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:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_k x_{tk} + u_t,\ t = 1,2,\dots,n, \tag{E.1}$$

where y_t is the dependent variable for observation t, and x_{tj}, $j = 2,3,\dots,k$, are the independent variables. Notice how our labeling convention here differs from the text: we call the intercept β_1 and let $\beta_2, \dots, \beta_k$ denote the slope parameters. This relabeling is not important, but it simplifies the matrix approach to multiple regression.

For each t, define a $1 \times k$ vector, $\boldsymbol{x}_t = (1, x_{t2}, \dots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)'$ be the $k \times 1$ vector of all parameters. Then, we can write (E.1) as

$$y_t = \boldsymbol{x}_t \boldsymbol{\beta} + u_t,\ t = 1,2,\dots,n. \tag{E.2}$$

[Some authors prefer to define $\boldsymbol{x}_t$ as a column vector, in which case, $\boldsymbol{x}_t$ is replaced with $\boldsymbol{x}_t'$ 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 $\boldsymbol{y}$ denote the $n \times 1$ vector of observations on y: the t^{th} element of $\boldsymbol{y}$ is y_t. Let $\boldsymbol{X}$ be the $n \times k$ vector of observations on the explanatory variables. In other words, the t^{th} row of $\mathbf{X}$ consists of the vector $\boldsymbol{x}_t$. Equivalently, the $(t,j)^{\text{th}}$ element of $\mathbf{X}$ is simply x_{tj}:

$$\underset{n \times k}{\mathbf{X}} \equiv \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \ldots & x_{1k} \\ 1 & x_{22} & x_{23} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n2} & x_{n3} & \ldots & x_{nk} \end{bmatrix}.$$

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

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

Remember, because $\mathbf{X}$ is $n \times k$ and $\boldsymbol{\beta}$ is $k \times 1$, $\mathbf{X}\boldsymbol{\beta}$ 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 $k \times 1$ parameter vector $\boldsymbol{b}$ as

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

The $k \times 1$ vector of ordinary least squares estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)'$, minimizes SSR($\boldsymbol{b}$) over all possible $k \times 1$ vectors $\boldsymbol{b}$. 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**

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

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

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

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

$$\begin{aligned} \sum_{t=1}^{n} (y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \ldots - \hat{\beta}_k x_{tk}) &= 0 \\ \sum_{t=1}^{n} x_{t2}(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \ldots - \hat{\beta}_k x_{tk}) &= 0 \\ &\vdots \\ \sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \ldots - \hat{\beta}_k x_{tk}) &= 0, \end{aligned}$$

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

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

or

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

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 $k \times k$ 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 $\hat{\boldsymbol{\beta}}$:

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

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($\mathbf{X}$) $= k$, which means that the columns of $\mathbf{X}$ must be linearly independent. This is the matrix version of MLR.4 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:

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

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

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

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

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

$$\mathbf{X}'\hat{\boldsymbol{u}} = \mathbf{0}. \tag{E.9}$$

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

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

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.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 $\boldsymbol{y}$ is an observed $n \times 1$ vector, $\mathbf{X}$ is an $n \times k$ observed matrix, and $\boldsymbol{u}$ is an $n \times 1$ vector of unobserved errors or disturbances.

ASSUMPTION E.2 (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,

$$\mathrm{E}(\boldsymbol{u}|\mathbf{X}) = \mathbf{0}. \tag{E.11}$$

This assumption is implied by MLR.3 under the random sampling assumption, MLR.2. In time series applications, Assumption E.2 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.2, we can condition on the x_{tj} when we compute the expected value of $\hat{\boldsymbol{\beta}}$.

ASSUMPTION E.3 (NO PERFECT COLLINEARITY)
The matrix $\mathbf{X}$ has rank k.

This is a careful statement of the assumption that rules out linear dependencies among the explanatory variables. Under Assumption E.3, $\mathbf{X}'\mathbf{X}$ is nonsingular, and so $\hat{\boldsymbol{\beta}}$ is unique and can be written as in (E.8).

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.3 and simple algebra to write

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

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

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

because $\mathrm{E}(\boldsymbol{u}|\mathbf{X}) = \mathbf{0}$ under Assumption E.2. 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) $\mathrm{Var}(u_t|\mathbf{X}) = \sigma^2$, $t = 1,2, \dots, n$. (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

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

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 $\boldsymbol{u}$ has **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,

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

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

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

Now, we use Assumption E.4 to get

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

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 hypotheses 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

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\boldsymbol{y}, \tag{E.15}$$

where $\mathbf{A}$ is an $n \times k$ 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 $\boldsymbol{y}$.) To see what further restrictions on $\mathbf{A}$ are needed, write

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\boldsymbol{u}. \tag{E.16}$$

Then,

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

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

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ for all } k \times 1 \text{ vectors } \boldsymbol{\beta}. \tag{E.17}$$

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

Next, from (E.16), we have

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

by Assumption E.4. Therefore,

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

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 $n \times k$ matrix $\mathbf{A}$. This establishes that the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. How is this significant? Let $\boldsymbol{c}$ be any $k \times 1$ vector and consider the linear combination $\boldsymbol{c}'\boldsymbol{\beta} = c_1\beta_1 + c_2\beta_2 + \ldots + c_k\beta_k$, which is a scalar. The unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$ are $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ and $\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$. But

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

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

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

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

where we have labeled the explanatory variables so that there are k total parameters, including the intercept.

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

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

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

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

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

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

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

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

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, $\boldsymbol{u}$ given $\mathbf{X}$ is distributed as multivariate normal with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}_n$: $\boldsymbol{u} \sim \mathrm{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$.

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 $\hat{\beta}$)

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

Under Assumptions E.1 through E.5,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k}, j = 1,2, \ldots, k.$$

PROOF: The proof requires several steps; the following statements are initially conditional on $\mathbf{X}$. First, by Theorem E.5, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}\ ) \sim \text{Normal}(0,1)$, 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 $\mathbf{X}$,

$$(n - k)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k}. \tag{E.18}$$

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

We also need to show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. But $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{u}$, and $\hat{\sigma}^2 = \boldsymbol{u}'\mathbf{M}\boldsymbol{u}/(n - k)$. 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 $\mathbf{M}\boldsymbol{u}$ are independent. Since $\hat{\sigma}^2$ is a function of $\mathbf{M}\boldsymbol{u}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also independent.

Finally, we can write

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

which is the ratio of a standard normal random variable and the square root of a $\chi^2_{n-k}/(n - k)$ random variable. We just showed that these are independent, and so, by definition of a t random variable, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has the t_{n-k} 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 $\mathbf{X}$): $\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semi-definite for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$; we no longer have to restrict our attention to estimators linear in $\boldsymbol{y}$.

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 $\mathrm{Normal}(\boldsymbol{x}_t\boldsymbol{\beta},\sigma^2)$. Because the y_t are independent conditional on $\mathbf{X}$, the likelihood function for the sample is obtained from the product of the densities:

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

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

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

For obtaining $\hat{\boldsymbol{\beta}}$, this is the same as minimizing $\sum_{t=1}^{n} (y_t - \boldsymbol{x}_t\boldsymbol{\beta})^2$—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, $\mathrm{SSR}/(n - k)$, 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.

SUMMARY

This appendix has provided a brief discussion 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), and Wooldridge (1999), can be consulted for details.

KEY TERMS

First Order Condition
Matrix Notation
Minimum Variance Unbiased
Scalar Variance-Covariance Matrix
Variance-Covariance Matrix of the OLS Estimator

PROBLEMS

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

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

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

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

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

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

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

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

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

(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 $\boldsymbol{y}$ on $\mathbf{Z}$. Show that $\tilde{y}_t = \hat{y}_t$, for all $t = 1,2, \dots, n$. How do the residuals from the two regressions compare?

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

(iv) Let the $\hat{\beta}_j$ be the OLS estimates from regressing y_t on $1, x_{t2}, \dots, x_{tk}$, and let the $\tilde{\beta}_j$ be the OLS estimates from the regression of y_t on $1, a_2x_{t2}, \dots, a_kx_{tk}$, where $a_j \neq 0$, $j = 2, \dots, k$. 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.