Appendix E

The Linear Regression Model in Matrix Form

This appendix derives various results for ordinary least squares estimation of the multiple linear regression model using matrix notation and matrix algebra (see Appendix D for a summary). The material presented here is much more advanced than that in the text.

E-1 The Model and Ordinary Least Squares Estimation

Throughout this appendix, we use the t subscript to index observations and an n to denote the sample size. It is useful to write the multiple linear regression model with k parameters as follows:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t, t = 1, 2, \ldots, n, \qquad [E.1]$$

where y_t is the dependent variable for observation t and $x_{tj}, j = 1, 2, \ldots, k$, are the independent variables. As usual, β_0 is the intercept and $\beta_1, \ldots, \beta_k$ denote the slope parameters.

For each t, define a $1 \times (k + 1)$ vector, $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ be the $(k + 1) \times 1$ vector of all parameters. Then, we can write (E.1) as

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t, t = 1, 2, \ldots, n. \qquad [E.2]$$

[Some authors prefer to define $\mathbf{x}_t$ as a column vector, in which case $\mathbf{x}_t$ is replaced with $\mathbf{x}_t'$ in (E.2). Mathematically, it makes more sense to define it as a row vector.] We can write (E.2) in full matrix notation by appropriately defining data vectors and matrices. Let $\mathbf{y}$ denote the $n \times 1$ vector of observations on y: the t^{th} element of $\mathbf{y}$ is y_t. Let $\mathbf{X}$ be the $n \times (k + 1)$ vector of observations on the explanatory variables. In other words, the t^{th} row of $\mathbf{X}$ consists of the vector $\mathbf{x}_t$. Written out in detail,

$$\underset{n \times (k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}.$$

Finally, let $\mathbf{u}$ be the $n \times 1$ vector of unobservable errors or disturbances. Then, we can write (E.2) for all n observations in **matrix notation**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{E.3}$$

Remember, because $\mathbf{X}$ is $n \times (k + 1)$ and $\boldsymbol{\beta}$ is $(k + 1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ is $n \times 1$.

Estimation of $\boldsymbol{\beta}$ proceeds by minimizing the sum of squared residuals, as in Section 3-2. Define the sum of squared residuals function for any possible $(k + 1) \times 1$ parameter vector $\mathbf{b}$ as

$$\text{SSR}(\mathbf{b}) \equiv \sum_{t=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2.$$

The $(k + 1) \times 1$ vector of ordinary least squares estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$, minimizes SSR($\mathbf{b}$) over all possible $(k + 1) \times 1$ vectors $\mathbf{b}$. This is a problem in multivariable calculus. For $\hat{\boldsymbol{\beta}}$ to minimize the sum of squared residuals, it must solve the **first order condition**

$$\partial \text{SSR}(\hat{\boldsymbol{\beta}})/\partial \mathbf{b} \equiv 0. \tag{E.4}$$

Using the fact that the derivative of $(y_t - \mathbf{x}_t\mathbf{b})^2$ with respect to $\mathbf{b}$ is the $1 \times (k + 1)$ vector $-2(y_t - \mathbf{x}_t\mathbf{b})\mathbf{x}_t$, (E.4) is equivalent to

$$\sum_{t=1}^{n} \mathbf{x}_t'(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}) \equiv \mathbf{0}. \tag{E.5}$$

(We have divided by −2 and taken the transpose.) We can write this first order condition as

$$\begin{aligned}
&\sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0 \\
&\sum_{t=1}^{n} x_{t1}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0 \\
&\quad \vdots \\
&\sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0,
\end{aligned}$$

which is identical to the first order conditions in equation (3.13). We want to write these in matrix form to make them easier to manipulate. Using the formula for partitioned multiplication in Appendix D, we see that (E.5) is equivalent to

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{E.6}$$

or

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{E.7}$$

It can be shown that (E.7) always has at least one solution. Multiple solutions do not help us, as we are looking for a unique set of OLS estimates given our data set. Assuming that the $(k + 1) \times (k + 1)$ symmetric matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, we can premultiply both sides of (E.7) by $(\mathbf{X}'\mathbf{X})^{-1}$ to solve for the OLS estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{E.8}$$

This is the critical formula for matrix analysis of the multiple linear regression model. The assumption that $\mathbf{X}'\mathbf{X}$ is invertible is equivalent to the assumption that rank$(\mathbf{X}) = (k + 1)$, which means that the columns of $\mathbf{X}$ must be linearly independent. This is the matrix version of MLR.3 in Chapter 3.

Before we continue, (E.8) warrants a word of warning. It is tempting to simplify the formula for $\hat{\boldsymbol{\beta}}$ as follows:

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

The flaw in this reasoning is that $\mathbf{X}$ is usually not a square matrix, so it cannot be inverted. In other words, we cannot write $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ unless $n = (k + 1)$, a case that virtually never arises in practice.

The $n \times 1$ vectors of OLS fitted values and residuals are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ respectively.}$$

From (E.6) and the definition of $\hat{\mathbf{u}}$, we can see that the first order condition for $\hat{\boldsymbol{\beta}}$ is the same as

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \qquad \text{[E.9]}$$

Because the first column of $\mathbf{X}$ consists entirely of ones, (E.9) implies that the OLS residuals always sum to zero when an intercept is included in the equation and that the sample covariance between each independent variable and the OLS residuals is zero. (We discussed both of these properties in Chapter 3.)

The sum of squared residuals can be written as

$$\text{SSR} = \sum_{t=1}^{n} \hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \qquad \text{[E.10]}$$

All of the algebraic properties from Chapter 3 can be derived using matrix algebra. For example, we can show that the total sum of squares is equal to the explained sum of squares plus the sum of squared residuals [see (3.27)]. The use of matrices does not provide a simpler proof than summation notation, so we do not provide another derivation.

The matrix approach to multiple regression can be used as the basis for a geometrical interpretation of regression. This involves mathematical concepts that are even more advanced than those we covered in Appendix D. [See Goldberger (1991) or Greene (1997).]

E-1a The Frisch-Waugh Theorem

In Section 3-2, we described a "partialling out" interpretation of the ordinary least squares estimates. We can establish the partialling out interpretation very generally using matrix notation. Partition the $n \times (k + 1)$ matrix $\mathbf{X}$ as

$$\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2),$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and includes the intercept—although that is not required for the result to hold—and $\mathbf{X}_2$ is $n \times k_2$. We still assume that $\mathbf{X}$ has rank $k + 1$, which means $\mathbf{X}_1$ has rank $k_1 + 1$ and $\mathbf{X}_2$ has rank k_2.

Consider the OLS estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ from the (long) regression

$$\mathbf{y} \text{ on } \mathbf{X}_1, \mathbf{X}_2.$$

As we know, the multiple regression coefficients on $\mathbf{X}_2$, $\hat{\boldsymbol{\beta}}_2$, generally differs from $\tilde{\boldsymbol{\beta}}_2$ from the regression $\mathbf{y}$ on $\mathbf{X}_2$. One way to describe the difference is to understand that we can obtain $\hat{\boldsymbol{\beta}}_2$ from a shorter regression, but first we must "partial out" $\mathbf{X}_1$ from $\mathbf{X}_2$. Consider the following two-step method:

(i) Regress (each column of) $\mathbf{X}_2$ on $\mathbf{X}_1$ and obtain the matrix of residuals, say $\ddot{\mathbf{X}}_2$. We can write $\ddot{\mathbf{X}}_2$ as

$$\ddot{\mathbf{X}}_2 = [\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2,$$

where $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ are $n \times n$ symmetric, idempotent matrices.

(ii) Regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ and call the $k_2 \times 1$ vector of coefficient $\ddot{\boldsymbol{\beta}}_2$.

The **Frisch-Waugh (FW) theorem** states that

$$\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2.$$

Importantly, the FW theorem generally says nothing about equality of the estimates from the long regression, $\hat{\boldsymbol{\beta}}_2$, and those from the short regression, $\tilde{\boldsymbol{\beta}}_2$. Usually $\hat{\boldsymbol{\beta}}_2 \neq \tilde{\boldsymbol{\beta}}_2$. However, if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ then $\ddot{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$, in which case $\ddot{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$; then $\hat{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$ follows from FW. It is also worth noting that we obtain $\hat{\boldsymbol{\beta}}_2$ if we also partial $\mathbf{X}_1$ out of $\mathbf{y}$. In other words, let $\ddot{\mathbf{y}}$ be the residuals from regressing $\mathbf{y}$ on $\mathbf{X}_1$, so that

$$\ddot{\mathbf{y}} = \mathbf{M}_1\mathbf{y}.$$

Then $\hat{\boldsymbol{\beta}}_2$ is obtained from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$. It is important to understand that it is not enough to only partial out $\mathbf{X}_1$ from $\mathbf{y}$. The important step is partialling out $\mathbf{X}_1$ from $\mathbf{X}_2$. Problem 6 at the end of this chapter asks you to derive the FW theorem and to investigate some related issues.

Another useful algebraic result is that when we regress $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ and save the residuals, say $\ddot{\mathbf{u}}$, these are identical to the OLS residuals from the original (long) regression:

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \ddot{\mathbf{u}} = \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2,$$

where we have used the FW result $\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2$. We do not obtain the original OLS residuals if we regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ (but we do obtain $\hat{\boldsymbol{\beta}}_2$).

Before the advent of powerful computers, the Frisch-Waugh result was sometimes used as a computational device. Today, the result is more of theoretical interest, and it is very helpful in understanding the mechanics of OLS. For example, recall that in Chapter 10 we used the FW theorem to establish that adding a time trend to a multiple regression is algebraically equivalent to first linearly detrending all of the explanatory variables before running the regression. The FW theorem also can be used in Chapter 14 to establish that the fixed effects estimator, which we introduced as being obtained from OLS on time-demeaned data, can also be obtained from the (long) dummy variable regression.

E-2 Finite Sample Properties of OLS

Deriving the expected value and variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ is facilitated by matrix algebra, but we must show some care in stating the assumptions.

Assumption E.1 **Linear in Parameters**

The model can be written as in (E.3), where $\mathbf{y}$ is an observed $n \times 1$ vector, $\mathbf{X}$ is an $n \times (k + 1)$ observed matrix, and $\mathbf{u}$ is an $n \times 1$ vector of unobserved errors or disturbances.

Assumption E.2 **No Perfect Collinearity**

The matrix $\mathbf{X}$ has rank $(k + 1)$.

This is a careful statement of the assumption that rules out linear dependencies among the explanatory variables. Under Assumption E.2, $\mathbf{X}'\mathbf{X}$ is nonsingular, so $\hat{\boldsymbol{\beta}}$ is unique and can be written as in (E.8).

Assumption E.3 **Zero Conditional Mean**

Conditional on the entire matrix $\mathbf{X}$, each error u_t has zero mean: $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots, n$.

In vector form, Assumption E.3 can be written as

$$\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \qquad [E.11]$$

This assumption is implied by MLR.4 under the random sampling assumption, MLR.2. In time series applications, Assumption E.3 imposes strict exogeneity on the explanatory variables, something discussed at length in Chapter 10. This rules out explanatory variables whose future values are correlated with u_t; in particular, it eliminates lagged dependent variables. Under Assumption E.3, we can condition on the x_{tj} when we compute the expected value of $\hat{\boldsymbol{\beta}}$.

THEOREM E.1 **UNBIASEDNESS OF OLS**

Under Assumptions E.1, E.2, and E.3, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

PROOF: Use Assumptions E.1 and E.2 and simple algebra to write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}, \qquad [E.12]\end{aligned}$$

where we use the fact that $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. Taking the expectation conditional on $\mathbf{X}$ gives

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta},\end{aligned}$$

because $\mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ under Assumption E.3. This argument clearly does not depend on the value of $\boldsymbol{\beta}$, so we have shown that $\hat{\boldsymbol{\beta}}$ is unbiased.

To obtain the simplest form of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, we impose the assumptions of homoskedasticity and no serial correlation.

Assumption E.4 **Homoskedasticity and No Serial Correlation**

(i) $\mathrm{Var}(u_t|\mathbf{X}) = \sigma^2$, $t = 1, 2, \ldots, n$. (ii) $\mathrm{Cov}(u_t, u_s|\mathbf{X}) = 0$, for all $t \neq s$. In matrix form, we can write these two assumptions as

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}_n, \qquad [E.13]$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Part (i) of Assumption E.4 is the homoskedasticity assumption: the variance of u_t cannot depend on any element of $\mathbf{X}$, and the variance must be constant across observations, t. Part (ii) is the no serial correlation assumption: the errors cannot be correlated across observations. Under random sampling, and in any other cross-sectional sampling schemes with independent observations, part (ii) of Assumption E.4 automatically holds. For time series applications, part (ii) rules out correlation in the errors over time (both conditional on $\mathbf{X}$ and unconditionally).

Because of (E.13), we often say that $\mathbf{u}$ has a **scalar variance-covariance matrix** when Assumption E.4 holds. We can now derive the **variance-covariance matrix of the OLS estimator**.

THEOREM E.2 **VARIANCE-COVARIANCE MATRIX OF THE OLS ESTIMATOR**

Under Assumptions E.1 through E.4,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad [E.14]$$

PROOF: From the last formula in equation (E.12), we have

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Now, we use Assumption E.4 to get

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Formula (E.14) means that the variance of $\hat{\beta}_j$ (conditional on $\mathbf{X}$) is obtained by multiplying σ^2 by the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. For the slope coefficients, we gave an interpretable formula in equation (3.51). Equation (E.14) also tells us how to obtain the covariance between any two OLS estimates: multiply σ^2 by the appropriate off-diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. In Chapter 4, we showed how to avoid explicitly finding covariances for obtaining confidence intervals and hypothesis tests by appropriately rewriting the model.

The Gauss-Markov Theorem, in its full generality, can be proven.

THEOREM E.3

GAUSS-MARKOV THEOREM

Under Assumptions E.1 through E.4, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator.

PROOF: Any other linear estimator of $\boldsymbol{\beta}$ can be written as

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \qquad \text{[E.15]}$$

where $\mathbf{A}$ is an $n \times (k+1)$ matrix. In order for $\tilde{\boldsymbol{\beta}}$ to be unbiased conditional on $\mathbf{X}$, $\mathbf{A}$ can consist of nonrandom numbers and functions of $\mathbf{X}$. (For example, $\mathbf{A}$ cannot be a function of $\mathbf{y}$.) To see what further restrictions on $\mathbf{A}$ are needed, write

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \qquad \text{[E.16]}$$

Then,

$$\begin{aligned}\text{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \text{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\text{E}(\mathbf{u}|\mathbf{X}) \text{ because } \mathbf{A} \text{ is a function of } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ because } \text{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}.\end{aligned}$$

For $\tilde{\boldsymbol{\beta}}$ to be an unbiased estimator of $\boldsymbol{\beta}$, it must be true that $\text{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ for all $(k+1) \times 1$ vectors $\boldsymbol{\beta}$, that is,

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ for all } (k+1) \times 1 \text{ vectors } \boldsymbol{\beta}. \qquad \text{[E.17]}$$

Because $\mathbf{A}'\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix, (E.17) holds if, and only if, $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. Equations (E.15) and (E.17) characterize the class of linear, unbiased estimators for $\boldsymbol{\beta}$.

Next, from (E.16), we have

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

by Assumption E.4. Therefore,

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ because } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A},\end{aligned}$$

where $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Because $\mathbf{M}$ is symmetric and idempotent, $\mathbf{A}'\mathbf{M}\mathbf{A}$ is positive semi-definite for any $n \times (k+1)$ matrix $\mathbf{A}$. This establishes that the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. Why is this important? Let $\mathbf{c}$ be any $(k+1) \times 1$ vector and consider the linear combination $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k$, which is a scalar. The unbiased estimators of $\mathbf{c}'\boldsymbol{\beta}$ are $\mathbf{c}'\hat{\boldsymbol{\beta}}$ and $\mathbf{c}'\tilde{\boldsymbol{\beta}}$. But

$$\text{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]\mathbf{c} \geq 0,$$

because $[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]$ is p.s.d. Therefore, when it is used for estimating any linear combination of $\boldsymbol{\beta}$, OLS yields the smallest variance. In particular, $\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}_j\mathbf{X})$ for any other linear, unbiased estimator of β_j.

The unbiased estimator of the error variance σ^2 can be written as

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-k-1),$$

which is the same as equation (3.56).

THEOREM E.4 **UNBIASEDNESS OF $\hat{\sigma}^2$**

Under Assumptions E.1 through E.4, $\hat{\sigma}^2$ is unbiased for σ^2: $\text{E}(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ for all $\sigma^2 > 0$.

PROOF: Write $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{M}\mathbf{X} = \mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}.$$

Because $\mathbf{u}'\mathbf{M}\mathbf{u}$ is a scalar, it equals its trace. Therefore,

$$\begin{aligned}\text{E}(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X}) &= \text{E}[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})|\mathbf{X}] = \text{E}[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}] \\ &= \text{tr}[\text{E}(\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}\text{E}(\mathbf{u}\mathbf{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n-k-1).\end{aligned}$$

The last equality follows from $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \text{tr}(\mathbf{I}_{k+1}) = n - (k+1) = n - k - 1$. Therefore,

$$\text{E}(\hat{\sigma}^2|\mathbf{X}) = \text{E}(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X})/(n-k-1) = \sigma^2.$$

E-3 Statistical Inference

When we add the final classical linear model assumption, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, which leads to the t and F distributions for the standard test statistics covered in Chapter 4.

Assumption E.5 **Normality of Errors**

Conditional on $\mathbf{X}$, the u_t are independent and identically distributed as Normal$(0,\sigma^2)$. Equivalently, $\mathbf{u}$ given $\mathbf{X}$ is distributed as multivariate normal with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}_n$: $u \sim \text{Normal}(0,\sigma^2\mathbf{I}_n)$.

Under Assumption E.5, each u_t is independent of the explanatory variables for all t. In a time series setting, this is essentially the strict exogeneity assumption.

THEOREM E.5

NORMALITY OF $\hat{\boldsymbol{\beta}}$

Under the classical linear model Assumptions E.1 through E.5, $\hat{\boldsymbol{\beta}}$ conditional on $\mathbf{X}$ is distributed as multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem E.5 is the basis for statistical inference involving $\boldsymbol{\beta}$. In fact, along with the properties of the chi-square, t, and F distributions that we summarized in Appendix D, we can use Theorem E.5 to establish that t statistics have a t distribution under Assumptions E.1 through E.5 (under the null hypothesis) and likewise for F statistics. We illustrate with a proof for the t statistics.

THEOREM E.6

DISTRIBUTION OF *t* STATISTIC

Under Assumptions E.1 through E.5,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k-1}, j = 0, 1, \ldots, k.$$

***PROOF*:** The proof requires several steps; the following statements are initially conditional on $\mathbf{X}$. First, by Theorem E.5, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$, where $\text{sd}(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$, and c_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Next, under Assumptions E.1 through E.5, conditional on $\mathbf{X}$,

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}. \qquad \text{[E.18]}$$

This follows because $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma)$, where $\mathbf{M}$ is the $n \times n$ symmetric, idempotent matrix defined in Theorem E.4. But $\mathbf{u}/\sigma \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ by Assumption E.5. It follows from Property 1 for the chi-square distribution in Appendix D that $(\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma) \sim \chi^2_{n-k-1}$ (because $\mathbf{M}$ has rank $n - k - 1$).

We also need to show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. But $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, and $\hat{\sigma}^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Now, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$ because $\mathbf{X}'\mathbf{M} = \mathbf{0}$. It follows, from Property 5 of the multivariate normal distribution in Appendix D, that $\hat{\boldsymbol{\beta}}$ and $\mathbf{M}\mathbf{u}$ are independent. Because $\hat{\sigma}^2$ is a function of $\mathbf{M}\mathbf{u}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also independent.

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

which is the ratio of a standard normal random variable and the square root of a $\chi^2_{n-k-1}/(n - k - 1)$ random variable. We just showed that these are independent, so, by definition of a t random variable, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has the t_{n-k-1} distribution. Because this distribution does not depend on $\mathbf{X}$, it is the unconditional distribution of $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ as well.

From this theorem, we can plug in any hypothesized value for β_j and use the t statistic for testing hypotheses, as usual.

Under Assumptions E.1 through E.5, we can compute what is known as the *Cramer-Rao* lower bound for the variance-covariance matrix of unbiased estimators of $\boldsymbol{\beta}$ (again conditional on $\mathbf{X}$) [see Greene (1997, Chapter 4)]. This can be shown to be $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which is exactly the variance-covariance matrix of the OLS estimator. This implies that $\hat{\boldsymbol{\beta}}$ is the **minimum variance unbiased estimator** of $\boldsymbol{\beta}$ (conditional on $\mathbf{X}$): $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semi-definite for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$; we no longer have to restrict our attention to estimators linear in $\mathbf{y}$.

It is easy to show that the OLS estimator is in fact the maximum likelihood estimator of $\boldsymbol{\beta}$ under Assumption E.5. For each t, the distribution of y_t given $\mathbf{X}$ is $\text{Normal}(\mathbf{x}_t\boldsymbol{\beta},\sigma^2)$. Because the y_t are

independent conditional on $\mathbf{X}$, the likelihood function for the sample is obtained from the product of the densities:

$$\prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2}\exp[-(y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)],$$

where Π denotes product. Maximizing this function with respect to $\boldsymbol{\beta}$ and σ^2 is the same as maximizing its natural logarithm:

$$\sum_{t=1}^{n} [-(1/2)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

For obtaining $\hat{\boldsymbol{\beta}}$, this is the same as minimizing $\sum_{t=1}^{n}(y_t - \mathbf{x}_t\boldsymbol{\beta})^2$—the division by $2\sigma^2$ does not affect the optimization—which is just the problem that OLS solves. The estimator of σ^2 that we have used, $\text{SSR}/(n - k)$, turns out not to be the MLE of σ^2; the MLE is SSR/n, which is a biased estimator. Because the unbiased estimator of σ^2 results in t and F statistics with exact t and F distributions under the null, it is always used instead of the MLE.

That the OLS estimator is the MLE under Assumption E.5 implies an interesting robustness property of the MLE based on the normal distribution. The reasoning is simple. We know that the OLS estimator is unbiased under Assumptions E.1 to E.3; normality of the errors is used nowhere in the proof, and neither is Assumption E.4. As the next section shows, the OLS estimator is also consistent without normality, provided the law of large numbers holds (as is widely true). These statistical properties of the OLS estimator imply that the MLE based on the normal log-likelihood function is robust to distributional specification: the distribution can be (almost) anything and yet we still obtain a consistent (and, under E.1 to E.3, unbiased) estimator. As discussed in Section 17-3, a maximum likelihood estimator obtained without assuming the distribution is correct is often called a **quasi-maximum likelihood estimator (QMLE).**

Generally, consistency of the MLE relies on having a correct distribution in order to conclude that it is consistent for the parameters. We have just seen that the normal distribution is a notable exception. There are some other distributions that share this property, including the Poisson distribution—as discussed in Section 17-3. Wooldridge (2010, Chapter 18) discusses some other useful examples.

E-4 Some Asymptotic Analysis

The matrix approach to the multiple regression model can also make derivations of asymptotic properties more concise. In fact, we can give general proofs of the claims in Chapter 11.

We begin by proving the consistency result of Theorem 11.1. Recall that these assumptions contain, as a special case, the assumptions for cross-sectional analysis under random sampling.

Proof of Theorem 11.1. As in Problem E.1 and using Assumption TS.1′ we write the OLS estimator as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'y_t\right) = \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'(\mathbf{x}_t\boldsymbol{\beta} + u_t)\right)\\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \qquad \text{[E.19]}\\ &= \boldsymbol{\beta} + \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right).\end{aligned}$$

Now, by the law of large numbers,

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ and } n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t \xrightarrow{p} \mathbf{0}, \qquad \text{[E.20]}$$

where $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$ is a $(k + 1) \times (k + 1)$ nonsingular matrix under Assumption TS.2′ and we have used the fact that $\mathrm{E}(\mathbf{x}_t'u_t) = 0$ under Assumption TS.3′. Now, we must use a matrix version of Property PLIM.1 in Appendix C. Namely, because $\mathbf{A}$ is nonsingular,

$$\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}. \qquad \text{[E.21]}$$

[Wooldridge (2010, Chapter 3) contains a discussion of these kinds of convergence results.] It now follows from (E.19), (E.20), and (E.21) that

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}.$$

This completes the proof.

Next, we sketch a proof of the asymptotic normality result in Theorem 11.2.

Proof of Theorem 11.2. From equation (E.19), we can write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \\ &= \mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) + o_p(1), \qquad \text{[E.22]}\end{aligned}$$

where the term "$o_p(1)$" is a remainder term that converges in probability to zero. This term is equal to $[(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t)^{-1} - \mathbf{A}^{-1}](n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. The term in brackets converges in probability to zero (by the same argument used in the proof of Theorem 11.1), while $(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$ is bounded in probability because it converges to a multivariate normal distribution by the central limit theorem. A well-known result in asymptotic theory is that the product of such terms converges in probability to zero. Further, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ inherits its asymptotic distribution from $\mathbf{A}^{-1}(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. See Wooldridge (2010, Chapter 3) for more details on the convergence results used in this proof.

By the central limit theorem, $n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t$ has an asymptotic normal distribution with mean zero and, say, $(k + 1) \times (k + 1)$ variance-covariance matrix $\mathbf{B}$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has an asymptotic multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. We now show that, under Assumptions TS.4′ and TS.5′, $\mathbf{B} = \sigma^2\mathbf{A}$. (The general expression is useful because it underlies heteroskedasticity-robust and serial correlation-robust standard errors for OLS, of the kind discussed in Chapter 12.) First, under Assumption TS.5′ $\mathbf{x}_t'u_t$ and $\mathbf{x}_s'u_s$ are uncorrelated for $t \neq s$. Why? Suppose $s < t$ for concreteness. Then, by the law of iterated expectations, $\mathrm{E}(\mathbf{x}_t'u_tu_s\mathbf{x}_s) = \mathrm{E}[\mathrm{E}(u_tu_s|\mathbf{x}_t'\mathbf{x}_s)\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[\mathrm{E}(u_tu_s|\mathbf{x}_t'\mathbf{x}_s)\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[0 \cdot \mathbf{x}_t'\mathbf{x}_s] = 0$. The zero covariances imply that the variance of the sum is the sum of the variances. But $\mathrm{Var}(\mathbf{x}_t'u_t) = \mathrm{E}(\mathbf{x}_t'u_tu_t\mathbf{x}_t) = \mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$. By the law of iterated expectations, $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t|\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2|\mathbf{x}_t)\mathbf{x}_t'\mathbf{x}_t] = \mathrm{E}[\sigma^2\mathbf{x}_t'\mathbf{x}_t] = \sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathbf{A}$, where we use $\mathrm{E}(u_t^2|\mathbf{x}_t) = \sigma^2$ under Assumptions TS.3′ and TS.4′. This shows that $\mathbf{B} = \sigma^2\mathbf{A}$, and so, under Assumptions TS.1′ to TS.5′, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} \text{Normal}(\mathbf{0},\sigma^2\mathbf{A}^{-1}). \qquad \text{[E.23]}$$

This completes the proof.

From equation (E.23), we treat $\hat{\boldsymbol{\beta}}$ as if it is approximately normally distributed with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2\mathrm{A}^{-1}/n$. The division by the sample size, n, is expected here: the approximation to the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ shrinks to zero at the rate $1/n$. When we replace

σ^2 with its consistent estimator, $\hat{\sigma}^2 = \text{SSR}/(n - k - 1)$, and replace A with its consistent estimator, $n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t = \mathbf{X}'\mathbf{X}/n$, we obtain an estimator for the asymptotic variance of $\hat{\boldsymbol{\beta}}$:

$$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.24]}$$

Notice how the two divisions by n cancel, and the right-hand side of (E.24) is just the usual way we estimate the variance matrix of the OLS estimator under the Gauss-Markov assumptions. To summarize, we have shown that, under Assumptions TS.1′ to TS.5′—which contain MLR.1 to MLR.5 as special cases—the usual standard errors and t statistics are asymptotically valid. It is perfectly legitimate to use the usual t distribution to obtain critical values and p-values for testing a single hypothesis. Interestingly, in the general setup of Chapter 11, assuming normality of the errors—say, u_t given $\mathbf{x}_t$, u_{t-1}, $\mathbf{x}_{t-1}$, …, u_1, $\mathbf{x}_1$ is distributed as Normal$(0, \sigma^2)$—does not necessarily help, as the t statistics would not generally have exact t statistics under this kind of normality assumption. When we do not assume strict exogeneity of the explanatory variables, exact distributional results are difficult, if not impossible, to obtain.

If we modify the argument above, we can derive a heteroskedasticity-robust, variance-covariance matrix. The key is that we must estimate $\text{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$ separately because this matrix no longer equals $\sigma^2\text{E}(\mathbf{x}_t'\mathbf{x}_t)$. But, if the $\hat{u}_t$ are the OLS residuals, a consistent estimator is

$$(n - k - 1)^{-1}\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t, \qquad \text{[E.25]}$$

where the division by $n - k - 1$ rather than n is a degrees of freedom adjustment that typically helps the finite sample properties of the estimator. When we use the expression in equation (E.25), we obtain

$$\widehat{\text{Avar}(\hat{\boldsymbol{\beta}})} = [n/(n - k - 1)](\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t\right)(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.26]}$$

The square roots of the diagonal elements of this matrix are the same heteroskedasticity-robust standard errors we obtained in Section 8-2 for the pure cross-sectional case. A matrix extension of the serial correlation- (and heteroskedasticity-) robust standard errors we obtained in Section 12-5 is also available, but the matrix that must replace (E.25) is complicated because of the serial correlation. See, for example, Hamilton (1994, Section 10-5).

E-4 Wald Statistics for Testing Multiple Hypotheses

Similar arguments can be used to obtain the asymptotic distribution of the **Wald statistic** for testing multiple hypotheses. Let $\mathbf{R}$ be a $q \times (k + 1)$ matrix, with $q \leq (k + 1)$. Assume that the q restrictions on the $(k + 1) \times 1$ vector of parameters, $\boldsymbol{\beta}$, can be expressed as H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{r}$ is a $q \times 1$ vector of known constants. Under Assumptions TS.1′ to TS.5′, it can be shown that, under H_0,

$$[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \overset{a}{\sim} \chi_q^2, \qquad \text{[E.27]}$$

where $\mathbf{A} = \text{E}(\mathbf{x}_t'\mathbf{x}_t)$, as in the proofs of Theorems 11.1 and 11.2. The intuition behind equation (E.25) is simple. Because $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is roughly distributed as Normal$(\mathbf{0},\sigma^2\mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is approximately Normal$(0,\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ by Property 3 of the multivariate normal distribution in Appendix D. Under H_0, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, so $\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \overset{a}{\sim} \text{Normal}(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ under H_0. By Property 3 of the chi-square distribution, $z'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}z \sim \chi_q^2$ if $z \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. To obtain the final result formally, we need to use an asymptotic version of this property, which can be found in Wooldridge (2010, Chapter 3).

Given the result in (E.25), we obtain a computable statistic by replacing $\mathbf{A}$ and σ^2 with their consistent estimators; doing so does not change the asymptotic distribution. The result is the so-called Wald statistic, which, after canceling the sample sizes and doing a little algebra, can be written as

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \qquad \text{[E.28]}$$

Under H_0, $W \overset{a}{\sim} \chi^2_q$, where we recall that q is the number of restrictions being tested. If $\hat{\sigma}^2 = \text{SSR}/(n - k - 1)$, it can be shown that W/q is exactly the F statistic we obtained in Chapter 4 for testing multiple linear restrictions. [See, for example, Greene (1997, Chapter 7).] Therefore, under the classical linear model assumptions TS.1 to TS.6 in Chapter 10, W/q has an exact $F_{q,n-k-1}$ distribution. Under Assumptions TS.1′ to TS.5′, we only have the asymptotic result in (E.26). Nevertheless, it is appropriate, and common, to treat the usual F statistic as having an approximate $F_{q,n-k-1}$ distribution.

A Wald statistic that is robust to heteroskedasticity of unknown form is obtained by using the matrix in (E.26) in place of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, and similarly for a test statistic robust to both heteroskedasticity and serial correlation. The robust versions of the test statistics cannot be computed via sums of squared residuals or R-squareds from the restricted and unrestricted regressions.

Summary

This appendix has provided a brief treatment of the linear regression model using matrix notation. This material is included for more advanced classes that use matrix algebra, but it is not needed to read the text. In effect, this appendix proves some of the results that we either stated without proof, proved only in special cases, or proved through a more cumbersome method of proof. Other topics—such as asymptotic properties, instrumental variables estimation, and panel data models—can be given concise treatments using matrices. Advanced texts in econometrics, including Davidson and MacKinnon (1993), Greene (1997), Hayashi (2000), and Wooldridge (2010), can be consulted for details.

Key Terms

First Order Condition
Frisch-Waugh (FW) theorem
Matrix Notation
Minimum Variance Unbiased Estimator
Scalar Variance-Covariance Matrix
Variance-Covariance Matrix of the OLS Estimator
Wald Statistic
Quasi-Maximum Likelihood Estimator (QMLE)

Problems

1 Let $\mathbf{x}_t$ be the $1 \times (k + 1)$ vector of explanatory variables for observation t. Show that the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n} \mathbf{x}_t'y_t\right).$$

Dividing each summation by n shows that $\hat{\boldsymbol{\beta}}$ is a function of sample averages.

2 Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimates.

(i) Show that for any $(k + 1) \times 1$ vector $\mathbf{b}$, we can write the sum of squared residuals as

$$\text{SSR}(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

{*Hint*: Write $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = [\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ and use the fact that $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.}

(ii) Explain how the expression for SSR(**b**) in part (i) proves that $\hat{\boldsymbol{\beta}}$ uniquely minimizes SSR(**b**) over all possible values of **b**, assuming **X** has rank $k + 1$.

3 Let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the regression of **y** on **X**. Let **A** be a $(k + 1) \times (k + 1)$ nonsingular matrix and define $\mathbf{z}_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Therefore, $\mathbf{z}_t$ is $1 \times (k + 1)$ and is a nonsingular linear combination of $\mathbf{x}_t$. Let **Z** be the $n \times (k + 1)$ matrix with rows $\mathbf{z}_t$. Let $\tilde{\boldsymbol{\beta}}$ denote the OLS estimate from a regression of **y** on **Z**.

(i) Show that $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.

(ii) Let $\hat{y}_t$ be the fitted values from the original regression and let $\tilde{y}_t$ be the fitted values from regressing **y** on **Z**. Show that $\tilde{y}_t = \hat{y}_t$, for all $t = 1, 2, \ldots, n$. How do the residuals from the two regressions compare?

(iii) Show that the estimated variance matrix for $\tilde{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1'}$, where $\hat{\sigma}^2$ is the usual variance estimate from regressing **y** on **X**.

(iv) Let the $\hat{\beta}_j$ be the OLS estimates from regressing y_t on $1, x_{t1}, \ldots, x_{tk}$, and let the $\tilde{\beta}_j$ be the OLS estimates from the regression of y_t on $1, a_1x_{t1}, \ldots, a_kx_{tk}$, where $a_i \neq 0$, $j = 1, \ldots, k$. Use the results from part (i) to find the relationship between the $\tilde{\beta}_j$ and the $\hat{\beta}_j$.

(v) Assuming the setup of part (iv), use part (iii) to show that $\text{se}(\tilde{\beta}_j) = \text{se}(\hat{\beta}_j)/|a_j|$.

(vi) Assuming the setup of part (iv), show that the absolute values of the t statistics for $\tilde{\beta}_j$ and $\hat{\beta}_j$ are identical.

4 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions, let **G** be a $(k + 1) \times (k + 1)$ nonsingular, nonrandom matrix, and define $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, so that $\boldsymbol{\delta}$ is also a $(k + 1) \times 1$ vector. Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimators and define $\hat{\boldsymbol{\delta}} = \mathbf{G}\hat{\boldsymbol{\beta}}$ as the OLS estimator of $\boldsymbol{\delta}$.

(i) Show that $\text{E}(\hat{\boldsymbol{\delta}}|\mathbf{X}) = \boldsymbol{\delta}$.

(ii) Find $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ in terms of σ^2, **X**, and **G**.

(iii) Use Problem E.3 to verify that $\hat{\boldsymbol{\delta}}$ and the appropriate estimate of $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ are obtained from the regression of **y** on $\mathbf{XG}^{-1}$.

(iv) Now, let **c** be a $(k + 1) \times 1$ vector with at least one nonzero entry. For concreteness, assume that $c_k \neq 0$. Define $\theta = \mathbf{c}'\boldsymbol{\beta}$, so that θ is a scalar. Define $\delta_j = \boldsymbol{\beta}_j, j = 0, 1, \ldots, k - 1$ and $\delta_k = \theta$. Show how to define a $(k + 1) \times (k + 1)$ nonsingular matrix **G** so that $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$. (*Hint*: Each of the first k rows of **G** should contain k zeros and a one. What is the last row?)

(v) Show that for the choice of **G** in part (iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & 1 & 0 \\ -c_0/c_k & -c_1/c_k & \cdot & \cdot & \cdot & -c_{k-1}/c_k & 1/c_k \end{bmatrix}$$

Use this expression for $\mathbf{G}^{-1}$ and part (iii) to conclude that $\hat{\theta}$ and its standard error are obtained as the coefficient on x_{tk}/c_k in the regression of

$$y_t \text{ on } [1 - (c_0/c_k)x_{tk}], [x_{t1} - (c_1/c_k)x_{tk}], \ldots, [x_{t,k-1} - (c_{k-1}/c_k)x_{tk}], x_{tk}/c_k, t = 1, \ldots, n.$$

This regression is exactly the one obtained by writing β_k in terms of θ and $\beta_0, \beta_1, \ldots, \beta_{k-1}$, plugging the result into the original model, and rearranging. Therefore, we can formally justify the trick we use throughout the text for obtaining the standard error of a linear combination of parameters.

5 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions and let $\hat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$. Let $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ be an $n \times (k + 1)$ matrix function of $\mathbf{X}$ and assume that $\mathbf{Z}'\mathbf{X}$[a $(k + 1) \times (k + 1)$ matrix] is nonsingular. Define a new estimator of $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

(i) Show that $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, so that $\tilde{\boldsymbol{\beta}}$ is also unbiased conditional on $\mathbf{X}$.

(ii) Find $\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Make sure this is a symmetric, $(k + 1) \times (k + 1)$ matrix that depends on $\mathbf{Z}$, $\mathbf{X}$, and σ^2.

(iii) Which estimator do you prefer, $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$? Explain.

6 Consider the setup of the Frisch-Waugh Theorem.

(i) Using partitioned matrices, show that the first order conditions $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ can be written as

$$\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$$
$$\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}.$$

(ii) Multiply the first set of equations by $\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and subtract the result from the second set of equations to show that

$$(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

where $\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Conclude that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\mathbf{y}.$$

(iii) Use part (ii) to show that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\ddot{y}.$$

(iv) Use the fact that $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ to show that the residuals $\ddot{\mathbf{u}}$ from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ are identical to the residuals $\hat{\mathbf{u}}$ from the regression $\mathbf{y}$ on $\mathbf{X}_1$, $\mathbf{X}_2$. [*Hint*: By definition and the FW theorem,

$$\ddot{\mathbf{u}} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2).$$

Now you do the rest.]

7 Suppose that the linear model, written in matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

satisfies Assumptions E.1, E.2, and E.3. Partition the model as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and $\mathbf{X}_2$ is $n \times k_2$.

(i) Consider the following proposal for estimating $\boldsymbol{\beta}_2$. First, regress $\mathbf{y}$ on $\mathbf{X}_1$ and obtain the residuals, say, $\ddot{\mathbf{y}}$. Then, regress $\ddot{\mathbf{y}}$ on $\mathbf{X}_2$ to get $\breve{\boldsymbol{\beta}}_2$. Show that $\breve{\boldsymbol{\beta}}_2$ is generally biased and show what the bias is. [You should find $\mathrm{E}(\breve{\boldsymbol{\beta}}_2|\mathbf{X})$ in terms of $\boldsymbol{\beta}_2$, $\mathbf{X}_2$, and the residual-making matrix $\mathbf{M}_1$.]

(ii) As a special case, write

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \beta_k\mathbf{X}_k + \mathbf{u},$$

where $\mathbf{X}_k$ is an $n \times 1$ vector on the variable x_{tk}. Show that

$$\mathrm{E}(\tilde{\beta}_k|\mathbf{X}) = \left(\frac{\mathrm{SSR}_k}{\sum_{t=1}^{n} x_{tk}^2}\right)\beta_k,$$

where SSR_k is the sum of squared residuals from regressing x_{tk} on 1, x_{t1}, x_{t2}, …, $x_{t,\,k-1}$. How come the factor multiplying β_k is never greater than one?

(iii) Suppose you know $\boldsymbol{\beta}_1$. Show that the regression $\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$ on $\mathbf{X}_2$ produces an unbiased estimator of $\boldsymbol{\beta}_2$ (conditional on $\mathbf{X}$).