Introduction to Econometrics

THIRD EDITION UPDATE

James H. Stock
Harvard University

Mark W. Watson
Princeton University

PEARSON

18 The Theory of Multiple Regression

This chapter provides an introduction to the theory of multiple regression analysis. The chapter has four objectives. The first is to present the multiple regression model in matrix form, which leads to compact formulas for the OLS estimator and test statistics. The second objective is to characterize the sampling distribution of the OLS estimator, both in large samples (using asymptotic theory) and in small samples (if the errors are homoskedastic and normally distributed). The third objective is to study the theory of efficient estimation of the coefficients of the multiple regression model and to describe generalized least squares (GLS), a method for estimating the regression coefficients efficiently when the errors are heteroskedastic and/or correlated across observations. The fourth objective is to provide a concise treatment of the asymptotic distribution theory of instrumental variables (IV) regression in the linear model, including an introduction to generalized method of moments (GMM) estimation in the linear IV regression model with heteroskedastic errors.

The chapter begins by laying out the multiple regression model and the OLS estimator in matrix form in Section 18.1. This section also presents the extended least squares assumptions for the multiple regression model. The first four of these assumptions are the same as the least squares assumptions of Key Concept 6.4 and underlie the asymptotic distributions used to justify the procedures described in Chapters 6 and 7. The remaining two extended least squares assumptions are stronger and permit us to explore in more detail the theoretical properties of the OLS estimator in the multiple regression model.

The next three sections examine the sampling distribution of the OLS estimator and test statistics. Section 18.2 presents the asymptotic distributions of the OLS estimator and *t*-statistic under the least squares assumptions of Key Concept 6.4. Section 18.3 unifies and generalizes the tests of hypotheses involving multiple coefficients presented in Sections 7.2 and 7.3, and provides the asymptotic distribution of the resulting *F*-statistic. In Section 18.4, we examine the exact sampling distributions of the OLS estimator and test statistics in the special case that the errors are homoskedastic and normally distributed. Although the assumption of homoskedastic normal errors is implausible in most econometric applications, the exact sampling distributions are of theoretical interest, and *p*-values computed using these distributions often appear in the output of regression software.

The next two sections turn to the theory of efficient estimation of the coefficients of the multiple regression model. Section 18.5 generalizes the Gauss–Markov theorem to multiple regression. Section 18.6 develops the method of generalized least squares (GLS).

The final section takes up IV estimation in the general IV regression model when the instruments are valid and strong. This section derives the asymptotic distribution of the TSLS estimator when the errors are heteroskedastic and provides expressions for the standard error of the TSLS estimator. The TSLS estimator is one of many possible GMM estimators, and this section provides an introduction to GMM estimation in the linear IV regression model. It is shown that the TSLS estimator is the efficient GMM estimator if the errors are homoskedastic.

Mathematical prerequisite. The treatment of the linear model in this chapter uses matrix notation and the basic tools of linear algebra and assumes that the reader has taken an introductory course in linear algebra. Appendix 18.1 reviews vectors, matrices, and the matrix operations used in this chapter. In addition, multivariate calculus is used in Section 18.1 to derive the OLS estimator.

18.1 The Linear Multiple Regression Model and OLS Estimator in Matrix Form

The linear multiple regression model and the OLS estimator can each be represented compactly using matrix notation.

The Multiple Regression Model in Matrix Notation

The population multiple regression model (Key Concept 6.2) is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \ldots, n. \tag{18.1}$$

To write the multiple regression model in matrix form, define the following vectors and matrices:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}, \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \tag{18.2}$$

so $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times (k+1)$, $\boldsymbol{U}$ is $n \times 1$, and $\boldsymbol{\beta}$ is $(k+1) \times 1$. Throughout we denote matrices and vectors by bold type. In this notation,

- $\boldsymbol{Y}$ is the $n \times 1$ dimensional vector of n observations on the dependent variable.
- $\boldsymbol{X}$ is the $n \times (k+1)$ dimensional matrix of n observations on the $k+1$ regressors (including the "constant" regressor for the intercept).
- The $(k+1) \times 1$ dimensional column vector $\boldsymbol{X}_i$ is the i^{th} observation on the $k+1$ regressors; that is, $\boldsymbol{X}_i' = (1 \; X_{1i} \ldots X_{ki})$, where $\boldsymbol{X}_i'$ denotes the transpose of $\boldsymbol{X}_i$.
- $\boldsymbol{U}$ is the $n \times 1$ dimensional vector of the n error terms.
- $\boldsymbol{\beta}$ is the $(k+1) \times 1$ dimensional vector of the $k+1$ unknown regression coefficients.

The multiple regression model in Equation (18.1) for the i^{th} observation, written using the vectors $\boldsymbol{\beta}$ and $\boldsymbol{X}_i$, is

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, i = 1, \ldots, n. \tag{18.3}$$

The Extended Least Squares Assumptions in the Multiple Regression Model

KEY CONCEPT

18.1

The linear regression model with multiple regressors is

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, i = 1, \ldots, n. \tag{18.4}$$

The extended least squares assumptions are

1. $E(u_i|\boldsymbol{X}_i) = 0$ (u_i has conditional mean zero);
2. $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.) draws from their joint distribution;
3. $\boldsymbol{X}_i$ and u_i have nonzero finite fourth moments;
4. $\boldsymbol{X}$ has full column rank (there is no perfect multicollinearity);
5. $\text{var}(u_i|\boldsymbol{X}_i) = \sigma_u^2$ (homoskedasticity); and
6. The conditional distribution of u_i given $\boldsymbol{X}_i$ is normal (normal errors).

In Equation (18.3), the first regressor is the "constant" regressor that always equals 1, and its coefficient is the intercept. Thus the intercept does not appear separately in Equation (18.3); rather, it is the first element of the coefficient vector $\boldsymbol{\beta}$.

Stacking all n observations in Equation (18.3) yields the multiple regression model in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}. \tag{18.5}$$

The Extended Least Squares Assumptions

The extended least squares assumptions for the multiple regressor model are the four least squares assumptions for the multiple regression model in Key Concept 6.4, plus the two additional assumptions of homoskedasticity and normally distributed errors. The assumption of homoskedasticity is used when we study the efficiency of the OLS estimator, and the assumption of normality is used when we study the exact sampling distribution of the OLS estimator and test statistics.

The extended least squares assumptions are summarized in Key Concept 18.1.

Except for notational differences, the first three assumptions in Key Concept 18.1 are identical to the first three assumptions in Key Concept 6.4.

The fourth assumption in Key Concepts 6.4 and 18.1 might appear different, but in fact they are the same: They are simply different ways of saying that there cannot be perfect multicollinearity. Recall that perfect multicollinearity arises when one regressor can be written as a perfect linear combination of the others. In the matrix notation of Equation (18.2), perfect multicollinearity means that one column of $\boldsymbol{X}$ is a perfect linear combination of the other columns of $\boldsymbol{X}$, but if this is true, then $\boldsymbol{X}$ does not have full column rank. Thus saying that $\boldsymbol{X}$ has rank $k + 1$, that is, rank equal to the number of columns of $\boldsymbol{X}$, is just another way to say that the regressors are not perfectly multicollinear.

The fifth least squares assumption in Key Concept 18.1 is that the error term is conditionally homoskedastic, and the sixth assumption is that the conditional distribution of u_i, given $\boldsymbol{X}_i$, is normal. These two assumptions are the same as the final two assumptions in Key Concept 17.1, except that they are now stated for multiple regressors.

Implications for the mean vector and covariance matrix of U. The least squares assumptions in Key Concept 18.1 imply simple expressions for the mean vector and covariance matrix of the conditional distribution of $\boldsymbol{U}$ given the matrix of regressors $\boldsymbol{X}$. (The mean vector and covariance matrix of a vector of random variables are defined in Appendix 18.2.) Specifically, the first and second assumptions in Key Concept 18.1 imply that $E(u_i|\boldsymbol{X}) = E(u_i|\boldsymbol{X}_i) = 0$ and that $\text{cov}(u_i, u_j|\boldsymbol{X}) = \boldsymbol{E(u_iu_j|X)} = E(u_iu_j|\boldsymbol{X}_i, \boldsymbol{X}_j) = E(u_i|\boldsymbol{X}_i)E(u_j|\boldsymbol{X}_j) = 0$ for $i \neq j$ (Exercise 17.7). The first, second, and fifth assumptions imply that $E(u_i^2|\boldsymbol{X}) = E(u_i^2|\boldsymbol{X}_i) = \sigma_u^2$. Combining these results, we have that

$$\text{under Assumptions \#1 and \#2, } E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n\text{, and} \tag{18.6}$$

$$\text{under Assumptions \#1, \#2, and \#5, } E(\boldsymbol{UU}'|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n, \tag{18.7}$$

where $\boldsymbol{0}_n$ is the n-dimensional vector of zeros and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Similarly, the first, second, fifth, and sixth assumptions in Key Concept 18.1 imply that the conditional distribution of the n-dimensional random vector $\boldsymbol{U}$, conditional on $\boldsymbol{X}$, is the multivariate normal distribution (defined in Appendix 18.2). That is,

$$\begin{array}{l}\text{under Assumptions \#1, \#2, \#5, and \#6, the} \\ \text{conditional distribution of } \boldsymbol{U} \text{ given } \boldsymbol{X} \text{ is } N(\boldsymbol{0}_n, \sigma_u^2\boldsymbol{I}_n).\end{array} \tag{18.8}$$

The OLS Estimator

The OLS estimator minimizes the sum of squared prediction mistakes, $\sum_{i=1}^n(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$ [Equation (6.8)]. The formula for the OLS estimator is obtained by taking the derivative of the sum of squared prediction mistakes with respect to each element of the coefficient vector, setting these derivatives to zero, and solving for the estimator $\hat{\boldsymbol{\beta}}$.

The derivative of the sum of squared prediction mistakes with respect to the j^{th} regression coefficient, b_j, is

$$\begin{aligned}&\frac{\partial}{\partial b_j}\sum_{i=1}^n(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2 \\ &= -2\sum_{i=1}^n X_{ji}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})\end{aligned} \tag{18.9}$$

for $j = 0, \ldots, k$, where, for $j = 0$, $X_{0i} = 1$ for all i. The derivative on the right-hand side of Equation (18.9) is the j^{th} element of the $k + 1$ dimensional vector, $-\boldsymbol{2X}'(\boldsymbol{Y} - \boldsymbol{Xb})$, where $\boldsymbol{b}$ is the $k + 1$ dimensional vector consisting of $b_0, \ldots, b_k$. There are $k + 1$ such derivatives, each corresponding to an element of $\boldsymbol{b}$. Combined, these yield the system of $k + 1$ equations that, when set to zero, constitute the first order conditions for the OLS estimator $\hat{\boldsymbol{\beta}}$. That is, $\hat{\boldsymbol{\beta}}$ solves the system of $k + 1$ equations

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}_{k+1}, \tag{18.10}$$

or, equivalently, $\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Solving the system of equations (18.10) yields the OLS estimator $\hat{\boldsymbol{\beta}}$ in matrix form:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{18.11}$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the inverse of the matrix $\boldsymbol{X}'\boldsymbol{X}$.

The role of "no perfect multicollinearity." The fourth least squares assumption in Key Concept 18.1 states that $\boldsymbol{X}$ has full column rank. In turn, this implies that the matrix $\boldsymbol{X}'\boldsymbol{X}$ has full rank, that is, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular. Because $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular, it is invertible. Thus the assumption that there is no perfect multicollinearity ensures that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, so Equation (18.10) has a unique solution and the formula in Equation (18.11) for the OLS estimator can actually be computed. Said differently, if $\boldsymbol{X}$ does *not* have full column rank, there is not a unique solution to Equation (18.10) and $\boldsymbol{X}'\boldsymbol{X}$ is singular. Therefore, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ cannot be computed and thus $\hat{\boldsymbol{\beta}}$ cannot be computed from Equation (18.11).

18.2 Asymptotic Distribution of the OLS Estimator and *t*-Statistic

If the sample size is large and the first four assumptions of Key Concept 18.1 are satisfied, then the OLS estimator has an asymptotic joint normal distribution, the heteroskedasticity-robust estimator of the covariance matrix is consistent, and the heteroskedasticity-robust OLS *t*-statistic has an asymptotic standard normal distribution. These results make use of the multivariate normal distribution (Appendix 18.2) and a multivariate extension of the central limit theorem.

The Multivariate Central Limit Theorem

The central limit theorem of Key Concept 2.7 applies to a one-dimensional random variable. To derive the *joint* asymptotic distribution of the elements of $\hat{\boldsymbol{\beta}}$, we need a multivariate central limit theorem that applies to vector-valued random variables.

The Multivariate Central Limit Theorem

KEY CONCEPT

18.2

Suppose that $\boldsymbol{W}_1, \dots, \boldsymbol{W}_n$ are i.i.d. m-dimensional random variables with mean vector $E(\boldsymbol{W}_i) = \boldsymbol{\mu}_{\boldsymbol{W}}$ and covariance matrix $E[(\boldsymbol{W}_i - \boldsymbol{\mu}_{\boldsymbol{W}})(\boldsymbol{W}_i - \boldsymbol{\mu}_{\boldsymbol{W}})'] = \boldsymbol{\Sigma}_{\boldsymbol{W}}$, where $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ is positive definite and finite. Let $\overline{\boldsymbol{W}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i$. Then $\sqrt{n}(\overline{\boldsymbol{W}} - \boldsymbol{\mu}_{\boldsymbol{W}}) \xrightarrow{d} N(\boldsymbol{0}_m, \boldsymbol{\Sigma}_{\boldsymbol{W}})$.

The multivariate central limit theorem extends the univariate central limit theorem to averages of observations on a vector-valued random variable, $\boldsymbol{W}$, where $\boldsymbol{W}$ is m-dimensional. The difference between the central limit theorems for a scalar as opposed to a vector-valued random variable is the conditions on the variances. In the scalar case in Key Concept 2.7, the requirement is that the variance is both nonzero and finite. In the vector case, the requirement is that the covariance matrix is both positive definite and finite. If the vector-valued random variable $\boldsymbol{W}$ has a finite positive definite covariance matrix, then $0 < \text{var}(\boldsymbol{c}'\boldsymbol{W}) < \infty$ for all nonzero m-dimensional vectors $\boldsymbol{c}$ (Exercise 18.3).

The multivariate central limit theorem that we will use is stated in Key Concept 18.2.

Asymptotic Normality of $\hat{\beta}$

In large samples, the OLS estimator has the multivariate normal asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}_{k+1}, \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}), \text{ where } \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} = \boldsymbol{Q}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}^{-1}, \quad (18.12)$$

where $\boldsymbol{Q}_{\boldsymbol{X}}$ is the $(k + 1) \times (k + 1)$-dimensional matrix of second moments of the regressors, that is, $\boldsymbol{Q}_{\boldsymbol{X}} = E(\boldsymbol{X}_i\boldsymbol{X}_i')$, and $\boldsymbol{\Sigma}_{\boldsymbol{V}}$ is the $(k + 1) \times (k + 1)$-dimensional covariance matrix of $\boldsymbol{V}_i = \boldsymbol{X}_i u_i$, that is, $\boldsymbol{\Sigma}_{\boldsymbol{V}} = E(\boldsymbol{V}_i\boldsymbol{V}_i')$. Note that the second least squares assumption in Key Concept 18.1 implies that $\boldsymbol{V}_i$, $i = 1, \dots, n$, are i.i.d.

Written in terms of $\hat{\boldsymbol{\beta}}$ rather than $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, the normal approximation in Equation (18.12) is

$$\hat{\boldsymbol{\beta}}, \text{ in large samples, is approximately distributed } N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})$$
$$\text{where } \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}/n = \boldsymbol{Q}_{\boldsymbol{X}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{Q}_{\boldsymbol{X}}^{-1}/n. \quad (18.13)$$

The covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ in Equation (18.13) is the covariance matrix of the approximate normal distribution of $\hat{\boldsymbol{\beta}}$, whereas $\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ in Equation (18.12) is the covariance matrix of the asymptotic normal distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. These two covariance matrices differ by a factor of n, depending on whether the OLS estimator is scaled by $\sqrt{n}$.

Derivation of Equation (18.12). To derive Equation (18.12), first use Equations (18.4) and (18.11) to write $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U})$ so that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}. \tag{18.14}$$

Thus $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$, so

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{U}}{\sqrt{n}}\right). \tag{18.15}$$

The derivation of Equation (18.12) involves arguing first that the “denominator” matrix in Equation (18.15), $\boldsymbol{X}'\boldsymbol{X}/n$, is consistent for $\boldsymbol{Q_X}$ and second that the “numerator” matrix, $\boldsymbol{X}'\boldsymbol{U}/\sqrt{n}$, obeys the multivariate central limit theorem in Key Concept 18.2. The details are given in Appendix 18.3.

Heteroskedasticity-Robust Standard Errors

The heteroskedasticity-robust estimator of $\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ is obtained by replacing the population moments in its definition [Equation (18.12)] by sample moments. Accordingly, the heteroskedasticity-robust estimator of the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is

$$\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{\hat{V}}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}, \text{ where } \hat{\boldsymbol{\Sigma}}_{\hat{V}} = \frac{1}{n-k-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\hat{u}_i^2, \tag{18.16}$$

The estimator $\hat{\boldsymbol{\Sigma}}_{\hat{V}}$ incorporates the same degrees-of-freedom adjustment that is in the *SER* for the multiple regression model (Section 6.4) to adjust for potential downward bias because of estimation of $k + 1$ regression coefficients.

The proof that $\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} \xrightarrow{p} \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ is conceptually similar to the proof, presented in Section 17.3, of the consistency of heteroskedasticity-robust standard errors for the single-regressor model.

Heteroskedasticity-robust standard errors. The heteroskedasticity-robust estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = n^{-1}\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}. \tag{18.17}$$

The heteroskedasticity-robust standard error for the j^{th} regression coefficient is the square root of the j^{th} diagonal element of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. That is, the heteroskedasticity-robust standard error of the j^{th} coefficient is

$$SE(\hat{\boldsymbol{\beta}}_j) = \sqrt{(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})_{jj}}, \tag{18.18}$$

where $(\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}})_{jj}$ is the (j, j) element of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$.

Confidence Intervals for Predicted Effects

Section 8.1 describes two methods for computing the standard error of predicted effects that involve changes in two or more regressors. There are compact matrix expressions for these standard errors and thus for confidence intervals for predicted effects.

Consider a change in the value of the regressors for the i^{th} observation from some initial value, say $\boldsymbol{X}_{i,0}$, to some new value, $\boldsymbol{X}_{i,0} + \boldsymbol{d}$, so that the change in $\boldsymbol{X}_i$ is $\Delta\boldsymbol{X}_i = \boldsymbol{d}$, where $\boldsymbol{d}$ is a $k + 1$ dimensional vector. This change in $\boldsymbol{X}$ can involve multiple regressors (that is, multiple elements of $\boldsymbol{X}_i$). For example, if two of the regressors are the value of an independent variable and its square, then $\boldsymbol{d}$ is the difference between the subsequent and initial values of these two variables.

The expected effect of this change in $\boldsymbol{X}_i$ is $\boldsymbol{d}'\boldsymbol{\beta}$, and the estimator of this effect is $\boldsymbol{d}'\hat{\boldsymbol{\beta}}$. Because linear combinations of normally distributed random variables are themselves normally distributed, $\sqrt{n}(\boldsymbol{d}'\hat{\boldsymbol{\beta}} - \boldsymbol{d}'\boldsymbol{\beta}) = \boldsymbol{d}'\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{d}'\Sigma_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\boldsymbol{d})$. Thus the standard error of this predicted effect is $(\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{d})^{1/2}$. A 95% confidence interval for this predicted effect is

$$\boldsymbol{d}'\hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{d}}. \tag{18.19}$$

Asymptotic Distribution of the *t*-Statistic

The *t*-statistic testing the null hypothesis that $\beta_j = \beta_{j,0}$, constructed using the heteroskedasticity-robust standard error in Equation (18.18), is given in Key Concept 7.1. The argument that this *t*-statistic has an asymptotic standard normal distribution parallels the argument given in Section 17.3 for the single-regressor model.

18.3 Tests of Joint Hypotheses

Section 7.2 considers tests of joint hypotheses that involve multiple restrictions, where each restriction involves a single coefficient, and Section 7.3 considers tests of a single restriction involving two or more coefficients. The matrix setup of

Section 18.1 permits a unified representation of these two types of hypotheses as linear restrictions on the coefficient vector, where each restriction can involve multiple coefficients. Under the first four least squares assumptions in Key Concept 18.1, the heteroskedasticity-robust OLS F-statistic testing these hypotheses has an $F_{q,\infty}$ asymptotic distribution under the null hypothesis.

Joint Hypotheses in Matrix Notation

Consider a joint hypothesis that is linear in the coefficients and imposes q restrictions, where $q \leq k + 1$. Each of these q restrictions can involve one or more of the regression coefficients. This joint null hypothesis can be written in matrix notation as

$$\boldsymbol{R\beta} = \boldsymbol{r}, \tag{18.20}$$

where $\boldsymbol{R}$ is a $q \times (k + 1)$ nonrandom matrix with full row rank and $\boldsymbol{r}$ is a nonrandom $q \times 1$ vector. The number of rows of $\boldsymbol{R}$ is q, which is the number of restrictions being imposed under the null hypothesis.

The null hypothesis in Equation (18.20) subsumes all the null hypotheses considered in Sections 7.2 and 7.3. For example, a joint hypothesis of the type considered in Section 7.2 is that $\beta_0 = 0, \beta_1 = 0, \ldots, \beta_{q-1} = 0$. To write this joint hypothesis in the form of Equation (18.20), set $\boldsymbol{R} = [\boldsymbol{I}_q \quad \boldsymbol{0}_{q\times(k+1-q)}]$ and $\boldsymbol{r} = \boldsymbol{0}_q$.

The formulation in Equation (18.20) also captures the restrictions of Section 7.3 involving multiple regression coefficients. For example, if $k = 2$, then the hypothesis that $\beta_1 + \beta_2 = 1$ can be written in the form of Equation (18.20) by setting $\boldsymbol{R} = [0 \quad 1 \quad 1]$, $\boldsymbol{r} = 1$, and $q = 1$.

Asymptotic Distribution of the *F*-Statistic

The heteroskedasticity-robust F-statistic testing the joint hypothesis in Equation (18.20) is

$$F = (\boldsymbol{R\hat{\beta}} - \boldsymbol{r})'[\boldsymbol{R\hat{\Sigma}_{\hat{\beta}}R'}]^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{r})/q. \tag{18.21}$$

If the first four assumptions in Key Concept 18.1 hold, then under the null hypothesis

$$F \xrightarrow{d} F_{q,\infty}. \tag{18.22}$$

This result follows by combining the asymptotic normality of $\hat{\boldsymbol{\beta}}$ with the consistency of the heteroskedasticity-robust estimator $\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$ of the covariance matrix. Specifically, first note that Equation (18.12) and Equation (18.74) in Appendix 18.2 imply that, under the null hypothesis, $\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) = \sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\boldsymbol{R}')$. It follows from Equation (18.77) that, under the null hypothesis, $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}\,(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) = [\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]'[\boldsymbol{R}\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\boldsymbol{R}']^{-1}[\sqrt{n}\boldsymbol{R}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \xrightarrow{d} \chi^2_q$. However, because $\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})} \xrightarrow{p} \boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}$, it follows from Slutsky's theorem that $[\sqrt{n}\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\boldsymbol{R}']^{-1}[\sqrt{n}\boldsymbol{R}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \xrightarrow{d} \chi^2_q$. or, equivalently (because $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}/n$), that $F \xrightarrow{d} \chi^2_q/q$, which is in turn distributed $F_{q,\infty}$.

Confidence Sets for Multiple Coefficients

As discussed in Section 7.4, an asymptotically valid confidence set for two or more elements of $\boldsymbol{\beta}$ can be constructed as the set of values that, when taken as the null hypothesis, are not rejected by the F-statistic. In principle, this set could be computed by repeatedly evaluating the F-statistic for many values of $\boldsymbol{\beta}$, but, as is the case with a confidence interval for a single coefficient, it is simpler to manipulate the formula for the test statistic to obtain an explicit formula for the confidence set.

Here is the procedure for constructing a confidence set for two or more of the elements of $\boldsymbol{\beta}$. Let $\boldsymbol{\delta}$ denote the q-dimensional vector consisting of the coefficients for which we wish to construct a confidence set. For example, if we are constructing a confidence set for the regression coefficients β_1 and β_2, then $q = 2$ and $\boldsymbol{\delta} = (\beta_1\,\beta_2)'$. In general, we can write $\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\beta}$, where the matrix $\boldsymbol{R}$ consists of zeros and ones [as discussed following Equation (18.20)]. The F-statistic testing the hypothesis that $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is $F = (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0)/q$, where $\hat{\boldsymbol{\delta}} = \boldsymbol{R}\hat{\boldsymbol{\beta}}$. A 95% confidence set for $\boldsymbol{\delta}$ is the set of values $\boldsymbol{\delta}_0$ that are not rejected by the F-statistic. That is, when $\boldsymbol{\delta} = \boldsymbol{R}\boldsymbol{\beta}$, a 95% confidence set for $\boldsymbol{\delta}$ is

$$\{\boldsymbol{\delta}\colon (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})'[\boldsymbol{R}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{R}']^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})/q \le c\}, \qquad (18.23)$$

where c is the 95th percentile (the 5% critical value) of the $F_{q,\infty}$ distribution.

The set in Equation (18.23) consists of all the points contained inside the ellipse determined when the inequality in Equation (18.23) is an equality (this is an ellipsoid when $q > 2$). Thus the confidence set for δ can be computed by solving Equation (18.23) for the boundary ellipse.

18.4 Distribution of Regression Statistics with Normal Errors

The distributions presented in Sections 18.2 and 18.3, which were justified by appealing to the law of large numbers and the central limit theorem, apply when the sample size is large. If, however, the errors are homoskedastic and normally distributed, conditional on $\boldsymbol{X}$, then the OLS estimator has a multivariate normal distribution in finite sample, conditional on $\boldsymbol{X}$. In addition, the finite sample distribution of the square of the standard error of the regression is proportional to the chi-squared distribution with $n - k - 1$ degrees of freedom, the homoskedasticity-only OLS t-statistic has a Student t distribution with $n - k - 1$ degrees of freedom, and the homoskedasticity-only F-statistic has an $F_{q,n-k-1}$ distribution. The arguments in this section employ some specialized matrix formulas for OLS regression statistics, which are presented first.

Matrix Representations of OLS Regression Statistics

The OLS predicted values, residuals, and sum of squared residuals have compact matrix representations. These representations make use of two matrices, $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$.

The matrices P_X and M_X. The algebra of OLS in the multivariate model relies on the two symmetric $n \times n$ matrices, $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$:

$$\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \text{ and} \tag{18.24}$$

$$\boldsymbol{M_X} = \boldsymbol{I_n} - \boldsymbol{P_X}. \tag{18.25}$$

A matrix $\boldsymbol{C}$ is idempotent if $\boldsymbol{C}$ is square and $\boldsymbol{CC} = \boldsymbol{C}$ (see Appendix 18.1). Because $\boldsymbol{P_X} = \boldsymbol{P_X P_X}$ and $\boldsymbol{M_X} = \boldsymbol{M_X M_X}$ (Exercise 18.5), and because $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric, $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric idempotent matrices.

The matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ have some additional useful properties (Exercise 18.5), which follow directly from the definitions in Equations (18.24) and (18.25):

$$\begin{gathered}\boldsymbol{P_X X} = \boldsymbol{X} \text{ and } \boldsymbol{M_X X} = \boldsymbol{0}_{n\times(k+1)};\\ \text{rank}(\boldsymbol{P_X}) = k + 1 \text{ and rank}(\boldsymbol{M_X}) = n - k - 1,\end{gathered} \tag{18.26}$$

where rank$(\boldsymbol{P_X})$ is the rank of $\boldsymbol{P_X}$.

The matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ can be used to decompose an n-dimensional vector $\boldsymbol{Z}$ into two parts: a part that is spanned by the columns of $\boldsymbol{X}$ and a part orthogonal to the columns of $\boldsymbol{X}$. In other words, $\boldsymbol{P_X Z}$ is the projection of $\boldsymbol{Z}$ onto the space spanned by the columns of $\boldsymbol{X}$, $\boldsymbol{M_X Z}$ is the part of $\boldsymbol{Z}$ orthogonal to the columns of $\boldsymbol{X}$, and $\boldsymbol{Z} = \boldsymbol{P_X Z} + \boldsymbol{M_X Z}$.

OLS predicted values and residuals. The matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ provide some simple expressions for OLS predicted values and residuals. The OLS predicted values, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and the OLS residuals, $\hat{\boldsymbol{U}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, can be expressed as follows (Exercise 18.5):

$$\hat{\boldsymbol{Y}} = \boldsymbol{P_X Y} \text{ and} \tag{18.27}$$

$$\hat{\boldsymbol{U}} = \boldsymbol{M_X Y} = \boldsymbol{M_X U}. \tag{18.28}$$

The expressions in Equations (18.27) and (18.28) provide a simple proof that the OLS residuals and predicted values are orthogonal, that is, Equation (4.37) holds: $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{U}} = \boldsymbol{Y}'\boldsymbol{P}_{\boldsymbol{X}}'\boldsymbol{M_X Y} = 0$, where the second equality follows from $\boldsymbol{P}_{\boldsymbol{X}}'\boldsymbol{M_X} = \boldsymbol{0}_{n\times n}$, which in turn follows from $\boldsymbol{M_X X} = \boldsymbol{0}_{n\times(k+1)}$ in Equation (18.26).

The standard error of the regression. The SER, defined in Section 4.3, is $s_{\hat{u}}$, where

$$s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{1}{n-k-1}\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}} = \frac{1}{n-k-1}\boldsymbol{U}'\boldsymbol{M_X U}, \tag{18.29}$$

where the final equality follows because $\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}} = (\boldsymbol{M_X U})'(\boldsymbol{M_X U}) = \boldsymbol{U}'\boldsymbol{M_X M_X U} = \boldsymbol{U}'\boldsymbol{M_X U}$ (because $\boldsymbol{M_X}$ is symmetric and idempotent).

Distribution of $\hat{\beta}$ with Normal Errors

Because $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ [Equation (18.14)] and because the distribution of $\boldsymbol{U}$ conditional on $\boldsymbol{X}$ is, by assumption, $N(\boldsymbol{0}_n, \sigma_u^2\boldsymbol{I}_n)$ [Equation (18.8)], the conditional distribution of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{X}$ is multivariate normal with mean $\boldsymbol{\beta}$. The covariance matrix of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, is $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}} = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'|\boldsymbol{X}] = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_u^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Accordingly, under all six assumptions in Key Concept 18.1, the finite-sample conditional distribution of $\hat{\boldsymbol{\beta}}$ given $\boldsymbol{X}$ is

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}}), \text{ where } \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}} = \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{18.30}$$

Distribution of $s_{\hat{u}}^2$

If all six assumptions in Key Concept 18.1 hold, then $s_{\hat{u}}^2$ has an exact sampling distribution that is proportional to a chi-squared distribution with $n - k - 1$ degrees of freedom:

$$s_{\hat{u}}^2 \sim \frac{\sigma_u^2}{n-k-1} \times \chi_{n-k-1}^2 \tag{18.31}$$

The proof of Equation (18.31) starts with Equation (18.29). Because $\boldsymbol{U}$ is normally distributed conditional on $\boldsymbol{X}$ and because $\boldsymbol{M_X}$ is a symmetric idempotent matrix, the quadratic form $\boldsymbol{U}'\boldsymbol{M_X}\boldsymbol{U}/\sigma_u^2$ has an exact chi-squared distribution with degrees of freedom equal to the rank of $\boldsymbol{M_X}$ [Equation (18.78) in Appendix 18.2]. From Equation (18.26), the rank of $\boldsymbol{M_X}$ is $n - k - 1$. Thus $\boldsymbol{U}'\boldsymbol{M_X}\boldsymbol{U}/\sigma_u^2$ has an exact χ_{n-k-1}^2 distribution, from which Equation (18.31) follows.

The degrees-of-freedom adjustment ensures that $s_{\hat{u}}^2$ is unbiased. The expectation of a random variable with a χ_{n-k-1}^2 distribution is $n - k - 1$; thus $E(\boldsymbol{U}'\boldsymbol{M_X}\boldsymbol{U}) = (n-k-1)\sigma_u^2$, so $E(s_{\hat{u}}^2) = \sigma_u^2$.

Homoskedasticity-Only Standard Errors

The homoskedasticity-only estimator $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ of the covariance matrix of $\hat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$, is obtained by substituting the sample variance $s_{\hat{u}}^2$ for the population variance σ_u^2 in the expression for $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}}$ in Equation (18.30). Accordingly,

$$\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = s_{\hat{u}}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad \text{(homoskedasticity-only).} \tag{18.32}$$

The estimator of the variance of the normal conditional distribution of $\hat{\beta}_j$, given $\boldsymbol{X}$, is the (j, j) element of $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. Thus the homoskedasticity-only standard error of $\hat{\beta}_j$ is the square root of the j^{th} diagonal element of $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. That is, the homoskedasticity-only standard error of $\hat{\beta}_j$ is

$$\widetilde{SE}(\hat{\beta}_j) = \sqrt{(\widetilde{\boldsymbol{\Sigma}}_{\hat{\beta}})_{jj}} \quad \text{(homoskedasticity-only).} \tag{18.33}$$

Distribution of the *t*-Statistic

Let $\widetilde{t}$ be the *t*-statistic testing the hypothesis $\beta_j = \beta_{j,0}$, constructed using the homoskedasticity-only standard error; that is, let

$$\widetilde{t} = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{(\widetilde{\boldsymbol{\Sigma}}_{\hat{\beta}})_{jj}}}. \tag{18.34}$$

Under all six of the extended least squares assumptions in Key Concept 18.1, the exact sampling distribution of $\widetilde{t}$ is the Student *t* distribution with $n - k - 1$ degrees of freedom; that is,

$$\widetilde{t} \sim t_{n-k-1}. \tag{18.35}$$

The proof of Equation (18.35) is given in Appendix 18.4.

Distribution of the *F*-Statistic

If all six least squares assumptions in Key Concept 18.1 hold, then the *F*-statistic testing the hypothesis in Equation (18.20), constructed using the homoskedasticity-only estimator of the covariance matrix, has an exact $F_{q,\,n-k-1}$ distribution under the null hypothesis.

The homoskedasticity-only F-statistic. The homoskedasticity-only *F*-statistic is similar to the heteroskedasticity-robust *F*-statistic in Equation (18.21), except that the homoskedasticity-only estimator $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ is used instead of the heteroskedasticity-robust estimator $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$. Substituting the expression $\widetilde{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = s_{\hat{u}}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ into the expression for the *F*-statistic in Equation (18.21) yields the homoskedasticity-only *F*-statistic testing the null hypothesis in Equation (18.20):

$$\widetilde{F} = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{s_{\hat{u}}^2}. \tag{18.36}$$

If all six assumptions in Key Concept 18.1 hold, then under the null hypothesis

$$\widetilde{F} \sim F_{q,n-k-1}. \tag{18.37}$$

The proof of Equation (18.37) is given in Appendix 18.4.

The F-statistic in Equation (18.36) is called the Wald version of the F-statistic (named after the statistician Abraham Wald). Although the formula for the homoskedastic-only F-statistic given in Equation (7.13) appears quite different from the formula for the Wald statistic in Equation (18.36), the homoskedastic-only F-statistic and the Wald F-statistic are two versions of the same statistic. That is, the two expressions are equivalent, a result shown in Exercise 18.13.

18.5 Efficiency of the OLS Estimator with Homoskedastic Errors

Under the Gauss–Markov conditions for multiple regression, the OLS estimator of $\boldsymbol{\beta}$ is efficient among all linear conditionally unbiased estimators; that is, the OLS estimator is BLUE.

The Gauss–Markov Conditions for Multiple Regression

The **Gauss–Markov conditions for multiple regression** are

$$\begin{aligned}&\text{(i) } E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n,\\ &\text{(ii) } E(\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{X}) = \sigma_u^2\boldsymbol{I}_n, \text{ and}\\ &\text{(iii) } \boldsymbol{X} \text{ has full column rank.}\end{aligned} \tag{18.38}$$

The Gauss–Markov conditions for multiple regression in turn are implied by the first five assumptions in Key Concept 18.1 [see Equations (18.6) and (18.7)]. The conditions in Equation (18.38) generalize the Gauss–Markov conditions for a single regressor model to multiple regression. [By using matrix notation, the second and third Gauss–Markov conditions in Equation (5.31) are collected into the single condition (ii) in Equation (18.38).]

Linear Conditionally Unbiased Estimators

We start by describing the class of linear unbiased estimators and by showing that OLS is in that class.

The class of linear conditionally unbiased estimators. An estimator of $\boldsymbol{\beta}$ is said to be linear if it is a linear function of $Y_1, \ldots, Y_n$. Accordingly, the estimator $\widetilde{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$ if it can be written in the form

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{Y}, \tag{18.39}$$

where $\boldsymbol{A}$ is an $n \times (k+1)$ dimensional matrix of weights that may depend on $\boldsymbol{X}$ and on nonrandom constants, but not on $\boldsymbol{Y}$.

An estimator is conditionally unbiased if the mean of its conditional sampling distribution, given $\boldsymbol{X}$, is $\boldsymbol{\beta}$. That is, $\widetilde{\boldsymbol{\beta}}$ is conditionally unbiased if $E(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}$.

The OLS estimator is linear and conditionally unbiased. Comparison of Equations (18.11) and (18.39) shows that the OLS estimator is linear in $\boldsymbol{Y}$; specifically, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{A}}'\boldsymbol{Y}$, where $\hat{\boldsymbol{A}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. To show that $\hat{\boldsymbol{\beta}}$ is conditionally unbiased, recall from Equation (18.14) that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$. Taking the conditional expectation of both sides of this expression yields, $E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta} + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}|\boldsymbol{X}] = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{\beta}$, where the final equality follows because $E(\boldsymbol{U}|\boldsymbol{X}) = 0$ by the first Gauss–Markov condition.

The Gauss–Markov Theorem for Multiple Regression

The **Gauss–Markov theorem for multiple regression** provides conditions under which the OLS estimator is efficient among the class of linear conditionally unbiased estimators. A subtle point arises, however, because $\hat{\boldsymbol{\beta}}$ is a vector and its "variance" is a covariance matrix. When the "variance" of an estimator is a matrix, just what does it mean to say that one estimator has a smaller variance than another?

The Gauss–Markov theorem handles this problem by comparing the variance of a candidate estimator of a *linear combination* of the elements of $\boldsymbol{\beta}$ to the variance of the corresponding linear combination of $\hat{\boldsymbol{\beta}}$. Specifically, let $\boldsymbol{c}$ be a $k+1$ dimensional vector and consider the problem of estimating the linear combination $\boldsymbol{c}'\boldsymbol{\beta}$ using the candidate estimator $\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}$ (where $\widetilde{\boldsymbol{\beta}}$ is a linear conditionally unbiased estimator) on the one hand and $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ on the other hand. Because $\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ are both scalars and are both linear conditionally unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$, it now makes sense to compare their variances.

The Gauss–Markov theorem for multiple regression says that the OLS estimator of $\boldsymbol{c}'\boldsymbol{\beta}$ is efficient; that is, the OLS estimator $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ has the smallest conditional variance of all linear conditionally unbiased estimators $\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}$. Remarkably, this is true no matter what the linear combination is. It is in this sense that the OLS estimator is BLUE in multiple regression.

The Gauss–Markov theorem is stated in Key Concept 18.3 and proven in Appendix 18.5.

KEY CONCEPT 18.3

Gauss–Markov Theorem for Multiple Regression

Suppose that the Gauss–Markov conditions for multiple regression in Equation (18.38) hold. Then the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. That is, let $\widetilde{\boldsymbol{\beta}}$ be a linear conditionally unbiased estimator of $\boldsymbol{\beta}$ and let $\boldsymbol{c}$ be a nonrandom $k + 1$ dimensional vector. Then $\text{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}|\boldsymbol{X}) \leq \text{var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})$ for every nonzero vector $\boldsymbol{c}$, where the inequality holds with equality for all $\boldsymbol{c}$ only if $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

18.6 Generalized Least Squares[1]

The assumption of i.i.d. sampling fits many applications. For example, suppose that Y_i and $\boldsymbol{X}_i$ correspond to information about individuals, such as their earnings, education, and personal characteristics, where the individuals are selected from a population by simple random sampling. In this case, because of the simple random sampling scheme, $(\boldsymbol{X}_i, Y_i)$ are necessarily i.i.d. Because $(\boldsymbol{X}_i, Y_i)$ and $(\boldsymbol{X}_j, Y_j)$ are independently distributed for $i \neq j$, u_i and u_j are independently distributed for $i \neq j$. This in turn implies that u_i and u_j are uncorrelated for $i \neq j$. In the context of the Gauss–Markov assumptions, the assumption that $E(\boldsymbol{UU}'|\boldsymbol{X})$ is diagonal therefore is appropriate if the data are collected in a way that makes the observations independently distributed.

Some sampling schemes encountered in econometrics do not, however, result in independent observations and instead can lead to error terms u_i that are correlated from one observation to the next. The leading example is when the data are sampled over time for the same entity, that is, when the data are time series data. As discussed in Section 15.3, in regressions involving time series data, many omitted factors are correlated from one period to the next, and this can result in regression error terms (which represent those omitted factors) that are correlated from one period of observation to the next. In other words, the error term in one period will not, in general, be distributed independently of the error term in the

[1] The GLS estimator was introduced in Section 15.5 in the context of distributed lag time series regression. This presentation here is a self-contained mathematical treatment of GLS that can be read independently of Section 15.5, but reading that section first will help to make these ideas more concrete.

next period. Instead, the error term in one period could be correlated with the error term in the next period.

The presence of correlated error terms creates two problems for inference based on OLS. First, *neither* the heteroskedasticity-robust nor the homoskedasticity-only standard errors produced by OLS provide a valid basis for inference. The solution to this problem is to use standard errors that are robust to both heteroskedasticity and correlation of the error terms across observations. This topic—heteroskedasticity- and autocorrelation-consistent (HAC) covariance matrix estimation—is the subject of Section 15.4 and we do not pursue it further here.

Second, if the error term is correlated across observations, then $E(\boldsymbol{UU}'|\boldsymbol{X})$ is not diagonal, the second Gauss–Markov condition in Equation (18.38) does not hold, and OLS is not BLUE. In this section we study an estimator, **generalized least squares (GLS)**, that is BLUE (at least asymptotically) when the conditional covariance matrix of the errors is no longer proportional to the identity matrix. A special case of GLS is weighted least squares, discussed in Section 17.5, in which the conditional covariance matrix is diagonal and the i^{th} diagonal element is a function of $\boldsymbol{X}_i$. Like WLS, GLS transforms the regression model so that the errors of the transformed model satisfy the Gauss–Markov conditions. The GLS estimator is the OLS estimator of the coefficients in the transformed model.

The GLS Assumptions

There are four assumptions under which GLS is valid. The first GLS assumption is that u_i has a mean of zero, conditional on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$; that is,

$$E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n. \tag{18.40}$$

This assumption is implied by the first two least squares assumptions in Key Concept 18.1; that is, if $E(u_i|\boldsymbol{X}_i) = 0$ and $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$, are i.i.d., then $E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n$. In GLS, however, we will not want to maintain the i.i.d. assumption; after all, one purpose of GLS is to handle errors that are correlated across observations. We discuss the significance of the assumption in Equation (18.40) after introducing the GLS estimator.

The second GLS assumption is that the conditional covariance matrix of $\boldsymbol{U}$ given $\boldsymbol{X}$ is some function of $\boldsymbol{X}$:

$$E(\boldsymbol{UU}'|\boldsymbol{X}) = \boldsymbol{\Omega}(\boldsymbol{X}), \tag{18.41}$$

where $\boldsymbol{\Omega}(\boldsymbol{X})$ is an $n \times n$ positive definite matrix-valued function of $\boldsymbol{X}$.

KEY CONCEPT 18.4

The GLS Assumptions

In the linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$, the GLS assumptions are

1. $E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n$;
2. $E(\boldsymbol{UU}'|\boldsymbol{X}) = \boldsymbol{\Omega}(\boldsymbol{X})$, where $\boldsymbol{\Omega}(\boldsymbol{X})$ is an $n \times n$ positive definite matrix that can depend on $\boldsymbol{X}$;
3. $\boldsymbol{X}_i$ and u_i satisfy suitable moment conditions; and
4. $\boldsymbol{X}$ has full column rank (there is no perfect multicollinearity).

There are two main applications of GLS that are covered by this assumption. The first is independent sampling with heteroskedastic errors, in which case $\boldsymbol{\Omega}(\boldsymbol{X})$ is a diagonal matrix with diagonal element $\lambda h(\boldsymbol{X}_i)$, where λ is a constant and h is a function. In this case, discussed in Section 17.5, GLS is WLS.

The second application is to homoskedastic errors that are serially correlated. In practice, in this case a model is developed for the serial correlation. For example, one model is that the error term is correlated with only its neighbor, so $\text{corr}(u_i, u_{i-1}) = \rho \neq 0$ but $\text{corr}(u_i, u_j) = 0$ if $|i - j| \geq 2$. In this case, $\boldsymbol{\Omega}(\boldsymbol{X})$ has σ_u^2 as its diagonal element, $\rho\sigma_u^2$ in the first off-diagonal, and zeros elsewhere. Thus $\boldsymbol{\Omega}(\boldsymbol{X})$ does not depend on $\boldsymbol{X}$, $\boldsymbol{\Omega}_{ii} = \sigma_u^2$, $\boldsymbol{\Omega}_{ij} = \rho\sigma_u^2$ for $|i - j| = 1$, and $\boldsymbol{\Omega}_{ij} = 0$ for $|i - j| > 1$. Other models for serial correlation, including the first order autoregressive model, are discussed further in the context of GLS in Section 15.5 (also see Exercise 18.8).

One assumption that has appeared on all previous lists of least squares assumptions for cross-sectional data is that $\boldsymbol{X}_i$ and u_i have nonzero, finite fourth moments. In the case of GLS, the specific moment assumptions needed to prove asymptotic results depend on the nature of the function $\boldsymbol{\Omega}(\boldsymbol{X})$, whether $\boldsymbol{\Omega}(\boldsymbol{X})$ is known or estimated, and the statistic under consideration (the GLS estimator, t-statistic, etc.). Because the assumptions are case- and model-specific, we do not present specific moment assumptions here, and the discussion of the large-sample properties of GLS assumes that such moment conditions apply for the relevant case at hand. For completeness, as the third GLS assumption, $\boldsymbol{X}_i$ and u_i are simply assumed to satisfy suitable moment conditions.

The fourth GLS assumption is that $\boldsymbol{X}$ has full column rank; that is, the regressors are not perfectly multicollinear.

The GLS assumptions are summarized in Key Concept 18.4.

We consider GLS estimation in two cases. In the first case, $\boldsymbol{\Omega}(\boldsymbol{X})$ is known. In the second case, the functional form of $\boldsymbol{\Omega}(\boldsymbol{X})$ is known up to some parameters that can be estimated. To simplify notation, we refer to the function $\boldsymbol{\Omega}(\boldsymbol{X})$ as the matrix $\boldsymbol{\Omega}$, so the dependence of $\boldsymbol{\Omega}$ on $\boldsymbol{X}$ is implicit.

GLS When $\boldsymbol{\Omega}$ Is Known

When $\boldsymbol{\Omega}$ is known, the GLS estimator uses $\boldsymbol{\Omega}$ to transform the regression model to one with errors that satisfy the Gauss–Markov conditions. Specifically, let $\boldsymbol{F}$ be a matrix square root of $\boldsymbol{\Omega}^{-1}$; that is, let $\boldsymbol{F}$ be a matrix that satisfies $\boldsymbol{F'F} = \boldsymbol{\Omega}^{-1}$ (see Appendix 18.1). A property of $\boldsymbol{F}$ is that $\boldsymbol{F\Omega F'} = \boldsymbol{I}_n$. Now premultiply both sides of Equation (18.4) by $\boldsymbol{F}$ to obtain

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{U}}, \tag{18.42}$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{FY}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{FX}$, and $\widetilde{\boldsymbol{U}} = \boldsymbol{FU}$.

The key insight of GLS is that, under the four GLS assumptions, the Gauss–Markov assumptions hold for the transformed regression in Equation (18.42). That is, by transforming all the variables by the inverse of the matrix square root of $\boldsymbol{\Omega}$, the regression errors in the transformed regression have a conditional mean of zero and a covariance matrix that equals the identity matrix. To show this mathematically, first note that $E(\widetilde{\boldsymbol{U}}|\widetilde{\boldsymbol{X}}) = E(\boldsymbol{FU}|\boldsymbol{FX}) = \boldsymbol{F}E(\boldsymbol{U}|\boldsymbol{FX}) = \boldsymbol{0}_n$ by the first GLS assumption [Equation (18.40)]. In addition, $E(\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}'|\widetilde{\boldsymbol{X}}) = E[(\boldsymbol{FU})(\boldsymbol{FU})'|\boldsymbol{FX}] = \boldsymbol{F}E(\boldsymbol{UU}'|\boldsymbol{FX})\boldsymbol{F}' = \boldsymbol{F\Omega F}' = \boldsymbol{I}_n$, where the second equality follows because $(\boldsymbol{FU})' = \boldsymbol{U}'\boldsymbol{F}'$ and the final equality follows from the definition of $\boldsymbol{F}$. It follows that the transformed regression model in Equation (18.42) satisfies the Gauss–Markov conditions in Key Concept 18.3.

The GLS estimator, $\widetilde{\boldsymbol{\beta}}^{GLS}$, is the OLS estimator of $\boldsymbol{\beta}$ in Equation (18.42); that is, $\widetilde{\boldsymbol{\beta}}^{GLS} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}})$. Because the transformed regression model satisfies the Gauss–Markov conditions, the GLS estimator is the best conditionally unbiased estimator that is linear in $\widetilde{\boldsymbol{Y}}$. But because $\widetilde{\boldsymbol{Y}} = \boldsymbol{FY}$ and $\boldsymbol{F}$ is (here) assumed to be known, and because $\boldsymbol{F}$ is invertible (because $\boldsymbol{\Omega}$ is positive definite), the class of estimators that are linear in $\widetilde{\boldsymbol{Y}}$ is the same as the class of estimators that are linear in $\boldsymbol{Y}$. Thus the OLS estimator of $\boldsymbol{\beta}$ in Equation (18.42) is also the best conditionally unbiased estimator among estimators that are linear in $\boldsymbol{Y}$. In other words, under the GLS assumptions, the GLS estimator is BLUE.

The GLS estimator can be expressed directly in terms of $\boldsymbol{\Omega}$, so in principle there is no need to compute the square root matrix $\boldsymbol{F}$. Because $\widetilde{\boldsymbol{X}} = \boldsymbol{FX}$ and $\widetilde{\boldsymbol{Y}} = \boldsymbol{FY}$, $\widetilde{\boldsymbol{\beta}}^{GLS} = (\boldsymbol{X}'\boldsymbol{F}'\boldsymbol{FX})^{-1}(\boldsymbol{X}'\boldsymbol{F}'\boldsymbol{FY})$. But $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{\Omega}^{-1}$, so

$$\widetilde{\boldsymbol{\beta}}^{GLS} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}). \tag{18.43}$$

In practice, $\boldsymbol{\Omega}$ is typically unknown, so the GLS estimator in Equation (18.43) typically cannot be computed and thus is sometimes called the **infeasible GLS** estimator. If, however, $\boldsymbol{\Omega}$ has a known functional form but the parameters of that function are unknown, then $\boldsymbol{\Omega}$ can be estimated and a feasible version of the GLS estimator can be computed.

GLS When $\boldsymbol{\Omega}$ Contains Unknown Parameters

If $\boldsymbol{\Omega}$ is a known function of some parameters that in turn can be estimated, then these estimated parameters can be used to calculate an estimator of the covariance matrix $\boldsymbol{\Omega}$. For example, consider the time series application discussed following Equation (18.41), in which $\boldsymbol{\Omega}(\boldsymbol{X})$ does not depend on $\boldsymbol{X}$, $\boldsymbol{\Omega}_{ii} = \sigma_u^2$, $\boldsymbol{\Omega}_{ij} = \rho\sigma_u^2$ for $|i - j| = 1$, and $\boldsymbol{\Omega}_{ij} = 0$ for $|i - j| > 1$. Then $\boldsymbol{\Omega}$ has two unknown parameters, σ_u^2 and ρ. These parameters can be estimated using the residuals from a preliminary OLS regression; specifically, σ_u^2 can be estimated by $s_{\hat{u}}^2$ and ρ can be estimated by the sample correlation between all neighboring pairs of OLS residuals. These estimated parameters can in turn be used to compute an estimator of $\boldsymbol{\Omega}$, $\hat{\boldsymbol{\Omega}}$.

In general, suppose that you have an estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. Then the GLS estimator based on $\hat{\boldsymbol{\Omega}}$ is

$$\hat{\boldsymbol{\beta}}^{GLS} = (\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{Y}). \tag{18.44}$$

The GLS estimator in Equation (18.44) is sometimes called the **feasible GLS** estimator because it can be computed if the covariance matrix contains some unknown parameters that can be estimated.

The Zero Conditional Mean Assumption and GLS

For the OLS estimator to be consistent, the first least squares assumption must hold; that is, $E(u_i|\boldsymbol{X}_i)$ must be zero. In contrast, the first GLS assumption is that $E(u_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = 0$. In other words, the first OLS assumption is that the error for the i^{th} observation has a conditional mean of zero, given the values of the regressors for that observation, whereas the first GLS assumption is that u_i has a conditional mean of zero, given the values of the regressors for *all* observations.

As discussed in Section 18.1, the assumptions that $E(u_i|\boldsymbol{X}_i) = 0$ and that sampling is i.i.d. together imply that $E(u_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = 0$. Thus, when sampling is i.i.d. so that GLS is WLS, the first GLS assumption is implied by the first least squares assumption in Key Concept 18.1.

When sampling is not i.i.d., however, the first GLS assumption is not implied by the assumption that $E(u_i|\boldsymbol{X}_i) = 0$; that is, the first GLS assumption is stronger. Although the distinction between these two conditions might seem slight, it can be very important in applications to time series data. This distinction is discussed in Section 15.5 in the context of whether the regressor is "past and present" exogenous or "strictly" exogenous; the assumption that $E(u_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = 0$ corresponds to strict exogeneity. Here, we discuss this distinction at a more general level using matrix notation. To do so, we focus on the case that $\boldsymbol{U}$ is homoskedastic, $\boldsymbol{\Omega}$ is known, and $\boldsymbol{\Omega}$ has nonzero off-diagonal elements.

The role of the first GLS assumption. To see the source of the difference between these assumptions, it is useful to contrast the consistency arguments for GLS and OLS.

We first sketch the argument for the consistency of the GLS estimator in Equation (18.43). Substituting Equation (18.4) into Equation (18.43), we have $\tilde{\boldsymbol{\beta}}^{GLS} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}/n)^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}/n)$. Under the first GLS assumption, $E(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}) = E[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}E(\boldsymbol{U}|\boldsymbol{X})] = \boldsymbol{0}_n$. If in addition the variance of $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}/n$ tends to zero and $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}/n \xrightarrow{p} \tilde{\boldsymbol{Q}}$, where $\tilde{\boldsymbol{Q}}$ is some invertible matrix, then $\tilde{\boldsymbol{\beta}}^{GLS} \xrightarrow{p} \boldsymbol{\beta}$. Critically, when $\boldsymbol{\Omega}$ has off-diagonal elements, the term $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U} = \sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{X}_i(\boldsymbol{\Omega}^{-1})_{ij}u_j$ involves products of $\boldsymbol{X}_i$ and u_j for different i, j, where $(\boldsymbol{\Omega}^{-1})_{ij}$ denotes the (i, j) element of $\boldsymbol{\Omega}^{-1}$. Thus, for $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}$ to have a mean of zero, it is not enough that $E(u_i|\boldsymbol{X}_i) = 0$; rather $E(u_i|\boldsymbol{X}_j)$ must equal zero for all i, j pairs corresponding to nonzero values of $(\boldsymbol{\Omega}^{-1})_{ij}$. Depending on the covariance structure of the errors, only some of or all the elements of $(\boldsymbol{\Omega}^{-1})_{ij}$ might be nonzero. For example, if u_i follows a first order autoregression (as discussed in Section 15.5), the only nonzero elements $(\boldsymbol{\Omega}^{-1})_{ij}$ are those for which $|i - j| \leq 1$. In general, however, all the elements of $\boldsymbol{\Omega}^{-1}$ can be nonzero, so in general for $\boldsymbol{X}'\Omega^{-1}\boldsymbol{U}/n \xrightarrow{p} \boldsymbol{0}_{(k+1)\times 1}$ (and thus for $\tilde{\boldsymbol{\beta}}^{GLS}$ to be consistent) we need that $E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_n$; that is, the first GLS assumption must hold.

In contrast, recall the argument that the OLS estimator is consistent. Rewrite Equation (18.14) as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X}/n)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i u_i$. If $E(u_i \,|\, \boldsymbol{X}_i) = 0$, then the term $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i u_i$ has mean zero, and if this term has a variance that tends to zero, it converges in probability to zero. If in addition $\boldsymbol{X}'\boldsymbol{X}/n \xrightarrow{p} \boldsymbol{Q}_X$, then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Is the first GLS assumption restrictive? The first GLS assumption requires that the errors for the i^{th} observation be uncorrelated with the regressors for all other observations. This assumption is dubious in some time series applications. This issue is discussed in Section 15.6 in the context of an empirical example, the relationship between the change in the price of a contract for future delivery of frozen orange concentrate and the weather in Florida. As explained there, the error term in the regression of price changes on the weather is plausibly uncorrelated with current and past values of the weather, so the first OLS assumption holds. However, this error term is plausibly correlated with future values of the weather, so the first GLS assumption does *not* hold.

This example illustrates a general phenomenon in economic time series data that arises when the value of a variable today is set in part based on expectations of the future: Those future expectations typically imply that the error term today depends on a forecast of the regressor tomorrow, which in turn is correlated with the actual value of the regressor tomorrow. For this reason, the first GLS assumption is in fact much stronger than the first OLS assumption. Accordingly, in some applications with economic time series data the GLS estimator is not consistent even though the OLS estimator is.

18.7 Instrumental Variables and Generalized Method of Moments Estimation

This section provides an introduction to the theory of instrumental variables (IV) estimation and the asymptotic distribution of IV estimators. It is assumed throughout that the IV regression assumptions in Key Concepts 12.3 and 12.4 hold and, moreover, that the instruments are strong. These assumptions apply to cross-sectional data with i.i.d. observations. Under certain conditions the results derived in this section are applicable to time series data as well, and the extension to time series data is briefly discussed at the end of this section. All asymptotic results in this section are developed under the assumption of strong instruments.

This section begins by presenting the IV regression model, the two stage least squares (TSLS) estimator, and its asymptotic distribution in the general case of heteroskedasticity, all in matrix form. It is next shown that, in the special case of homoskedasticity, the TSLS estimator is asymptotically efficient among the class of IV estimators in which the instruments are linear combinations of the exogenous variables. Moreover, the J-statistic has an asymptotic chi-squared distribution in which the degrees of freedom equal the number of overidentifying restrictions. This section concludes with a discussion of efficient IV estimation and the test of overidentifying restrictions when the errors are heteroskedastic—a situation in which the efficient IV estimator is known as the efficient generalized method of moments (GMM) estimator.

The IV Estimator in Matrix Form

In this section, we let $\boldsymbol{X}$ denote the $n \times (k + r + 1)$ matrix of the regressors in the equation of interest, so $\boldsymbol{X}$ contains the included endogenous regressors (the X's in Key Concept 12.1) *and* the included exogenous regressors (the W's in Key Concept 12.1). That is, in the notation of Key Concept 12.1, the i^{th} row of $\boldsymbol{X}$ is $\boldsymbol{X}_i' = (1 \quad X_{1i} \quad X_{2i} \quad \ldots \quad X_{ki} \quad W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$. Also, let $\boldsymbol{Z}$ denote the $n \times (m + r + 1)$ matrix of all the exogenous regressors, both those included in the equation of interest (the W's) *and* those excluded from the equation of interest (the instruments). That is, in the notation of Key Concept 12.1, the i^{th} row of $\boldsymbol{Z}$ is $\boldsymbol{Z}_i' = (1 \quad Z_{1i} \quad Z_{2i} \quad \ldots \quad Z_{mi} \quad W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$.

With this notation, the IV regression model of Key Concept 12.1, written in matrix form, is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}, \tag{18.45}$$

where $\boldsymbol{U}$ is the $n \times 1$ vector of errors in the equation of interest, with i^{th} element u_i.

The matrix $\boldsymbol{Z}$ consists of all the exogenous regressors, so under the IV regression assumptions in Key Concept 12.4,

$$E(\boldsymbol{Z}_i u_i) = \boldsymbol{0} \qquad \text{(instrument exogeneity)}. \tag{18.46}$$

Because there are k included endogenous regressors, the first stage regression consists of k equations.

The TSLS estimator. The TSLS estimator is the instrumental variables estimator in which the instruments are the predicted values of $\boldsymbol{X}$ based on OLS estimation of the first stage regression. Let $\hat{\boldsymbol{X}}$ denote this matrix of predicted values so that the i^{th} row of $\hat{\boldsymbol{X}}$ is $(\hat{X}_{1i} \quad \hat{X}_{2i} \quad \ldots \quad \hat{X}_{ki} \quad W_{1i} \quad W_{2i} \quad \ldots \quad W_{ri})$, where $\hat{X}_{1i}$ is the predicted value from the regression of X_{1i} on $\boldsymbol{Z}$, and so forth. Because the W's are contained in $\boldsymbol{Z}$, the predicted value from a regression of W_{1i} on $\boldsymbol{Z}$ is just W_{1i}, and so forth, so $\hat{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$, where $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ [see Equation (18.27)]. Accordingly, the TSLS estimator is

$$\hat{\boldsymbol{\beta}}^{TSLS} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y}. \tag{18.47}$$

Because $\hat{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$, $\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} = \boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$, and $\hat{\boldsymbol{X}}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}$, the TSLS estimator can be rewritten as

$$\hat{\boldsymbol{\beta}}^{TSLS} = (\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}. \tag{18.48}$$

Asymptotic Distribution of the TSLS Estimator

Substituting Equation (18.45) into Equation (18.48), rearranging, and multiplying by $\sqrt{n}$ yields the expression for the centered and scaled TSLS estimator:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}^{TSLS} - \boldsymbol{\beta}) &= \left(\frac{\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X}}{n}\right)^{-1}\frac{\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{U}}{\sqrt{n}} \\ &= \left[\frac{\boldsymbol{X}'\boldsymbol{Z}}{n}\left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}\right)^{-1}\frac{\boldsymbol{Z}'\boldsymbol{X}}{n}\right]^{-1}\left[\frac{\boldsymbol{X}'\boldsymbol{Z}}{n}\left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}\right)^{-1}\frac{\boldsymbol{Z}'\boldsymbol{U}}{\sqrt{n}}\right], \end{aligned} \qquad (18.49)$$

where the second equality uses the definition of $\boldsymbol{P_Z}$. Under the IV regression assumptions, $\boldsymbol{X}'\boldsymbol{Z}/n \xrightarrow{p} \boldsymbol{Q_{XZ}}$ and $\boldsymbol{Z}'\boldsymbol{Z}/n \xrightarrow{p} \boldsymbol{Q_{ZZ}}$, where $\boldsymbol{Q_{XZ}} = E(\boldsymbol{X}_i\boldsymbol{Z}_i')$ and $\boldsymbol{Q_{ZZ}} = E(\boldsymbol{Z}_i\boldsymbol{Z}_i')$. In addition, under the IV regression assumptions, $\boldsymbol{Z}_i u_i$ is i.i.d. with mean zero [Equation (18.46)] and a nonzero finite variance, so its sum, divided by $\sqrt{n}$, satisfies the conditions of the central limit theorem and

$$\boldsymbol{Z}'\boldsymbol{U}/\sqrt{n} \xrightarrow{d} \boldsymbol{\Psi}_{ZU}, \text{ where } \boldsymbol{\Psi_{ZU}} \sim N(\boldsymbol{0}, \boldsymbol{H}), \boldsymbol{H} = E(\boldsymbol{Z}_i\boldsymbol{Z}_i'u_i^2) \qquad (18.50)$$

and $\boldsymbol{\Psi_{ZU}}$ is $(m + r + 1) \times 1$.

Application of Equation (18.50) and of the limits $\boldsymbol{X}'\boldsymbol{Z}/n \xrightarrow{p} \boldsymbol{Q_{XZ}}$ and $\boldsymbol{Z}'\boldsymbol{Z}/n \xrightarrow{p} \boldsymbol{Q_{ZZ}}$ to Equation (18.49) yields the result that, under the IV regression assumptions, the TSLS estimator is asymptotically normally distributed:

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}^{TSLS} - \boldsymbol{\beta}) \xrightarrow{d} (\boldsymbol{Q_{XZ}}\boldsymbol{Q_{ZZ}^{-1}}\boldsymbol{Q_{ZX}})^{-1}\boldsymbol{Q_{XZ}}\boldsymbol{Q_{ZZ}^{-1}}\Psi_{\boldsymbol{ZU}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}^{TSLS}), \qquad (18.51)$$

where

$$\boldsymbol{\Sigma}^{TSLS} = (\boldsymbol{Q_{XZ}}\boldsymbol{Q_{ZZ}^{-1}}\boldsymbol{Q_{ZX}})^{-1}\boldsymbol{Q_{XZ}}\boldsymbol{Q_{ZZ}^{-1}}\,\boldsymbol{H}\boldsymbol{Q_{ZZ}^{-1}}\,\boldsymbol{Q_{ZX}}\,(\boldsymbol{Q_{XZ}}\boldsymbol{Q_{ZZ}^{-1}}\boldsymbol{Q_{ZX}})^{-1}, \qquad (18.52)$$

where $\boldsymbol{H}$ is defined in Equation (18.50).

Standard errors for TSLS. The formula in Equation (18.52) is daunting. Nevertheless, it provides a way to estimate $\boldsymbol{\Sigma}^{TSLS}$ by substituting sample moments for the population moments. The resulting variance estimator is

$$\hat{\boldsymbol{\Sigma}}^{TSLS} = (\hat{\boldsymbol{Q}}_{\boldsymbol{XZ}}\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{Q}}_{\boldsymbol{ZX}})^{-1}\hat{\boldsymbol{Q}}_{\boldsymbol{XZ}}\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{H}}\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{Q}}_{\boldsymbol{ZX}}\,(\hat{\boldsymbol{Q}}_{\boldsymbol{XZ}}\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{Q}}_{\boldsymbol{ZX}})^{-1}, \qquad (18.53)$$

where $\hat{\boldsymbol{Q}}_{\boldsymbol{XZ}} = \boldsymbol{X}'\boldsymbol{Z}/n$, $\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}} = \boldsymbol{Z}'\boldsymbol{Z}/n$, $\hat{\boldsymbol{Q}}_{\boldsymbol{ZX}} = \boldsymbol{Z}'\boldsymbol{X}/n$, and

$$\hat{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i\,\hat{u}_i^2, \text{ where } \hat{\boldsymbol{U}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{TSLS} \qquad (18.54)$$

so that $\hat{\boldsymbol{U}}$ is the vector of TSLS residuals and where $\hat{u}_i$ is the i^{th} element of that vector (the TSLS residual for the i^{th} observation).

The TSLS standard errors are the square roots of the diagonal elements of $\hat{\boldsymbol{\Sigma}}^{TSLS}/n$.

Properties of TSLS When the Errors Are Homoskedastic

If the errors are homoskedastic, then the TSLS estimator is asymptotically efficient among the class of IV estimators in which the instruments are linear combinations of the rows of $\boldsymbol{Z}$. This result is the IV counterpart to the Gauss–Markov theorem and constitutes an important justification for using TSLS.

The TSLS distribution under homoskedasticity. If the errors are homoskedastic, that is, if $E(u_i^2|\boldsymbol{Z}_i) = \sigma_u^2$, then $\boldsymbol{H} = E(\boldsymbol{Z}_i\boldsymbol{Z}_i'u_i^2) = E[E(\boldsymbol{Z}_i\boldsymbol{Z}_i'u_i^2\,|\,\boldsymbol{Z}_i)] = E[\boldsymbol{Z}_i\boldsymbol{Z}_i'E(u_i^2\,|\,\boldsymbol{Z}_i)] = \boldsymbol{Q}_{\boldsymbol{ZZ}}\sigma_u^2$. In this case, the variance of the asymptotic distribution of the TSLS estimator in Equation (18.52) simplifies to

$$\boldsymbol{\Sigma}^{TSLS} = (\boldsymbol{Q}_{\boldsymbol{XZ}}\boldsymbol{Q}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{Q}_{\boldsymbol{ZX}})^{-1}\sigma_u^2 \quad \text{(homoskedasticity only).} \qquad (18.55)$$

The homoskedasticity-only estimator of the TSLS variance matrix is

$$\widetilde{\boldsymbol{\Sigma}}^{TSLS} = (\hat{\boldsymbol{Q}}_{\boldsymbol{XZ}}\hat{\boldsymbol{Q}}_{\boldsymbol{ZZ}}^{-1}\hat{\boldsymbol{Q}}_{\boldsymbol{ZX}})^{-1}\hat{\sigma}_u^2, \text{ where } \hat{\sigma}_u^2 = \frac{\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}}{n-k-r-1} \quad \text{(homoskedasticity only),} \qquad (18.56)$$

and the homoskedasticity-only TSLS standard errors are the square root of the diagonal elements of $\widetilde{\boldsymbol{\Sigma}}^{TSLS}/n$.

The class of IV estimators that use linear combinations of Z. The class of IV estimators that use linear combinations of $\boldsymbol{Z}$ as instruments can be generated in two equivalent ways. Both start with the same moment equation: Under the assumption of instrument exogeneity, the errors $\boldsymbol{U} = \boldsymbol{Y} - \boldsymbol{X\beta}$ are uncorrelated with the exogenous regressors; that is, at the true value of $\boldsymbol{\beta}$, Equation (18.46) implies that

$$E[(\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{Z}] = 0. \qquad (18.57)$$

Equation (18.57) constitutes a system of $m + r + 1$ equations involving the $k + r + 1$ unknown elements of $\boldsymbol{\beta}$. When $m > k$, these equations are redundant, in the sense that all are satisfied at the true value of $\boldsymbol{\beta}$. When these population moments are replaced by their sample moments, the system of equations $(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{Z} = 0$ can be solved for $\boldsymbol{b}$ when there is exact identification $(m = k)$. This value of $\boldsymbol{b}$ is the IV estimator of $\boldsymbol{\beta}$. However, when there is overidentification $(m > k)$, the system of equations typically cannot all be satisfied by the same value of $\boldsymbol{b}$ because of sampling variation—there are more equations than unknowns—and in general this system does not have a solution.

The first approach to the problem of estimating $\boldsymbol{\beta}$ when there is overidentification is to trade off the desire to satisfy each equation by minimizing a quadratic form involving all the equations. Specifically, let $\boldsymbol{A}$ be an $(m + r + 1) \times (m + r + 1)$ symmetric positive semidefinite weight matrix and let $\hat{\boldsymbol{\beta}}_{\mathbf{A}}^{IV}$ denote the estimator that minimizes

$$\min_b(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{ZAZ}'(\boldsymbol{Y} - \boldsymbol{Xb}). \tag{18.58}$$

The solution to this minimization problem is found by taking the derivative of the objective function with respect to $\boldsymbol{b}$, setting the result equal to zero, and rearranging. Doing so yields $\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}^{IV}$, the IV estimator based on the weight matrix $\boldsymbol{A}$:

$$\hat{\boldsymbol{\beta}}_{\mathbf{A}}^{IV} = (\boldsymbol{X}'\boldsymbol{ZAZ}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{ZAZ}'\boldsymbol{Y}. \tag{18.59}$$

Comparison of Equations (18.59) and (18.48) shows that TSLS is the IV estimator with $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. That is, TSLS is the solution of the minimization problem in Equation (18.58) with $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

The calculations leading to Equations (18.51) and (18.52), applied to $\hat{\boldsymbol{\beta}}_{\mathbf{A}}^{IV}$, show that

$$\begin{gathered}\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathbf{A}}^{IV} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{A}}^{IV}), \text{ where} \\ \boldsymbol{\Sigma}_{A}^{IV} = (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{AHA}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}.\end{gathered} \tag{18.60}$$

The second way to generate the class of IV estimators that use linear combinations of $\boldsymbol{Z}$ is to consider IV estimators in which the instruments are $\boldsymbol{ZB}$, where $\boldsymbol{B}$ is an $(m + r + 1) \times (k + r + 1)$ matrix with full row rank. Then the system of $(k + r + 1)$ equations, $(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{ZB} = 0$, can be solved uniquely for the $(k + r + 1)$ unknown elements of $\boldsymbol{b}$. Solving these equations for $\boldsymbol{b}$ yields $\hat{\boldsymbol{\beta}}^{IV} = (\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{B}'\boldsymbol{Z}'\boldsymbol{Y})$, and substitution of $\boldsymbol{B} = \boldsymbol{AZ}'\boldsymbol{X}$ into this expression yields Equation (18.59). Thus the two approaches to defining IV estimators that are linear combinations of the instruments yield the same family of IV estimators. It is conventional to work with the first approach, in which the IV estimator solves the quadratic minimization problem in Equation (18.58), and that is the approach taken here.

Asymptotic efficiency of TSLS under homoskedasticity. If the errors are homoskedastic, then $\boldsymbol{H} = \boldsymbol{Q}_{ZZ}\sigma_u^2$ and the expression for $\boldsymbol{\Sigma}_{\boldsymbol{A}}^{IV}$ in Equation (18.60) becomes

$$\boldsymbol{\Sigma}_{\boldsymbol{A}}^{IV} = (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZZ}\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2. \tag{18.61}$$

To show that TSLS is asymptotically efficient among the class of estimators that are linear combinations of $\boldsymbol{Z}$ when the errors are homoskedastic, we need to show that, under homoskedasticity,

$$\boldsymbol{c}'\boldsymbol{\Sigma}_{\boldsymbol{A}}^{IV}\boldsymbol{c} \geq \boldsymbol{c}'\boldsymbol{\Sigma}^{TSLS}\boldsymbol{c} \tag{18.62}$$

for all positive semidefinite matrices $\boldsymbol{A}$ and all $(k + r + 1) \times 1$ vectors $\boldsymbol{c}$, where $\boldsymbol{\Sigma}^{TSLS} = (\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2$ [Equation (18.55)]. The inequality (18.62), which is proven in Appendix 18.6, is the same efficiency criterion as is used in the multivariate Gauss–Markov theorem in Key Concept 18.3. Consequently, TSLS is the efficient IV estimator under homoskedasticity, among the class of estimators in which the instruments are linear combinations of $\boldsymbol{Z}$.

The J-statistic under homoskedasticity. The J-statistic (Key Concept 12.6) tests the null hypothesis that all the overidentifying restrictions hold against the alternative that some or all of them do not hold.

The idea of the J-statistic is that, if the overidentifying restrictions hold, u_i will be uncorrelated with the instruments and thus a regression of $\boldsymbol{U}$ on $\boldsymbol{Z}$ will have population regression coefficients that all equal zero. In practice, $\boldsymbol{U}$ is not observed, but it can be estimated by the TSLS residuals $\hat{\boldsymbol{U}}$, so a regression of $\hat{\boldsymbol{U}}$ on $\boldsymbol{Z}$ should yield statistically insignificant coefficients. Accordingly, the TSLS J-statistic is the homoskedasticity-only F-statistic testing the hypothesis that the coefficients on $\boldsymbol{Z}$ are all zero, in the regression of $\hat{\boldsymbol{U}}$ on $\boldsymbol{Z}$, multiplied by $(m + r + 1)$ so that the F-statistic is in its asymptotic chi-squared form.

An explicit formula for the J-statistic can be obtained using Equation (7.13) for the homoskedasticity-only F-statistic. The unrestricted regression is the regression of $\hat{\boldsymbol{U}}$ on the $m + r + 1$ regressors $\boldsymbol{Z}$, and the restricted regression has no regressors. Thus, in the notation of Equation (7.13), $SSR_{unrestricted} = \hat{\boldsymbol{U}}'\boldsymbol{M}_{\boldsymbol{Z}}\hat{\boldsymbol{U}}$ and $SSR_{restricted} = \hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}$, so $SSR_{restricted} - SSR_{unrestricted} = \hat{\boldsymbol{U}}'\hat{\boldsymbol{U}} - \hat{\boldsymbol{U}}'\boldsymbol{M}_{\boldsymbol{Z}}\hat{\boldsymbol{U}} = \hat{\boldsymbol{U}}'\boldsymbol{P}_{\boldsymbol{Z}}\hat{\boldsymbol{U}}$ and the J-statistic is

$$J = \frac{\hat{\boldsymbol{U}}'\boldsymbol{P}_{\boldsymbol{Z}}\hat{\boldsymbol{U}}}{\hat{\boldsymbol{U}}'\boldsymbol{M}_{\boldsymbol{Z}}\hat{\boldsymbol{U}}/(n - m - r - 1)}. \tag{18.63}$$

The method for computing the J-statistic described in Key Concept 12.6 entails testing only the hypothesis that the coefficients on the excluded instruments are zero. Although these two methods have different computational steps, they produce identical J-statistics (Exercise 18.14).

It is shown in Appendix 18.6 that, under the null hypothesis that $E(u_i\boldsymbol{Z}_i) = 0$,

$$J \xrightarrow{d} \chi^2_{m-k}. \tag{18.64}$$

Generalized Method of Moments Estimation in Linear Models

If the errors are heteroskedastic, then the TSLS estimator is no longer efficient among the class of IV estimators that use linear combinations of $\boldsymbol{Z}$ as instruments. The efficient estimator in this case is known as the efficient generalized method of moments (GMM) estimator. In addition, if the errors are heteroskedastic, then the J-statistic as defined in Equation (18.63) no longer has a chi-squared distribution. However, an alternative formulation of the J-statistic, constructed using the efficient GMM estimator, does have a chi-squared distribution with $m - k$ degrees of freedom.

These results parallel the results for the estimation of the usual regression model with exogenous regressors and heteroskedastic errors: If the errors are heteroskedastic, then the OLS estimator is not efficient among estimators that are linear in $\boldsymbol{Y}$ (the Gauss–Markov conditions are not satisfied) and the homoskedasticity-only F-statistic no longer has an F distribution, even in large samples. In the regression model with exogenous regressors and heteroskedasticity, the efficient estimator is weighted least squares; in the IV regression model with heteroskedasticity, the efficient estimator uses a different weighting matrix than TSLS, and the resulting estimator is the efficient GMM estimator.

GMM estimation. **Generalized method of moments (GMM)** estimation is a general method for the estimation of the parameters of linear or nonlinear models, in which the parameters are chosen to provide the best fit to multiple equations, each of which sets a sample moment to zero. These equations, which in the context of GMM are called moment conditions, typically cannot all be satisfied simultaneously. The GMM estimator trades off the desire to satisfy each of the equations by minimizing a quadratic objective function.

In the linear IV regression model with exogenous variables $\boldsymbol{Z}$, the class of GMM estimators consists of all the estimators that are solutions to the quadratic minimization problem in Equation (18.58). Thus the class of GMM estimators based on the full set of instruments $\boldsymbol{Z}$ with different-weight matrices $\boldsymbol{A}$ is the same as the class of IV estimators in which the instruments are linear combinations of $\boldsymbol{Z}$.

In the linear IV regression model, GMM is just another name for the class of estimators we have been studying—that is, estimators that solve Equation (18.58).

The asymptotically efficient GMM estimator. Among the class of GMM estimators, the **efficient GMM** estimator is the GMM estimator with the smallest asymptotic variance matrix [where the smallest variance matrix is defined as in Equation (18.62)]. Thus the result in Equation (18.62) can be restated as saying that TSLS is the efficient GMM estimator in the linear model when the errors are homoskedastic.

To motivate the expression for the efficient GMM estimator when the errors are heteroskedastic, recall that when the errors are homoskedastic, $\boldsymbol{H}$ [the variance matrix of $\boldsymbol{Z}_i u_i$; see Equation (18.50)] equals $\boldsymbol{Q_{ZZ}}\sigma_u^2$, and the asymptotically efficient weight matrix is obtained by setting $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, which yields the TSLS estimator. In large samples, using the weight matrix $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ is equivalent to using $\boldsymbol{A} = (\boldsymbol{Q_{ZZ}}\sigma_u^2)^{-1} = \boldsymbol{H}^{-1}$. This interpretation of the TSLS estimator suggests that, by analogy, the efficient IV estimator under heteroskedasticity can be obtained by setting $\boldsymbol{A} = \boldsymbol{H}^{-1}$ and solving

$$\min_b(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'\boldsymbol{Z}\boldsymbol{H}^{-1}\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}). \tag{18.65}$$

This analogy is correct: The solution to the minimization problem in Equation (18.65) is the efficient GMM estimator. Let $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ denote the solution to the minimization problem in Equation (18.65). By Equation (18.59), this estimator is

$$\widetilde{\boldsymbol{\beta}}^{Eff.GMM} = (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{H}^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{H}^{-1}\boldsymbol{Z}'\boldsymbol{Y}. \tag{18.66}$$

The asymptotic distribution of $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ is obtained by substituting $\boldsymbol{A} = \boldsymbol{H}^{-1}$ into Equation (18.60) and simplifying; thus

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}^{Eff.GMM} - \boldsymbol{\beta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}^{Eff.GMM}),$$
$$\text{where } \boldsymbol{\Sigma}^{Eff.GMM} = (\boldsymbol{Q_{XZ}}\boldsymbol{H}^{-1}\boldsymbol{Q_{ZX}})^{-1}. \tag{18.67}$$

The result that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ is the efficient GMM estimator is proven by showing that $\boldsymbol{c}'\boldsymbol{\Sigma}_A^{IV}\boldsymbol{c} \geq \boldsymbol{c}'\boldsymbol{\Sigma}^{Eff.GMM}\boldsymbol{c}$ for all vectors $\boldsymbol{c}$, where $\boldsymbol{\Sigma}_A^{IV}$ is given in Equation (18.60). The proof of this result is given in Appendix 18.6.

Feasible efficient GMM estimation. The GMM estimator defined in Equation (18.66) is not a feasible estimator because it depends on the unknown variance matrix $\boldsymbol{H}$. However, a feasible efficient GMM estimator can be computed by substituting a consistent estimator of $\boldsymbol{H}$ into the minimization problem of Equation (18.65) or, equivalently, by substituting a consistent estimator of $\boldsymbol{H}$ into the formula for $\hat{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (18.66).

The efficient GMM estimator can be computed in two steps. In the first step, estimate $\boldsymbol{\beta}$ using any consistent estimator. Use this estimator of $\boldsymbol{\beta}$ to compute the residuals from the equation of interest, and then use these residuals to compute an estimator of $\boldsymbol{H}$. In the second step, use this estimator of $\boldsymbol{H}$ to estimate the optimal weight matrix $\boldsymbol{H}^{-1}$ and to compute the efficient GMM estimator. To be concrete, in the linear IV regression model, it is natural to use the TSLS estimator in the first step and to use the TSLS residuals to estimate $\boldsymbol{H}$. If TSLS is used in the first step, then the feasible efficient GMM estimator computed in the second step is

$$\hat{\boldsymbol{\beta}}^{Eff.GMM} = (\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{H}}^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{H}}^{-1}\boldsymbol{Z}'\boldsymbol{Y}, \tag{18.68}$$

where $\hat{\boldsymbol{H}}$ is given in Equation (18.54).

Because $\hat{\boldsymbol{H}} \xrightarrow{p} \boldsymbol{H}, \sqrt{n}(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \tilde{\boldsymbol{\beta}}^{Eff.GMM}) \xrightarrow{p} 0$ (Exercise 18.12), and

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}^{Eff.GMM}), \tag{18.69}$$

where $\boldsymbol{\Sigma}^{Eff.GMM} = (\boldsymbol{Q}_{XZ}\boldsymbol{H}^{-1}\boldsymbol{Q}_{ZX})^{-1}$ [Equation (18.67)]. That is, the feasible two-step estimator $\hat{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (18.68) is, asymptotically, the efficient GMM estimator.

The heteroskedasticity-robust J-statistic. The **heteroskedasticity-robust *J*-statistic**, also known as the **GMM *J*-statistic**, is the counterpart of the TSLS-based *J*-statistic, computed using the efficient GMM estimator and weight function. That is, the GMM *J*-statistic is given by

$$J^{GMM} = (\boldsymbol{Z}'\hat{\boldsymbol{U}}^{GMM})'\hat{\boldsymbol{H}}^{-1}(\boldsymbol{Z}'\hat{\boldsymbol{U}}^{GMM})/n, \tag{18.70}$$

where $\hat{\boldsymbol{U}}^{GMM} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{Eff.GMM}$ are the residuals from the equation of interest, estimated by (feasible) efficient GMM, and $\hat{\boldsymbol{H}}^{-1}$ is the weight matrix used to compute $\hat{\boldsymbol{\beta}}^{Eff.GMM}$.

Under the null hypothesis $E(\boldsymbol{Z}_i u_i) = \boldsymbol{0}$, $J^{GMM} \xrightarrow{d} \chi^2_{m-k}$ (see Appendix 18.6).

GMM with time series data. The results in this section were derived under the IV regression assumptions for cross–sectional data. In many applications, however, these results extend to time series applications of IV regression and GMM. Although a formal mathematical treatment of GMM with time series data is beyond the scope of this book (for such a treatment, see Hayashi, 2000, Chapter 6), we nevertheless will summarize the key ideas of GMM estimation with time series data. This summary assumes familiarity with the material in Chapters 14 and 15. For this discussion, it is assumed that the variables are stationary.

It is useful to distinguish between two types of applications: applications in which the error term u_t is serially correlated and applications in which u_t is serially uncorrelated. If the error term u_t is serially correlated, then the asymptotic distribution of the GMM estimator continues to be normally distributed, but the formula for $\boldsymbol{H}$ in Equation (18.50) is no longer correct. Instead, the correct expression for $\boldsymbol{H}$ depends on the autocovariances of $\boldsymbol{Z}_t u_t$ and is analogous to the formula given in Equation (15.14) for the variance of the OLS estimator when the error term is serially correlated. The efficient GMM estimator is still constructed using a consistent estimator of $\boldsymbol{H}$; however, that consistent estimator must be computed using the HAC methods discussed in Chapter 15.

If the error term u_t is not serially correlated, then HAC estimation of $\boldsymbol{H}$ is unnecessary and the formulas presented in this section all extend to time series GMM applications. In modern applications to finance and macroeconometrics, it is common to encounter models in which the error term represents an unexpected or unforecastable disturbance, in which case the model implies that u_t is serially uncorrelated. For example, consider a model with a single included endogenous variable and no included exogenous variables so that the equation of interest is $Y_t = \beta_0 + \beta_1 X_t + u_t$. Suppose that an economic theory implies that u_t is unpredictable given past information. Then the theory implies the moment condition

$$E(u_t | Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \ldots) = 0, \tag{18.71}$$

where Z_{t-1} is the lagged value of some other variable. The moment condition in Equation (18.71) implies that all the lagged variables $Y_{t-1}, X_{t-1}, Z_{t-1}, Y_{t-2}, X_{t-2}, Z_{t-2}, \ldots$ are candidates for being valid instruments (they satisfy the exogeneity condition). Moreover, because $u_{t-1} = Y_{t-1} - \beta_0 - \beta_1 X_{t-1}$, the moment condition in Equation (18.71) is equivalent to $E(u_t | u_{t-1}, X_{t-1}, Z_{t-1}, u_{t-2}, X_{t-2}, Z_{t-2}, \ldots) = 0$. Because u_t is serially uncorrelated, HAC estimation of $\boldsymbol{H}$ is unnecessary. The theory of GMM presented in this section, including efficient GMM estimation and the GMM J-statistic, therefore applies directly to time series applications with moment conditions of the form in Equation (18.71), under the hypothesis that the moment condition in Equation (18.71) is, in fact, correct.

Summary

1. The linear multiple regression model in matrix form is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$, where $\boldsymbol{Y}$ is the $n \times 1$ vector of observations on the dependent variable, $\boldsymbol{X}$ is the $n \times (k+1)$ matrix of n observations on the $k + 1$ regressors (including a constant), $\boldsymbol{\beta}$ is the $k + 1$ vector of unknown parameters, and $\boldsymbol{U}$ is the $n \times 1$ vector of error terms.

2. The OLS estimator is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Under the first four least squares assumptions in Key Concept 18.1, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed. If in addition the errors are homoskedastic, then the conditional variance of $\hat{\boldsymbol{\beta}}$ is $\text{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
3. General linear restrictions on $\boldsymbol{\beta}$ can be written as the q equations $\boldsymbol{R\beta} = \boldsymbol{r}$, and this formulation can be used to test joint hypotheses involving multiple coefficients or to construct confidence sets for elements of $\boldsymbol{\beta}$.
4. When the regression errors are i.i.d. and normally distributed, conditional on $\boldsymbol{X}$, $\boldsymbol{\beta}$ has an exact normal distribution and the homoskedasticity-only t- and F-statistics have exact $t_{n-k-1}$and $F_{q,n-k-1}$ distributions, respectively.
5. The Gauss–Markov theorem says that, if the errors are homoskedastic and conditionally uncorrelated across observations and if $E(u_i|\boldsymbol{X}) = 0$, the OLS estimator is efficient among linear conditionally unbiased estimators (that is, OLS is BLUE).
6. If the error covariance matrix $\boldsymbol{\Omega}$ is not proportional to the identity matrix, and if $\boldsymbol{\Omega}$ is known or can be estimated, then the GLS estimator is asymptotically more efficient than OLS. However, GLS requires that, in general, u_i be uncorrelated with *all* observations on the regressors, not just with $\boldsymbol{X}_i$, as is required by OLS, an assumption that must be evaluated carefully in applications.
7. The TSLS estimator is a member of the class of GMM estimators of the linear model. In GMM, the coefficients are estimated by making the sample covariance between the regression error and the exogenous variables as small as possible—specifically, by solving $\min_b[(\boldsymbol{Y} - \boldsymbol{Xb})'\boldsymbol{Z}]\boldsymbol{A}[\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{Xb})]$, where $\boldsymbol{A}$ is a weight matrix. The asymptotically efficient GMM estimator sets $\boldsymbol{A} = [E(\boldsymbol{Z}_i\boldsymbol{Z}_i'u_i^2)]^{-1}$. When the errors are homoskedastic, the asymptotically efficient GMM estimator in the linear IV regression model is TSLS.

Key Terms

Gauss–Markov conditions for multiple regression (720)
Gauss–Markov theorem for multiple regression (721)
generalized least squares (GLS) (723)
infeasible GLS (726)
feasible GLS (726)
generalized method of moments (GMM) (734)
efficient GMM (735)

heteroskedasticity-robust
 J-statistic (736)
GMM *J*-statistic (736)
mean vector (750)
covariance matrix (750)

Review the Concepts

18.1 A researcher studying the relationship between earnings and gender for a group of workers specifies the regression model $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + u_i$, where X_{1i} is a binary variable that equals 1 if the i^{th} person is a female and X_{2i} is a binary variable that equals 1 if the i^{th} person is a male. Write the model in the matrix form of Equation (18.2) for a hypothetical set of $n = 5$ observations. Show that the columns of $\boldsymbol{X}$ are linearly dependent so that $\boldsymbol{X}$ does not have full rank. Explain how you would respecifiy the model to eliminate the perfect multicollinearity.

18.2 You are analyzing a linear regression model with 500 observations and one regressor. Explain how you would construct a confidence interval for β_1 if:

a. Assumptions #1 through #4 in Key Concept 18.1 are true, but you think Assumption #5 or #6 might not be true.

b. Assumptions #1 through #5 are true, but you think Assumption #6 might not be true. (Give two ways to construct the confidence interval.)

c. Assumptions #1 through #6 are true.

18.3 Suppose that Assumptions #1 through #5 in Key Concept 18.1 are true but that Assumption #6 is not. Does the result in Equation (18.31) hold? Explain.

18.4 Can you compute the BLUE estimator of $\boldsymbol{\beta}$ if Equation (18.41) holds and you do not know $\boldsymbol{\Omega}$? What if you know $\boldsymbol{\Omega}$?

18.5 Construct an example of a regression model that satisfies the assumption $E(u_i \mid \boldsymbol{X}_i) = 0$ but for which $E(\boldsymbol{U} \mid \boldsymbol{X}) \neq \boldsymbol{0}_n$.

Exercises

18.1 Consider the population regression of test scores against income and the square of income in Equation (8.1).

a. Write the regression in Equation (8.1) in the matrix form of Equation (18.5). Define $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{\beta}$.

b. Explain how to test the null hypothesis that the relationship between test scores and income is linear against the alternative that it is quadratic. Write the null hypothesis in the form of Equation (18.20). What are $\boldsymbol{R}$, $\boldsymbol{r}$, and q?

18.2 Suppose that a sample of $n = 20$ households has the sample means and sample covariances below for a dependent variable and two regressors:

		Sample Covariances		
	Sample Means	Y	X_1	X_2
Y	6.39	0.26	0.22	0.32
X_1	7.24		0.80	0.28
X_2	4.00			2.40

a. Calculate the OLS estimates of β_0, β_1, and β_2. Calculate $s_{\hat{u}}^2$. Calculate the R^2 of the regression.

b. Suppose that all six assumptions in Key Concept 18.1 hold. Test the hypothesis that $\beta_1 = 0$ at the 5% significance level.

18.3 Let $\boldsymbol{W}$ be an $m \times 1$ vector with covariance matrix $\boldsymbol{\Sigma_W}$, where $\boldsymbol{\Sigma_W}$ is finite and positive definite. Let $\boldsymbol{c}$ be a nonrandom $m \times 1$ vector and let $Q = \boldsymbol{c}'\boldsymbol{W}$.

a. Show that $\text{var}(Q) = \boldsymbol{c}'\boldsymbol{\Sigma_W c}$.

b. Suppose that $\boldsymbol{c} \neq \boldsymbol{0}_m$. Show that $0 < \text{var}(Q) < \infty$.

18.4 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ from Chapter 4 and assume that the least squares assumptions in Key Concept 4.3 hold.

a. Write the model in the matrix form given in Equations (18.2) and (18.4).

b. Show that Assumptions #1 through #4 in Key Concept 18.1 are satisfied.

c. Use the general formula for $\hat{\boldsymbol{\beta}}$ in Equation (18.11) to derive the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ given in Key Concept 4.2.

d. Show that the (1, 1) element of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ in Equation (18.13) is equal to the expression for $\sigma_{\hat{\beta}_0}^2$ given in Key Concept 4.4.

18.5 Let $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ be as defined in Equations (18.24) and (18.25).

a. Prove that $\boldsymbol{P_X M_X} = \boldsymbol{0}_{n\times n}$ and that $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are idempotent.

b. Derive Equations (18.27) and (18.28).

c. Show that rank($\boldsymbol{P_X}$) $= k + 1$ and rank($\boldsymbol{M_X}$) $= n - k - 1$. [*Hint:* First solve Exercise 18.10 and then use the fact that trace($\boldsymbol{AB}$) = trace($\boldsymbol{BA}$) for conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.]

18.6 Consider the regression model in matrix form, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{W\gamma} + \boldsymbol{U}$, where $\boldsymbol{X}$ is an $n \times k_1$ matrix of regressors and $\boldsymbol{W}$ is an $n \times k_2$ matrix of regressors. Then, as shown in Exercise 18.17, the OLS estimator $\hat{\boldsymbol{\beta}}$ can be expressed

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M_W X})^{-1}(\boldsymbol{X}'\boldsymbol{M_W Y}).$$

Now let $\hat{\beta}_1^{BV}$ be the "binary variable" fixed effects estimator computed by estimating Equation (10.11) by OLS and let $\hat{\beta}_1^{DM}$ be the "de-meaning" fixed effects estimator computed by estimating Equation (10.14) by OLS, in which the entity-specific sample means have been subtracted from X and Y. Use the expression for $\hat{\boldsymbol{\beta}}$ given above to prove that $\hat{\beta}_1^{BV} = \hat{\beta}_1^{DM}$. [*Hint*: Write Equation (10.11) using a full set of fixed effects, $D1_i, D2_i, \ldots, Dn_i$ and no constant term. Include all of the fixed effects in $\boldsymbol{W}$. Write out the matrix $\boldsymbol{M_W X}$.]

18.7 Consider the regression model $Y_i = \beta_1 X_i + \beta_2 W_i + u_i$, where for simplicity the intercept is omitted and all variables are assumed to have a mean of zero. Suppose that X_i is distributed independently of (W_i, u_i) but W_i and u_i might be correlated and let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the OLS estimators for this model. Show that

a. Whether or not W_i and u_i are correlated, $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

b. If W_i and u_i are correlated, then $\hat{\beta}_2$ is inconsistent.

c. Let $\hat{\beta}_1^r$ be the OLS estimator from the regression of Y on X (the restricted regression that excludes W). Will $\hat{\beta}_1$ have a smaller asymptotic variance than $\hat{\beta}_1^r$, allowing for the possibility that W_i and u_i are correlated? Explain.

18.8 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $u_1 = \tilde{u}_1$ and $u_i = 0.5u_{i-1} + \tilde{u}_i$ for $i = 2, 3, \ldots, n$. Suppose that $\tilde{u}_i$ are i.i.d. with mean 0 and variance 1 and are distributed independently of X_j for all i and j.

a. Derive an expression for $E(\boldsymbol{UU}') = \boldsymbol{\Omega}$.

b. Explain how to estimate the model by GLS without explicitly inverting the matrix $\boldsymbol{\Omega}$. (*Hint:* Transform the model so that the regression errors are $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n$.)

18.9 This exercise shows that the OLS estimator of a subset of the regression coefficients is consistent under the conditional mean independence assumption stated in Appendix 7.2. Consider the multiple regression model in matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{W\gamma} + \boldsymbol{U}$, where $\boldsymbol{X}$ and $\boldsymbol{W}$ are, respectively, $n \times k_1$ and $n \times k_2$ matrices of regressors. Let $\boldsymbol{X}_i'$ and $\boldsymbol{W}_i'$ denote the i^{th} rows of $\boldsymbol{X}$ and $\boldsymbol{W}$ [as in Equation (18.3)]. Assume that (i) $E(u_i|\boldsymbol{X}_i, \boldsymbol{W}_i) = \boldsymbol{W}_i'\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a $k_2 \times 1$ vector of unknown parameters; (ii) $(\boldsymbol{X}_i, \boldsymbol{W}_i, Y_i)$ are i.i.d.; (iii) $(\boldsymbol{X}_i, \boldsymbol{W}_i, u_i)$ have four finite, nonzero moments; and (iv) there is no perfect multicollinearity. These are Assumptions #1 through #4 of Key Concept 18.1, with the conditional mean independence assumption (i) replacing the usual conditional mean zero assumption.

a. Use the expression for $\hat{\boldsymbol{\beta}}$ given in Exercise 18.6 to write $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (n^{-1}\boldsymbol{X}'\boldsymbol{M_W X})^{-1}(n^{-1}\boldsymbol{X}'\boldsymbol{M_W U})$.

b. Show that $n^{-1}\boldsymbol{X}'\boldsymbol{M_W X} \xrightarrow{p} \boldsymbol{\Sigma_{XX}} - \boldsymbol{\Sigma_{XW}\Sigma_{WW}^{-1}\Sigma_{WX}}$, where $\boldsymbol{\Sigma_{XX}} = E(\boldsymbol{X}_i\boldsymbol{X}_i')$, $\boldsymbol{\Sigma_{XW}} = E(\boldsymbol{X}_i\boldsymbol{W}_i')$, and so forth. [The matrix $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$ if $\boldsymbol{A}_{n,ij} \xrightarrow{p} \boldsymbol{A}_{ij}$ for all i, j, where $\boldsymbol{A}_{n,ij}$ and $\boldsymbol{A}_{ij}$ are the (i, j) elements of $\boldsymbol{A}_n$ and $\boldsymbol{A}$.]

c. Show that assumptions (i) and (ii) imply that $E(\boldsymbol{U}|\boldsymbol{X}, \boldsymbol{W}) = \boldsymbol{W\delta}$.

d. Use (c) and the law of iterated expectations to show that $n^{-1}\boldsymbol{X}'\boldsymbol{M_W U} \xrightarrow{p} \boldsymbol{0}_{k_1 \times 1}$.

e. Use (a) through (d) to conclude that, under conditions (i) through (iv), $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

18.10 Let $\boldsymbol{C}$ be a symmetric idempotent matrix.

a. Show that the eigenvalues of $\boldsymbol{C}$ are either 0 or 1. (*Hint:* Note that $\boldsymbol{Cq} = \gamma\boldsymbol{q}$ implies $0 = \boldsymbol{Cq} - \gamma\boldsymbol{q} = \boldsymbol{CCq} - \gamma\boldsymbol{q} = \gamma\boldsymbol{Cq} - \gamma\boldsymbol{q} = \gamma^2\boldsymbol{q} - \gamma\boldsymbol{q}$ and solve for $\boldsymbol{\gamma}$.)

b. Show that trace($\boldsymbol{C}$) = rank($\boldsymbol{C}$).

c. Let $\boldsymbol{d}$ be an $n \times 1$ vector. Show that $\boldsymbol{d}'\boldsymbol{Cd} \geq 0$.

18.11 Suppose that $\boldsymbol{C}$ is an $n \times n$ symmetric idempotent matrix with rank r and let $\boldsymbol{V} \sim \boldsymbol{N}(\boldsymbol{0}_n, \boldsymbol{I}_n)$.

a. Show that $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}'$, where $\boldsymbol{A}$ is $n \times r$ with $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_r$. (*Hint:* $\boldsymbol{C}$ is positive semidefinite and can be written as $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$, as explained in Appendix 18.1.)

b. Show that $\boldsymbol{A}'\boldsymbol{V} \sim N(\boldsymbol{0}_r, \boldsymbol{I}_r)$.

c. Show that $\boldsymbol{V}'\boldsymbol{C}\boldsymbol{V} \sim \chi^2_{r.}$

18.12 **a.** Show that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ is the efficient GMM estimator—that is, that $\widetilde{\boldsymbol{\beta}}^{Eff.GMM}$ in Equation (18.66) is the solution to Equation (18.65).

b. Show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^{Eff.GMM} - \widetilde{\boldsymbol{\beta}}^{Eff.GMM}) \xrightarrow{p} 0$.

c. Show that $J^{GMM} \xrightarrow{d} \chi^2_{m-k}$.

18.13 Consider the problem of minimizing the sum of squared residuals, subject to the constraint that $\boldsymbol{Rb} = \boldsymbol{r}$, where $\boldsymbol{R}$ is $q \times (k + 1)$ with rank q. Let $\widetilde{\boldsymbol{\beta}}$ be the value of $\boldsymbol{b}$ that solves the constrained minimization problem.

a. Show that the Lagrangian for the minimization problem is $L(\boldsymbol{b}, \boldsymbol{\gamma}) = (\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb}) + \boldsymbol{\gamma}'(\boldsymbol{Rb} - \boldsymbol{r})$, where $\boldsymbol{\gamma}$ is a $q \times 1$ vector of Lagrange multipliers.

b. Show that $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$.

c. Show that $(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) - (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$.

d. Show that $\widetilde{F}$ in Equation (18.36) is equivalent to the homoskedasticity-only F-statistic in Equation (7.13).

18.14 Consider the regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$. Partition $\boldsymbol{X}$ as $[\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and $\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1' \quad \boldsymbol{\beta}_2']'$, where $\boldsymbol{X}_1$ has k_1 columns and $\boldsymbol{X}_2$ has k_2 columns. Suppose that $\boldsymbol{X}_2'\boldsymbol{Y} = \boldsymbol{0}_{k_2 \times 1}$. Let $\boldsymbol{R} = [\boldsymbol{I}_{k_1} \quad \boldsymbol{0}_{k_1 \times k_2}]$.

a. Show that $\hat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = (\boldsymbol{R}\hat{\boldsymbol{\beta}})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}})$.

b. Consider the regression described in Equation (12.17). Let $\boldsymbol{W} = [\boldsymbol{1} \quad \boldsymbol{W}_1 \quad \boldsymbol{W}_2 \quad \ldots \quad \boldsymbol{W}_r]$, where $\boldsymbol{1}$ is an $n \times 1$ vector of ones, $\boldsymbol{W}_1$ is the $n \times 1$ vector with i^{th} element W_{1i}, and so forth. Let $\hat{\boldsymbol{U}}^{TSLS}$ denote the vector of two-stage least squares residuals.

i. Show that $\boldsymbol{W}'\hat{\boldsymbol{U}}^{TSLS} = 0$.

ii. Show that the method for computing the J-statistic described in Key Concept 12.6 (using a homoskedasticity-only F-statistic) and the formula in Equation (18.63) produce the same value for the J-statistic. [*Hint:* Use the results in (a), (b, i), and Exercise 18.13.]

18.15 (Consistency of clustered standard errors.) Consider the panel data model $Y_{it} = \beta X_{it} + \alpha_i + u_{it}$, where all variables are scalars. Assume that Assumptions #1, #2, and #4 in Key Concept 10.3 hold and strengthen Assumption #3 so that X_{it} and u_{it} have eight nonzero finite moments. Let $\boldsymbol{M} = \boldsymbol{I}_T - T^{-1}\boldsymbol{\iota\iota}'$, where $\boldsymbol{\iota}$ is a $T \times 1$ vector of ones. Also let $\boldsymbol{Y}_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{iT})'$, $\boldsymbol{X}_i = (X_{i1} \quad X_{i2} \quad \cdots \quad X_{iT})'$, $\boldsymbol{u}_i = (u_{i1} \quad u_{i2} \quad \cdots \quad u_{iT})'$, $\widetilde{\boldsymbol{Y}}_i = \boldsymbol{MY}_i$, $\widetilde{\boldsymbol{X}}_i = \boldsymbol{MX}_i$, and $\widetilde{\boldsymbol{u}}_i = \boldsymbol{Mu}_i$. For the asymptotic calculations in this problem, suppose that T is fixed and $n \longrightarrow \infty$.

a. Show that the fixed effects estimator of β from Section 10.3 can be written as $\hat{\beta} = (\Sigma_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\widetilde{\boldsymbol{X}}_i)^{-1}\Sigma_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\widetilde{\boldsymbol{Y}}_i$.

b. Show that $\hat{\beta} - \beta = (\Sigma_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\widetilde{\boldsymbol{X}}_i)^{-1}\Sigma_{i=1}^{n}\widetilde{\boldsymbol{X}}_i'\boldsymbol{u}_i$. (*Hint:* $\boldsymbol{M}$ is idempotent.)

c. Let $Q_{\widetilde{X}} = T^{-1}E(\widetilde{\boldsymbol{X}}_i'\widetilde{\boldsymbol{X}}_i)$ and $\hat{Q}_{\widetilde{X}} = \frac{1}{nT}\Sigma_{i=1}^{n}\Sigma_{t=1}^{T}\widetilde{X}_{it}^2$. Show that $\hat{Q}_{\widetilde{X}} \xrightarrow{p} Q_{\widetilde{X}}$.

d. Let $\eta_i = \widetilde{\boldsymbol{X}}_i'\boldsymbol{u}_i/\sqrt{T}$ and $\sigma_\eta^2 = \text{var}(\eta_i)$. Show that $\sqrt{\frac{1}{n}}\Sigma_{i=1}^{n}\eta_i \xrightarrow{d} N(0, \sigma_\eta^2)$.

e. Use your answers to (b) through (d) to prove Equation (10.25); that is, show that $\sqrt{nT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma_\eta^2/Q_{\widetilde{X}}^2)$.

f. Let $\widetilde{\sigma}_{\eta,clustered}^2$ be the infeasible clustered variance estimator, computed using the true errors instead of the residuals so that $\widetilde{\sigma}_{\eta,clustered}^2 = \frac{1}{nT}\Sigma_{i=1}^{n}(\widetilde{\boldsymbol{X}}_i'\boldsymbol{u}_i)^2$. Show that $\widetilde{\sigma}_{\eta,clustered}^2 \xrightarrow{p} \sigma_\eta^2$.

g. Let $\hat{\widetilde{\boldsymbol{u}}}_1 = \widetilde{\boldsymbol{Y}}_i - \hat{\beta}\widetilde{\boldsymbol{X}}_i$ and $\hat{\sigma}_{\eta,\,clustered}^2 = \frac{n}{n-1}\frac{1}{nT}\Sigma_{i=1}^{n}(\widetilde{\boldsymbol{X}}_i'\hat{\widetilde{\boldsymbol{u}}}_i)^2$ [this is Equation (10.27) in matrix form]. Show that $\hat{\sigma}_{\eta,\,clustered}^2 \xrightarrow{p} \sigma_\eta^2$. [*Hint:* Use an argument like that used in Equation (17.16) to show that $\hat{\sigma}_{\eta,\,clustered}^2 - \widetilde{\sigma}_{\eta,\,clustered}^2 \xrightarrow{p} 0$ and then use your answer to (f).]

18.16 This exercise takes up the problem of missing data discussed in Section 9.2. Consider the regression model $Y_i = X_i\beta + u_i$, $i = 1, \ldots, n$, where all variables are scalars and the constant term/intercept is omitted for convenience.

a. Suppose that the least squares assumptions in Key Concept 4.3 are satisfied. Show that the least squares estimator of β is unbiased and consistent.

b. Now suppose that some of the observations are missing. Let I_i denote a binary random variable that indicates the nonmissing observations; that is, $I_i = 1$ if observation i is not missing and $I_i = 0$ if observation i is missing. Assume that $\{I_i, X_i, u_i\}$ are i.i.d.

i. Show that the OLS estimator can be written as

$$\hat{\beta} = \left(\sum_{i=1}^{n} I_i X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} I_i X_i Y_i\right) = \beta + \left(\sum_{i=1}^{n} I_i X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} I_i X_i u_i\right).$$

ii. Suppose that data are missing, "completely at random," in the sense that $\Pr(I_i = 1|X_i, u_i) = p$, where p is a constant. Show that $\hat{\beta}$ is unbiased and consistent.

iii. Suppose that the probability that the i^{th} observation is missing depends of X_i, but not on u_i; that is, $\Pr(I_i = 1|X_i, u_i) = p(X_i)$. Show that $\hat{\beta}$ is unbiased and consistent.

iv. Suppose that the probability that the i^{th} observation is missing depends on both X_i and u_i; that is, $\Pr(I_i = 1|X_i, u_i) = p(X_i, u_i)$. Is $\hat{\beta}$ unbiased? Is $\hat{\beta}$ consistent? Explain.

c. Suppose that $\beta = 1$ and that X_i and u_i are mutually independent standard normal random variables [so that both X_i and u_i are distributed $N(0, 1)$]. Suppose that $I_i = 1$ when $Y_i \geq 0$, but $I_i = 0$ when $Y_i < 0$. Is $\hat{\beta}$ unbiased? Is $\hat{\beta}$ consistent? Explain.

18.17 Consider the regression model in matrix form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{W\gamma} + \boldsymbol{U}$, where $\boldsymbol{X}$ and $\boldsymbol{W}$ are matrices of regressors and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of unknown regression coefficients. Let $\widetilde{\boldsymbol{X}} = \boldsymbol{M_W X}$ and $\widetilde{\boldsymbol{Y}} = \boldsymbol{M_W Y}$, where $\boldsymbol{M_W} = \boldsymbol{I} - \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}$.

a. Show that the OLS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be written as

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{W} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{W}'\boldsymbol{Y} \end{bmatrix}$$

b. Show that

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{X} & \boldsymbol{W}'\boldsymbol{W} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{M_W X})^{-1} & -(\boldsymbol{X}'\boldsymbol{M_W X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1} \\ -(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M_W X})^{-1} & (\boldsymbol{W}'\boldsymbol{W})^{-1} + (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M_W X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1} \end{bmatrix}.$$

(*Hint:* Show that the product of the two matrices is equal to the identity matrix.)

c. Show that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{Y}$.

d. The Frisch–Waugh theorem (Appendix 6.2) says that $\hat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}$. Use the result in (c) to prove the Frisch–Waugh theorem.

APPENDIX

18.1 Summary of Matrix Algebra

This appendix summarizes vectors, matrices, and the elements of matrix algebra used in Chapter 1. The purpose of this appendix is to review some concepts and definitions from a course in linear algebra, not to replace such a course.

Definitions of Vectors and Matrices

A **vector** is a collection of n numbers or elements, collected either in a column (a **column vector**) or in a row (a **row vector**). The n-dimensional column vector $\boldsymbol{b}$ and the n-dimensional row vector $\boldsymbol{c}$ are

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \text{ and } \boldsymbol{c} = [c_1 \quad c_2 \quad \cdots \quad c_n],$$

where b_1 is the first element of $\boldsymbol{b}$ and in general b_i is the i^{th} element of $\boldsymbol{b}$.

Throughout, a boldface symbol denotes a vector or matrix.

A **matrix** is a collection, or an array, of numbers or elements in which the elements are laid out in columns and rows. The dimension of a matrix is $n \times m$, where n is the number of rows and m is the number of columns. The $n \times m$ matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

where a_{ij} is the (i, j) element of $\boldsymbol{A}$, that is, a_{ij} is the element that appears in the i^{th} row and j^{th} column. An $n \times m$ matrix consists of n row vectors or, alternatively, of m column vectors.

To distinguish one-dimensional numbers from vectors and matrices, a one-dimensional number is called a **scalar**.

Types of Matrices

Square, symmetric, and diagonal matrices. A matrix is said to be **square** if the number of rows equals the number of columns. A square matrix is said to be **symmetric** if its (i, j) element equals its (j, i) element. A **diagonal** matrix is a square matrix in which all the off-diagonal elements equal zero; that is, if the square matrix $\boldsymbol{A}$ is diagonal, then $a_{ij} = 0$ for $i \neq j$.

Special matrices. An important matrix is the **identity matrix**, $\boldsymbol{I}_n$, which is an $n \times n$ diagonal matrix with ones on the diagonal. The **null matrix**, $\mathbf{0}_{n \times m}$, is the $n \times m$ matrix with all elements equal to zero.

The transpose. The **transpose** of a matrix switches the rows and the columns. That is, the transpose of a matrix turns the $n \times m$ matrix $\boldsymbol{A}$ into the $m \times n$ matrix, which is denoted by $\boldsymbol{A}'$, where the (i, j) element of $\boldsymbol{A}$ becomes the (j, i) element of $\boldsymbol{A}'$; said differently, the transpose of the matrix $\boldsymbol{A}$ turns the rows of $\boldsymbol{A}$ into the columns of $\boldsymbol{A}'$. If a_{ij} is the (i, j) element of $\boldsymbol{A}$, then $\boldsymbol{A}'$ (the transpose of $\boldsymbol{A}$) is

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}.$$

The transpose of a vector is a special case of the transpose of a matrix. Thus the transpose of a vector turns a column vector into a row vector; that is, if $\boldsymbol{b}$ is an $n \times 1$ column vector, then its transpose is the $1 \times n$ row vector

$$\boldsymbol{b}' = [b_1 \quad b_2 \quad \cdots \quad b_n].$$

The transpose of a row vector is a column vector.

Elements of Matrix Algebra: Addition and Multiplication

Matrix addition. Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ that have the same dimensions (for example, that are both $n \times m$) can be added together. The sum of two matrices is the sum of their elements; that is, if $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$, then $c_{ij} = a_{ij} + b_{ij}$. A special case of matrix addition is vector addition: If $\boldsymbol{a}$ and $\boldsymbol{b}$ are both $n \times 1$ column vectors, then their sum $\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b}$ is the element-wise sum; that is, $c_i = a_i + b_i$.

Vector and matrix multiplication. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two $n \times 1$ column vectors. Then the product of the transpose of $\boldsymbol{a}$ (which is itself a row vector) with $\boldsymbol{b}$ is $\boldsymbol{a}'\boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$. Applying this definition with $\boldsymbol{b} = \boldsymbol{a}$ yields $\boldsymbol{a}'\boldsymbol{a} = \sum_{i=1}^{n} a_i^2$.

Similarly, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be multiplied together if they are conformable—that is, if the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$. Specifically, suppose that $\boldsymbol{A}$ has dimension $n \times m$ and $\boldsymbol{B}$ has dimension $m \times r$. Then the product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is an $n \times r$ matrix, $\boldsymbol{C}$; that is, $\boldsymbol{C} = \boldsymbol{AB}$, where the (i, j) element of $\boldsymbol{C}$ is $c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$. Said differently, the (i, j) element of $\boldsymbol{AB}$ is the product of multiplying the row vector that is the i^{th} row of $\boldsymbol{A}$ with the column vector that is the j^{th} column of $\boldsymbol{B}$.

The product of a scalar d with the matrix $\boldsymbol{A}$ has the (i, j) element da_{ij}; that is, each element of $\boldsymbol{A}$ is multiplied by the scalar d.

Some useful properties of matrix addition and multiplication. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices. Then:

a. $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$;

b. $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$;

c. $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$;

d. If $\boldsymbol{A}$ is $n \times m$, then $\boldsymbol{AI}_m = \boldsymbol{A}$ and $\boldsymbol{I}_n\boldsymbol{A} = \boldsymbol{A}$;

e. $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$;

f. $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$; and

g. $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$.

In general, matrix multiplication does not commute; that is, in general $\boldsymbol{AB} \neq \boldsymbol{BA}$, although there are some special cases in which matrix multiplication commutes; for example, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are both $n \times n$ diagonal matrices, then $\boldsymbol{AB} = \boldsymbol{BA}$.

Matrix Inverse, Matrix Square Roots, and Related Topics

The matrix inverse. Let $\boldsymbol{A}$ be a square matrix. Assuming that it exists, the **inverse** of the matrix $\boldsymbol{A}$ is defined as the matrix for which $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$. If in fact the inverse matrix $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}$ is said to be **invertible** or **nonsingular**. If both $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible, then $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

Positive definite and positive semidefinite matrices. Let $\boldsymbol{V}$ be an $n \times n$ square matrix. Then $\boldsymbol{V}$ is **positive definite** if $\boldsymbol{c}'\boldsymbol{Vc} > 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{c}$. Similarly, $\boldsymbol{V}$ is **positive semidefinite** if $\boldsymbol{c}'\boldsymbol{Vc} \geq 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{c}$. If $\boldsymbol{V}$ is positive definite, then it is invertible.

Linear independence. The $n \times 1$ vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are **linearly independent** if there do not exist nonzero scalars c_1 and c_2 such that $c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 = \boldsymbol{0}_{n\times 1}$. More generally, the set of k vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k$ are linearly independent if there do not exist nonzero scalars $c_1, c_2, \ldots, c_k$ such that $c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_k\boldsymbol{a}_k = \boldsymbol{0}_{n\times 1}$.

The rank of a matrix. The **rank** of the $n \times m$ matrix $\boldsymbol{A}$ is the number of linearly independent columns of $\boldsymbol{A}$. The rank of $\boldsymbol{A}$ is denoted rank($\boldsymbol{A}$). If the rank of $\boldsymbol{A}$ equals the number of columns of $\boldsymbol{A}$, then $\boldsymbol{A}$ is said to have full column rank. If the $n \times m$ matrix $\boldsymbol{A}$ has full column rank, then there does not exist a nonzero $m \times 1$ vector $\boldsymbol{c}$ such that $\boldsymbol{Ac} = \boldsymbol{0}_{n\times1}$. If $\boldsymbol{A}$ is $n \times n$ with $\text{rank}(\boldsymbol{A}) = n$, then $\boldsymbol{A}$ is nonsingular. If the $n \times m$ matrix $\boldsymbol{A}$ has full column rank, then $\boldsymbol{A}'\boldsymbol{A}$ is nonsingular.

The matrix square root. Let $\boldsymbol{V}$ be an $n \times n$ square symmetric positive definite matrix. The matrix square root of $\boldsymbol{V}$ is defined to be an $n \times n$ matrix $\boldsymbol{F}$ such that $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{V}$. The matrix square root of a positive definite matrix will always exist, but it is not unique. The matrix square root has the property that $\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F}' = \boldsymbol{I}_n$. In addition, the matrix square root of a positive definite matrix is invertible, so $\boldsymbol{F}'^{-1}\boldsymbol{V}\boldsymbol{F}^{-1} = \boldsymbol{I}_n$.

Eigenvalues and eigenvectors. Let $\boldsymbol{A}$ be an $n \times n$ matrix. If the $n \times 1$ vector $\boldsymbol{q}$ and the scalar λ satisfy $\boldsymbol{Aq} = \lambda\boldsymbol{q}$, where $\boldsymbol{q}'\boldsymbol{q} = 1$, then λ is an **eigenvalue** of $\boldsymbol{A}$, and $\boldsymbol{q}$ is the **eigenvector** of $\boldsymbol{A}$ associated with that eigenvalue. An $n \times n$ matrix has n eigenvalues, which need not take on distinct values, and n eigenvectors.

If $\boldsymbol{V}$ is an $n \times n$ symmetric positive definite matrix, then all the eigenvalues of $\boldsymbol{V}$ are positive real numbers, and all the eigenvectors of $\boldsymbol{V}$ are real. Also, $\boldsymbol{V}$ can be written in terms of its eigenvalues and eigenvectors as $\boldsymbol{V} = \boldsymbol{Q\Lambda Q}'$, where $\boldsymbol{\Lambda}$ is a diagonal $n \times n$ matrix with diagonal elements that equal the eigenvalues of $\boldsymbol{V}$, and $\boldsymbol{Q}$ is an $n \times n$ matrix consisting of the eigenvectors of $\boldsymbol{V}$, arranged so that the i^{th} column of $\boldsymbol{Q}$ is the eigenvector corresponding to the eigenvalue that is the i^{th} diagonal element of $\boldsymbol{\Lambda}$. The eigenvectors are orthonormal, so $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_n$.

Idempotent matrices. A matrix $\boldsymbol{C}$ is idempotent if $\boldsymbol{C}$ is square and $\boldsymbol{CC} = \boldsymbol{C}$. If $\boldsymbol{C}$ is an $n \times n$ idempotent matrix that is also symmetric, then $\boldsymbol{C}$ is positive semidefinite and $\boldsymbol{C}$ has r eigenvalues that equal 1 and $n - r$ eigenvalues that equal 0, where $r = \text{rank}(\boldsymbol{C})$ (Exercise 18.10).

APPENDIX 18.2 Multivariate Distributions

This appendix collects various definitions and facts about distributions of vectors of random variables. We start by defining the mean and covariance matrix of the n-dimensional random variable $\boldsymbol{V}$. Next we present the multivariate normal distribution. Finally, we summarize some facts about the distributions of linear and quadratic functions of jointly normally distributed random variables.

The Mean Vector and Covariance Matrix

The first and second moments of an $m \times 1$ vector of random variables, $\boldsymbol{V} = (V_1 \quad V_2 \quad \cdots \quad V_m)'$, are summarized by its mean vector and covariance matrix.

Because $\boldsymbol{V}$ is a vector, the vector of its means—that is, its **mean vector**—is $E(\boldsymbol{V}) = \boldsymbol{\mu}_V$. The i^{th} element of the mean vector is the mean of the i^{th} element of $\boldsymbol{V}$.

The **covariance matrix** of $\boldsymbol{V}$ is the matrix consisting of the variance $\text{var}(V_i)$, $i = 1, \ldots, m$, along the diagonal and the (i, j) off-diagonal elements $\text{cov}(V_i, V_j)$. In matrix form, the covariance matrix $\boldsymbol{\Sigma}_V$ is

$$\boldsymbol{\Sigma}_V = E[(\boldsymbol{V} - \boldsymbol{\mu}_V)(\boldsymbol{V} - \boldsymbol{\mu}_V)'] = \begin{bmatrix} \text{var}(V_1) & \cdots & \text{cov}(V_1, V_m) \\ \vdots & \ddots & \vdots \\ \text{cov}(V_m, V_1) & \cdots & \text{var}(V_m) \end{bmatrix}. \quad (18.72)$$

The Multivariate Normal Distribution

The $m \times 1$ vector random variable $\boldsymbol{V}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}_V$ and covariance matrix $\boldsymbol{\Sigma}_V$ if it has the joint probability density function

$$f(\boldsymbol{V}) = \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma}_V)}} \exp\left[-\frac{1}{2}(\boldsymbol{V} - \boldsymbol{\mu}_V)'\boldsymbol{\Sigma}_V^{-1}(\boldsymbol{V} - \boldsymbol{\mu}_V)\right], \quad (18.73)$$

where $\det(\boldsymbol{\Sigma}_V)$ is the determinant of the matrix $\boldsymbol{\Sigma}_V$. The multivariate normal distribution is denoted $N(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$.

An important fact about the multivariate normal distribution is that if two jointly normally distributed random variables are uncorrelated (equivalently, have a block-diagonal covariance matrix), then they are independently distributed. That is, let $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ be jointly normally distributed random variables with respective dimensions $m_1 \times 1$ and $m_2 \times 1$. Then if $\text{cov}(\boldsymbol{V}_1, \boldsymbol{V}_2) = E[(\boldsymbol{V}_1 - \boldsymbol{\mu}_{V_1})(\boldsymbol{V}_2 - \boldsymbol{\mu}_{V_2})'] = \boldsymbol{0}_{m_1 \times m_2}$, $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are independent.

If $\{V_i\}$ are i.i.d. $N(0, \sigma_v^2)$, then $\boldsymbol{\Sigma}_V = \sigma_v^2 \boldsymbol{I}_m$, and the multivariate normal distribution simplifies to the product of m univariate normal densities.

Distributions of Linear Combinations and Quadratic Forms of Normal Random Variables

Linear combinations of multivariate normal random variables are themselves normally distributed, and certain quadratic forms of multivariate normal random variables have a chi-squared distribution. Let $\boldsymbol{V}$ be an $m \times 1$ random variable distributed $N(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$, let $\boldsymbol{A}$

and $\boldsymbol{B}$ be nonrandom $a \times m$ and $b \times m$ matrices, and let $\boldsymbol{d}$ be a nonrandom $a \times 1$ vector. Then

$$\boldsymbol{d} + \boldsymbol{AV} \text{ is distributed } N(\boldsymbol{d} + \boldsymbol{A\mu_V}, \boldsymbol{A\Sigma_V A}'); \tag{18.74}$$

$$\text{cov}(\boldsymbol{AV}, \boldsymbol{BV}) = \boldsymbol{A\Sigma_V B}'; \tag{18.75}$$

$$\text{if } \boldsymbol{A\Sigma_V B}' = \boldsymbol{0}_{a\times b}, \text{ then } \boldsymbol{AV} \text{ and } \boldsymbol{BV} \text{ are independently distributed; and} \tag{18.76}$$

$$(\boldsymbol{V} - \boldsymbol{\mu_V})'\boldsymbol{\Sigma_V^{-1}}(\boldsymbol{V} - \boldsymbol{\mu_V}) \text{ is distributed } \chi^2_m. \tag{18.77}$$

Let $\boldsymbol{U}$ be an m-dimensional multivariate standard normal random variable with distribution $N(\boldsymbol{0}, \boldsymbol{I}_m)$. If $\boldsymbol{C}$ is symmetric and idempotent, then

$$\boldsymbol{U'CU} \text{ has a } \chi^2_r \text{ distribution, where } r = \text{rank}(\boldsymbol{C}). \tag{18.78}$$

Equation (18.78) is proven as Exercise 18.11.

APPENDIX

18.3 Derivation of the Asymptotic Distribution of $\hat{\beta}$

This appendix provides the derivation of the asymptotic normal distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ given in Equation (18.12). An implication of this result is that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

First consider the "denominator" matrix $\boldsymbol{X'X}/n = \frac{1}{n}\Sigma_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'$ in Equation (18.15). The (j, l) element of this matrix is $\frac{1}{n}\Sigma_{i=1}^n X_{ji}X_{li}$. By the second assumption in Key Concept 18.1, $\boldsymbol{X}_i$ is i.i.d., so $X_{ji}X_{li}$ is i.i.d. By the third assumption in Key Concept 18.1, each element of $\boldsymbol{X}_i$ has four moments, so, by the Cauchy–Schwarz inequality (Appendix 17.2), $X_{ji}X_{li}$ has two moments. Because $X_{ji}X_{li}$ is i.i.d. with two moments, $\frac{1}{n}\Sigma_{i=1}^n X_{ji}X_{li}$ obeys the law of large numbers, so $\frac{1}{n}\Sigma_{i=1}^n X_{ji}X_{li} \xrightarrow{p} E(X_{ji}X_{li})$. This is true for all the elements of $\boldsymbol{X'X}/n$, so $\boldsymbol{X'X}/n \xrightarrow{p} E(\boldsymbol{X}_i\boldsymbol{X}_i') = \boldsymbol{Q_X}$.

Next consider the "numerator" matrix in Equation (18.15), $\boldsymbol{X'U}/\sqrt{n} = \sqrt{\frac{1}{n}}\Sigma_{i=1}^n \boldsymbol{V}_i$, where $\boldsymbol{V}_i = \boldsymbol{X}_i u_i$. By the first assumption in Key Concept 18.1 and the law of iterated expectations, $E(\boldsymbol{V}_i) = E[\boldsymbol{X}_i E(u_i|\boldsymbol{X}_i)] = \boldsymbol{0}_{k+1}$. By the second least squares assumption, $\boldsymbol{V}_i$ is i.i.d. Let $\boldsymbol{c}$ be a finite $k + 1$ dimensional vector. By the Cauchy–Schwarz inequality, $E[(\boldsymbol{c}'\boldsymbol{V}_i)^2] = E[(\boldsymbol{c}'\boldsymbol{X}_i u_i)^2] = E[(\boldsymbol{c}'\boldsymbol{X}_i)^2(u_i)^2] \leq \sqrt{E[(\boldsymbol{c}'\boldsymbol{X}_i)^4]E(u_i^4)}$, which is finite by the third least squares assumption. This is true for every such vector $\boldsymbol{c}$, so $E(\boldsymbol{V}_i\boldsymbol{V}_i') = \boldsymbol{\Sigma_V}$ is finite and, we assume, positive definite. Thus the multivariate central limit theorem of Key Concept 18.2 applies to $\sqrt{\frac{1}{n}}\Sigma_{i=1}^n \boldsymbol{V}_i = \frac{1}{\sqrt{n}}\boldsymbol{X'U}$; that is,

$$\frac{1}{\sqrt{n}}\boldsymbol{X'U} \xrightarrow{d} N(\boldsymbol{0}_{k+1}, \boldsymbol{\Sigma_V}). \tag{18.79}$$

The result in Equation (18.12) follows from Equations (18.15) and (18.79), the consistency of $\boldsymbol{X}'\boldsymbol{X}/n$, the fourth least squares assumption (which ensures that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists), and Slutsky's theorem.

APPENDIX

18.4 Derivations of Exact Distributions of OLS Test Statistics with Normal Errors

This appendix presents the proofs of the distributions under the null hypothesis of the homoskedasticity-only *t*-statistic in Equation (18.35) and the homoskedasticity-only *F*-statistic in Equation (18.37), assuming that all six assumptions in Key Concept 18.1 hold.

Proof of Equation (18.35)

If (i) Z has a standard normal distribution, (ii) W has a χ_m^2 distribution, and (iii) Z and W are independently distributed, then the random variable $Z/\sqrt{W/m}$ has the *t*-distribution with m degrees of freedom (Appendix 17.1). To put $\widetilde{t}$ in this form, notice that $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = (s_{\hat{u}}^2/\sigma_u^2)\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}}$. Then rewrite Equation (18.34) as

$$\widetilde{t} = \frac{(\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}}{\sqrt{W/(n-k-1)}}, \tag{18.80}$$

where $W = (n-k-1)(s_{\hat{u}}^2/\sigma_u^2)$, and let $Z = (\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}$ and $m = n-k-1$. With these definitions, $\widetilde{t} = Z/\sqrt{W/m}$. Thus, to prove the result in Equation (18.35), we must show (i) through (iii) for these definitions of Z, W, and m.

i. An implication of Equation (18.30) is that, under the null hypothesis, $Z = (\hat{\beta}_j - \beta_{j,0})/\sqrt{(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}|\boldsymbol{X}})_{jj}}$ has an exact standard normal distribution, which shows (i).
ii. From Equation (18.31), W is distributed as χ_{n-k-1}^2, which shows (ii).
iii. To show (iii), it must be shown that $\hat{\beta}_j$ and $s_{\hat{u}}^2$ are independently distributed.

From Equations (18.14) and (18.29), $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $s_{\hat{u}}^2 = (\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U})'(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U})/(n-k-1)$. Thus $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $s_{\hat{u}}^2$ are independent if $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are independent. Both $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are linear combinations of $\boldsymbol{U}$, which has an $N(\boldsymbol{0}_{n\times 1}, \sigma_u^2\boldsymbol{I}_n)$ distribution, conditional on $\boldsymbol{X}$. But because $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{0}_{n\times(k+1)}$ [Equation (18.26)], it follows that $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{U}$ and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{U}$ are independently distributed [Equation (18.76)]. Consequently, under all six assumptions in Key Concept 18.1,

$$\hat{\boldsymbol{\beta}} \text{ and } s_{\hat{u}}^2 \text{ are independently distributed,} \tag{18.81}$$

which shows (iii) and thus proves Equation (18.35).

Proof of Equation (18.37)

The F_{n_1,n_2} distribution is the distribution of $(W_1/n_1)/(W_2/n_2)$, where (i) W_1 is distributed $\chi^2_{n_1}$; (ii) W_2 is distributed $\chi^2_{n_2}$; and (iii) W_1 and W_2 are independently distributed (Appendix 17.1). To express $\widetilde{F}$ in this form, let $W_1 = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma^2_u]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ and $W_2 = (n - k - 1)s^2_{\hat{u}}/\sigma^2_u$ Substitution of these definitions into Equation (18.36) shows that $\widetilde{F} = (W_1/q)/[W_2/(n - k - 1)]$. Thus, by the definition of the F distribution, $\widetilde{F}$ has an $F_{q,\,n-k-1}$ distribution if (i) through (iii) hold with $n_1 = q$ and $n_2 = n - k - 1$.

i. Under the null hypothesis, $\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r} = \boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Because $\hat{\boldsymbol{\beta}}$ has the conditional normal distribution in Equation (18.30) and because $\boldsymbol{R}$ is a nonrandom matrix, $\boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is distributed $N(\boldsymbol{0}_{q\times 1}, \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma^2_u)$, conditional on $\boldsymbol{X}$. Thus, by Equation (18.77) in Appendix 18.2, $(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{R}'\sigma^2_u]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})$ is distributed χ^2_q, proving (i).

ii. Requirement (ii) is shown in Equation (18.31).

iii. It has already been shown that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $s^2_{\hat{u}}$ are independently distributed [Equation (18.81)]. It follows that $\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}$ and $s^2_{\hat{u}}$ are independently distributed, which in turn implies that W_1 and W_2 are independently distributed, proving (iii) and completing the proof.

APPENDIX

18.5 Proof of the Gauss–Markov Theorem for Multiple Regression

This appendix proves the Gauss–Markov theorem (Key Concept 18.3) for the multiple regression model. Let $\widetilde{\boldsymbol{\beta}}$ be a linear conditionally unbiased estimator of $\boldsymbol{\beta}$ so that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{Y}$ and $E(\widetilde{\boldsymbol{\beta}}\,|\,\boldsymbol{X}) = \boldsymbol{\beta}$, where $\boldsymbol{A}$ is an $n \times (k + 1)$ matrix that can depend on $\boldsymbol{X}$ and nonrandom constants. We show that $\mathrm{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) \le \mathrm{var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}})$ for all $k + 1$ dimensional vectors $\boldsymbol{c}$, where the inequality holds with equality only if $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

Because $\widetilde{\boldsymbol{\beta}}$ is linear, it can be written as $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{Y} = \boldsymbol{A}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}) = (\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta} + \boldsymbol{A}'\boldsymbol{U}$. By the first Gauss–Markov condition, $E(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}_{n\times 1}$, so $E(\widetilde{\boldsymbol{\beta}}\,|\boldsymbol{X}) = (\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta}$, but because $\widetilde{\boldsymbol{\beta}}$ is conditionally unbiased, $E(\widetilde{\boldsymbol{\beta}}\,|\,\boldsymbol{X}) = \boldsymbol{\beta} = (\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta}$, which implies that $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_{k+1}$. Thus $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{A}'\boldsymbol{U}$, so $\mathrm{var}(\widetilde{\boldsymbol{\beta}}\,|\boldsymbol{X}) = \mathrm{var}(\boldsymbol{A}'\boldsymbol{U}|\boldsymbol{X}) = E(\boldsymbol{A}'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{A}|\boldsymbol{X}) = \boldsymbol{A}'E(\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{X})\boldsymbol{A} = \boldsymbol{\sigma}^2_u\boldsymbol{A}'\boldsymbol{A}$, where the third equality follows because $\boldsymbol{A}$ can depend on $\boldsymbol{X}$ but not $\boldsymbol{U}$, and the final equality follows from the second Gauss–Markov condition. That is, if $\widetilde{\boldsymbol{\beta}}$ is linear and unbiased, then under the Gauss–Markov conditions,

$$\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_{k+1} \text{ and } \mathrm{var}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2_u\boldsymbol{A}'\boldsymbol{A}. \tag{18.82}$$

The results in Equation (18.82) also apply to $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{A} = \hat{\boldsymbol{A}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists by the third Gauss–Markov condition.

Now let $\boldsymbol{A} = \hat{\boldsymbol{A}} + \boldsymbol{D}$ so that $\boldsymbol{D}$ is the difference between the matrices $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$. Note that $\hat{\boldsymbol{A}}'\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ [by Equation (18.82)] and $\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, so $\hat{\boldsymbol{A}}'\boldsymbol{D} = \hat{\boldsymbol{A}}'(\boldsymbol{A} - \hat{\boldsymbol{A}}) = \hat{\boldsymbol{A}}'\boldsymbol{A} - \hat{\boldsymbol{A}}'\hat{\boldsymbol{A}} = \boldsymbol{0}_{(k+1)\times(k+1)}$. Substituting $\boldsymbol{A} = \hat{\boldsymbol{A}} + \boldsymbol{D}$ into the formula for the conditional variance in Equation (18.82) yields

$$\begin{aligned} \operatorname{var}(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) &= \sigma_u^2(\hat{\boldsymbol{A}} + \boldsymbol{D})'(\hat{\boldsymbol{A}} + \boldsymbol{D}) \\ &= \sigma_u^2[\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}} + \hat{\boldsymbol{A}}'\boldsymbol{D} + \boldsymbol{D}'\hat{\boldsymbol{A}} + \boldsymbol{D}'\boldsymbol{D}] \\ &= \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1} + \sigma_u^2\boldsymbol{D}'\boldsymbol{D}, \end{aligned} \qquad (18.83)$$

where the final equality uses the facts $\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{A}}'\boldsymbol{D} = \boldsymbol{0}_{(k+1)\times(k+1)}$.

Because $\operatorname{var}(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}) = \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, Equations (18.82) and (18.83) imply that $\operatorname{var}(\tilde{\boldsymbol{\beta}}\,|\boldsymbol{X}) - \operatorname{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_u^2\boldsymbol{D}'\boldsymbol{D}$. The difference between the variances of the two estimators of the linear combination $\boldsymbol{c}'\boldsymbol{\beta}$ thus is

$$\operatorname{var}(\boldsymbol{c}'\tilde{\boldsymbol{\beta}}|\boldsymbol{X}) - \operatorname{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_u^2\,\boldsymbol{c}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{c} \geq \boldsymbol{0}. \qquad (18.84)$$

The inequality in Equation (18.84) holds for all linear combinations $\boldsymbol{c}'\boldsymbol{\beta}$, and the inequality holds with equality for all nonzero $\boldsymbol{c}$ only if $\boldsymbol{D} = \boldsymbol{0}_{n\times(k+1)}$—that is, if $\boldsymbol{A} = \hat{\boldsymbol{A}}$ or, equivalently, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. Thus $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ has the smallest variance of all linear conditionally unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$; that is, the OLS estimator is BLUE.

APPENDIX 18.6 Proof of Selected Results for IV and GMM Estimation

The Efficiency of TSLS Under Homoskedasticity [Proof of Equation (18.62)]

When the errors u_i are homoskedastic, the difference between $\boldsymbol{\Sigma}_A^{IV}$ [Equation (18.61)] and $\boldsymbol{\Sigma}^{TSLS}$ [Equation (18.55)] is given by

$$\begin{aligned} &\boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS} = (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZZ}\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2 - (\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2 \\ &= (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}[\boldsymbol{Q}_{ZZ} - \boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}]\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2, \end{aligned} \qquad (18.85)$$

where the second term in brackets in the second equality follows from $(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX} = \boldsymbol{I}_{(k+r+1)}$. Let $\boldsymbol{F}$ be the matrix square root of $\boldsymbol{Q}_{ZZ}$, so $\boldsymbol{Q}_{ZZ} = \boldsymbol{F}'\boldsymbol{F}$ and $\boldsymbol{Q}_{ZZ}^{-1} = \boldsymbol{F}^{-1}\boldsymbol{F}^{-1'}$. [The latter equality follows from noting that $(\boldsymbol{F}'\boldsymbol{F})^{-1} = \boldsymbol{F}^{-1}\boldsymbol{F}'^{-1}$ and $\boldsymbol{F}'^{-1} = \boldsymbol{F}^{-1'}$.] Then the final expression in Equation (18.85) can be rewritten to yield

$$\begin{aligned} \boldsymbol{\Sigma}_A^{IV} - \boldsymbol{\Sigma}^{TSLS} = {} & (\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{F}'[\boldsymbol{I} - \boldsymbol{F}^{-1'}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{F}^{-1}\boldsymbol{F}^{-1'}\boldsymbol{Q}_{ZX})^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{F}^{-1}] \\ & \times \boldsymbol{F}\boldsymbol{A}\boldsymbol{Q}_{ZX}(\boldsymbol{Q}_{XZ}\boldsymbol{A}\boldsymbol{Q}_{ZX})^{-1}\sigma_u^2, \end{aligned} \qquad (18.86)$$