > 0$ then $A^{-1} = \boldsymbol{H}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}'$.

- If $A \geq 0$ and $\text{rank}(\boldsymbol{A}) = r < k$ then $A^{-} = \boldsymbol{H}\boldsymbol{\Lambda}^{-}\boldsymbol{H}'$ where A^{-} is the Moore-Penrose generalized inverse, and $\boldsymbol{\Lambda}^{-} = \text{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_k^{-1}, 0, ..., 0\right)$
- If $\boldsymbol{A} > 0$ we can find a matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}'$. We call $\boldsymbol{B}$ a **matrix square root** of $\boldsymbol{A}$. The matrix $\boldsymbol{B}$ need not be unique. One way to construct $\boldsymbol{B}$ is to use the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}$ is diagonal, and then set $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$.

A square matrix $\boldsymbol{A}$ is **idempotent** if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. If $\boldsymbol{A}$ is idempotent and symmetric then all its characteristic roots equal either zero or one and is thus positive semi-definite. To see this, note that we can write $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}$ is orthogonal and $\boldsymbol{\Lambda}$ contains the r (real) characteristic roots. Then

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}' = \boldsymbol{H}\boldsymbol{\Lambda}^2\boldsymbol{H}'.$$

By the uniqueness of the characteristic roots, we deduce that $\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}$ and $\lambda_i^2 = \lambda_i$ for $i = 1, ..., r$. Hence they must equal either 0 or 1. It follows that the spectral decomposition of idempotent $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{k-r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}' \tag{A.4}$$

with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$. Additionally, $\text{tr}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.

A.9 Matrix Calculus

Let $\boldsymbol{x} = (x_1, ..., x_k)$ be $k \times 1$ and $g(\boldsymbol{x}) = g(x_1, ..., x_k) : \mathbb{R}^k \to \mathbb{R}$. The vector derivative is

$$\frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}$$

and

$$\frac{\partial}{\partial \boldsymbol{x}'} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}.$$

Some properties are now summarized.

- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}'\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{a}) = \boldsymbol{a}$
- $\frac{\partial}{\partial \boldsymbol{x}'}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$
- $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A} + \boldsymbol{A}'$

A.10 Kronecker Products and the Vec Operator

Let $\boldsymbol{A} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_n]$ be $m \times n$. The **vec** of $\boldsymbol{A}$, denoted by $\text{vec}(\boldsymbol{A})$, is the $mn \times 1$ vector

$$\text{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and let $\boldsymbol{B}$ be any matrix. The **Kronecker product** of $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted $\boldsymbol{A} \otimes \boldsymbol{B}$, is the matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

- $(\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C}$
- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
- $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes C$
- $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$
- $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$
- If $\boldsymbol{A}$ is $m \times m$ and $\boldsymbol{B}$ is $n \times n$, $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = (\det(\boldsymbol{A}))^n (\det(\boldsymbol{B}))^m$
- $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$
- If $\boldsymbol{A} > 0$ and $\boldsymbol{B} > 0$ then $\boldsymbol{A} \otimes \boldsymbol{B} > 0$
- $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$
- $\operatorname{tr}(\boldsymbol{ABCD}) = \operatorname{vec}(\boldsymbol{D}')' (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$

A.11 Vector and Matrix Norms

The **Euclidean norm** of an $m \times 1$ vector $\boldsymbol{a}$ is

$$\begin{aligned} \|\boldsymbol{a}\| &= \left(\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} \\ &= \left(\sum_{i=1}^{m} a_i^2\right)^{1/2}. \end{aligned}$$

The **Euclidean norm** of an $m \times n$ matrix $\boldsymbol{A}$ is

$$\begin{aligned} \|\boldsymbol{A}\| &= \|\operatorname{vec}(\boldsymbol{A})\| \\ &= \operatorname{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{1/2} \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}. \end{aligned}$$

A useful calculation is for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\left\|\boldsymbol{a}\boldsymbol{b}'\right\| = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\|$$

and in particular

$$\left\|\boldsymbol{a}\boldsymbol{a}'\right\| = \|\boldsymbol{a}\|^2 \tag{A.5}$$

Some useful inequalities are now given:

Schwarz Inequality: For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\left|\boldsymbol{a}'\boldsymbol{b}\right| \leq \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| . \tag{A.6}$$

Schwarz Matrix Inequality: For any $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\left\|\boldsymbol{A}'\boldsymbol{B}\right\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| . \tag{A.7}$$

Triangle Inequality: For any $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\| . \tag{A.8}$$

Proof of Schwarz Inequality: First, suppose that $\|\boldsymbol{b}\| = 0$. Then $\boldsymbol{b} = \boldsymbol{0}$ and both $|\boldsymbol{a}'\boldsymbol{b}| = 0$ and $\|\boldsymbol{a}\| \, \|\boldsymbol{b}\| = 0$ so the inequality is true. Second, suppose that $\|\boldsymbol{b}\| > 0$ and define $\boldsymbol{c} = \boldsymbol{a} - \boldsymbol{b}\left(\boldsymbol{b}'\boldsymbol{b}\right)^{-1} \boldsymbol{b}'\boldsymbol{a}$. Since $\boldsymbol{c}$ is a vector, $\boldsymbol{c}'\boldsymbol{c} \geq 0$. Thus

$$0 \leq \boldsymbol{c}'\boldsymbol{c} = \boldsymbol{a}'\boldsymbol{a} - \left(\boldsymbol{a}'\boldsymbol{b}\right)^2 / \left(\boldsymbol{b}'\boldsymbol{b}\right) .$$

Rearranging, this implies that

$$\left(\boldsymbol{a}'\boldsymbol{b}\right)^2 \leq \left(\boldsymbol{a}'\boldsymbol{a}\right)\left(\boldsymbol{b}'\boldsymbol{b}\right) .$$

Taking the square root of each side yields the result.

Proof of Schwarz Matrix Inequality: Partition $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_n]$. Then by partitioned matrix multiplication, the definition of the matrix Euclidean norm and the Schwarz inequality

$$\begin{aligned}
\left\|\boldsymbol{A}'\boldsymbol{B}\right\| &= \left\| \begin{array}{ccc} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\| \\
&\leq \left\| \begin{array}{ccc} \|\boldsymbol{a}_1\| \, \|\boldsymbol{b}_1\| & \|\boldsymbol{a}_1\| \, \|\boldsymbol{b}_2\| & \cdots \\ \|\boldsymbol{a}_2\| \, \|\boldsymbol{b}_1\| & \|\boldsymbol{a}_2\| \, \|\boldsymbol{b}_2\| & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\| \\
&= \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \|\boldsymbol{a}_i\|^2 \, \|\boldsymbol{b}_j\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^{n} \|\boldsymbol{a}_i\|^2 \right)^{1/2} \left(\sum_{i=1}^{n} \|\boldsymbol{b}_i\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{a}_{ji}^2 \right)^{1/2} \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \|\boldsymbol{b}_{ji}\|^2 \right)^{1/2} \\
&= \|\boldsymbol{A}\| \, \|\boldsymbol{B}\|
\end{aligned}$$

Proof of Triangle Inequality: Let $\boldsymbol{a} = \mathrm{vec}\,(\boldsymbol{A})$ and $\boldsymbol{b} = \mathrm{vec}\,(\boldsymbol{B})$. Then by the definition of the matrix norm and the Schwarz Inequality

$$\begin{aligned}
\|\boldsymbol{A} + \boldsymbol{B}\|^2 &= \|\boldsymbol{a} + \boldsymbol{b}\|^2 \\
&= \boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{a}'\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{b} \\
&\leq \boldsymbol{a}'\boldsymbol{a} + 2\left|\boldsymbol{a}'\boldsymbol{b}\right| + \boldsymbol{b}'\boldsymbol{b} \\
&\leq \|\boldsymbol{a}\|^2 + 2\|\boldsymbol{a}\| \, \|\boldsymbol{b}\| + \|\boldsymbol{b}\|^2 \\
&= \left(\|\boldsymbol{a}\| + \|\boldsymbol{b}\|\right)^2 \\
&= \left(\|\boldsymbol{A}\| + \|\boldsymbol{B}\|\right)^2
\end{aligned}$$

Appendix B

Probability

B.1 Foundations

The set S of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $S = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $S = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of S, including S itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint. Furthermore, if the sets $A_1, A_2, ...$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ is called a **partition** of S.

The following are elementary set operations:

Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$.

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Complement: $A^c = \{x : x \notin A\}$.

The following are useful properties of set operations.

Communtatitivity: $A \cup B = B \cup A$; $\quad A \cap B = B \cap A$.

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$; $\quad A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $\quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $\quad (A \cap B)^c = A^c \cup B^c$.

A **probability function** assigns probabilities (numbers between 0 and 1) to events A in S. This is straightforward when S is countable; when S is uncountable we must be somewhat more careful. A set $\mathcal{B}$ is called a **sigma algebra** (or Borel field) if $\emptyset \in \mathcal{B}$, $A \in \mathcal{B}$ implies $A^c \in \mathcal{B}$, and $A_1, A_2, ... \in \mathcal{B}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. A simple example is $\{\emptyset, S\}$ which is known as the trivial sigma algebra. For any sample space S, let $\mathcal{B}$ be the smallest sigma algebra which contains all of the open sets in S. When S is countable, $\mathcal{B}$ is simply the collection of all subsets of S, including $\emptyset$ and S. When S is the real line, then $\mathcal{B}$ is the collection of all open and closed intervals. We call $\mathcal{B}$ the sigma algebra associated with S. We only define probabilities for events contained in $\mathcal{B}$.

We now can give the axiomatic definition of probability. Given S and $\mathcal{B}$, a probability function $\mathbb{P}$ satisfies $\mathbb{P}(S) = 1$, $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$, and if $A_1, A_2, ... \in \mathcal{B}$ are pairwise disjoint, then $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Some important properties of the probability function include the following

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

- $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- Bonferroni's Inequality: $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$
- Boole's Inequality: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

For some elementary probability models, it is useful to have simple rules to count the number of objects in a set. These counting rules are facilitated by using the binomial coefficients which are defined for nonnegative integers n and r, $n \geq r$, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

When counting the number of objects in a set, there are two important distinctions. Counting may be **with replacement** or **without replacement**. Counting may be **ordered** or **unordered**. For example, consider a lottery where you pick six numbers from the set 1, 2, ..., 49. This selection is without replacement if you are not allowed to select the same number twice, and is with replacement if this is allowed. Counting is ordered or not depending on whether the sequential order of the numbers is relevant to winning the lottery. Depending on these two distinctions, we have four expressions for the number of objects (possible arrangements) of size r from n objects.

	Without Replacement	With Replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the lottery example, if counting is unordered and without replacement, the number of potential combinations is $\binom{49}{6} = 13,983,816$.

If $\mathbb{P}(B) > 0$ the **conditional probability** of the event A given the event B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

For any B, the conditional probability function is a valid probability function where S has been replaced by B. Rearranging the definition, we can write

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$$

which is often quite useful. We can say that the occurrence of B has no information about the likelihood of event A when $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, in which case we find

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \tag{B.1}$$

We say that the events A and B are **statistically independent** when (B.1) holds. Furthermore, we say that the collection of events $A_1, ..., A_k$ are **mutually independent** when for any subset $\{A_i : i \in I\}$,

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

Theorem 1 *(Bayes' Rule). For any set B and any partition $A_1, A_2, ...$ of the sample space, then for each $i = 1, 2, ...$*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\,\mathbb{P}(A_j)}$$

B.2 Random Variables

A **random variable** X is a function from a sample space S into the real line. This induces a new sample space – the real line – and a new probability function on the real line. Typically, we denote random variables by uppercase letters such as X, and use lower case letters such as x for potential values and realized values. (This is in contrast to the notation adopted for most of the textbook.) For a random variable X we define its **cumulative distribution function** (CDF) as

$$F(x) = \mathbb{P}(X \leq x). \tag{B.2}$$

Sometimes we write this as $F_X(x)$ to denote that it is the CDF of X. A function $F(x)$ is a CDF if and only if the following three properties hold:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
2. $F(x)$ is nondecreasing in x
3. $F(x)$ is right-continuous

We say that the random variable X is **discrete** if $F(x)$ is a step function. In the latter case, the range of X consists of a countable set of real numbers $\tau_1, ..., \tau_r$. The probability function for X takes the form

$$\mathbb{P}(X = \tau_j) = \pi_j, \qquad j = 1, ..., r \tag{B.3}$$

where $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{r} \pi_j = 1$.

We say that the random variable X is **continuous** if $F(x)$ is continuous in x. In this case $\mathbb{P}(X = \tau) = 0$ for all $\tau \in R$ so the representation (B.3) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (PDF)

$$f(x) = \frac{d}{dx}F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u)du$$

and

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f(u)du.$$

These expressions only make sense if $F(x)$ is differentiable. While there are examples of continuous random variables which do not possess a PDF, these cases are unusual and are typically ignored.

A function $f(x)$ is a PDF if and only if $f(x) \geq 0$ for all $x \in R$ and $\int_{-\infty}^{\infty} f(x)dx$.

B.3 Expectation

For any measurable real function g, we define the **mean** or **expectation** $\mathbb{E}g(X)$ as follows. If X is discrete,

$$\mathbb{E}g(X) = \sum_{j=1}^{r} g(\tau_j)\pi_j,$$

and if X is continuous

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The latter is well defined and finite if

$$\int_{-\infty}^{\infty} |g(x)|\, f(x)dx < \infty. \tag{B.4}$$

If (B.4) does not hold, evaluate

$$\begin{aligned} I_1 &= \int_{g(x)>0} g(x)f(x)dx \\ I_2 &= -\int_{g(x)<0} g(x)f(x)dx \end{aligned}$$

If $I_1 = \infty$ and $I_2 < \infty$ then we define $\mathbb{E}g(X) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}g(X) = -\infty$. If both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}g(X)$ is undefined.

Since $\mathbb{E}(a + bX) = a + b\mathbb{E}X$, we say that expectation is a linear operator.

For $m > 0$, we define the $m'th$ **moment** of X as $\mathbb{E}X^m$ and the $m'th$ **central moment** as $\mathbb{E}(X - \mathbb{E}X)^m$.

Two special moments are the **mean** $\mu = \mathbb{E}X$ and **variance** $\sigma^2 = \mathbb{E}(X - \mu)^2 = \mathbb{E}X^2 - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = \text{var}(X)$. For example, this allows the convenient expression $\text{var}(a + bX) = b^2 \text{var}(X)$.

The **moment generating function** (MGF) of X is

$$M(\lambda) = \mathbb{E}\exp(\lambda X).$$

The MGF does not necessarily exist. However, when it does and $\mathbb{E}|X|^m < \infty$ then

$$\frac{d^m}{d\lambda^m} M(\lambda)\bigg|_{\lambda=0} = \mathbb{E}(X^m)$$

which is why it is called the moment generating function.

More generally, the **characteristic function** (CF) of X is

$$C(\lambda) = \mathbb{E}\exp(\mathrm{i}\lambda X)$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. The CF always exists, and when $\mathbb{E}|X|^m < \infty$

$$\frac{d^m}{d\lambda^m} C(\lambda)\bigg|_{\lambda=0} = \mathrm{i}^m \mathbb{E}(X^m).$$

The L^p **norm**, $p \geq 1$, of the random variable X is

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}.$$

B.4 Gamma Function

The gamma function is defined for $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x).$$

It satisfies the property

$$\Gamma(1 + \alpha) = \Gamma(\alpha)\alpha$$

so for positive integers n,

$$\Gamma(n) = (n - 1)!$$

Special values include

$$\Gamma(1) = 1$$

and

$$\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}.$$

Sterling's formula is an expansion for the its logarithm

$$\log\Gamma(\alpha) = \frac{1}{2}\log(2\pi) + \left(\alpha - \frac{1}{2}\right)\log\alpha - z + \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1260\alpha^5} + \cdots$$

B.5 Common Distributions

For reference, we now list some important discrete distribution function.

Bernoulli

$$\begin{aligned} \mathbb{P}(X=x) &= p^x(1-p)^{1-x}, \qquad x=0,1; \qquad 0\le p\le 1 \\ \mathbb{E}X &= p \\ \operatorname{var}(X) &= p(1-p) \end{aligned}$$

Binomial

$$\begin{aligned} \mathbb{P}(X=x) &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x=0,1,...,n; \qquad 0\le p\le 1 \\ \mathbb{E}X &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Geometric

$$\begin{aligned} \mathbb{P}(X=x) &= p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1 \\ \mathbb{E}X &= \frac{1}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

Multinomial

$$\begin{aligned} \mathbb{P}(X_1=x_1, X_2=x_2, ..., X_m=x_m) &= \frac{n!}{x_1!x_2!\cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}, \\ x_1+\cdots+x_m &= n; \\ p_1+\cdots+p_m &= 1 \\ \mathbb{E}X_i &= p_i \\ \operatorname{var}(X_i) &= np_i(1-p_i) \\ \operatorname{cov}(X_i, X_j) &= -np_ip_j \end{aligned}$$

Negative Binomial

$$\begin{aligned} \mathbb{P}(X=x) &= \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r (1-p)^{x-1}, \qquad x=0,1,2,...; \qquad 0\le p\le 1 \\ \mathbb{E}X &= \frac{r(1-p)}{p} \\ \operatorname{var}(X) &= \frac{r(1-p)}{p^2} \end{aligned}$$

Poisson

$$\begin{aligned} \mathbb{P}(X=x) &= \frac{\exp(-\lambda)\lambda^x}{x!}, \qquad x=0,1,2,..., \qquad \lambda>0 \\ \mathbb{E}X &= \lambda \\ \operatorname{var}(X) &= \lambda \end{aligned}$$

We now list some important continuous distributions.

Beta

$$\begin{aligned} f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \le x \le 1; \qquad \alpha > 0,\ \beta > 0 \\ \mu &= \frac{\alpha}{\alpha+\beta} \\ \operatorname{var}(X) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \end{aligned}$$

Cauchy

$$\begin{aligned} f(x) &= \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty \\ \mathbb{E}X &= \infty \\ \operatorname{var}(X) &= \infty \end{aligned}$$

Exponential

$$\begin{aligned} f(x) &= \frac{1}{\theta}\exp\left(\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \theta > 0 \\ \mathbb{E}X &= \theta \\ \operatorname{var}(X) &= \theta^2 \end{aligned}$$

Logistic

$$\begin{aligned} f(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2}, \qquad -\infty < x < \infty; \\ \mathbb{E}X &= 0 \\ \operatorname{var}(X) &= \frac{\pi^2}{3} \end{aligned}$$

Lognormal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right), \qquad 0 \le x < \infty; \qquad \sigma > 0 \\ \mathbb{E}X &= \exp\left(\mu+\sigma^2/2\right) \\ \operatorname{var}(X) &= \exp\left(2\mu+2\sigma^2\right) - \exp\left(2\mu+\sigma^2\right) \end{aligned}$$

Pareto

$$\begin{aligned} f(x) &= \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha \le x < \infty, \qquad \alpha > 0, \qquad \beta > 0 \\ \mathbb{E}X &= \frac{\beta\alpha}{\beta-1}, \qquad \beta > 1 \\ \operatorname{var}(X) &= \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \qquad \beta > 2 \end{aligned}$$

Uniform

$$\begin{aligned} f(x) &= \frac{1}{b-a}, \qquad a \le x \le b \\ \mathbb{E}X &= \frac{a+b}{2} \\ \operatorname{var}(X) &= \frac{(b-a)^2}{12} \end{aligned}$$

Weibull

$$\begin{aligned} f(x) &= \frac{\gamma}{\beta} x^{\gamma-1} \exp\left(-\frac{x^\gamma}{\beta}\right), \qquad 0 \le x < \infty; \qquad \gamma > 0,\ \beta > 0 \\ \mathbb{E}X &= \beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right) \\ \text{var}(X) &= \beta^{2/\gamma} \left(\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right) \end{aligned}$$

Gamma

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \alpha > 0,\ \theta > 0 \\ \mathbb{E}X &= \alpha\theta \\ \text{var}(X) &= \alpha\theta^2 \end{aligned}$$

Chi-Square

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} \exp\left(-\frac{x}{2}\right), \qquad 0 \le x < \infty; \qquad r > 0 \\ \mathbb{E}X &= r \\ \text{var}(X) &= 2r \end{aligned}$$

Normal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty; \qquad -\infty < \mu < \infty,\ \sigma^2 > 0 \\ \mathbb{E}X &= \mu \\ \text{var}(X) &= \sigma^2 \end{aligned}$$

Student t

$$\begin{aligned} f(x) &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty; \qquad r > 0 \\ \mathbb{E}X &= 0 \text{ if } r > 1 \\ \text{var}(X) &= \frac{r}{r-2} \text{ if } r > 2 \end{aligned}$$

B.6 Multivariate Random Variables

A pair of bivariate random variables (X, Y) is a function from the sample space into $\mathbb{R}^2$. The joint CDF of (X, Y) is

$$F(x, y) = \mathbb{P}(X \le x, Y \le y).$$

If F is continuous, the joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

For a Borel measurable set $A \in R^2$,

$$\mathbb{P}((X < Y) \in A) = \int \int_A f(x, y) dx dy$$

For any measurable function $g(x, y)$,

$$\mathbb{E}g(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \lim_{y \to \infty} F(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dy dx \end{aligned}$$

so the **marginal density** of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

The random variables X and Y are defined to be **independent** if $f(x, y) = f_X(x) f_Y(y)$. Furthermore, X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x, y) = g(x)h(y)$.

If X and Y are independent, then

$$\begin{aligned} \mathbb{E}\left(g(X)h(Y)\right) &= \int \int g(x)h(y)f(y, x) dy dx \\ &= \int \int g(x)h(y)f_Y(y)f_X(x) dy dx \\ &= \int g(x)f_X(x) dx \int h(y)f_Y(y) dy \\ &= \mathbb{E}g\left(X\right) \mathbb{E}h\left(Y\right). \end{aligned} \tag{B.5}$$

if the expectations exist. For example, if X and Y are independent then

$$\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y.$$

Another implication of (B.5) is that if X and Y are independent and $Z = X + Y$, then

$$\begin{aligned} M_Z(\lambda) &= \mathbb{E} \exp\left(\lambda\left(X + Y\right)\right) \\ &= \mathbb{E}\left(\exp\left(\lambda X\right) \exp\left(\lambda Y\right)\right) \\ &= \mathbb{E} \exp\left(\lambda' X\right) \mathbb{E} \exp\left(\lambda' Y\right) \\ &= M_X(\lambda) M_Y(\lambda). \end{aligned} \tag{B.6}$$

The covariance between X and Y is

$$\operatorname{cov}(X, Y) = \sigma_{XY} = \mathbb{E}\left(\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

The correlation between X and Y is

$$\operatorname{corr}\left(X, Y\right) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_x \sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that

$$|\rho_{XY}| \leq 1. \tag{B.7}$$

The correlation is a measure of linear dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $\mathbb{E}X = 0$ and $\mathbb{E}X^3 = 0$, then $\operatorname{cov}(X, X^2) = 0$.

A useful fact is that

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y).$$

An implication is that if X and Y are independent, then

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $\boldsymbol{X} = (X_1, ..., X_k)'$ is a function from S to $\mathbb{R}^k$. Let $\boldsymbol{x} = (x_1, ..., x_k)'$ denote a vector in $\mathbb{R}^k$. (In this Appendix, we use bold to denote vectors. Bold capitals $\boldsymbol{X}$ are random vectors and bold lower case $\boldsymbol{x}$ are nonrandom vectors. Again, this is in distinction to the notation used in the bulk of the text) The vector $\boldsymbol{X}$ has the distribution and density functions

$$\begin{aligned} F(\boldsymbol{x}) &= \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) \\ f(\boldsymbol{x}) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(\boldsymbol{x}). \end{aligned}$$

For a measurable function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^s$, we define the expectation

$$\mathbb{E}\boldsymbol{g}(\boldsymbol{X}) = \int_{\mathbb{R}^k} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

where the symbol $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}$$

and $k \times k$ covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu}) (\boldsymbol{X} - \boldsymbol{\mu})' \right) \\ &= \mathbb{E}\boldsymbol{X}\boldsymbol{X}' - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

If the elements of $\boldsymbol{X}$ are mutually independent, then $\boldsymbol{\Sigma}$ is a diagonal matrix and

$$\operatorname{var}\left(\sum_{i=1}^{k} \boldsymbol{X}_i \right) = \sum_{i=1}^{k} \operatorname{var}(\boldsymbol{X}_i)$$

B.7 Conditional Distributions and Expectation

The **conditional density** of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

if $f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$. One way to derive this expression from the definition of conditional probability is

$$\begin{aligned}
f_{Y|X}\left(y \mid \boldsymbol{x}\right) &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \mathbb{P}\left(Y \leq y \mid \boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon\right) \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\mathbb{P}\left(\{Y \leq y\} \cap \{\boldsymbol{x} \leq \boldsymbol{X} \leq \boldsymbol{x} + \varepsilon\}\right)}{\mathbb{P}(\boldsymbol{x} \leq X \leq \boldsymbol{x} + \varepsilon)} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{F(\boldsymbol{x} + \varepsilon, y) - F(\boldsymbol{x}, y)}{F_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon) - F_X(\boldsymbol{x})} \\
&= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\frac{\partial}{\partial x} F(\boldsymbol{x} + \varepsilon, y)}{f_{\boldsymbol{X}}(\boldsymbol{x} + \varepsilon)} \\
&= \frac{\frac{\partial^2}{\partial x \partial y} F(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}.
\end{aligned}$$

The **conditional mean** or **conditional expectation** is the function

$$m(\boldsymbol{x}) = \mathbb{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) dy.$$

The conditional mean $m(\boldsymbol{x})$ is a function, meaning that when $\boldsymbol{X}$ equals $\boldsymbol{x}$, then the expected value of Y is $m(\boldsymbol{x})$.

Similarly, we define the conditional variance of Y given $\boldsymbol{X} = \boldsymbol{x}$ as

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= \operatorname{var}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
&= \mathbb{E}\left(\left(Y - m(\boldsymbol{x})\right)^2 \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
&= \mathbb{E}\left(Y^2 \mid X = \boldsymbol{x}\right) - m(\boldsymbol{x})^2.
\end{aligned}$$

Evaluated at $\boldsymbol{x} = \boldsymbol{X}$, the conditional mean $m(\boldsymbol{X})$ and conditional variance $\sigma^2(\boldsymbol{X})$ are random variables, functions of $\boldsymbol{X}$. We write this as $\mathbb{E}(Y \mid \boldsymbol{X}) = m(\boldsymbol{X})$ and $\operatorname{var}\left(Y \mid \boldsymbol{X}\right) = \sigma^2(\boldsymbol{X})$. For example, if $\mathbb{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \alpha + \beta' \boldsymbol{x}$, then $\mathbb{E}\left(Y \mid \boldsymbol{X}\right) = \alpha + \beta' \boldsymbol{X}$, a transformation of $\boldsymbol{X}$.

The following are important facts about conditional expectations.

Simple Law of Iterated Expectations:

$$\mathbb{E}\left(\mathbb{E}\left(Y \mid \boldsymbol{X}\right)\right) = \mathbb{E}\left(Y\right) \tag{B.8}$$

Proof:

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left(Y \mid \boldsymbol{X}\right)\right) &= \mathbb{E}\left(m(\boldsymbol{X})\right) \\
&= \int_{-\infty}^{\infty} m(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) f_{\boldsymbol{X}}(\boldsymbol{x}) dy d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f\left(y, \boldsymbol{x}\right) dy d\boldsymbol{x} \\
&= \mathbb{E}(Y).
\end{aligned}$$

Law of Iterated Expectations:

$$\mathbb{E}\left(\mathbb{E}\left(Y \mid \boldsymbol{X}, \boldsymbol{Z}\right) \mid \boldsymbol{X}\right) = \mathbb{E}\left(Y \mid \boldsymbol{X}\right) \tag{B.9}$$

Conditioning Theorem. For any function $g(\boldsymbol{x})$,

$$\mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X}\right) = g\left(\boldsymbol{X}\right)\mathbb{E}\left(Y \mid \boldsymbol{X}\right) \tag{B.10}$$

Proof: Let

$$\begin{aligned} h(\boldsymbol{x}) &= \mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= \int_{-\infty}^{\infty} g(\boldsymbol{x}) y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) m(\boldsymbol{x}) \end{aligned}$$

where $m(\boldsymbol{x}) = \mathbb{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right)$. Thus $h(\boldsymbol{X}) = g(\boldsymbol{X})m(\boldsymbol{X})$, which is the same as $\mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X}\right) = g\left(\boldsymbol{X}\right)\mathbb{E}\left(Y \mid \boldsymbol{X}\right)$.

B.8 Transformations

Suppose that $\boldsymbol{X} \in \mathbb{R}^k$ with continuous distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ and density $f_{\boldsymbol{X}}(\boldsymbol{x})$. Let $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ where $\boldsymbol{g}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one, differentiable, and invertible. Let $\boldsymbol{h}(\boldsymbol{y})$ denote the inverse of $\boldsymbol{g}(\boldsymbol{x})$. The **Jacobian** is

$$J(\boldsymbol{y}) = \det\left(\frac{\partial}{\partial \boldsymbol{y}'}\boldsymbol{h}(\boldsymbol{y})\right).$$

Consider the univariate case $k = 1$. If $g(x)$ is an increasing function, then $g(X) \le Y$ if and only if $X \le h(Y)$, so the distribution function of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(g(X) \le y\right) \\ &= \mathbb{P}\left(X \le h(Y)\right) \\ &= F_X\left(h(Y)\right). \end{aligned}$$

Taking the derivative, the density of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

If $g(x)$ is a decreasing function, then $g(X) \le Y$ if and only if $X \ge h(Y)$, so

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(g(X) \le y\right) \\ &= 1 - \mathbb{P}\left(X \ge h(Y)\right) \\ &= 1 - F_X\left(h(Y)\right) \end{aligned}$$

and the density of Y is

$$f_Y(y) = -f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

We can write these two cases jointly as

$$f_Y(y) = f_X\left(h(Y)\right)\left|J(y)\right|. \tag{B.11}$$

This is known as the **change-of-variables** formula. This same formula (B.11) holds for $k > 1$, but its justification requires deeper results from analysis.

As one example, take the case $X \sim U[0,1]$ and $Y = -\log(X)$. Here, $g(x) = -\log(x)$ and $h(y) = \exp(-y)$ so the Jacobian is $J(y) = -\exp(y)$. As the range of X is $[0,1]$, that for Y is $[0,\infty)$. Since $f_X\left(x\right) = 1$ for $0 \le x \le 1$ (B.11) shows that

$$f_Y(y) = \exp(-y), \qquad 0 \le y \le \infty,$$

an exponential density.

B.9 Normal and Related Distributions

The **standard normal** density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

It is conventional to write $X \sim \mathrm{N}(0,1)$, and to denote the standard normal density function by $\phi(x)$ and its distribution function by $\Phi(x)$. The latter has no closed-form solution. The normal density has all moments finite. Since it is symmetric about zero all odd moments are zero. By iterated integration by parts, we can also show that $\mathbb{E}X^2 = 1$ and $\mathbb{E}X^4 = 3$. In fact, for any positive integer m, $\mathbb{E}X^{2m} = (2m-1)!! = (2m-1)\cdot(2m-3)\cdots 1$. Thus $\mathbb{E}X^4 = 3$, $\mathbb{E}X^6 = 15$, $\mathbb{E}X^8 = 105$, and $\mathbb{E}X^{10} = 945$.

If Z is standard normal and $X = \mu + \sigma Z$, then using the change-of-variables formula, X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

which is the **univariate normal density**. The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$.

For $\boldsymbol{x} \in \mathbb{R}^k$, the **multivariate normal density** is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \qquad \boldsymbol{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of the distribution are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and it is conventional to write $X \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The MGF and CF of the multivariate normal are $\exp\left(\boldsymbol{\lambda}'\boldsymbol{\mu} + \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2\right)$ and $\exp\left(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{\mu} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2\right)$, respectively.

If $\boldsymbol{X} \in \mathbb{R}^k$ is multivariate normal and the elements of $\boldsymbol{X}$ are mutually uncorrelated, then $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_j^2\}$ is a diagonal matrix. In this case the density function can be written as

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi)^{k/2} \sigma_1 \cdots \sigma_k} \exp\left(-\left(\frac{(x_1-\mu_1)^2/\sigma_1^2 + \cdots + (x_k-\mu_k)^2/\sigma_k^2}{2}\right)\right) \\ &= \prod_{j=1}^{k} \frac{1}{(2\pi)^{1/2}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \end{aligned}$$

which is the product of marginal univariate normal densities. This shows that if $\boldsymbol{X}$ is multivariate normal with uncorrelated elements, then they are mutually independent.

Theorem B.9.1 *If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ with $\boldsymbol{B}$ an invertible matrix, then $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.*

Theorem B.9.2 *Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. Then $Q = \boldsymbol{X}'\boldsymbol{X}$ is distributed chi-square with r degrees of freedom, written χ_r^2.*

Theorem B.9.3 *If $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{A})$ with $\boldsymbol{A} > 0$, $q \times q$, then $\boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z} \sim \chi_q^2$.*

Theorem B.9.4 *Let $Z \sim \mathrm{N}(0,1)$ and $Q \sim \chi_r^2$ be independent. Then $T_r = Z/\sqrt{Q/r}$ is distributed as student's t with r degrees of freedom.*

Proof of Theorem B.9.1. By the change-of-variables formula, the density of $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ is

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma}_Y)^{1/2}} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{\mu}_Y)' \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} (\boldsymbol{y}-\boldsymbol{\mu}_Y)}{2}\right), \qquad \boldsymbol{y} \in \mathbb{R}^k.$$

where $\boldsymbol{\mu}_Y = \boldsymbol{a}+\boldsymbol{B}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_Y = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'$, where we used the fact that $\det(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')^{1/2} = \det(\boldsymbol{\Sigma})^{1/2}\det(\boldsymbol{B})$. ■

Proof of Theorem B.9.2. First, suppose a random variable Q is distributed chi-square with r degrees of freedom. It has the MGF

$$\mathbb{E}\exp(tQ) = \int_0^\infty \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} x^{r/2-1} \exp(tx) \exp(-x/2)\, dy = (1-2t)^{-r/2}$$

where the second equality uses the fact that $\int_0^\infty y^{a-1}\exp(-by)\,dy = b^{-a}\Gamma(a)$, which can be found by applying change-of-variables to the gamma function. Our goal is to calculate the MGF of $Q = \boldsymbol{X}'\boldsymbol{X}$ and show that it equals $(1-2t)^{-r/2}$, which will establish that $Q \sim \chi^2_r$.

Note that we can write $Q = \boldsymbol{X}'\boldsymbol{X} = \sum_{j=1}^r Z_j^2$ where the Z_j are independent $\mathrm{N}(0,1)$. The distribution of each of the Z_j^2 is

$$\begin{aligned}
\mathbb{P}\left(Z_j^2 \le y\right) &= 2\mathbb{P}\left(0 \le Z_j \le \sqrt{y}\right) \\
&= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \int_0^y \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} s^{-1/2} \exp\left(-\frac{s}{2}\right) ds
\end{aligned}$$

using the change–of-variables $s = x^2$ and the fact $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Thus the density of Z_j^2 is

$$f_1(x) = \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} x^{-1/2} \exp\left(-\frac{x}{2}\right)$$

which is the χ^2_1 and by our above calculation has the MGF of $\mathbb{E}\exp\left(tZ_j^2\right) = (1-2t)^{-1/2}$.

Since the Z_j^2 are mutually independent, (B.6) implies that the MGF of $Q = \sum_{j=1}^r Z_j^2$ is $\left[(1-2t)^{-1/2}\right]^r = (1-2t)^{-r/2}$, which is the MGF of the χ^2_r density as desired. ■

Proof of Theorem B.9.3. The fact that $\boldsymbol{A} > 0$ means that we can write $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}'$ where $\boldsymbol{C}$ is non-singular. Then $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}$ and

$$\boldsymbol{C}^{-1}\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{C}'\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_q\right).$$

Thus

$$\boldsymbol{Z}'\boldsymbol{\mathrm{A}}^{-1}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{Z} = \left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right)'\left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right) \sim \chi^2_q.$$

■

Proof of Theorem B.9.4. Using the simple law of iterated expectations, T_r has distribution

function

$$\begin{aligned} F(x) &= \mathbb{P}\left(\frac{Z}{\sqrt{Q/r}} \leq x\right) \\ &= \mathbb{E}\left\{Z \leq x\sqrt{\frac{Q}{r}}\right\} \\ &= \mathbb{E}\left[\mathbb{P}\left(Z \leq x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\ &= \mathbb{E}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \end{aligned}$$

Thus its density is

$$\begin{aligned} f(x) &= \mathbb{E}\frac{d}{dx}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \\ &= \mathbb{E}\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\ &= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{qx^2}{2r}\right)\right)\sqrt{\frac{q}{r}}\left(\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q/2\right)\right)dq \\ &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} \end{aligned}$$

which is that of the student t with r degrees of freedom. ■

Appendix C

Asymptotic Theory

C.1 Inequalities

The following inequalities are frequently used in asymptotic distribution theory.

Jensen's Inequality. If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is convex, then for any random variable x for which $\mathbb{E}|x| < \infty$ and $\mathbb{E}|g(x)| < \infty$,

$$g(\mathbb{E}(x)) \leq \mathbb{E}(g(x)). \tag{C.1}$$

Expectation Inequality. For any random variable x for which $\mathbb{E}|x| < \infty$,

$$|\mathbb{E}(x)| \leq \mathbb{E}|x|. \tag{C.2}$$

Cauchy-Schwarz Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^2\right)^{1/2}\left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^2\right)^{1/2}. \tag{C.3}$$

Holder's Inequality. If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^p\right)^{1/p}\left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^q\right)^{1/q}. \tag{C.4}$$

Minkowski's Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\left(\mathbb{E}\left\|\boldsymbol{X}+\boldsymbol{Y}\right\|^p\right)^{1/p} \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^p\right)^{1/p} + \left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^p\right)^{1/p} \tag{C.5}$$

Markov's Inequality. For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\mathbb{P}(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}g(\boldsymbol{x}). \tag{C.6}$$

Proof of Jensen's Inequality. Let $a + bu$ be the tangent line to $g(u)$ at $u = \mathbb{E}x$. Since $g(u)$ is convex, tangent lines lie below it. So for all u, $g(u) \geq a + bu$ yet $g(\mathbb{E}x) = a + b\mathbb{E}x$ since the curve is tangent at $\mathbb{E}x$. Applying expectations, $\mathbb{E}g(x) \geq a + b\mathbb{E}x = g(\mathbb{E}x)$, as stated. ■

Proof of Expecation Inequality. Follows from an application of Jensen's Inequality, noting that the function $g(u) = |u|$ is convex. ■

Proof of Holder's Inequality. Since $\frac{1}{p} + \frac{1}{q} = 1$ an application of Jensen's Inequality shows that for any real a and b

$$\exp\left[\frac{1}{p}a + \frac{1}{q}b\right] \leq \frac{1}{p}\exp(a) + \frac{1}{q}\exp(b).$$

Setting $u = \exp(a)$ and $v = \exp(b)$ this implies

$$u^{1/p} v^{1/q} \leq \frac{u}{p} + \frac{v}{q}$$

and this inequality holds for any $u > 0$ and $v > 0$.

Set $u = \|\boldsymbol{X}\|^p / \mathbb{E}\|\boldsymbol{X}\|^p$ and $v = \|\boldsymbol{Y}\|^q / \mathbb{E}\|\boldsymbol{Y}\|^q$. Note that $\mathbb{E}u = \mathbb{E}v = 1$. By the matrix Schwarz Inequality (A.7), $\|\boldsymbol{X}'\boldsymbol{Y}\| \leq \|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|$. Thus

$$\begin{aligned}
\frac{\mathbb{E}\|\boldsymbol{X}'\boldsymbol{Y}\|}{(\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p}(\mathbb{E}\|\boldsymbol{Y}\|^q)^{1/q}} &\leq \frac{\mathbb{E}(\|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|)}{(\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p}(\mathbb{E}\|\boldsymbol{Y}\|^q)^{1/q}} \\
&= \mathbb{E}\left(u^{1/p} v^{1/q}\right) \\
&\leq \mathbb{E}\left(\frac{u}{p} + \frac{v}{q}\right) \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1,
\end{aligned}$$

which is (C.4). ■

Proof of Minkowski's Inequality. Note that by rewriting, using the triangle inequality (A.8), and then Holder's Inequality to the two expectations

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{X} + \boldsymbol{Y}\|^p &= \mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\|\,\|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) \\
&\leq \mathbb{E}\left(\|\boldsymbol{X}\|\,\|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) + \mathbb{E}\left(\|\boldsymbol{Y}\|\,\|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) \\
&\leq (\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p}\,\mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^{q(p-1)}\right)^{1/q} + (\mathbb{E}\|\boldsymbol{Y}\|^p)^{1/p}\,\mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^{q(p-1)}\right)^{1/q} \\
&= \left((\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p} + (\mathbb{E}\|\boldsymbol{Y}\|^p)^{1/p}\right)\mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^p\right)^{(p-1)/p}
\end{aligned}$$

where the second equality picks q to satisfy $1/p + 1/q = 1$, and the final equality uses this fact to make the substitution $q = p/(p-1)$ and then collects terms. Dividing both sides by $\mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^p\right)^{(p-1)/p}$, we obtain (C.5). ■

Proof of Markov's Inequality. Let f denote the density function of $\boldsymbol{x}$. Then

$$\begin{aligned}
\mathbb{P}(g(\boldsymbol{x}) \geq \alpha) &= \int_{\{\boldsymbol{u} g(\boldsymbol{u}) \geq \alpha\}} f(\boldsymbol{u}) d\boldsymbol{u} \\
&\leq \int_{\{\boldsymbol{u} g(\boldsymbol{u}) \geq \alpha\}} \frac{g(\boldsymbol{u})}{\alpha} f(\boldsymbol{u}) d\boldsymbol{u} \\
&\leq \alpha^{-1} \int_{-\infty}^{\infty} g(\boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u} \\
&= \alpha^{-1} \mathbb{E}(g(\boldsymbol{x}))
\end{aligned}$$

the first inequality using the region of integration $\{g(\boldsymbol{u}) > \alpha\}$. ■

C.2 Convergence in Distribution

Let $\boldsymbol{z}_n$ be a random vector with distribution $F_n(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{z}_n \leq \boldsymbol{u})$. We say that $\boldsymbol{z}_n$ **converges in distribution** to $\boldsymbol{z}$ as $n \to \infty$, denoted $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$, where $\boldsymbol{z}$ has distribution $F(\boldsymbol{u}) = \mathbb{P}(\boldsymbol{z} \leq \boldsymbol{u})$, if for all $\boldsymbol{u}$ at which $F(\boldsymbol{u})$ is continuous, $F_n(\boldsymbol{u}) \to F(\boldsymbol{u})$ as $n \to \infty$.

Theorem C.2.1 ***Central Limit Theorem (CLT).*** *If $\boldsymbol{x}_i \in \mathbb{R}^k$ is iid and $\mathbb{E}x_{ji}^2 < \infty$ for $j = 1, ..., k$, then as $n \to \infty$*

$$\sqrt{n}\left(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right).$$

where $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{x}_i$ and $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)'$.

Proof: The moment bound $\mathbb{E}\boldsymbol{x}_{ji}^2 < \infty$ is sufficient to guarantee that the elements of $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are well defined and finite. Without loss of generality, it is sufficient to consider the case $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Omega} = \boldsymbol{I}_k$.

For $\boldsymbol{\lambda} \in \mathbb{R}^k$, let $C\left(\boldsymbol{\lambda}\right) = \mathbb{E}\exp\left(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{x}_i\right)$ denote the characteristic function of $\boldsymbol{x}_i$ and set $c\left(\boldsymbol{\lambda}\right) = \log C(\boldsymbol{\lambda})$. Then observe

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\lambda}}C(\boldsymbol{\lambda}) &= \mathrm{i}\mathbb{E}\left(\boldsymbol{x}_i\exp\left(i\boldsymbol{\lambda}'\boldsymbol{x}_i\right)\right) \\
\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda}) &= \mathrm{i}^2\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\exp\left(i\boldsymbol{\lambda}'\boldsymbol{x}_i\right)\right)
\end{aligned}$$

so when evaluated at $\boldsymbol{\lambda} = \mathbf{0}$

$$\begin{aligned}
C(\mathbf{0}) &= 1 \\
\frac{\partial}{\partial\boldsymbol{\lambda}}C(\mathbf{0}) &= i\mathbb{E}\left(\boldsymbol{x}_i\right) = \mathbf{0} \\
\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\mathbf{0}) &= -\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) = -\boldsymbol{I}_k.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
c_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) &= \frac{\partial}{\partial\boldsymbol{\lambda}}c(\boldsymbol{\lambda}) = C(\boldsymbol{\lambda})^{-1}\frac{\partial}{\partial\boldsymbol{\lambda}}C(\boldsymbol{\lambda}) \\
c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}) &= \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}c(\boldsymbol{\lambda}) = C(\boldsymbol{\lambda})^{-1}\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda}) - C(\boldsymbol{\lambda})^{-2}\frac{\partial}{\partial\boldsymbol{\lambda}}C\left(\boldsymbol{\lambda}\right)\frac{\partial}{\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda})
\end{aligned}$$

so when evaluated at $\boldsymbol{\lambda} = \mathbf{0}$

$$\begin{aligned}
c(\mathbf{0}) &= 0 \\
c_{\lambda}(\mathbf{0}) &= \mathbf{0} \\
c_{\lambda\lambda}(\mathbf{0}) &= -\boldsymbol{I}_k.
\end{aligned}$$

By a second-order Taylor series expansion of $c(\boldsymbol{\lambda})$ about $\boldsymbol{\lambda} = \mathbf{0}$,

$$c(\boldsymbol{\lambda}) = c(\mathbf{0}) + c_{\lambda}(\mathbf{0})'\boldsymbol{\lambda} + \frac{1}{2}\boldsymbol{\lambda}'c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*)\boldsymbol{\lambda} = \frac{1}{2}\boldsymbol{\lambda}'c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*)\boldsymbol{\lambda} \tag{C.7}$$

where $\boldsymbol{\lambda}^*$ lies on the line segment joining $\mathbf{0}$ and $\boldsymbol{\lambda}$.

We now compute $C_n(\boldsymbol{\lambda}) = E\exp\left(\mathrm{i}\boldsymbol{\lambda}'\sqrt{n}\overline{\boldsymbol{x}}_n\right)$ the characteristic function of $\sqrt{n}\overline{\boldsymbol{x}}_n$. By the properties of the exponential function, the independence of the $\boldsymbol{x}_i$, the definition of $c(\boldsymbol{\lambda})$ and (C.7)

$$\begin{aligned}
\log C_n(\boldsymbol{\lambda}) &= \log\mathbb{E}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\lambda}'\boldsymbol{x}_i\right) \\
&= \log\mathbb{E}\prod_{i=1}^{n}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\boldsymbol{\lambda}'\boldsymbol{x}_i\right) \\
&= \log\prod_{i=1}^{n}\mathbb{E}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\boldsymbol{\lambda}'\boldsymbol{x}_i\right) \\
&= nc\left(\frac{\boldsymbol{\lambda}}{\sqrt{n}}\right) \\
&= \frac{1}{2}\boldsymbol{\lambda}'c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}_n)\boldsymbol{\lambda}
\end{aligned}$$

where $\boldsymbol{\lambda}_n \to \mathbf{0}$ lies on the line segment joining $\mathbf{0}$ and $\boldsymbol{\lambda}/\sqrt{n}$. Since $c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}_n) \to c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\mathbf{0}) = -\boldsymbol{I}_k$, we see that as $n \to \infty$,

$$C_n(\boldsymbol{\lambda}) \to \exp\left(-\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{\lambda}\right)$$

the characteristic function of the $\mathrm{N}(\mathbf{0}, \boldsymbol{I}_k)$ distribution. This is sufficient to establish the theorem. ■

C.3 Asymptotic Transformations

Theorem C.3.1 ***Continuous Mapping Theorem 1*** *(CMT). If $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$ as $n \to \infty$ and $g(\cdot)$ is continuous at $\boldsymbol{c}$, then $g(\boldsymbol{z}_n) \xrightarrow{p} g(\boldsymbol{c})$ as $n \to \infty$.*

Proof: Since g is continuous at $\boldsymbol{c}$, for all $\varepsilon > 0$ we can find a $\delta > 0$ such that if $\|\boldsymbol{z}_n - \boldsymbol{c}\| < \delta$ then $|g(\boldsymbol{z}_n) - g(\boldsymbol{c})| \leq \varepsilon$. Recall that $A \subset B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$. Thus $\mathbb{P}(|g(\boldsymbol{z}_n) - g(\boldsymbol{c})| \leq \varepsilon) \geq \mathbb{P}(\|\boldsymbol{z}_n - \boldsymbol{c}\| < \delta) \to 1$ as $n \to \infty$ by the assumption that $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$. Hence $g(\boldsymbol{z}_n) \xrightarrow{p} g(\boldsymbol{c})$ as $n \to \infty$.

Theorem C.3.2 ***Continuous Mapping Theorem 2.*** *If $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ as $n \to \infty$ and $g(\cdot)$ is continuous, then $g(\boldsymbol{z}_n) \xrightarrow{d} g(\boldsymbol{z})$ as $n \to \infty$.*

Theorem C.3.3 ***Delta Method:*** *If $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}$ is $m \times 1$ and $\boldsymbol{\Sigma}$ is $m \times m$, and $g(\boldsymbol{\theta}) : \mathbb{R}^m \to \mathbb{R}^k$, $k \leq m$, then*

$$\sqrt{n}(g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathrm{N}\left(0, g_{\boldsymbol{\theta}}\boldsymbol{\Sigma}g_{\boldsymbol{\theta}}'\right)$$

where $g_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'}g(\boldsymbol{\theta})$ and $g_{\boldsymbol{\theta}} = g_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$.

Proof: By a vector Taylor series expansion, for each element of g,

$$g_j(\boldsymbol{\theta}_n) = g_j(\boldsymbol{\theta}_0) + g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^*)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0)$$

where $\boldsymbol{\theta}_{nj}^*$ lies on the line segment between $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_0$ and therefore converges in probability to $\boldsymbol{\theta}_0$. It follows that $a_{jn} = g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^*) - g_{j\boldsymbol{\theta}} \xrightarrow{p} 0$. Stacking across elements of g, we find

$$\sqrt{n}(g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta}_0)) = (g_{\boldsymbol{\theta}} + a_n)\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{d} g_{\boldsymbol{\theta}}\,\mathrm{N}(0, \boldsymbol{\Sigma}) = \mathrm{N}\left(0, g_{\boldsymbol{\theta}}\boldsymbol{\Sigma}g_{\boldsymbol{\theta}}'\right).$$

Appendix D

Maximum Likelihood

If the distribution of $\boldsymbol{y}_i$ is $F(\boldsymbol{y}, \boldsymbol{\theta})$ where F is a known distribution function and $\boldsymbol{\theta} \in \Theta$ is an unknown $m \times 1$ vector, we say that the distribution is **parametric** and that $\boldsymbol{\theta}$ is the **parameter** of the distribution F. The space Θ is the set of permissible value for $\boldsymbol{\theta}$. In this setting the **method of maximum likelihood** is the appropriate technique for estimation and inference on $\boldsymbol{\theta}$.

If the distribution F is continuous then the density of $\boldsymbol{y}_i$ can be written as $f(\boldsymbol{y}, \boldsymbol{\theta})$ and the joint density of a random sample $(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ is

$$f_n(\boldsymbol{y}_1, ..., \boldsymbol{y}_n, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\theta}).$$

The **likelihood** of the sample is this joint density evaluated at the observed sample values, viewed as a function of $\boldsymbol{\theta}$. The **log-likelihood** function is its natural log

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i, \boldsymbol{\theta}).$$

If the distribution F is discrete, the likelihood and log-likelihood are constructed by setting $f(\boldsymbol{y}, \boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{y}_i = \boldsymbol{y}, \boldsymbol{\theta})$.

Define the **Hessian**

$$\boldsymbol{\mathcal{H}} = -\mathbb{E}\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\boldsymbol{y}_i, \boldsymbol{\theta}_0) \tag{D.1}$$

and the **outer product** matrix

$$\boldsymbol{\Omega} = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i, \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i, \boldsymbol{\theta}_0)'\right). \tag{D.2}$$

Two important features of the likelihood are

Theorem D.0.4

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} \log f(\boldsymbol{y}_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0} \tag{D.3}$$

$$\boldsymbol{\mathcal{H}} = \boldsymbol{\Omega} \equiv \mathcal{I}_0 \tag{D.4}$$

The matrix $\mathcal{I}_0$ is called the **information**, and the equality (D.4) is often called the **information matrix equality**.

Theorem D.0.5 ***Cramer-Rao Lower Bound.*** *If $\tilde{\theta}$ is an unbiased estimator of $\theta \in \mathbb{R}$, then* $\text{var}(\tilde{\theta}) \geq (n\mathcal{I}_0)^{-1}$.

The Cramer-Rao Theorem gives a lower bound for estimation. However, the restriction to unbiased estimators means that the theorem has little direct relevance for finite sample efficiency.

The **maximum likelihood estimator** or **MLE** $\hat{\boldsymbol{\theta}}$ is the parameter value which maximizes the likelihood (equivalently, which maximizes the log-likelihood). We can write this as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \log L(\boldsymbol{\theta}).$$

In some simple cases, we can find an explicit expression for $\hat{\boldsymbol{\theta}}$ as a function of the data, but these cases are rare. More typically, the MLE $\hat{\boldsymbol{\theta}}$ must be found by numerical methods.

Why do we believe that the MLE $\hat{\boldsymbol{\theta}}$ is estimating the parameter $\boldsymbol{\theta}$? Observe that when standardized, the log-likelihood is a sample average

$$\frac{1}{n}\log L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}\right) \xrightarrow{p} \mathbb{E}\log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}\right).$$

As the MLE $\hat{\boldsymbol{\theta}}$ maximizes the left-hand-side, we can see that it is an estimator of the maximizer of the right-hand-side. The first-order condition for the latter problem is

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}\right)$$

which holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ by (D.3). In fact, under conventional regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent for this value, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$.

Theorem D.0.6 *Under regularity conditions,* $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathcal{I}_0^{-1}\right)$.

Thus in large samples, the approximate variance of the MLE is $(nI_0)^{-1}$ which is the Cramer-Rao lower bound. Thus in large samples the MLE has approximately the best possible variance. Therefore the MLE is called **asymptotically efficient**.

Typically, to estimate the asymptotic variance of the MLE we use an estimate based on the Hessian formula (D.1)

$$\widehat{\boldsymbol{\mathcal{H}}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}\right) \tag{D.5}$$

We then set $\widehat{\mathcal{I}}_0^{-1} = \widehat{\boldsymbol{\mathcal{H}}}^{-1}$. Asymptotic standard errors for $\hat{\boldsymbol{\theta}}$ are then the square roots of the diagonal elements of $n^{-1}\widehat{\mathcal{I}}_0^{-1}$.

Sometimes a parametric density function $f(\boldsymbol{y}, \boldsymbol{\theta})$ is used to approximate the true unknown density $f(\boldsymbol{y})$, but it is not literally believed that the model $f(\boldsymbol{y}, \boldsymbol{\theta})$ is necessarily the true density. In this case, we refer to $\log L(\boldsymbol{\theta})$ as a **quasi-likelihood** and the its maximizer $\hat{\boldsymbol{\theta}}$ as a **quasi-mle** or **QMLE.**

In this case there is not a "true" value of the parameter $\boldsymbol{\theta}$. Instead we define the **pseudo-true** value $\boldsymbol{\theta}_0$ as the maximizer of

$$\mathbb{E}\log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}\right) = \int f\left(\boldsymbol{y}\right) \log f\left(\boldsymbol{y}, \boldsymbol{\theta}\right) d\boldsymbol{y}$$

which is the same as the minimizer of

$$KLIC = \int f\left(\boldsymbol{y}\right) \log\left(\frac{f(\boldsymbol{y})}{f\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}\right) d\boldsymbol{y}$$

the Kullback-Leibler information distance between the true density $f(\boldsymbol{y})$ and the parametric density $f(\boldsymbol{y}, \boldsymbol{\theta})$. Thus the QMLE $\boldsymbol{\theta}_0$ is the value which makes the parametric density "closest" to the true value according to this measure of distance. The QMLE is consistent for the pseudo-true value, but

has a different covariance matrix than in the pure MLE case, since the information matrix equality (D.4) does not hold. A minor adjustment to Theorem (D.0.6) yields the asymptotic distribution of the QMLE:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right), \qquad \boldsymbol{V}=\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\Omega}\boldsymbol{\mathcal{H}}^{-1}$$

The moment estimator for $\boldsymbol{V}$ is

$$\hat{\boldsymbol{V}}=\widehat{\boldsymbol{\mathcal{H}}}^{-1}\hat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\mathcal{H}}}^{-1}$$

where $\widehat{\boldsymbol{\mathcal{H}}}$ is given in (D.5) and

$$\hat{\boldsymbol{\Omega}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}\right)\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}\right)'.$$

Asymptotic standard errors (sometimes called qmle standard errors) are then the square roots of the diagonal elements of $n^{-1}\hat{\boldsymbol{V}}$.

Proof of Theorem D.0.4. To see (D.3),

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \frac{\partial}{\partial\boldsymbol{\theta}}\int\log f\left(\boldsymbol{y},\boldsymbol{\theta}\right)f\left(\boldsymbol{y},\boldsymbol{\theta}_0\right)d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\
&= \int\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{y},\boldsymbol{\theta}\right)\frac{f\left(\boldsymbol{y},\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\int f\left(\boldsymbol{y},\boldsymbol{\theta}\right)d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\
&= \frac{\partial}{\partial\boldsymbol{\theta}}1\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=\mathbf{0}.
\end{aligned}$$

Similarly, we can show that

$$\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}\right)=\mathbf{0}.$$

By direction computation,

$$\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right) &= \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}-\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)'}{f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)^2}\\
&= \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)}-\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_0\right)'.
\end{aligned}$$

Taking expectations yields (D.4). ■

Proof of Theorem D.0.5. Let $\boldsymbol{Y}=(\boldsymbol{y}_1,...,\boldsymbol{y}_n)$ be the sample, and set

$$S=\frac{\partial}{\partial\theta}\log f_n\left(\mathbf{Y},\theta_0\right)=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f\left(\boldsymbol{y}_i,\theta_0\right)$$

which by Theorem (D.0.4) has mean zero and variance nH. Write the estimator $\tilde{\theta}=\tilde{\theta}\left(\mathbf{Y}\right)$ as a function of the data. Since $\tilde{\theta}$ is unbiased for any θ,

$$\theta=\mathbb{E}\tilde{\theta}=\int\tilde{\theta}\left(\mathbf{Y}\right)f\left(\mathbf{Y},\theta\right)d\mathbf{Y}.$$

Differentiating with respect to θ and evaluating at θ_0 yields

$$1=\int\tilde{\theta}\left(\mathbf{Y}\right)\frac{\partial}{\partial\theta}f\left(\mathbf{Y},\theta\right)d\mathbf{Y}=\int\tilde{\theta}\left(\mathbf{Y}\right)\frac{\partial}{\partial\theta}\log f\left(\mathbf{Y},\theta\right)f\left(\mathbf{Y},\theta_0\right)d\mathbf{Y}=\mathbb{E}\left(\tilde{\theta}S\right).$$

By the Cauchy-Schwarz inequality

$$1 = \left|\mathbb{E}\left(\tilde{\theta}S\right)\right|^2 \leq \text{var}\left(S\right)\text{var}\left(\tilde{\theta}\right)$$

so

$$\text{var}\left(\tilde{\theta}\right) \geq \frac{1}{\text{var}\left(S\right)} = \frac{1}{n\mathcal{H}}.$$

■

Proof of Theorem D.0.6 Taking the first-order condition for maximization of $\log L(\boldsymbol{\theta})$, and making a first-order Taylor series expansion,

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}\right) \\ &\simeq \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_0\right) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_n\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right), \end{aligned}$$

where $\boldsymbol{\theta}_n$ lies on a line segment joining $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. (Technically, the specific value of $\boldsymbol{\theta}_n$ varies by row in this expansion.) Rewriting this equation, we find

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_n\right)\right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_0\right)\right).$$

Since $\frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_0\right)$ is mean-zero with covariance matrix $\boldsymbol{\Omega}$, an application of the CLT yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_0\right) \xrightarrow{d} \text{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right).$$

The analysis of the sample Hessian is somewhat more complicated due to the presence of $\boldsymbol{\theta}_n$. Let $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}\right)$. If it is continuous in $\boldsymbol{\theta}$, then since $\boldsymbol{\theta}_n \xrightarrow{p} \boldsymbol{\theta}_0$ we find $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n) \xrightarrow{p} \boldsymbol{\mathcal{H}}$ and so

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_n\right) &= \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i, \boldsymbol{\theta}_n\right) - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n)\right) + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n) \\ &\xrightarrow{p} \boldsymbol{\mathcal{H}} \end{aligned}$$

by an application of a uniform WLLN. Together,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \boldsymbol{\mathcal{H}}^{-1}\text{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right) = \text{N}\left(\mathbf{0}, \boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\Omega}\boldsymbol{\mathcal{H}}^{-1}\right) = \text{N}\left(\mathbf{0}, \boldsymbol{\mathcal{H}}^{-1}\right),$$

the final equality using Theorem D.0.4 . ■

Appendix E

Numerical Optimization

Many econometric estimators are defined by an optimization problem of the form

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, Q(\boldsymbol{\theta}) \tag{E.1}$$

where the parameter is $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ and the criterion function is $Q(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}$. For example NLLS, GLS, MLE and GMM estimators take this form. In most cases, $Q(\boldsymbol{\theta})$ can be computed for given $\boldsymbol{\theta}$, but $\hat{\boldsymbol{\theta}}$ is not available in closed form. In this case, numerical methods are required to obtain $\hat{\boldsymbol{\theta}}$.

E.1 Grid Search

Many optimization problems are either one dimensional ($m = 1$) or involve one-dimensional optimization as a sub-problem (for example, a line search). In this context grid search may be employed.

Grid Search. Let $\Theta = [a, b]$ be an interval. Pick some $\varepsilon > 0$ and set $G = (b - a)/\varepsilon$ to be the number of gridpoints. Construct an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and find the gridpoint which yields the smallest value of the criterion, which is $\boldsymbol{\theta}(\hat{\jmath})$ where $\hat{\jmath} = \operatorname{argmin}_{0\le j\le G} Q(\boldsymbol{\theta}(j))$. This value $\boldsymbol{\theta}(\hat{\jmath})$ is the gridpoint estimate of $\hat{\boldsymbol{\theta}}$**.** If the grid is sufficiently fine to capture small oscillations in $Q(\boldsymbol{\theta})$, the approximation error is bounded by ε, that is, $\left|\boldsymbol{\theta}(\hat{\jmath}) - \hat{\boldsymbol{\theta}}\right| \le \varepsilon$. Plots of $Q(\boldsymbol{\theta}(j))$ against $\boldsymbol{\theta}(j)$ can help diagnose errors in grid selection. This method is quite robust but potentially costly.

Two-Step Grid Search. The gridsearch method can be refined by a two-step execution. For an error bound of ε pick G so that $G^2 = (b - a)/\varepsilon$ For the first step define an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and let $\hat{\jmath} = \operatorname{argmin}_{0\le j\le G} Q(\boldsymbol{\theta}(j))$. For the second step define an equally spaced grid on $[\boldsymbol{\theta}(\hat{\jmath} - 1), \boldsymbol{\theta}(\hat{\jmath} + 1)]$ with G gridpoints, which is $\{\boldsymbol{\theta}'(k) = \boldsymbol{\theta}(\hat{\jmath} - 1) + 2k(b - a)/G^2 : k = 0, ..., G\}$. Let $\hat{k} = \operatorname{argmin}_{0\le k\le G} Q(\boldsymbol{\theta}'(k))$. The estimate of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}\left(\hat{k}\right)$. The advantage of the two-step method over a one-step grid search is that the number of function evaluations has been reduced from $(b-a)/\varepsilon$ to $2\sqrt{(b - a)/\varepsilon}$ which can be substantial. The disadvantage is that if the function $Q(\boldsymbol{\theta})$ is irregular, the first-step grid may not bracket $\hat{\boldsymbol{\theta}}$ which thus would be missed.

E.2 Gradient Methods

Gradient Methods are iterative methods which produce a sequence $\boldsymbol{\theta}_i : i = 1, 2, ...$ which are designed to converge to $\hat{\boldsymbol{\theta}}$. All require the choice of a starting value $\boldsymbol{\theta}_1$, and all require the

computation of the **gradient** of $Q(\boldsymbol{\theta})$

$$\boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta})$$

and some require the **Hessian**

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}).$$

If the functions $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ are not analytically available, they can be calculated numerically. Take the $j'th$ element of $\boldsymbol{g}(\boldsymbol{\theta})$. Let δ_j be the $j'th$ unit vector (zeros everywhere except for a one in the $j'th$ row). Then for ε small

$$g_j(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon) - Q(\boldsymbol{\theta})}{\varepsilon}.$$

Similarly,

$$g_{jk}(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_j \varepsilon) + Q(\boldsymbol{\theta})}{\varepsilon^2}$$

In many cases, numerical derivatives can work well but can be computationally costly relative to analytic derivatives. In some cases, however, numerical derivatives can be quite unstable.

Most gradient methods are a variant of **Newton's method** which is based on a quadratic approximation. By a Taylor's expansion for $\boldsymbol{\theta}$ close to $\hat{\boldsymbol{\theta}}$

$$0 = \boldsymbol{g}(\hat{\boldsymbol{\theta}}) \simeq \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)$$

which implies

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta})^{-1} \boldsymbol{g}(\boldsymbol{\theta}).$$

This suggests the iteration rule

$$\hat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where

One problem with Newton's method is that it will send the iterations in the wrong direction if $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)$ is not positive definite. One modification to prevent this possibility is quadratic hill-climbing which sets

$$\hat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m \right)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where α_i is set just above the smallest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta}_i)$ if $\boldsymbol{H}(\boldsymbol{\theta})$ is not positive definite.

Another productive modification is to add a scalar **steplength** λ_i. In this case the iteration rule takes the form

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{D}_i \boldsymbol{g}_i \lambda_i \tag{E.2}$$

where $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{\theta}_i)$ and $\boldsymbol{D}_i = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1}$ for Newton's method and $D_i = \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m \right)^{-1}$ for quadratic hill-climbing.

Allowing the steplength to be a free parameter allows for a line search, a one-dimensional optimization. To pick λ_i write the criterion function as a function of λ

$$Q(\lambda) = Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$$

a one-dimensional optimization problem. There are two common methods to perform a line search. A **quadratic approximation** evaluates the first and second derivatives of $Q(\lambda)$ with respect to λ, and picks λ_i as the value minimizing this approximation. The **half-step** method considers the sequence $\lambda = 1, 1/2, 1/4, 1/8, \ldots$. Each value in the sequence is considered and the criterion $Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$ evaluated. If the criterion has improved over $Q(\boldsymbol{\theta}_i)$, use this value, otherwise move to the next element in the sequence.

Newton's method does not perform well if $Q(\boldsymbol{\theta})$ is irregular, and it can be quite computationally costly if $\boldsymbol{H}(\boldsymbol{\theta})$ is not analytically available. These problems have motivated alternative choices for the weight matrix D_i. These methods are called **Quasi-Newton** methods. Two popular methods are do to Davidson-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS).

Let

$$\begin{aligned} \Delta \boldsymbol{g}_i &= \boldsymbol{g}_i - \boldsymbol{g}_{i-1} \\ \Delta \boldsymbol{\theta}_i &= \boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1} \end{aligned}$$

and . The DFP method sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i \Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i}.$$

The BFGS methods sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} - \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{\theta}_i'}{\left(\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i\right)^2} \Delta g_i' \boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i + \frac{\Delta \boldsymbol{\theta}_i \Delta \boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1} \Delta \boldsymbol{g}_i \Delta \boldsymbol{\theta}_i'}{\Delta \boldsymbol{\theta}_i' \Delta \boldsymbol{g}_i}.$$

For any of the gradient methods, the iterations continue until the sequence has converged in some sense. This can be defined by examining whether $|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}|$, $|Q(\boldsymbol{\theta}_i) - Q(\boldsymbol{\theta}_{i-1})|$ or $|g(\boldsymbol{\theta}_i)|$ has become small.

E.3 Derivative-Free Methods

All gradient methods can be quite poor in locating the global minimum when $Q(\boldsymbol{\theta})$ has several local minima. Furthermore, the methods are not well defined when $Q(\boldsymbol{\theta})$ is non-differentiable. In these cases, alternative optimization methods are required. One example is the **simplex method** of Nelder-Mead (1965).

A more recent innovation is the method of **simulated annealing (SA).** For a review see Goffe, Ferrier, and Rodgers (1994). The SA method is a sophisticated random search. Like the gradient methods, it relies on an iterative sequence. At each iteration, a random variable is drawn and added to the current value of the parameter. If the resulting criterion is decreased, this new value is accepted. If the criterion is increased, it may still be accepted depending on the extent of the increase and another randomization. The latter property is needed to keep the algorithm from selecting a local minimum. As the iterations continue, the variance of the random innovations is shrunk. The SA algorithm stops when a large number of iterations is unable to improve the criterion. The SA method has been found to be successful at locating global minima. The downside is that it can take considerable computer time to execute.

Bibliography

[1] Abadir, Karim M. and Jan R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.

[2] Aitken, A.C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[3] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[4] Anderson, T.W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46-63.

[5] Andrews, Donald W. K. (1988): "Laws of large numbers for dependent non-identically distributed random variables,' *Econometric Theory,* 4, 458-467.

[6] Andrews, Donald W. K. (1991), "Asymptotic normality of series estimators for nonparameric and semiparametric regression models," *Econometrica*, 59, 307-345.

[7] Andrews, Donald W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.

[8] Andrews, Donald W. K. and Moshe Buchinsky: (2000): "A three-step method for choosing the number of bootstrap replications," *Econometrica*, 68, 23-51.

[9] Andrews, Donald W. K. and Werner Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica,* 62, 1383-1414.

[10] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.

[11] Basmann, R. L. (1957): "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* 25, 77-83.

[12] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators, *Econometrica,* 62, 657-681.

[13] Billingsley, Patrick (1968): *Convergence of Probability Measures.* New York: Wiley.

[14] Billingsley, Patrick (1995): *Probability and Measure,* 3rd Edition, New York: Wiley.

[15] Bose, A. (1988): "Edgeworth correction by bootstrap in autoregressions," *Annals of Statistics*, 16, 1709-1722.

[16] Breusch, T.S. and A.R. Pagan (1979): "The Lagrange multiplier test and its application to model specification in econometrics," *Review of Economic Studies*, 47, 239-253.

[17] Brown, B. W. and Whitney K. Newey (2002): "GMM, efficient bootstrapping, and improved inference ," *Journal of Business and Economic Statistics.*

[18] Carlstein, E. (1986): "The use of subseries methods for estimating the variance of a general statistic from a stationary time series," *Annals of Statistics*, 14, 1171-1179.

[19] Casella, George and Roger L. Berger (2002): *Statistical Inference*, 2nd Edition, Duxbury Press.

[20] Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[21] Choi, I. and P.C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations," *Journal of Econometrics*, 51, 113-150.

[22] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.

[23] Cragg, John (1992): "Quasi-Aitken Estimation for Heterskedasticity of Unknown Form" *Journal of Econometrics*, 54, 179-201.

[24] Davidson, James (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

[25] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application*. Cambridge University Press.

[26] Dickey, D.A. and W.A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.

[27] Donald Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[28] Dufour, J.M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.

[29] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.

[30] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[31] Efron, Bradley and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.

[32] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[33] Engle, Robert F. and Clive W. J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.

[34] Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1, 1-4.

[35] Frisch, R. and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.

[36] Gallant, A. Ronald and D.W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[37] Gallant, A. Ronald and Halbert White (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

[38] Galton, Francis (1886): "Regression Towards Mediocrity in Hereditary Stature," The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[39] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

[40] Goffe, W.L., G.D. Ferrier and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65-99.

[41] Gauss, K.F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.

[42] Granger, Clive W. J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.

[43] Granger, Clive W. J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

[44] Granger, Clive W. J. and Timo Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[45] Gregory, A. and M. Veall (1985): "On formulating Wald tests of nonlinear restrictions," *Econometrica*, 53, 1465-1468,

[46] Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica*, supplement, 12.

[47] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527,

[48] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

[49] Hall, P. (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.

[50] Hall, P. and J.L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.

[51] Hahn, J. (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.

[52] Hamilton, James D. (1994) *Time Series Analysis*.

[53] Hansen, Bruce E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.

[54] Hansen, Bruce E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.

[55] Hansen, Bruce E. (2006): "Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, edited by Dean Corbae, Steven N. Durlauf and Bruce E. Hansen. Cambridge University Press.

[56] Hansen, Lars Peter (1982): "Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

[57] Hansen, Lars Peter, John Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.

[58] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[59] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

[60] Horowitz, Joel (2001): "The Bootstrap," *Handbook of Econometrics, Vol. 5*, J.J. Heckman and E.E. Leamer, eds., Elsevier Science, 3159-3228.

[61] Imbens, G.W. (1997): "One step estimators for over-identified generalized method of moments models," Review of Economic Studies, 64, 359-383.

[62] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.

[63] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters*, 6, 255-259.

[64] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

[65] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

[66] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.

[67] Johansen, S. and K. Juselius (1992): "Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the UK," *Journal of Econometrics*, 53, 211-244.

[68] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.

[69] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..

[70] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

[71] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.

[72] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159-178.

[73] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.

[74] Lehmann, E.L. and George Casella (1998): *Theory of Point Estimation*, 2nd Edition, Springer.

[75] Lehmann, E.L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses*, 3rd Edition, Springer.

[76] Li, Qi and Jeffrey Racine (2007) *Nonparametric Econometrics*.

[77] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.

[78] MacKinnon, James G. (1990): "Critical values for cointegration," in Engle, R.F. and C.W. Granger (eds.) *Long-Run Economic Relationships: Readings in Cointegration*, Oxford, Oxford University Press.

[79] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[80] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.

[81] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.

[82] Nelder, J. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308-313.

[83] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.

[84] Newey, Whitney K. and Daniel L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, (eds.) *Handbook of Econometrics*, vol. IV, 2111-2245, North Holland: Amsterdam.

[85] Newey, Whitney K. and Kenneth D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[86] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

[87] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.

[88] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083,

[89] Phillips, Peter C.B. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181-240.

[90] Phillips, Peter C.B. and Sam Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.

[91] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[92] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[93] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

[94] Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

[95] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.

[96] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.

[97] Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.

[98] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica, 2*6, 393-415.

[99] Shao, Jun (2003): *Mathematical Statistics*, 2nd edition, Springer.

[100] Sheather, S.J. and M.C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

[101] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.

[102] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[103] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.

[104] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica, 48*, 1-48.

[105] Staiger, D. and James H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica,* 65, 557-586.

[106] Stock, James H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica,* 55, 1035-1056.

[107] Stock, James H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics,* 28, 435-460.

[108] Stock, James H. and Jonathan H. Wright (2000): "GMM with weak identification," *Econometrica,* 68, 1055-1096.

[109] Theil, H. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

[110] Theil, H. (1971): *Principles of Econometrics*, New York: Wiley.

[111] Tobin, James (1958): "Estimation of relationships for limited dependent variables," *Econometrica, 2*6, 24-36.

[112] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[113] Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.

[114] Wang, J. and E. Zivot (1998): "Inference on structural parameters in instrumental variables regression with weak instruments," *Econometrica,* 66, 1389-1404.

[115] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[116] White, Halbert (1984): *Asymptotic Theory for Econometricians,* Academic Press.

[117] Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

[118] Zellner, Arnold. (1962): "An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-368.