The Multiple Regression Model

The **multiple regression model** extends the single variable regression model of Chapter 4 to include additional variables as regressors. This model permits estimating the effect on Y_i of changing one variable (X_{1i}) while holding the other regressors (X_{2i}, X_{3i}, and so forth) constant.

The Multiple Regression Model

Key Concept 5.2

The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \; i = 1, \ldots, n. \qquad (5.7)$$

where:

- Y_i is i^{th} observation on the dependent variable; $X_{1i}, X_{2i}, \ldots, X_{ki}$ are the i^{th} observations on each of the k regressors; and u_i is the error term.
- The population regression line is the relationship that holds between Y and the X's on average in the population:

$$E(Y|X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

- β_1 is the slope coefficient on X_1, β_2 is the coefficient on X_2, etc. The coefficient β_1 is the expected change in Y_i resulting from changing X_{1i} by one unit, holding constant $X_{2i}, \ldots, X_{ki}$. The coefficients on the other X's are interpreted similarly.
- The intercept β_0 is the expected value of Y when all the X's equal zero. The intercept can be thought of as the coefficient on a regressor, X_{0i}, that equals one for all i.

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

Key Concept 5.3

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction mistakes $\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}, \; i = 1, \ldots, n, \text{ and} \qquad (5.11)$$

$$\hat{u}_i = Y_i - \hat{Y}_i, \; i = 1, \ldots, n. \qquad (5.12)$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term, u_i.

The Least Squares Assumptions in the Multiple Regression Model

Key Concept 5.4

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \; i = 1, \ldots, n, \text{ where:}$$

1. u_i has conditional mean zero given $X_{1i}, X_{2i}, \ldots, X_{ki}$, that is, $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$;
2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ are independently and identically distributed (i.i.d.) draws from their joint distribution;
3. $(X_{1i}, X_{2i}, \ldots, X_{ki}, u_i)$ have nonzero finite fourth moments; and
4. there is no perfect multicollinearity.

R^2 and $\overline{R}^2$: What They Tell You—and What They Don't

Key Concept 5.8

The R^2 and $\overline{R}^2$ tell you whether the regressors are good at predicting, or "explaining," the values of the dependent variable in the sample of data on hand. If the R^2 (or $\overline{R}^2$) is nearly one, then the regressors produce good predictions of the dependent variable in that sample, in the sense that the variance of the OLS residual is small compared to the variance of the dependent variable. If the R^2 (or $\overline{R}^2$) is nearly zero, the opposite is true.

The R^2 and $\overline{R}^2$ do NOT tell you whether:

1. an included variable is statistically significant;
2. the regressors are a true cause of the movements in the dependent variable;
3. there is omitted variable bias; or
4. you have chosen the most appropriate set of regressors.

Omitted Variable Bias in Multiple Regression

Key Concept 5.9

Omitted variable bias is the bias in the OLS estimator that arises when one or more included regressors are correlated with an omitted variable. For omitted variable bias to arise, two things must be true:

1. at least one of the included regressors must be correlated with the omitted variable; and
2. the omitted variable must be a determinant of the dependent variable, Y.

Instrumental Variables Regression

Instrumental variables (IV) regression is a general way to obtain a consistent estimator of the unknown coefficients of the population regression function when the regressor, X, is correlated with the error term, u. To understand how IV regression works, think of the variation in X as having two parts: one part that, for whatever reason, is correlated with u (this is the part that causes the problems), and a second part that is uncorrelated with u. If you had information that allowed you to isolate the second part, then you could focus on those variations in X that are uncorrelated with u and disregard the variations in X that bias the OLS estimates. This is, in fact, what IV regression does. The information about the movements in X that are uncorrelated with u is gleaned from one or more additional variables, called **instrumental variables** or simply **instruments**. Instrumental variables regression uses these additional variables as tools or "instruments" to isolate the movements in X that are uncorrelated with u, which in turn permit consistent estimation of the regression coefficients.

Key Concept 10.1

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \quad (10.12)$$

$i = 1, \ldots, n$, where:

- Y_i is the dependent variable;
- u_i is the error term, which represents measurement error and/or omitted factors;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown regression coefficients;
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$); they are underidentified if $m < k$; and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

Key Concept 10.2

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (10.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s):** Regress X_{1i} on the instrumental variables ($Z_{1i}, \ldots, Z_{mi}$) and the included exogenous variables ($W_{1i}, \ldots, W_{ri}$) using OLS. Compute the predicted values from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.
2. **Second-stage regression:** Regress Y_i on the predicted values of the endogenous variables ($\hat{X}_{1i}, \ldots, \hat{X}_{ki}$) and the included exogenous variables ($W_{1i}, \ldots, W_{ri}$) using OLS. The TSLS estimators $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in modern econometric software.

Key Concept 10.3

The Two Conditions for Valid Instruments

A set of m instruments $Z_{1i}, \ldots, Z_{mi}$ must satisfy the following two conditions to be valid:

1. Instrument Relevance

- *In general*, let $\hat{X}_{1i}^*$ be the predicted value of X_{1i} from the population regression of X_{1i} on the instruments (Z's) and the included exogenous regressors (W's), and let "1" denote a regressor that takes on the value "1" for all observations (its coefficient is the intercept). Then $(\hat{X}_{1i}^*, \ldots, \hat{X}_{ki}^*, W_{1i}, \ldots, W_{ri}, 1)$ are not perfectly multicollinear.
- *If there is only one* X, then at least one Z must enter the population regression of X on the Z's and the W's.

2. Instrument Exogeneity

The instruments are uncorrelated with the error term, that is, $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$.