Lecture 1
Review of Linear Models

1 General Linear Model without Endogeneity

A general linear model is defined as follows:

$$y = X\beta + u,\ E(u|X) = 0, \quad (1)$$

where $y \in R^n$, X is a $n \times k$ matrix, $u \in R^n$. Notice that the restriction $E(u|X) = 0$ is indispensable in the definition of the model. Without this restriction, the equation is just tautological. The conditional mean zero restriction, however, may be replaced with other restrictions on the error term.[1]

We outline several examples of linear models:

1. One of the simplest and most basic statistical models is the simple measurement error model. For example, measure the length of a football field n times $y = (y_1, ..., y_n)'$. The linear model is $y = \mathbf{1}_n\beta + u$, $E(u) = 0$, where $\mathbf{1}_n$ is a column vector of ones of length n and β is a scalar.

2. A more complicated measurement error model arises when the observations are not observations of the same object. For example, consider y_i is crop yield, fertilizer of type 1 is used on fields $i = 1, ..., n_1$ and fertilizer of type 2 is used on fields $i = n_1 + 1, ..., n$. The model is

$$\begin{aligned} y_i &= \beta_1 + u_i,\ i = 1, ..., n_1 \\ y_i &= \beta_2 + u_i,\ i = n_1 + 1, ..., n. \\ & E(u_i|i) = 0,\ i = 1, ..., n \end{aligned} \quad (2)$$

   The model can be written in the general form:

$$y = X\beta + u,\ E(u|X) = 0, \quad (3)$$

   where $\beta = (\beta_1, \beta_2)'$ and $X = \begin{pmatrix} 1 & 0 \\ ... & ... \\ 1 & 0 \\ 0 & 1 \\ ... & ... \\ 0 & 1 \end{pmatrix}$.

[1] For example, it is replaced with a conditional quantile zero restriction in quantile regression models.

3. Regression model: y is a dependent variable and X contains the explanatory variables

4. Seemingly unrelated regression model (SUR). We have m regression equations

$$\begin{aligned} y_1 &= X_1\beta_1 + u_1 \\ &... \\ y_m &= X_m\beta_m + u_2, \end{aligned} \tag{4}$$

where $y_j \in R^n$, $X_j \in R^{n\times k_j}$ and $u_j \in R^n$, $E(u_j|X_1, ..., X_m) = 0$. This can be written in the general form:

$$y = X\beta + u,\ E(u|X) = 0, \tag{5}$$

where $y = \begin{pmatrix} y_1 \\ ... \\ y_m \end{pmatrix}$, $X = \begin{pmatrix} X_1 & & 0 \\ & ... & \\ 0 & & X_m \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_1 \\ ... \\ \beta_m \end{pmatrix}$.

5. Multivariate regression model: a SUR model when $X_1 = X_2 = ... = X_m$.

2 Least Square Estimation

Given the general model $y = X\beta + u$, $E(u|X) = 0$, the typical estimator for the true value of β is obtained by minimizing the squared deviation from the true y and the predicted y: $X\beta$. The least square (LS) estimator is defined as:

$$\hat{\beta}_{OLS} = \min_{\beta} ||y - X\beta||^2, \tag{6}$$

where $||\cdot||$ is the euclidean norm: $||a|| = \sqrt{a'a}$. The least square problem has an explicitly solution:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y. \tag{7}$$

We call the least square predicted y: $\hat{y} = X\hat{\beta}_{OLS} = X(X'X)^{-1}X'Y$ the projection of y on the linear subspace spanned by X, i.e. $L = R(X)$. Consequently, we call $P_X = X(X'X)^{-1}X'$ the projection matrix.[2]

Before discussing the properties of the LS estimator, let's formalize the assumptions below. In the following assumptions, LN stands for "linearity", CM stands for "conditional mean", VA stands for "variace" and FR stands for "full rank".

Assumption LN. $y = X\beta_0 + u$, where β_0 is the true value of the parameter β.

[2] The matrix $M_X = I - P_X$ is the projection matrix of the orthogonal space of X.

Assumption CM. $E(u|X) = 0$.

Assumption VA. $E(uu'|X) = \sigma^2 I$, where I is the $n \times n$ identity matrix.

Assumption FR. $rank(X) = \text{col}(X) = k$.

Gauss-Markov Theorem. Suppose Assumptions LN, CM, VA, and FR hold. Then the LS estimator defined in (6) is the best linear unbiased estimator (BLUE) of β_0, where "best" means "minimum variance", "linear" means "linear in y" and "unbiased" means $E(\hat{\beta}_{OLS}|X) = \beta_0$.

Proof. First, $\hat{\beta}_{OLS}$ is clearly linear in y by (7), and it also is unbiased because $E\left(\hat{\beta}_{OLS}|X\right) = E((X'X)^{-1} X' (X\beta_0 + u)|X) = \beta_0 + E((X'X)^{-1} X'u|X) = \beta_0 + (X'X)^{-1} X'E(u|X) = \beta_0$.

Let $\hat{\beta}^*$ be another linear unbiased estimator of β_0. Then there exists a matrix A (which is a function of X) such that $\hat{\beta}^* = Ay$ and

$$\beta_0 = E\left(\hat{\beta}^*|X\right) = E\left(Ay|X\right) = E(AX\beta_0 + Du|X) = AX\beta_0. \tag{8}$$

Because β_0 can be any k-vector, (8) implies that

$$AX = I. \tag{9}$$

Then, the variance of $\hat{\beta}^*$ is

$$\begin{aligned} Var(\hat{\beta}^*|X) &= E\left[(Ay - \beta_0)(Ay - \beta_0)'|X\right] \\ &= E(Auu'A'|X) \\ &= \sigma^2 AA' \\ &= \sigma^2 (C + (X'X)^{-1}X')(C + (X'X)^{-1}X')' \\ &= \sigma^2 CC' + \sigma^2 CX(X'X)^{-1} + \sigma^2 (X'X)^{-1}X'C' + \sigma^2 \left(X'X\right)^{-1} \\ &= \sigma^2 CC' + \sigma^2 \left(X'X\right)^{-1} \\ &\geq \sigma^2 (X'X)^{-1}, \end{aligned} \tag{10}$$

where $C = A - (X'X)^{-1}X'$, the second equality holds by (9), the third equality holds by Assumption VA, and the last equality holds because $CX = 0$ by (9). The inequality holds because CC' is positive semi-definite and it is nonzero unless $C = 0$, in which case $A = (X'X)^{-1}X$ and $\hat{\beta}^* = \hat{\beta}_{OLS}$. ■

Comments on the Gauss-Markov Theorem.

1. The G-M theorem is a relatively strong justification for OLS under relatively minimal assumptions.

2. The G-M theorem rules out biased estimators. Sometimes allowing a little bit bias can reduce variance by a lot.

3. The G-M theorem rules out nonlinear estimators. Many maximum likelihood estimators are not linear in y (even if the linearity assumption LN is satisfied). That's one reason we need to study nonlinear methods.

4. The G-M theorem requires homoskedasticity and no auto-correlation: $Var(u|X) = \sigma^2 I$. Homoskedasticity is often violated in pooled cross-section data and many large cross-sectional data. Feasible Generalized Least Square may be used in those cases (or not). No-auto-correlation is often violated in time series data.

Assumption VA is typically violated in the SUR model (example 4) and the multivariate regression model (example 5). A weaker version of homoskedasticity: $Var(u|X) = \Sigma_m \otimes I_n$ is often suitable for these two types of models. Under the weaker assumption, the G-M theorem does not apply. However, since estimating Σ_m is easy, a feasible generalized least squared estimator can be shown to be efficient.

5. The G-M theorem requires full rank of X. If there is multicollinearity in X, we may modify the model a little bit and still obtain BLUE estimator by OLS. Suppose rank(X)$< k$. Then, β_0 is not identified. However, β_0 is identified under restriction $R\beta = 0$, where R is a matrix that satisfy: $\text{rank}\left(\begin{pmatrix} X \\ R \end{pmatrix}\right) = k$. We say β_0 is identified under restriction $R\beta = 0$ when there is a unique β_0 that satisfies the restriction such that $y = X\beta_0 + u$ and $E(u|X) = 0$. With this matrix R, modify the original model as

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \beta + \begin{pmatrix} u \\ 0 \end{pmatrix}. \tag{11}$$

Then a BLUE estimator for the β_0 that satisfies $R\beta_0 = 0$ can be obtained by running OLS of $\begin{pmatrix} y \\ 0 \end{pmatrix}$ on $\begin{pmatrix} X \\ R \end{pmatrix}$.

6. The G-M theorem requires linearity (Assumption LN). If linearity is not satisfied, the OLS estimator is only a linear projection.

7. The G-M theorem is a finite sample property of the LS estimator. Most of modern econometrics relies on large sample analysis (asymptotics).

3 Least Squared Estimation With Linear Constraints

Sometimes prior information is available in the form of a linear constraint: $Q\beta = r$, where Q is a $q \times k$ dimensional known nonsingular matrix and r is a $q \times 1$ known vector. Using this information in estimation will give better estimators – estimators with smaller variance.

The constraint least squared estimator of β is

$$\begin{aligned}\tilde{\beta}_n &= \arg\min_{\beta \in R^k} ||y - X\beta||^2 \\ & \quad s.t.\ Q\beta = r.\end{aligned}$$

One could solve the constraint optimization problem directly. But there is a more clever way to get $\tilde{\beta}_n$.

Find a $k \times (k-q)$ matrix R such that $R'Q = 0$ and (Q, R) is nonsingular. Such a R can always be found. Let $A = (Q, R)'$. Then write

$$\begin{aligned}y &= X(A^{-1}A)\beta + u \\ &= Z\gamma + u,\end{aligned}$$

where $Z = XA^{-1} = (XQ(Q'Q)^{-1}, XR(R'R)^{-1})$ and $\gamma = (\gamma_1^{'}, \gamma_2^{'})^{'}$, $\gamma_1 = Q'\beta$ and $\gamma_2 = R'\beta$. Notice that the linear constraint restricts γ_1 to be r and imposes no restriction on γ_2.

Impose the constraint and rewrite the model further to

$$y - XQ(Q'Q)^{-1}r = XR(R'R)^{-1}\gamma_2 + u. \tag{12}$$

Least squared estimator for γ_2 is

$$\begin{aligned}\hat{\gamma}_2 &= ((R'R)^{-1}R'X'XR(R'R)^{-1})^{-1}(R'R)^{-1}R'X'y \\ &\quad -((R'R)^{-1}R'X'XR(R'R)^{-1})^{-1}(R'R)^{-1}R'X'XQ(Q'Q)^{-1}r \\ &= R'R(R'X'XR)^{-1}R'X'y - R'R(R'X'XR)^{-1}R'X'XQ(Q'Q)^{-1}r.\end{aligned}$$

And

$$\begin{aligned}\bar{\beta}_n &= A^{-1}(r', \hat{\gamma}_2')' \\ &= Q(Q'Q)^{-1}r + R(R'R)^{-1}\hat{\gamma}_2 \\ &= Q(Q'Q)^{-1}r + R(R'X'XR)^{-1}R'X'y - R(R'X'XR)^{-1}R'X'XQ(Q'Q)^{-1}r. \end{aligned} \tag{13}$$

This turns out to be the same as $\tilde{\beta}_n$ if $X'X$ is invertible (see Amemiya, 1985).

The estimator $\bar{\beta}_n$ is the best linear unbiased estimator of β_0 under Assumptions LN, CM and VA and the linear constraint $Q\beta = r$. This is because $\hat{\gamma}_2$ is the BLUE estimator for $\gamma_{2,0}$ in the model (12).

4 Testing in Linear Models

Testing is based on the asymptotic distribution of the LS estimators. In order to derive the asymptotic distributions, we need the law of large numbers (LLN), a central limit theorem (CLT) and a law that allows us to operate the asymptotic limits of different components of an estimator.

The weak law of large numbers (WLLN). If X_1,...,X_n,... are iid with finite mean EX, then $\bar{X}_n \to_p EX$ as $n \to \infty$.

The central limit theorem (CLT). If $X_1, ..., X_n$ are iid with $EX_iX_i' < \infty$, then $\sqrt{n}(\bar{X}_n - EX) \to_d Z \sim N(0, Var(X_1))$ as $n \to \infty$.

Generalized Slutsky Theorem. If $Y_n \to_p c$, $W_n \to_d W$, and $g(c, W)$ is continuous in both arguments with probability one, then $g(Y_n, W_n) \to_d g(c, W)$ as $n \to \infty$.

Now we can easily derive the asymptotic distribution of $\hat{\beta}_{OLS}$.

First we derive the consistency of $\hat{\beta}_{OLS}$. Observe that

$$\begin{aligned}
\hat{\beta}_{OLS} - \beta_0 &= (X'X)^{-1}X'u \\
&= \left(\frac{X'X}{n}\right)^{-1} \frac{X'u}{n} \\
&= \left(n^{-1}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(n^{-1}\sum_{i=1}^{n} x_i u_i\right), \qquad (14)
\end{aligned}$$

where $X = \begin{pmatrix} x_1' \\ ... \\ x_n' \end{pmatrix}$.

As we can see, if $Ex_ix_i' < \infty$, we have $n^{-1}\sum_{i=1}^{n} x_i x_i' \to_p Ex_ix_i'$ by WLLN. If we have $Ex_iu_i = 0$, then $n^{-1}\sum_{i=1}^{n} x_i u_i \to_p 0$ by WLLN. If in addition Ex_ix_i' is invertible, by the generalized slutsky theorem, we immediately have: $\hat{\beta}_{OLS}$ is consistent (i.e. $\hat{\beta}_{OLS} - \beta_0 \to_p 0$) under the following two assumptions:

Assumption FR2. $Ex_ix_i' < \infty$ is invertible.

Assumption M0. $Ex_iu_i = 0$.

Assumption FR2 is the large sample version of the Ass. FR. Assumption M0 is slightly weaker than Assumption CM.

In order to derive the asymptotic distribution of $\hat{\beta}_{OLS}$, observe that

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta_0\right) = \left(n^{-1}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(n^{-1/2}\sum_{i=1}^{n} x_i u_i\right). \qquad (15)$$

If $Eu_i^2 x_i x_i' = E(\sigma^2(x_i)x_i x_i')$, then by CLT, $n^{-1/2}\sum_{i=1}^n x_i u_i \to_d N(0, E(\sigma^2(x_i)x_i x_i'))$. We can then apply the generalized slutsky theorem again and conclude:

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta_0\right) \to_d N(0, (Ex_i x_i)^{-1}E(\sigma^2(x_i)x_i x_i')(Ex_i x_i)^{-1}). \tag{16}$$

This result holds under Assumptions FR2, M0 and the additional assumption below:

Assumption VA2. $Eu_i^2 x_i x_i' = E(\sigma^2(x_i)x_i x_i')$ for some function $\sigma^2(x_i) > 0$.

Usually we assume homoskedasticity: $\sigma^2(x_i) = \sigma^2$. Under homoskedasticity,

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta_0\right) \to_d N(0, \sigma^2(Ex_i x_i)^{-1}). \tag{17}$$

Now suppose that we want to test the hypothesis: $H_0 : a'\beta = a'\beta_0$ for some $a \in R^k$ (e.g. $a = (1, 0, ..., 0)'$). A **t test** can be formulated using the above asymptotic distribution result:

$$T_n = \frac{\sqrt{n}(a'\hat{\beta}_{OLS} - a'\beta_0)}{\hat{\sigma}_n a'\left(n^{-1}\sum_{i=1}^n x_i x_i'\right)^{-1} a} \to_d N(0, 1), \tag{18}$$

where $\hat{\sigma}_n$ is a consistent estimator of σ.

The **t-test** rejects H_0 if $|T_n| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

We may also test **multiple hypotheses** at the same time: $H_0 : R\beta = r$ where R is a $m \times k$ matrix and r is a m-vector. In this case, a **Wald test** may be formulated also using the asymptotic distribution result above:

$$\begin{aligned} W_n &= n\hat{\sigma}_n^{-2}\left(R\hat{\beta}_{OLS} - r\right)'\left[R\left(n^{-1}\sum_{i=1}^n x_i x_i'\right)^{-1} R'\right]^{-1}\left(R\hat{\beta}_{OLS} - r\right) \\ &\to_d \chi_m^2. \end{aligned} \tag{19}$$

The Wald test rejects H_0 if $W_n > \chi_m^2(1-\alpha)$, where $\chi_m^2(1-\alpha)$ is the $1-\alpha$ quantile of the chi-squared distribution with m degrees of freedom.

The Wald test is the same as the F-test when $r = 0$. (Homework question: prove this.)