CHAPTER 7
The General Linear Model: The Basics

7.1 Introduction

Consider the following regression equation

$$y = X\beta + u \tag{7.1}$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ : \\ Y_n \end{bmatrix}; X = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ : & : & : & : \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ : \\ \beta_k \end{bmatrix}; u = \begin{bmatrix} u_1 \\ u_2 \\ : \\ u_n \end{bmatrix}$$

with n denoting the number of observations and k the number of variables in the regression, with $n > k$. In this case, y is a column vector of dimension $(n \times 1)$ and X is a matrix of dimension $(n \times k)$. Each column of X denotes a variable and each row of X denotes an observation on these variables. If y is log(wage) as in the empirical example in Chapter 4, see Table 4.1 then the columns of X contain a column of ones for the constant (usually the first column), weeks worked, years of full time experience, years of education, sex, race, marital status, etc.

7.2 Least Squares Estimation

Least squares minimizes the residual sum of squares where the residuals are given by $e = y - X\widehat{\beta}$ and $\widehat{\beta}$ denotes a guess on the regression parameters β. The residual sum of squares

$$RSS = \textstyle\sum_{i=1}^n e_i^2 = e'e = (y - X\beta)'(y - X\beta) = y'y - y'X\beta - \beta'X'y + \beta'X'X\beta$$

The last four terms are scalars as can be verified by their dimensions. It is essential that the reader keep track of the dimensions of the matrices used. This will insure proper multiplication, addition, subtraction of matrices and help the reader obtain the right answers. In fact the middle two terms are the same because the transpose of a scalar is a scalar. For a quick review of some matrix properties, see the Appendix to this chapter. Differentiating the RSS with respect to β one gets

$$\partial RSS/\partial\beta = -2X'y + 2X'X\beta \tag{7.2}$$

where use is made of the following two rules of differentiating matrices. The first is that $\partial a'b/\partial b = a$ and the second is

$$\partial(b'Ab)/\partial b = (A + A')b = 2Ab$$

where the last equality holds if A is a symmetric matrix. In the RSS equation a is $y'X$ and A is $X'X$. The first-order condition for minimization equates the expression in (7.2) to zero. This yields

$$X'X\beta = X'y \tag{7.3}$$

B.H. Baltagi, *Econometrics*, Springer Texts in Business and Economics, DOI 10.1007/978-3-642-20059-5_7,

which is known as the OLS normal equations. As long as X is of full column rank , i.e., of rank k, then $X'X$ is nonsingular and the solution to the above equations is $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$. Full column rank means that no column of X is a perfect linear combination of the other columns. In other words, no variable in the regression can be obtained from a linear combination of the other variables. Otherwise, at least one of the OLS normal equations becomes redundant. This means that we only have $(k-1)$ linearly independent equations to solve for k unknown β's. This yields no solution for $\widehat{\beta}_{OLS}$ and we say that $X'X$ is singular. $X'X$ is the sum of squares cross product matrix (SSCP). If it has a column of ones then it will contain the sums, the sum of squares, and the cross-product sum between any two variables

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n X_{i2} & \ldots & \sum_{i=1}^n X_{ik} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i2}^2 & \ldots & \sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n X_{ik} & \sum_{i=1}^n X_{ik}X_{i2} & \ldots & \sum_{i=1}^n X_{ik}^2 \end{bmatrix}$$

Of course y could be added to this matrix as another variable which will generate $X'y$ and $y'y$ automatically for us, i.e., the column pertaining to the variable y will generate $\sum_{i=1}^n y_i$, $\sum_{i=1}^n X_{i1}y_i, \ldots, \sum_{i=1}^n X_{ik}y_i$, and $\sum_{i=1}^n y_i^2$. To see this, let

$$Z = [y, X] \quad \text{then} \quad Z'Z = \begin{bmatrix} y'y & y'X \\ X'y & X'X \end{bmatrix}$$

This matrix summarizes the data and we can compute any regression of one variable in Z on any subset of the remaining variables in Z using only $Z'Z$. Denoting the least squares residuals by $e = y - X\widehat{\beta}_{OLS}$, the OLS normal equations given in (7.3) can be written as

$$X'(y - X\widehat{\beta}_{OLS}) = X'e = 0 \tag{7.4}$$

Note that if the regression includes a constant, the first column of X will be a vector of ones and the first equation of (7.4) becomes $\sum_{i=1}^n e_i = 0$. This proves the well known result that if there is a constant in the regression, the OLS residuals sum to zero. Equation (7.4) also indicates that the regressor matrix X is orthogonal to the residuals vector e. This will become clear when we define e in terms of the orthogonal projection matrix on X. This representation allows another interpretation of OLS as a method of moments estimator which was considered in Chapter 2. This follows from the classical assumptions where X satisfies $E(X'u) = 0$. The sample counterpart of this condition yields $X'e/n = 0$. These are the OLS normal equations and therefore, yield the OLS estimates without minimizing the residual sums of squares.

Since data in economics are not generated using experiments like the physical sciences, the X's are stochastic and we only observe one realization of this data. Consider for example, annual observations for GNP, money supply, unemployment rate, etc. One cannot repeat draws for this data in the real world or fix the X's to generate new y's (unless one is performing a Monte Carlo study). So we have to condition on the set of X's observed, see Chapter 5.

Classical Assumptions: $u \sim (0, \sigma^2 I_n)$ which means that (i) each disturbance u_i has zero mean, (ii) constant variance, and (iii) u_i and u_j for $i \neq j$ are not correlated. The u's are known as spherical disturbances. Also, (iv) the conditional expectation of u given X is zero, $E(u/X) = 0$. Note that the conditioning here is with respect to *every* regressor in X and for *all* observations $i = 1, 2, \ldots n$. In other words, it is conditional on all the elements of the matrix X. Using

(7.1), this implies that $E(y/X) = X\beta$ is *linear* in β, $\text{var}(u_i/X) = \sigma^2$ and $\text{cov}(u_i, u_j/X) = 0$. Additionally, we assume that plim $X'X/n$ is finite and positive definite and plim $X'u/n = 0$ as $n \to \infty$.

Given these classical assumptions, and conditioning on the X's observed, it is easy to show that $\widehat{\beta}_{OLS}$ is unbiased for β. In fact using (7.1) one can write

$$\widehat{\beta}_{OLS} = \beta + (X'X)^{-1}X'u \tag{7.5}$$

Taking expectations, conditioning on the X's, and using assumptions (i) and (iv), one attains the unbiasedness result. Furthermore, one can derive the variance-covariance matrix of $\widehat{\beta}_{OLS}$ from (7.5) since

$$\text{var}(\widehat{\beta}_{OLS}) = E(\widehat{\beta}_{OLS} - \beta)(\widehat{\beta}_{OLS} - \beta)' = E(X'X)^{-1}X'uu'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{7.6}$$

this uses assumption (iv) along with the fact that $E(uu') = \sigma^2 I_n$. This variance-covariance matrix is $(k \times k)$ and gives the variances of the $\widehat{\beta}_i$'s across the diagonal and the pairwise covariances of say $\widehat{\beta}_i$ and $\widehat{\beta}_j$ off the diagonal. The next theorem shows that among all linear unbiased estimators of $c'\beta$, it is $c'\widehat{\beta}_{OLS}$ which has the smallest variance. This is known as the Gauss-Markov Theorem.

Theorem 1: Consider the linear estimator $a'y$ for $c'\beta$, where both a and c are arbitrary vectors of constants. If $a'y$ is unbiased for $c'\beta$ then $\text{var}(a'y) \geq \text{var}(c'\widehat{\beta}_{OLS})$.

Proof: For $a'y$ to be unbiased for $c'\beta$ it must follow from (7.1) that $E(a'y) = a'X\beta + E(a'u) = a'X\beta = c'\beta$ which means that $a'X = c'$. Also, $\text{var}(a'y) = E(a'y - c'\beta)(a'y - c'\beta)' = E(a'uu'a) = \sigma^2 a'a$. Comparing this variance with that of $c'\widehat{\beta}_{OLS}$, one gets $\text{var}(a'y) -$ $\text{var}(c'\widehat{\beta}_{OLS}) = \sigma^2 a'a - \sigma^2 c'(X'X)^{-1}c$. But, $c' = a'X$, therefore this difference becomes $\sigma^2[a'a - a'P_X a] = \sigma^2 a'\bar{P}_X a$ where P_X is a projection matrix on the X-plane defined as $X(X'X)^{-1}X'$ and $\bar{P}_X$ is defined as $I_n - P_X$. In fact, $P_X y = X\widehat{\beta}_{OLS} = \widehat{y}$ and $\bar{P}_X y = y - P_X y = y - \widehat{y} = e$. So that $\widehat{y}$ projects the vector y on the X-plane and e is the projection of y on the plane orthogonal to X or perpendicular to X, see Figure 7.1. Both P_X and $\bar{P}_X$ are idempotent which means that the above difference $\sigma^2 a'\bar{P}_X a$ is greater or equal to zero since $\bar{P}_X$ is positive semi-definite. To see this, define $z = \bar{P}_X a$, then the above difference is equal to $\sigma^2 z'z \geq 0$.

The implications of the theorem are important. It means for example, that for the choice of $c' = (1, 0, \ldots, 0)$ one can pick $\beta_1 = c'\beta$ for which the best linear unbiased estimator would be $\widehat{\beta}_{1,OLS} = c'\widehat{\beta}_{OLS}$. Similarly any β_j can be chosen by using $c' = (0, \ldots, 1, \ldots, 0)$ which has 1 in the j-th position and zero elsewhere. Again, the BLUE of $\beta_j = c'\beta$ is $\widehat{\beta}_{j,OLS} = c'\widehat{\beta}_{OLS}$. Furthermore, any linear combination of these β's such as their sum $\sum_{j=1}^{k} \beta_j$ which corresponds to $c' = (1, 1, \ldots, 1)$ has the sum $\sum_{j=1}^{k} \widehat{\beta}_{j,OLS}$ as its BLUE.

The disturbance variance σ^2 is unknown and has to be estimated. Note that $E(u'u) = E(\text{tr}(uu')) = \text{tr}(E(uu')) = \text{tr}(\sigma^2 I_n) = n\sigma^2$, so that $u'u/n$ seems like a natural unbiased estimator for σ^2. However, u is not observed and is estimated by the OLS residuals e. It is therefore, natural to investigate $E(e'e)$. In what follows, we show that $s^2 = e'e/(n-k)$ is an unbiased estimator for σ^2. To prove this, we need the fact that

$$e = y - X\widehat{\beta}_{OLS} = y - X(X'X)^{-1}X'y = \bar{P}_X y = \bar{P}_X u \tag{7.7}$$

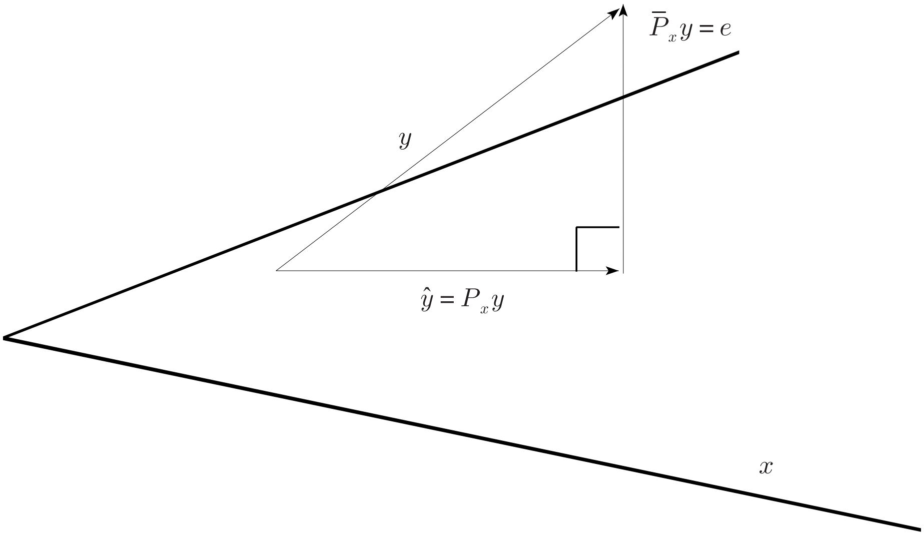

Figure 7.1 The Orthogonal Decomposition of y

where the last equality follows from the fact that $\bar{P}_X X = 0$. Hence,

$$\begin{aligned} E(e'e) &= E(u'\bar{P}_X u) = E(\text{tr}\{u'\bar{P}_X u\}) = E(\text{tr}\{uu'\bar{P}_X\}) \\ &= \text{tr}(\sigma^2 \bar{P}_X) = \sigma^2 \text{tr}(\bar{P}_X) = \sigma^2(n-k) \end{aligned}$$

where the second equality follows from the fact that the trace of a scalar is a scalar. The third equality from the fact that $\text{tr}(ABC) = \text{tr}(CAB)$. The fourth equality from the fact that $E(\mathit{trace}) = \mathit{trace}\{E(.)\}$, and $E(uu') = \sigma^2 I_n$. The last equality from the fact that

$$\begin{aligned} \text{tr}(\bar{P}_X) &= \text{tr}(I_n) - \text{tr}(P_X) = n - \text{tr}(X(X'X)^{-1}X') \\ &= n - \text{tr}(X'X(X'X)^{-1}) = n - \text{tr}(I_k) = n - k. \end{aligned}$$

Hence, an unbiased estimator of $\text{var}(\widehat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$ is given by $s^2(X'X)^{-1}$.

So far we have shown that $\widehat{\beta}_{OLS}$ is BLUE. It can also be shown that it is consistent for β. In fact, taking probability limits of (7.5) as $n \to \infty$, one gets

$$\text{plim}(\widehat{\beta}_{OLS}) = \text{plim}(\beta) + \text{plim}(X'X/n)^{-1}(X'u/n) = \beta$$

The first equality uses the fact that the plim of a sum is the sum of the plims. The second equality follows from assumption 1 and the fact that plim of a product is the product of plims.

7.3 Partitioned Regression and the Frisch-Waugh-Lovell Theorem

In Chapter 4, we studied a useful property of least squares which allows us to interpret multiple regression coefficients as simple regression coefficients. This was called the residualing interpretation of multiple regression coefficients. In general, this property applies whenever the k regressors given by X can be separated into two sets of variables X_1 and X_2 of dimension $(n \times k_1)$ and $(n \times k_2)$ respectively, with $X = [X_1, X_2]$ and $k = k_1 + k_2$. The regression in equation (7.1) becomes a partitioned regression given by

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u \tag{7.8}$$

One may be interested in the least squares estimates of β_2 corresponding to X_2, but one has to control for the presence of X_1 which may include seasonal dummy variables or a time trend, see Frisch and Waugh (1933) and Lovell (1963)[1].

The OLS normal equations from (7.8) are as follows:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_{1,OLS} \\ \widehat{\beta}_{2,OLS} \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \tag{7.9}$$

These can be solved by partitioned inversion of the matrix on the left, see the Appendix to this chapter, or by solving two equations in two unknowns. Problem 2 asks the reader to verify that

$$\widehat{\beta}_{2,OLS} = (X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y \tag{7.10}$$

where $\bar{P}_{X_1} = I_n - P_{X_1}$ and $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$. $\bar{P}_{X_1}$ is the orthogonal projection matrix of X_1 and $\bar{P}_{X_1}X_2$ generates the least squares residuals of each column of X_2 regressed on all the variables in X_1. In fact, if we denote by $\widetilde{X}_2 = \bar{P}_{X_1}X_2$ and $\widetilde{y} = \bar{P}_{X_1}y$, then (7.10) can be written as

$$\widehat{\beta}_{2,OLS} = (\widetilde{X}_2'\widetilde{X}_2)^{-1}\widetilde{X}_2'\widetilde{y} \tag{7.11}$$

using the fact that $\bar{P}_{X_1}$ is idempotent. This implies that $\widehat{\beta}_{2,OLS}$ can be obtained from the regression of $\widetilde{y}$ on $\widetilde{X}_2$. In words, the residuals from regressing y on X_1 are in turn regressed upon the residuals from each column of X_2 regressed on all the variables in X_1. This was illustrated in Chapter 4 with some examples. Following Davidson and MacKinnon (1993) we denote this result more formally as the Frisch-Waugh-Lovell (FWL) Theorem. In fact, if we premultiply (7.8) by $\bar{P}_{X_1}$ and use the fact that $\bar{P}_{X_1}X_1 = 0$, one gets

$$\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\beta_2 + \bar{P}_{X_1}u \tag{7.12}$$

The FWL Theorem states that: (1) The least squares *estimates* of β_2 from equations (7.8) and (7.12) are numerically identical and (2) The least squares *residuals* from equations (7.8) and (7.12) are identical.

Using the fact that $\bar{\mathrm{P}}_{X_1}$ is idempotent, it immediately follows that, OLS on (7.12) yields $\widehat{\beta}_{2,OLS}$ as given by equation (7.10). Alternatively, one can start from equation (7.8) and use the result that

$$y = P_Xy + \bar{P}_Xy = X\widehat{\beta}_{OLS} + \bar{P}_Xy = X_1\widehat{\beta}_{1,OLS} + X_2\widehat{\beta}_{2,OLS} + \bar{P}_Xy \tag{7.13}$$

where $P_X = X(X'X)^{-1}X'$ and $\bar{P}_X = I_n - P_X$. Premultiplying equation (7.13) by $X_2'\bar{P}_{X_1}$ and using the fact that $\bar{P}_{X_1}X_1 = 0$, one gets

$$X_2'\bar{P}_{X_1}y = X_2'\bar{P}_{X_1}X_2\widehat{\beta}_{2,OLS} + X_2'\bar{P}_{X_1}\bar{P}_Xy \tag{7.14}$$

But, $P_{X_1}P_X = P_{X_1}$. Hence, $\bar{P}_{X_1}\bar{P}_X = \bar{P}_X$. Using this fact along with $\bar{P}_XX = \bar{P}_X[X_1, X_2] = 0$, the last term of equation (7.14) drops out yielding the result that $\widehat{\beta}_{2,OLS}$ from (7.14) is identical to the expression in (7.10). Note that no partitioned inversion was used in this proof. This proves part (1) of the FWL Theorem.

Also, premultiplying equation (7.13) by $\bar{P}_{X_1}$ and using the fact that $\bar{P}_{X_1}\bar{P}_X = \bar{P}_X$, one gets

$$\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\widehat{\beta}_{2,OLS} + \bar{P}_X y \tag{7.15}$$

Now $\widehat{\beta}_{2,OLS}$ was shown to be numerically identical to the least squares estimate obtained from equation (7.12). Hence, the first term on the right hand side of equation (7.15) must be the fitted values from equation (7.12). Since the dependent variables are the same in equations (7.15) and (7.12), $\bar{P}_X y$ in equation (7.15) must be the least squares residuals from regression (7.12). But, $\bar{P}_X y$ is the least squares residuals from regression (7.8). Hence, the least squares residuals from regressions (7.8) and (7.12) are numerically identical. This proves part (2) of the FWL Theorem.

Several applications of the FWL Theorem will be given in this book. Problem 2 shows that if X_1 is the vector of ones indicating the presence of a constant in the regression, then regression (7.15) is equivalent to running $(y_i - \bar{y})$ on the set of variables in X_2 expressed as deviations from their respective sample means. Problem 3 shows that the FWL Theorem can be used to prove that including a dummy variable for one of the observations in the regression is equivalent to omitting that observation from the regression.

7.4 Maximum Likelihood Estimation

In Chapter 2, we introduced the method of maximum likelihood estimation which is based on specifying the distribution we are sampling from and writing the joint density of our sample. This joint density is then referred to as the likelihood function because it gives for a given set of parameters specifying the distribution, the probability of obtaining the observed sample. See Chapter 2 for several examples. For the regression equation, specifying the distribution of the disturbances in turn specifies the likelihood function. These disturbances could be Poisson, Exponential, Normal, etc. Once this distribution is chosen, the likelihood function is maximized and the MLE of the regression parameters are obtained. Maximum likelihood estimators are desirable because they are (1) consistent under fairly general conditions,[2] (2) asymptotically normal, (3) asymptotically efficient and (4) invariant to reparameterizations of the model[3]. Some of the undesirable properties of MLE are that (1) it requires explicit distributional assumptions on the disturbances, and (2) their finite sample properties can be quite different from their asymptotic properties. For example, MLE can be biased even though they are consistent, and their covariance estimates can be misleading for small samples. In this section, we derive the MLE under normality of the disturbances.

The Normality Assumption: $u \sim N(0, \sigma^2 I_n)$**.** This additional assumption allows us to derive distributions of estimators and other random variables. This is important for constructing confidence intervals and tests of hypotheses. In fact using (7.5) one can easily see that $\widehat{\beta}_{OLS}$ is a linear combination of the u's. But, a linear combination of normal random variables is itself a normal random variable. Hence, $\widehat{\beta}_{OLS}$ is $N(\beta, \sigma^2(X'X)^{-1})$. Similarly y is $N(X\beta, \sigma^2 I_n)$ and e is $N(0, \sigma^2\bar{P}_X)$. Moreover, we can write the joint probability density function of the u's as $f(u_1, u_2, \ldots, u_n; \sigma^2) = (1/2\pi\sigma^2)^{n/2}\exp(-u'u/2\sigma^2)$. To get the likelihood function we make the transformation $u = y - X\beta$ and note that the Jacobian of the transformation is one. Hence

$$f(y_1, y_2, \ldots, y_n; \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\beta)'(y - X\beta)/2\sigma^2\} \tag{7.16}$$

Taking the log of this likelihood, we get

$$\log L(\beta,\sigma^2) = -(n/2)\log(2\pi\sigma^2) - (y-X\beta)'(y-X\beta)/2\sigma^2 \tag{7.17}$$

Maximizing this likelihood with respect to β and σ^2 one gets the maximum likelihood estimators (MLE). Let $\theta=\sigma^2$ and $Q=(y-X\beta)'(y-X\beta)$, then

$$\begin{aligned}\frac{\partial logL(\beta,\theta)}{\partial\beta} &= \frac{2X'y-2X'X\beta}{2\theta}\\ \frac{\partial logL(\beta,\theta)}{\partial\theta} &= \frac{Q}{2\theta^2}-\frac{n}{2\theta}\end{aligned}$$

Setting these first-order conditions equal to zero, one gets

$$\widehat{\beta}_{MLE}=\widehat{\beta}_{OLS} \quad \text{and} \quad \widehat{\theta}=\widehat{\sigma}^2_{MLE}=Q/n=RSS/n=e'e/n.$$

Intuitively, only the second term in the log likelihood contains β and that term (without the negative sign) has already been minimized with respect to β in (7.2) giving us the OLS estimator. Note that $\widehat{\sigma}^2_{MLE}$ differs from s^2 only in the degrees of freedom. It is clear that $\widehat{\beta}_{MLE}$ is unbiased for β while $\widehat{\sigma}^2_{MLE}$ is not unbiased for σ^2. Substituting these MLE's into (7.17) one gets the maximum value of $\log L$ which is

$$\begin{aligned}\log L(\widehat{\beta}_{MLE},\widehat{\sigma}^2_{MLE}) &= -(n/2)\log(2\pi\widehat{\sigma}^2_{MLE})-e'e/2\widehat{\sigma}^2_{MLE}\\ &= -(n/2)\log(2\pi)-(n/2)\log(e'e/n)-n/2\\ &= \text{constant}-(n/2)\log(e'e).\end{aligned}$$

In order to get the Cramér-Rao lower bound for the unbiased estimators of β and σ^2 one first computes the information matrix

$$I(\beta,\sigma^2)=-E\begin{bmatrix}\partial^2\log L/\partial\beta\partial\beta' & \partial^2\log L/\partial\beta\partial\sigma^2\\ \partial^2\log L/\partial\sigma^2\partial\beta' & \partial^2\log L/\partial\sigma^2\partial\sigma^2\end{bmatrix} \tag{7.18}$$

Recall, that $\theta=\sigma^2$ and $Q=(y-X\beta)'(y-X\beta)$. It is easy to show (see problem 4) that

$$\frac{\partial^2\log L(\beta,\theta)}{\partial\beta\partial\theta}=\frac{1}{2\theta^2}\frac{\partial Q}{\partial\beta} \quad \text{and} \quad \frac{\partial^2\log L(\beta,\theta)}{\partial\theta\partial\beta}=\frac{-X'(y-X\beta)}{\theta^2}$$

Therefore,

$$E\left(\frac{\partial^2\log L(\beta,\theta)}{\partial\theta\partial\beta}\right)=\frac{-E(X'u)}{\theta^2}=0$$

Also

$$\frac{\partial^2\log L(\beta,\theta)}{\partial\beta\partial\beta'}=\frac{-X'X}{\theta} \quad \text{and} \quad \frac{\partial^2\log L(\beta,\theta)}{\partial\theta^2}=\frac{-4Q}{4\theta^3}+\frac{2n}{4\theta^2}=\frac{-Q}{\theta^3}+\frac{n}{2\theta^2}$$

so that

$$E\left(\frac{\partial^2\log L(\beta,\theta)}{\partial\theta^2}\right)=\frac{-n\theta}{\theta^3}+\frac{n}{2\theta^2}=\frac{-2n+n}{2\theta^2}=\frac{-n}{2\theta^2}$$

using the fact that $E(Q) = n\sigma^2 = n\theta$. Hence,

$$I(\beta, \sigma^2) = \begin{bmatrix} X'X/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix} \tag{7.19}$$

The information matrix is block-diagonal between β and σ^2. This is an important property for regression models with normal disturbances. It implies that the Cramér-Rao lower bound is

$$I^{-1}(\beta, \sigma^2) = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{bmatrix} \tag{7.20}$$

Note that $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$ attains the Cramér-Rao lower bound. Under normality, $\widehat{\beta}_{OLS}$ is MVU (minimum variance unbiased). This is best among all unbiased estimators not only *linear* unbiased estimators. By assuming more (in this case normality) we get more (MVU rather than BLUE)[4].

Problem 5 derives the variance of s^2 under normality of the disturbances. This is found to be $2\sigma^4/(n-k)$. This means that s^2 does not attain the Cramér-Rao lower bound. However, following the theory of complete sufficient statistics one can show that both $\widehat{\beta}_{OLS}$ and s^2 are MVU for their respective parameters and therefore both are small sample efficient. Note also that $\widehat{\sigma}^2_{MLE}$ is biased, therefore it is not meaningful to compare its variance to the Cramér-Rao lower bound. There is a trade-off between bias and variance in estimating σ^2. Problem 6 looks at all estimators of σ^2 of the type $e'e/r$ and derives r such that the mean squared error (MSE) is minimized. The choice of r turns out to be $(n-k+2)$.

We found the distribution of $\widehat{\beta}_{OLS}$, now we derive the distribution of s^2. In order to do that we need a result from matrix algebra, which is stated without proof, see Graybill (1961).

Lemma 1: For every symmetric idempotent matrix A of rank r, there exists an orthogonal matrix P such that $P'AP = J_r$ where J_r is a diagonal matrix with the first r elements equal to one and the rest equal to zero.

We use this lemma to show that the RSS/σ^2 is a chi-squared with $(n-k)$ degrees of freedom. To see this note that $e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2$ and that $\bar{P}_X$ is symmetric and idempotent of rank $(n-k)$. Using the lemma there exists a matrix P such that $P'\bar{P}_X P = J_{n-k}$ is a diagonal matrix with the first $(n-k)$ elements on the diagonal equal to 1 and the last k elements equal to zero. Now make the change of variable $v = P'u$. This makes $v \sim N(0, \sigma^2 I_n)$ since the v's are linear combinations of the u's and $P'P = I_n$. Replacing u by v in RSS/σ^2 we get

$$v'P\bar{P}_X Pv/\sigma^2 = v'J_{n-k}v/\sigma^2 = \textstyle\sum_{i=1}^{n-k} v_i^2/\sigma^2$$

where the last sum is only over $i = 1, 2, \ldots, n-k$. But, the v's are independent identically distributed $N(0, \sigma^2)$, hence v_i^2/σ^2 is the square of a standardized $N(0,1)$ random variable which is distributed as a χ^2_1. Moreover, the sum of independent χ^2 random variables is a χ^2 random variable with degrees of freedom equal to the sum of the respective degrees of freedom. Hence, RSS/σ^2 is distributed as χ^2_{n-k}.

The beauty of the above result is that it applies to all quadratic forms $u'Au$ where A is symmetric and idempotent. We will use this result again in the test of hypotheses section.

7.5 Prediction

Let us now predict T_o periods ahead. Those new observations are assumed to satisfy (7.1). In other words

$$y_o = X_o\beta + u_o \tag{7.21}$$

What is the Best Linear Unbiased Predictor (BLUP) of $E(y_o)$? From (7.21), $E(y_o) = X_o\beta$ which is a linear combination of the β's. Using the Gauss-Markov result $\widehat{y}_o = X_o\widehat{\beta}_{OLS}$ is BLUE for $X_o\beta$ and the variance of this predictor of $E(y_o)$ is $X_o\text{var}(\widehat{\beta}_{OLS})X_o' = \sigma^2 X_o(X'X)^{-1}X_o'$. But, what if we are interested in the predictor for y_o? The best predictor of u_o is zero, so the predictor for y_o is still $\widehat{y}_o$ but its MSE is

$$\begin{aligned} E(\widehat{y}_o - y_o)(\widehat{y}_o - y_o)' &= E\{X_o(\widehat{\beta}_{OLS} - \beta) - u_o\}\{X_o(\widehat{\beta}_{OLS} - \beta) - u_o\}' \\ &= X_o\text{var}(\widehat{\beta}_{OLS})X_o' + \sigma^2 I_{T_o} - 2\text{cov}\{X_o(\widehat{\beta}_{OLS} - \beta), u_o\} \\ &= \sigma^2 X_o(X'X)^{-1}X_o' + \sigma^2 I_{T_o} \end{aligned} \tag{7.22}$$

the last equality follows from the fact that $(\widehat{\beta}_{OLS} - \beta) = (X'X)^{-1}X'u$ and u_o have zero covariance. The latter holds because u_o and u have zero covariance. Intuitively this says that the future T_o disturbances are not correlated with the current sample disturbances.

Therefore, the predictor of the average consumption of a \$20,000 income household is the same as the predictor of consumption of a specific household whose income is \$20,000. The difference is not in the predictor itself but in the MSE attached to it. The latter MSE being larger.

Salkever (1976) suggested a simple way to compute these forecasts and their standard errors. The basic idea is to augment the usual regression in (7.1) with a matrix of observation-specific dummies, i.e., a dummy variable for each period where we want to forecast:

$$\begin{bmatrix} y \\ y_o \end{bmatrix} = \begin{bmatrix} X & 0 \\ X_o & I_{T_o} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} u \\ u_o \end{bmatrix} \tag{7.23}$$

or

$$y^* = X^*\delta + u^* \tag{7.24}$$

where $\delta' = (\beta', \gamma')$. X^* has in its second part a matrix of dummy variables one for each of the T_o periods for which we are forecasting. Since these T_o observations do not serve in the estimation, problem 7 asks the reader to verify that OLS on (7.23) yields $\widehat{\delta}' = (\widehat{\beta}', \widehat{\gamma}')$ where $\widehat{\beta} = (X'X)^{-1}X'y$, $\widehat{\gamma} = y_o - \widehat{y}_o$, and $\widehat{y}_o = X_o\widehat{\beta}$. In other words, OLS on (7.23) yields the OLS estimate of β without the T_o observations, and the coefficients of the T_o dummies are the forecast errors. This also means that the first n residuals are the usual OLS residuals $e = y - X\widehat{\beta}$ based on the first n observations, whereas the next T_o residuals are all zero. Therefore, $s^{*2} = s^2 = e'e/(n-k)$, and the variance covariance matrix of $\widehat{\delta}$ is given by

$$s^2(X^{*\prime}X^*)^{-1} = s^2 \begin{bmatrix} (X'X)^{-1} & \\ & [I_{T_o} + X_o(X'X)^{-1}X_o'] \end{bmatrix} \tag{7.25}$$

and the off-diagonal elements are of no interest. This means that the regression package gives the estimated variance of $\widehat{\beta}$ and the estimated variance of the forecast error in one stroke. Note that if the forecasts rather than the forecast errors are needed, one can replace y_o by zero, and I_{T_o} by $-I_{T_o}$ in (7.23). The resulting estimate of γ will be $\widehat{y}_o = X_o\widehat{\beta}$, as required. The variance of this forecast will be the same as that given in (7.25), see problem 7.

7.6 Confidence Intervals and Test of Hypotheses

We start by constructing a confidence interval for any linear combination of β, say $c'\beta$. We know that $c'\widehat{\beta}_{OLS} \sim N(c'\beta, \sigma^2 c'(X'X)^{-1}c)$ and it is a scalar. Hence,

$$z_{obs} = (c'\widehat{\beta}_{OLS} - c'\beta)/\sigma(c'(X'X)^{-1}c)^{1/2} \tag{7.26}$$

is a standardized $N(0,1)$ random variable. Replacing σ by s is equivalent to dividing z_{obs} by the square root of a χ^2 random variable divided by its degrees of freedom. The latter random variable is $(n-k)s^2/\sigma^2 = RSS/\sigma^2$ which was shown to be a χ^2_{n-k}. Problem 8 shows that z_{obs} and RSS/σ^2 are independent. This means that

$$t_{obs} = (c'\widehat{\beta}_{OLS} - c'\beta)/s(c'(X'X)^{-1}c)^{1/2} \tag{7.27}$$

is a $N(0,1)$ random variable divided by the square root of an independent $\chi^2_{n-k}/(n-k)$. This is a t-statistic with $(n-k)$ degrees of freedom. Hence, a $100(1-\alpha)\%$ confidence interval for $c'\beta$ is

$$c'\widehat{\beta}_{OLS} \pm t_{\alpha/2} s(c'(X'X)^{-1}c)^{1/2} \tag{7.28}$$

Example: Let us say we are predicting one year ahead so that $T_o = 1$ and x_o is a $(1 \times k)$ vector of next year's observations on the exogenous variables. The $100(1-\alpha)$ confidence interval for next year's forecast of y_o will be $\widehat{y}_o \pm t_{\alpha/2} s(1 + x'_o(X'X)^{-1}x_o)^{1/2}$. Similarly (7.28) allows us to construct confidence intervals or test any single hypothesis on any single β_j (again by picking c to have 1 in its j-th position and zero elsewhere). In this case we get the usual t-statistic reported in any regression package. More importantly, this allows us to test any hypothesis concerning any linear combination of the β's, e.g., testing that the sum of coefficients of input variables in a Cobb-Douglas production function is equal to one. This is known as a test for constant returns to scale, see Chapter 4.

7.7 Joint Confidence Intervals and Test of Hypotheses

We have learned how to test any single hypothesis involving any linear combination of the β's. But what if we are interested in testing two or three or more hypotheses involving linear combinations of the β's. For example, testing that $\beta_2 = \beta_4 = 0$, i.e., that variables X_2 and X_4 are not significant in the model. This can be written as $c'_2\beta = c'_4\beta = 0$ where c'_j is a row vector of zeros with a one in the j-th position. In order to test these two hypotheses simultaneously, we rearrange these restrictions on the β's in matrix form $R\beta = 0$ where $R' = [c_2, c_4]$. In a similar fashion, we can rearrange g restrictions on the β's into this matrix R which will now be of dimension $(g \times k)$. Also these restrictions need not be of the form $R\beta = 0$ and can be of the more general form $R\beta = r$ where r is a $(g \times 1)$ vector of constants. For example, $\beta_1 + \beta_2 = 1$ and $3\beta_3 + 2\beta_4 = 5$ are two such restrictions. Since $R\beta$ is a collection of linear combinations of the β's, the BLUE of these is $R\widehat{\beta}_{OLS}$ and the latter is distributed $N(R\beta, \sigma^2 R(X'X)^{-1}R')$. Standardization of the form encountered with the scalar $c'\beta$ gives us the following:

$$(R\widehat{\beta}_{OLS} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - R\beta)/\sigma^2 \tag{7.29}$$

rather than divide by the variance we multiply by its inverse, and since we divided by the variance rather than the standard deviation we square the numerator which means in vector form premultiplying by its transpose. Problem 9 replaces the matrix R by the vector c' and shows that (7.29) reduces to the square of the z-statistic observed in (7.26). This also proves that the resulting statistic is distributed as χ_1^2. But, what is the distribution of (7.29)? The trick is to write it in terms of the original disturbances, i.e.,

$$u'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'u/\sigma^2 \tag{7.30}$$

where $(R\widehat{\beta}_{OLS} - R\beta)$ is replaced by $R(X'X)^{-1}X'u$. Note that (7.30) is quadratic in the disturbances u of the form $u'Au/\sigma^2$. Problem 10 shows that A is symmetric and idempotent and of rank g. Applying the same proof as given below lemma 1 we get the result that (7.30) is distributed as χ_g^2. Again σ^2 is unobserved, so we divide by $(n-k)s^2/\sigma^2$ which is χ_{n-k}^2. This becomes a ratio of two χ^2's random variables. If we divide the numerator and denominator χ^2's by their respective degrees of freedom and prove that they are independent (see problem 11) the resulting statistic

$$(R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/gs^2 \tag{7.31}$$

is distributed under the null $R\beta = r$ as an $F(g, n-k)$.

7.8 Restricted MLE and Restricted Least Squares

Maximizing the likelihood function given in (7.16) subject to $R\beta = r$ is equivalent to minimizing the residual sum of squares subject to $R\beta = r$. Forming the Lagrangian function

$$\Psi(\beta, \mu) = (y - X\beta)'(y - X\beta) + 2\mu'(R\beta - r) \tag{7.32}$$

and differentiating with respect to β and μ one gets

$$\partial\Psi(\beta, \mu)/\partial\beta = -2X'y + 2X'X\beta + 2R'\mu = 0 \tag{7.33}$$

$$\partial\Psi(\beta, \mu)/\partial\mu = 2(R\beta - r) = 0 \tag{7.34}$$

Solving for μ, we premultiply (7.33) by $R(X'X)^{-1}$ and use (7.34)

$$\widehat{\mu} = [R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \tag{7.35}$$

Substituting (7.35) in (7.33) we get

$$\widehat{\beta}_{RLS} = \widehat{\beta}_{OLS} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \tag{7.36}$$

The restricted least squares estimator of β differs from that of the unrestricted OLS estimator by the second term in (7.36) with the term in parentheses showing the extent to which the unrestricted OLS estimator satisfies the constraint. Problem 12 shows that $\widehat{\beta}_{RLS}$ is biased unless the restriction $R\beta = r$ is satisfied. However, its variance is always less than that of $\widehat{\beta}_{OLS}$. This brings in the trade-off between bias and variance and the MSE criteria which was discussed in Chapter 2.

The Lagrange Multiplier estimator $\widehat{\mu}$ is distributed $N(0, \sigma^2[R(X'X)^{-1}R']^{-1})$ under the null hypothesis. Therefore, to test $\mu = 0$, we use

$$\widehat{\mu}'[R(X'X)^{-1}R']\widehat{\mu}/\sigma^2 = (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \tag{7.37}$$

Since μ measures the cost of imposing the restriction $R\beta = r$, it is no surprise that the right hand side of (7.37) was already encountered in (7.29) and is distributed as χ_g^2.

7.9 Likelihood Ratio, Wald and Lagrange Multiplier Tests

Before we go into the derivations of these three classical tests for the null hypothesis $H_0; R\beta = r$, it is important for the reader to review the intuitive graphical explanation of these tests given in Chapter 2.

The Likelihood Ratio test of $H_0; R\beta = r$ is based upon the ratio $\lambda = \max\ell_r/\max\ell_u$, where $\max\ell_u$ and $\max\ell_r$ are the maximum values of the unrestricted and restricted likelihoods, respectively. Let us assume for simplicity that σ^2 is *known*, then

$$\max\ell_u = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\widehat{\beta}_{MLE})'(y - X\widehat{\beta}_{MLE})/2\sigma^2\}$$

where $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$. Denoting the unrestricted residual sum of squares by URSS, we have

$$\max\ell_u = (1/2\pi\sigma^2)^{n/2}\exp\{-URSS/2\sigma^2\}$$

Similarly, $\max\ell_r$ is given by

$$\max\ell_r = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\widehat{\beta}_{RMLE})'(y - X\widehat{\beta}_{RMLE})/2\sigma^2\}$$

where $\widehat{\beta}_{RMLE} = \widehat{\beta}_{RLS}$. Denoting the restricted residual sum of squares by RRSS, we have

$$\max\ell_r = (1/2\pi\sigma^2)^{n/2}\exp\{-RRSS/2\sigma^2\}$$

Therefore, $-2\log\lambda = (RRSS - URSS)/\sigma^2$. Let us find the relationship between these residual sums of squares.

$$\begin{aligned} e_r &= y - X\widehat{\beta}_{RLS} = y - X\widehat{\beta}_{OLS} - X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) = e - X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) \qquad (7.38)\\ e_r'e_r &= e'e + (\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS})'X'X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) \end{aligned}$$

where e_r denotes the restricted residuals and $e_r'e_r$ the RRSS. The cross-product terms drop out because $X'e = 0$. Substituting the value of $(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS})$ from (7.36) into (7.38), we get:

$$RRSS - URSS = (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \qquad (7.39)$$

It is now clear that $-2\log\lambda$ is the right hand side of (7.39) divided by σ^2. In fact, this Likelihood Ratio (LR) statistic is the same as that given in (7.37) and (7.29). Under the null hypothesis $R\beta = r$, this was shown to be a χ^2_g.

The Wald test of $R\beta = r$ is based upon the unrestricted estimator and the extent of which it satisfies the restriction. More formally, if $r(\beta) = 0$ denote the vector of g restrictions on β and $R(\widehat{\beta}_{MLE})$ denotes the $(g \times k)$ matrix of partial derivatives $\partial r(\beta)/\partial\beta'$ evaluated at $\widehat{\beta}_{MLE}$, then the Wald statistic is given by

$$W = r(\widehat{\beta}_{MLE})'[R(\widehat{\beta}_{MLE})I(\widehat{\beta}_{MLE})^{-1}R(\widehat{\beta}_{MLE})']^{-1}r(\widehat{\beta}_{MLE}) \qquad (7.40)$$

where $I(\beta) = -E(\partial^2\log L/\partial\beta\partial\beta')$. In this case, $r(\beta) = R\beta - r, R(\widehat{\beta}_{MLE}) = R$ and $I(\widehat{\beta}_{MLE}) = (X'X)/\sigma^2$ as seen in (7.19). Therefore,

$$W = (R\widehat{\beta}_{MLE} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{MLE} - r)/\sigma^2 \qquad (7.41)$$

which is the same as the LR statistic[5].

The Lagrange Multiplier test is based upon the restricted estimator. In section 7.8, we derived the restricted estimator and the estimated Lagrange Multiplier $\widehat{\mu}$. The Lagrange Multiplier μ is the cost or shadow price of imposing the restrictions $R\beta = r$. If these restrictions are true, one would expect the estimated Lagrange Multiplier $\widehat{\mu}$ to have mean zero. Therefore, a test for the null hypothesis that $\mu = 0$, is called the LM test and the corresponding test statistic is given in equation (7.37). Alternatively, one can derive the LM test as a score' test based on the score or the first derivative of the log-likelihood function i.e., $S(\beta) = \partial \text{log}L/\partial\beta$. The score is zero for the unrestricted MLE, and the score test is based upon the departure of $S(\beta)$, evaluated at the restricted estimator $\widehat{\beta}_{RMLE}$, from zero. In this case, the score form of the LM statistic is given by

$$LM = S(\widehat{\beta}_{RMLE})'I(\widehat{\beta}_{RMLE})^{-1}S(\widehat{\beta}_{RMLE}) \tag{7.42}$$

For our model, $S(\beta) = (X'y - X'X\beta)/\sigma^2$ and from equation (7.36) we have

$$\begin{aligned} S(\widehat{\beta}_{RMLE}) &= X'(y - X\widehat{\beta}_{RMLE})/\sigma^2 \\ &= \{X'y - X'X\widehat{\beta}_{OLS} + R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)\}/\sigma^2 \\ &= R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \end{aligned}$$

Using (7.20), one gets $I^{-1}(\widehat{\beta}_{RMLE}) = \sigma^2(X'X)^{-1}$. Therefore, the score form of the LM test becomes

$$\begin{aligned} LM &= (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \\ &= (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \end{aligned} \tag{7.43}$$

This is numerically identical to the LM test derived in equation (7.37) and to the W and LR statistics derived above. Note that S$(\widehat{\beta}_{RMLE}) = R'\widehat{\mu}/\sigma^2$ from (7.35), so it is clear why the Score and the Lagrangian Multiplier tests are identical.

The score form of the LM test can also be obtained as a by-product of an artificial regression. In fact, $S(\beta)$ evaluated as $\widehat{\beta}_{RMLE}$ is given by

$$S(\widehat{\beta}_{RMLE}) = X'(y - X\widehat{\beta}_{RMLE})/\sigma^2$$

where $y - X\widehat{\beta}_{RMLE}$ is the vector of restricted residuals. If H_0 is true, then this converges asymptotically to u and the asymptotic variance of the vector of scores becomes $(X'X)/\sigma^2$. The score test is then based upon

$$(y - X\widehat{\beta}_{RMLE})'X(X'X)^{-1}X'(y - X\widehat{\beta}_{RMLE})/\sigma^2 \tag{7.44}$$

This expression is the explained sum of squares from the artificial regression of $(y - X\widehat{\beta}_{RMLE})/\sigma$ on X. To see that this is exactly identical to the LM test in equation (7.37), recall from equation (7.33) that $R'\widehat{\mu} = X'(y - X\widehat{\beta}_{RMLE})$ and substituting this expression for $R'\widehat{\mu}$ on the left hand side of equation (7.37) we get equation (7.44). In practice, σ^2 is estimated by $\widetilde{s}^2$ the Mean Square Error of the restricted regression. This is an example of the Gauss-Newton Regression which will be discussed in Chapter 8.

An alternative approach to testing H_0, is to estimate the restricted and unrestricted models and compute the following F-statistic

$$F_{obs} = \frac{(RRSS - URSS)/g}{URSS/(n-k)} \tag{7.45}$$

This statistic is known in the econometric literature as the Chow (1960) test and was encountered in Chapter 4. Note that from equation (7.39), if we divide the numerator by σ^2 we get a χ^2_g statistic divided by its degrees of freedom. Also, using the fact that $(n-k)s^2/\sigma^2$ is χ^2_{n-k}, the denominator divided by σ^2 is a χ^2_{n-k} statistic divided by its degrees of freedom. Problem 11 shows independence of the numerator and denominator and completes the proof that F_{obs} is distributed $F(g, n-k)$ under H_0.

Chow's (1960) Test for Regression Stability

Chow (1960) considered the problem of testing the equality of two sets of regression coefficients

$$y_1 = X_1\beta_1 + u_1 \quad \text{and} \quad y_2 = X_2\beta_2 + u_2 \tag{7.46}$$

where X_1 is $n_1 \times k$ and X_2 is $n_2 \times k$ with n_1 and $n_2 > k$. In this case, the unrestricted regression can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.47}$$

under the null hypothesis $H_0; \beta_1 = \beta_2 = \beta$, the restricted model becomes

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.48}$$

The URSS and the RRSS are obtained from these two regressions by stacking the $n_1 + n_2$ observations. It is easy to show that the $URSS = e_1'e_1 + e_2'e_2$ where e_1 is the OLS residuals from y_1 on X_1 and e_2 is the OLS residuals from y_2 on X_2. In other words, the URSS is the sum of two residual sums of squares from the separate regressions, see problem 13. The Chow F-statistic given in equation (7.45) has k and $(n_1 + n_2 - 2k)$ degrees of freedom, respectively. Equivalently, one can obtain this Chow F-statistic from running

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta_1 + \begin{bmatrix} 0 \\ X_2 \end{bmatrix} (\beta_2 - \beta_1) + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.49}$$

Note that the second set of explanatory variables whose coefficients are $(\beta_2 - \beta_1)$ are interaction variables obtained by multiplying each independent variable in equation (7.48) by a dummy variable, say D_2, that takes on the value 1 if the observation is from the second regression and 0 if it is from the first regression. A test for H_0; $\beta_1 = \beta_2$ becomes a joint test of significance for the coefficients of these interaction variables. Gujarati (1970) points out that this dummy variable approach has the additional advantage of giving the estimates of $(\beta_2 - \beta_1)$ and their t-statistics. If the Chow F-test rejects stability, these individual interaction dummy variable coefficients may point to the source of instability. Of course, one has to be careful with the interpretation of these individual t-statistics, after all they can all be insignificant with the joint F-statistic still being significant, see Maddala (1992).

In case one of the two regressions does not have sufficient observations to estimate a separate regression say $n_2 < k$, then one can proceed by running the regression on the full data set to get the RRSS. This is the restricted model because the extra n_2 observations are assumed to be generated by the same regression as the first n_1 observations. The URSS is the residual sums of squares based only on the longer period (n_1 observations). In this case, the Chow F-statistic given in equation (7.45) has n_2 and $n_1 - k$ degrees of freedom, respectively. This is known as Chow's *predictive test* since it tests whether the shorter n_2 observations are different from their predictions using the model with the longer n_1 observations. This predictive test can be performed with dummy variables as follows: Introduce n_2 observation specific dummies, one for each of the observations in the second regression. Test the joint significance of these n_2 dummy variables. Salkever's (1976) result applies and each dummy variable will have as its estimated coefficient the prediction error with its corresponding standard error and its t-statistic. Once, again, the individual dummies may point out possible outliers, but it is their joint significance that is under question.

The W, LR and LM Inequality

We have shown that the $LR = W = LM$ for linear restrictions if the log-likelihood is quadratic. However, this is not necessarily the case for more general situations. In fact, in the next chapter where we consider more general variance covariance structure on the disturbances, estimating this variance-covariance matrix destroys this equality and may lead to conflict in hypotheses testing as noted by Berndt and Savin (1977). In this case, $W \geq LR \geq LM$. See also the problems at the end of this chapter. The LR, W and LM tests are based on the *efficient* MLE. When *consistent* rather than *efficient* estimators are used, an alternative way of constructing the score-type test is known as Neyman's $C(\alpha)$. For details, see Bera and Permaratne (2001).

Although, these three tests are asymptotically equivalent, one test may be more convenient than another for a particular problem. For example, when the model is linear but the restriction is nonlinear, the unrestricted model is easier to estimate than the restricted model. So the Wald test suggests itself in that it relies only on the unrestricted estimator. Unfortunately, the Wald test has a drawback that the LR and LM test do not have. In finite samples, the Wald test is not invariant to testing two algebraically equivalent formulations of the nonlinear restriction. This fact has been pointed out in the econometric literature by Gregory and Veall (1985, 1986) and Lafontaine and White (1986). In what follows, we review some of Gregory and Veall's (1985) findings:

Consider the linear regression with two regressors

$$y_t = \beta_o + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \tag{7.50}$$

where the u_t's are IIN$(0, \sigma^2)$, and the nonlinear restriction $\beta_1\beta_2 = 1$. Two algebraically equivalent formulation of the null hypothesis are: H^A; $r^A(\beta) = \beta_1 - 1/\beta_2 = 0$, and H^B; $r^B(\beta) = \beta_1\beta_2 - 1 = 0$. The unrestricted maximum likelihood estimator is $\widehat{\beta}_{OLS}$ and the Wald statistic given in (7.40) is

$$W = r(\widehat{\beta}_{OLS})'[R(\widehat{\beta}_{OLS})\widehat{V}(\widehat{\beta}_{OLS})R'(\widehat{\beta}_{OLS})]^{-1}r(\widehat{\beta}_{OLS}) \tag{7.51}$$

where $\widehat{V}(\widehat{\beta}_{OLS})$ is the usual estimated variance-covariance matrix of $\widehat{\beta}_{OLS}$. Problem 19 asks the reader to verify that the Wald statistics corresponding to H_A and H_B using (7.51) are

$$W^A = (\widehat{\beta}_1\widehat{\beta}_2 - 1)^2/(\widehat{\beta}_2^2 v_{11} + 2v_{12} + v_{22}/\widehat{\beta}_2^2) \tag{7.52}$$

and

$$W^B = (\widehat{\beta}_1\widehat{\beta}_2 - 1)^2/(\widehat{\beta}_2^2 v_{11} + 2\widehat{\beta}_1\widehat{\beta}_2 v_{12} + \widehat{\beta}_1^2 v_{22}) \tag{7.53}$$

where the v_{ij}'s are the elements of $\widehat{V}(\widehat{\beta}_{OLS})$ for $i, j = 0, 1, 2$. These Wald statistics are clearly not identical, and other algebraically equivalent formulations of the null hypothesis can be generated with correspondingly different Wald statistics. Monte Carlo experiments were performed with 1000 replications on the model given in (7.50) with various values for β_1 and β_2, and for a sample size $n = 20, 30, 50, 100, 500$. The experiments were run when the null hypothesis is true and when it is false. For $n = 20$ and $\beta_1 = 10$, $\beta_2 = 0.1$, so that H_0 is satisfied, W^A rejects the null when it is true 293 times out of a 1000, while W^B rejects the null 65 times out of a 1000. At the 5% level one would expect 50 rejections with a 95% confidence interval $[36, 64]$. Both W^A and W^B reject too often but W^A performs worse than W^B. When n is increased to 500, W^A rejects 78 times while W^B rejects 39 times out of a 1000. W^A still rejects too often although its performance is better than that for $n = 20$, while W^B performs well and is within the 95% confidence region. When $n = 20$, $\beta_1 = 1$ and $\beta_2 = 0.5$, so that H_0 is *not* satisfied, W^A rejects the null when it is false 65 times out of a 1000 whereas W^B rejects it 584 times out of a 1000. For $n = 500$, both test statistics reject the null in 1000 out of 1000 times. Even in cases where the empirical sizes of the tests appear similar, see Table 1 of Gregory and Veall (1985), in particular the case where $\beta_1 = \beta_2 = 1$, Gregory and Veall find that W^A and W^B are in conflict about 5% of the time for $n = 20$, and this conflict drops to 0.5% at $n = 500$. Problem 20 asks the reader to derive four Wald statistics corresponding to four algebraically equivalent formulations of the *common factor* restriction analyzed by Hendry and Mizon (1978). Gregory and Veall (1986) give Monte Carlo results on the performance of these Wald statistics for various sample sizes. Once again they find conflict among these tests even when their empirical sizes appear to be similar. Also, the differences among the Wald statistics are much more substantial, and persist even when n is as large as 500.

Lafontaine and White (1985) consider a simple regression

$$y = \alpha + \beta x + \gamma z + u$$

where y is log of per capita consumption of textiles, x is log of per capita real income and z is log of relative prices of textiles, with the data taken from Theil (1971, p. 102). The estimated equation is:

$$\begin{aligned}\widehat{y} = \;& 1.37 \;\; + 1.14x \; - \; 0.83z \\ & (0.31) \quad (0.16) \quad (0.04)\end{aligned}$$

with $\widehat{\sigma}^2 = 0.0001833$, and $n = 17$, with standard errors shown in parentheses. Consider the null hypothesis H_0; $\beta = 1$. Algebraically equivalent formulations of H_0 are H_k; $\beta^k = 1$ for any exponent k. Applying (7.40) with $r(\beta) = \beta^k - 1$ and $R(\beta) = k\beta^{k-1}$, one gets the Wald statistic

$$W_k = (\widehat{\beta}^k - 1)^2/[(k\widehat{\beta}^{k-1})^2 V(\widehat{\beta})] \tag{7.54}$$

where $\widehat{\beta}$ is the OLS estimate of β and $V(\widehat{\beta})$is its corresponding estimated variance. For every k, W_k has a limiting χ^2_1 distribution under H_0. The critical values are $\chi^2_{1,.05} = 3.84$ and $F^{.05}_{1,14} = 4.6$. The latter is an exact distribution test for $\beta = 1$ under H_0. Lafontaine and White (1985) try different integer exponents ($\pm k$) where $k = 1, 2, 3, 6, 10, 20, 40$. Using $\widehat{\beta} = 1.14$ and $V(\widehat{\beta}) = (0.16)^2$ one gets $W_{-20} = 24.56$, $W_1 = 0.84$, and $W_{20} = 0.12$. The authors conclude that one could get any Wald statistic desired by choosing an appropriate exponent. Since $\beta > 1$, W_k is inversely related to k. So, we can find a W_k that exceeds the critical values given by the χ^2 and F distributions. In fact, W_{-20} leads to rejection whereas W_1 and W_{20} do not reject H_0.

For testing nonlinear restrictions, the Wald test is easy to compute. However, it has a serious problem in that it is not invariant to the way the null hypothesis is formulated. In this case, the score test may be difficult to compute, but Neyman's $C(\alpha)$ test is convenient to use and provide the invariance that is needed, see Dagenais and Dufour (1991).

Notes

1. For example, in a time-series setting, including the time trend in the multiple regression is equivalent to detrending each variable first, by residualing out the effect of time, and then running the regression on these residuals.

2. Two exceptions noted in Davidson and MacKinnon (1993) are the following: One, if the model is not identified asymptotically. For example, $y_t = \beta(1/t) + u_t$ for $t = 1, 2, \ldots, T$, will have $(1/t)$ tend to zero as $T \to \infty$. This means that as the sample size increase, there is no information on β. Two, if the number of parameters in the model increase as the sample size increase. For example, the fixed effects model in panel data discussed in Chapter 12.

3. If the MLE of β is $\widehat{\beta}_{MLE}$, then the MLE of $(1/\beta)$ is $(1/\widehat{\beta}_{MLE})$. Note that this invariance property implies that MLE cannot be in general unbiased. For example, even if $\widehat{\beta}_{MLE}$ is unbiased for β, by the above reparameterization, $(1/\widehat{\beta}_{MLE})$ is not unbiased for $(1/\beta)$.

4. If the distribution of disturbances is not normal, then OLS is still BLUE as long as the assumptions underlying the Gauss-Markov Theorem are satisfied. The MLE in this case will be in general more efficient than OLS as long as the distribution of the errors is correctly specified.

5. Using the Taylor Series approximation of $r(\widehat{\beta}_{MLE})$ around the true parameter vector β, one gets $r(\widehat{\beta}_{MLE}) \simeq r(\beta) + R(\beta)(\widehat{\beta}_{MLE} - \beta)$. Under the null hypothesis, $r(\beta) = 0$ and the var$[r(\widehat{\beta}_{MLE})] \simeq$ $R(\beta)$ var$(\widehat{\beta}_{MLE})R'(\beta)$.

Problems

1. *Invariance of the Fitted Values and Residuals to Nonsingular Transformations of the Independent Variables.* Post-multiply the independent variables in (7.1) by a nonsingular transformation C, so that $X^* = XC$.

   (a) Show that $P_{X^*} = P_X$ and $\bar{P}_{X^*} = \bar{P}_X$. Conclude that the regression of y on X has the same fitted values and the same residuals as the regression of y on X^*.

   (b) As an application of these results, suppose that every X was multiplied by a constant, say, a change in the units of measurement. Would the fitted values or residuals change when we rerun this regression?

   (c) Suppose that X contains two regressors X_1 and X_2 each of dimension $n \times 1$. If we run the regression of y on $(X_1 - X_2)$ and $(X_1 + X_2)$, will this yield the same fitted values and the same residuals as the original regression of y on X_1 and X_2?

2. *The FWL Theorem.*

   (a) Using partitioned inverse results from the Appendix, show that the solution to (7.9) yields $\widehat{\beta}_{2,OLS}$ given in (7.10).

(b) Alternatively, write (7.9) as a system of two equations in two unknowns $\widehat{\beta}_{1,OLS}$ and $\widehat{\beta}_{2,OLS}$. Solve, by eliminating $\widehat{\beta}_{1,OLS}$ and show that the resulting solution is given by (7.10).

(c) Using the FWL Theorem, show that if $X_1 = \iota_n$ a vector of ones indicating the presence of the constant in the regression, and X_2 is a set of economic variables, then (i) $\widehat{\beta}_{2,OLS}$ can be obtained by running $y_i - \bar{y}$ on the set of variables in X_2 expressed as deviations from their respective sample means. (ii) The least squares estimate of the constant $\widehat{\beta}_{1,OLS}$ can be retrieved as $\bar{y} - \bar{X}_2'\widehat{\beta}_{2,OLS}$ where $\bar{X}_2' = \iota_n' X_2/n$ is the vector of sample means of the independent variables in X_2.

3. Let $y = X\beta + D_i\gamma + u$ where y is $n \times 1$, X is $n \times k$ and D_i is a dummy variable that takes the value 1 for the i-th observation and 0 otherwise. Using the FWL Theorem, prove that the least squares estimates of β and γ from this regression are $\widehat{\beta}_{OLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$ and $\widehat{\gamma}_{OLS} = y_i - x_i'\widehat{\beta}_{OLS}$, where X^* denotes the X matrix without the i-th observation and y^* is the y vector without the i-th observation and (y_i, x_i') denotes the i-th observation on the dependent and independent variables. This means that $\widehat{\gamma}_{OLS}$ is the forecasted OLS residual from the regression of y^* on X^* for the i-th observation which was essentially excluded from the regression by the inclusion of the dummy variable D_i.

4. *Maximum Likelihood Estimation.* Given the log-likelihood in (7.17),

(a) Derive the first-order conditions for maximization and show that $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$ and that $\widehat{\sigma}^2_{MLE} = RSS/n$.

(b) Calculate the second derivatives given in (7.18) and verify that the information matrix reduces to (7.19).

5. Given that $u \sim N(0, \sigma^2 I_n)$, we showed that $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$. Use this fact to prove that,

(a) s^2 is unbiased for σ^2.

(b) $\text{var}(s^2) = 2\sigma^4/(n-k)$. **Hint:** $E(\chi^2_r) = r$ and $\text{var}(\chi^2_r) = 2r$.

6. Consider all estimators of σ^2 of the type $\widetilde{\sigma}^2 = e'e/r = u'\bar{P}_X u/r$ with $u \sim N(0, \sigma^2 I_n)$.

(a) Find $E(\widehat{\sigma}^2_{MLE})$ and the $\text{bias}(\widehat{\sigma}^2_{MLE})$.

(b) Find $\text{var}(\widehat{\sigma}^2_{MLE})$ and the $\text{MSE}(\widehat{\sigma}^2_{MLE})$.

(c) Compute $\text{MSE}(\widetilde{\sigma}^2)$ and minimize it with respect to r. Compare with the MSE of s^2 and $\widehat{\sigma}^2_{MLE}$.

7. *Computing Forecasts and Forecast Standard Errors Using a Regression Package.* This is based on Salkever (1976). From equations (7.23) and (7.24), show that

(a) $\widehat{\delta}'_{OLS} = (\widehat{\beta}'_{OLS}, \widehat{\gamma}'_{OLS})$ where $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$, and $\widehat{\gamma}_{OLS} = y_o - X_o\widehat{\beta}_{OLS}$. **Hint:** Set up the OLS normal equations and solve two equations in two unknowns. Alternatively, one can use the FWL Theorem to residual out the additional T_o dummy variables.

(b) $e^*_{OLS} = (e'_{OLS}, 0')'$ and $s^{*2} = s^2$.

(c) $s^{*2}(X^{*\prime}X^*)^{-1}$ is given by the expression in (7.25). **Hint:** Use partitioned inverse.

(d) Replace y_o by 0 and I_{T_o} by $-I_{T_o}$ in (7.23) and show that $\widehat{\gamma} = \widehat{y}_o = X_o\widehat{\beta}_{OLS}$ whereas all the results in parts (a), (b) and (c) remain the same.

8. (a) Show that $\text{cov}(\widehat{\beta}_{OLS}, e) = 0$. (Since both random variables are normally distributed, this proves their independence).

(b) Show that $\widehat{\beta}_{OLS}$ and s^2 are independent. **Hint:** A linear (Bu) and quadratic $(u'Au)$ forms in normal random variables are independent if $BA = 0$. See Graybill (1961) Theorem 4.17.

9. (a) Show that if one replaces R by c' in (7.29) one gets the square of the z-statistic given in (7.26).

   (b) Show that when we replace σ^2 by s^2, the χ_1^2 statistic given in part (a) becomes the square of a t-statistic which is distributed as $F(1, n-K)$. **Hint:** The square of a $N(0,1)$ is χ_1^2. Also the ratio of two independent χ^2 random variables divided by their degrees of freedom is an F-statistic with these corresponding degrees of freedom, see Chapter 2.

10. (a) Show that the matrix A defined in (7.30) by $u'Au/\sigma^2$ is symmetric, idempotent and of rank g.

    (b) Using the same proof given below lemma 1, show that (7.30) is χ_g^2.

11. (a) Show that the two quadratic forms $s^2 = u'\bar{P}_X u/(n-k)$ and that given in (7.30) are independent. **Hint:** Two positive semi-definite quadratic forms $u'Au$ and $u'Bu$ are independent if and only if $AB = 0$, see Graybill (1961) Theorem 4.10.

    (b) Conclude that (7.31) is distributed as an $F(g, n-k)$.

12. *Restricted Least Squares.*

    (a) Show that $\widehat{\beta}_{RLS}$ given by (7.36) is biased unless $R\beta = r$.

    (b) Show that the var$(\widehat{\beta}_{RLS})$ = var$(A(X'X)^{-1}X'u)$ where

    $$A = I_K - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R.$$

    Prove that $A^2 = A$, but $A' \neq A$. Conclude that

    $$\text{var}(\widehat{\beta}_{RLS}) = \sigma^2 A(X'X)^{-1}A' = \sigma^2\{(X'X)^{-1} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}.$$

    (c) Show that var$(\widehat{\beta}_{OLS})-$ var$(\widehat{\beta}_{RLS})$ is a positive semi-definite matrix.

13. *The Chow Test.*

    (a) Show that OLS on (7.47) yields OLS on each equation separately in (7.46). In other words, $\widehat{\beta}_{1,OLS} = (X_1'X_1)^{-1}X_1'y_1$ and $\widehat{\beta}_{2,OLS} = (X_2'X_2)^{-1}X_2'y_2$.

    (b) Show that the residual sum of squares for equation (7.47) is given by $RSS_1 + RSS_2$, where RSS_i is the residual sum of squares from running y_i on X_i for $i = 1, 2$.

    (c) Show that the Chow F-statistic can be obtained from (7.49) by testing for the joint significance of H_o; $\beta_2 - \beta_1 = 0$.

14. Suppose we would like to test H_o; $\beta_2 = 0$ in the following unrestricted model given also in (7.8)

    $$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

    (a) Using the FWL Theorem, show that the URSS is identical to the residual sum of squares obtained from $\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\beta_2 + \bar{P}_{X_1}u$. Conclude that

    $$URSS = y'\bar{P}_X y = y'\bar{P}_{X_1}y - y'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y.$$

    (b) Show that the numerator of the F-statistic for testing H_o; $\beta_2 = 0$ which is given in (7.45), is $y'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y/k_2$.

    Substituting $y = X_1\beta_1 + u$ under the null hypothesis, show that the above expression reduces to $u'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}u/k_2$.

(c) Let $v = X_2'\bar{P}_{X_1}u$, show that if $u \sim \text{IIN}(0,\sigma^2)$ then $v \sim N(0,\sigma^2 X_2'\bar{P}_{X_1}X_2)$. Conclude that the numerator of the F-statistic given in part (b) when divided by σ^2 can be written as $v'[\text{var}(v)]^{-1}v/k_2$ where $v'[\text{var}(v)]^{-1}v$ is distributed as $\chi^2_{k_2}$under H_o. **Hint**: See the discussion below lemma 1.

(d) Using the result that $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$ where s^2 is the $URSS/(n-k)$, show that the F-statistic given by (7.45) is distributed as $F(k_2, n-k)$ under H_o. **Hint:** You need to show that $u'\bar{P}_X u$ is independent of the quadratic term given in part (b), see problem 11.

(e) Show that the Wald Test for H_o; $\beta_2 = 0$, given in (7.41), reduces in this case to $W = \widehat{\beta}_2'[R(X'X)^{-1}R']^{-1}\widehat{\beta}_2/s^2$ were $R = [0, I_{k_2}]$, $\widehat{\beta}_2$ denotes the OLS or equivalently the MLE of β_2 from the unrestricted model and s^2 is the corresponding estimate of σ^2 given by $URSS/(n-k)$. Using partitioned inversion or the FWL Theorem, show that the numerator of W is k_2 times the expression in part (b).

(f) Show that the score form of the LM statistic, given in (7.42) and (7.44), can be obtained as the explained sum of squares from the artificial regression of the restricted residuals $(y - X_1\widehat{\beta}_{1,RLS})$ deflated by $\widetilde{s}$ on the matrix of regressors X. In this case, $\widetilde{s}^2 = RRSS/(n-k_1)$ is the Mean Square Error of the restricted regression. In other words, obtain the explained sum of squares from regressing $\bar{P}_{X_1}y/\widetilde{s}$ on X_1 and X_2.

15. *Iterative Estimation in Partitioned Regression Models.* This is based on Fiebig (1995). Consider the partitioned regression model given in (7.8) and let X_2 be a single regressor, call it x_2 of dimension $n \times 1$ so that β_2 is a scalar. Consider the following strategy for estimating β_2: Estimate β_1 from the shortened regression of y on X_1. Regress the residuals from this regression on x_2 to yield $b_2^{(1)}$.

(a) Prove that $b_2^{(1)}$ is biased.

Now consider the following iterative strategy for re-estimating β_2:

Re-estimate β_1 by regressing $y - x_2 b_2^{(1)}$ on X_1 to yield $b_1^{(1)}$. Next iterate according to the following scheme:

$$b_1^{(j)} = (X_1'X_1)^{-1}X_1'(y - x_2 b_2^{(j)})$$
$$b_2^{(j+1)} = (x_2'x_2)^{-1}x_2'(y - X_1 b_1^{(j)}), \qquad j = 1, 2, ...$$

(b) Determine the behavior of the bias of $b_2^{(j+1)}$ as j increases.

(c) Show that as j increases $b_2^{(j+1)}$ converges to the estimator of β_2 obtained by running OLS on (7.8).

16. *Maddala (1992, pp. 120–127).* Consider the simple linear regression

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, 2, \ldots, n.$$

where α and β are scalars and $u_i \sim \text{IIN}(0,\sigma^2)$. For H_o; $\beta = 0$,

(a) Derive the Likelihood Ratio (LR) statistic and show that it can be written as $n\text{log}[1/(1-r^2)]$ where r^2 is the square of the correlation coefficient between X and y.

(b) Derive the Wald (W) statistic for testing H_o; $\beta = 0$. Show that it can be written as $nr^2/(1-r^2)$. This is the square of the usual t-statistic on $\widehat{\beta}$ with $\widehat{\sigma}^2_{MLE} = \sum_{i=1}^n e_i^2/n$ used instead of s^2 in estimating σ^2. $\widehat{\beta}$ is the unrestricted MLE which is OLS in this case, and the e_i's are the usual least squares residuals.

(c) Derive the Lagrange Multiplier (LM) statistic for testing H_o; $\beta = 0$. Show that it can be written as nr^2. This is the square of the usual t-statistic on $\widehat{\beta}$ with $\widetilde{\sigma}^2_{RMLE} = \sum_{i=1}^n (Y_i - \overline{Y})^2/n$ used instead of s^2 in estimating σ^2. The $\widetilde{\sigma}^2_{RMLE}$ is restricted MLE of σ^2 (i.e., imposing H_o and maximizing the likelihood with respect to σ^2).

(d) Show that $LM/n = (W/n)/[1 + (W/n)]$, and $LR/n = \log[1 + (W/n)]$. Using the following inequality $x \geq \log(1 + x) \geq x/(1 + x)$, conclude that $W \geq LR \geq LM$. **Hint:** Use $x = W/n$.

(e) For the cigarette consumption data given in Table 3.2, compute the W, LR, LM for the simple regression of log$C$ on log$P$ and demonstrate the above inequality given in part (d) for testing that the price elasticity is zero?

17. *Engle (1984, pp. 785–786).* Consider a set of T independent observations on a Bernoulli random variable which takes on the values $y_t = 1$ with probability θ, and $y_t = 0$ with probability $(1 - \theta)$.

(a) Derive the log-likelihood function, the MLE of θ, the score $S(\theta)$, and the information $I(\theta)$.

(b) Compute the LR, W and LM test statistics for testing H_o; $\theta = \theta_o$, versus H_A; $\theta \neq \theta_o$ for $\theta \epsilon (0, 1)$.

18. *Engle (1984, pp. 787–788).* Consider the linear regression model

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

given in (7.8), where $u \sim N(0, \sigma^2 I_T)$.

(a) Write down the log-likelihood function, find the MLE of β and σ^2.

(b) Write down the score $S(\beta)$ and show that the information matrix is block-diagonal between β and σ^2.

(c) Derive the W, LR and LM test statistics in order to test H_o; $\beta_1 = \beta_1^o$, versus H_A; $\beta_1 \neq \beta_1^o$, where β_1 is say the first k_1 elements of β. Show that if $X = [X_1, X_2]$, then

$$\begin{aligned}
W &= (\beta_1^o - \widehat{\beta}_1)'[X_1'\bar{P}_{X_2}X_1](\beta_1^o - \widehat{\beta}_1)/\widehat{\sigma}^2 \\
LM &= \widetilde{u}'X_1[X_1'\bar{P}_{X_2}X_1]^{-1}X_1'\widetilde{u}/\widetilde{\sigma}^2 \\
LR &= T\ \log(\widetilde{u}'\widetilde{u}/\widehat{u}'\widehat{u})
\end{aligned}$$

where $\widehat{u} = y - X\widehat{\beta}$, $\widetilde{u} = y - X\widetilde{\beta}$ and $\widehat{\sigma}^2 = \widehat{u}'\widehat{u}/T$, $\widetilde{\sigma}^2 = \widetilde{u}'\widetilde{u}/T$. $\widehat{\beta}$ is the unrestricted MLE, whereas $\widetilde{\beta}$ is the restricted MLE.

(d) Using the above results, show that

$$\begin{aligned}
W &= T(\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u})/\widehat{u}'\widehat{u} \\
LM &= T(\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u})/\widetilde{u}'\widetilde{u}
\end{aligned}$$

Also, that $LR = T\ \log[1 + (W/T)]$; $LM = W/[1 + (W/T)]$; and $(T - k)W/Tk_1 \sim F_{k_1, T-k}$ under H_o. As in problem 16, we use the inequality $x \geq \log(1 + x) \geq x/(1 + x)$ to conclude that $W \geq LR \geq LM$. **Hint:** Use $x = W/T$. However, it is important to note that all the test statistics are monotonic functions of the F-statistic and exact tests for each would produce identical critical regions.

(e) For the cigarette consumption data given in Table 3.2, run the following regression:

$$\text{log}C = \alpha + \beta\text{log}P + \gamma\text{log}Y + u$$

compute the W, LR, LM given in part (c) for the null hypothesis H_o; $\beta = -1$.

(f) Compute the Wald statistics for H_o^A; $\beta = -1$, H_o^B; $\beta^5 = -1$ and H_o^C; $\beta^{-5} = -1$. How do these statistics compare?

19. *Gregory and Veall (1985).* Using equation (7.51) and the two formulations of the null hypothesis H^A and H^B given below (7.50), verify that the Wald statistics corresponding to these two formulations are those given in (7.52) and (7.53), respectively.

20. *Gregory and Veall (1986).* Consider the dynamic equation

$$y_t = \rho y_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + u_t$$

where $|\rho| < 1$, and $u_t \sim \text{NID}(0, \sigma^2)$. Note that for this equation to be the Cochrane-Orcutt transformation

$$y_t - \rho y_{t-1} = \beta_1(x_t - \rho x_{t-1}) + u_t$$

the following nonlinear restriction must be satisfied $-\beta_1\rho = \beta_2$ called the *common factor* restriction by Hendry and Mizon (1978). Now consider the following four formulations of this restriction H^A; $\beta_1\rho + \beta_2 = 0$; H^B; $\beta_1 + (\beta_2/\rho) = 0$; H^C; $\rho + (\beta_2/\beta_1) = 0$ and H^D; $(\beta_1\rho/\beta_2) + 1 = 0$.

(a) Using equation (7.51) derive the four Wald statistics corresponding to the four formulations of the null hypothesis.

(b) Apply these four Wald statistics to the equation relating real personal consumption expenditures to real disposable personal income in the U.S. over the post World War II period 1959–2007, see Table 5.3.

21. *Effect of Additional Regressors on R^2.* This problem was considered in non-matrix form in Chapter 4, problem 4. Regress y on X_1 which is $T \times K_1$ and compute SSE_1. Add X_2 which is $T \times K_2$ so that the number of regressors in now $K = K_1 + K_2$. Regress y on $X = [X_1, X_2]$ and get SSE_2. Show that $SSE_2 \leq SSE_1$. Conclude that the corresponding R-squares satisfy $R_2^2 \geq R_1^2$. **Hint:** Show that $P_X - P_{X_1}$ is a positive semi-definite matrix.

References

Additional readings for the material covered in this chapter can be found in Davidson and MacKinnon (1993), Kelejian and Oates (1989), Maddala (1992), Fomby, Hill and Johnson (1984), Greene (1993), Johnston (1984), Judge et al. (1985) and Theil (1971). These econometrics texts were cited earlier. Other references cited in this chapter are the following:

Bera A.K. and G. Permaratne (2001), "General Hypothesis Testing," Chapter 2 in Baltagi, B.H. (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Berndt, E.R. and N.E. Savin (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrica*, 45: 1263–1278.

Buse, A.(1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36: 153–157.

Chow, G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28: 591–605.

Dagenais, M.G. and J. M. Dufour (1991), "Invariance, Nonlinear Models, and Asymptotic Tests," *Econometrica*, 59: 1601–1615.

Engle, R.F. (1984), "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," In: Griliches, Z. and M.D. Intrilligator (eds) *Handbook of Econometrics* (North-Holland: Amsterdam).

Fiebig, D.G. (1995), "Iterative Estimation in Partitioned Regression Models," *Econometric Theory*, Problem 95.5.1, 11:1177.

Frisch, R., and F.V. Waugh (1933), “Partial Time Regression as Compared with Individual Trends,” *Econometrica*, 1: 387–401.

Graybill, F.A.(1961), *An Introduction to Linear Statistical Models*, Vol. 1 (McGraw-Hill: New York).

Gregory, A.W. and M.R. Veall (1985), “Formulating Wald Tests of Nonlinear Restrictions,” *Econometrica*, 53: 1465–1468.

Gregory, A.W. and M.R. Veall (1986), “Wald Tests of Common Factor Restrictions,” *Economics Letters*, 22: 203–208.

Gujarati, D. (1970), “Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Generalization,” *The American Statistician*, 24: 50–52.

Hendry, D.F. and G.E. Mizon (1978), “Serial Correlation as a Convenient Simplification, Not as a Nuisance: A Comment on A Study of the Demand for Money by the Bank of England,” *Economic Journal*, 88: 549–563.

Lafontaine, F. and K.J. White (1986), “Obtaining Any Wald Statistic You Want,” *Economics Letters*, 21: 35–40.

Lovell, M.C. (1963), “Seasonal Adjustment of Economic Time Series,” *Journal of the American Statistical Association*, 58: 993–1010.

Salkever, D. (1976), “The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals,” *Journal of Econometrics*, 4: 393–397.

Appendix
Some Useful Matrix Properties

This book assumes that the reader has encountered matrices before, and knows how to add, subtract and multiply conformable matrices. In addition, that the reader is familiar with the transpose, trace, rank, determinant and inverse of a matrix. Unfamiliar readers should consult standard texts like Bellman (1970) or Searle (1982). The purpose of this Appendix is to review some useful matrix properties that are used in the text and provide easy access to these properties. Most of these properties are given without proof.

Starting with Chapter 7, our data matrix X is organized such that it has n rows and k columns, so that each row denotes an observation on k variables and each column denotes n observations on one variable. This matrix is of dimension $n \times k$. The rank of an $n \times k$ matrix is always less than or equal to its smaller dimension. Since $n > k$, the rank $(X) \leq k$. When there is no perfect multicollinearity among the variables in X, this matrix is said to be of full column rank k. In this case, $X'X$, the matrix of cross-products is of dimension $k \times k$. It is square, symmetric and of full rank k. This uses the fact that the rank$(X'X) = \text{rank}(X) = k$. Therefore, $(X'X)$ is nonsingular and the inverse $(X'X)^{-1}$ exists. This is needed for the computation of *Ordinary Least Squares*. In fact, for least squares to be feasible, X should be of full column rank k and no variable in X should be a perfect linear combination of the other variables in X. If we write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

where x_i' denotes the i-th observation, in the data, then $X'X = \sum_{i=1}^{n} x_i x_i'$ where x_i is a column vector of dimension $k \times 1$.

An important and widely encountered matrix is the *Identity matrix* which will be denoted by I_n and subscripted by its dimension n. This is a square $n \times n$ matrix whose diagonal elements are all equal to one and its off diagonal elements are all equal to zero. Also, $\sigma^2 I_n$ will be a familiar *scalar covariance matrix*, with every diagonal element equal to σ^2 reflecting *homoskedasticity* or equal variances (see Chapter 5), and zero covariances or no *serial correlation* (see Chapter 5). Let

$$\Omega = \ \text{diag}[\sigma_i^2] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

be an $(n \times n)$ *diagonal* matrix with the i-th diagonal element equal to σ_i^2 for $i = 1, 2, \ldots, n$. This matrix will be encountered under heteroskedasticity, see Chapter 9. Note that $\text{tr}(\Omega) = \sum_{i=1}^{n} \sigma_i^2$ is the sum of its diagonal elements. Also, $\text{tr}(I_n) = n$ and $\text{tr}(\sigma^2 I_n) = n\sigma^2$. Another useful matrix is the *projection matrix* $P_X = X(X'X)^{-1}X'$ which is of dimension $n \times n$. This matrix is encountered in Chapter 7. If y denotes the $n \times 1$ vector of observations on the dependent variable, then $P_X y$ generates the predicted values $\widehat{y}$ from the least squares regression of y on X. This matrix P_X is symmetric and *idempotent*. This means that $P'_X = P_X$ and $P_X^2 = P_X P_X = P_X$ as can be easily verified. Some of the properties of idempotent matrices is that their rank is equal to their trace. Hence, $\text{rank}(P_X) = \text{tr}(P_X) = \text{tr}[X(X'X)^{-1}X'] = \text{tr}[X'X(X'X^{-1})] = \text{tr}(I_k) = k$.

Here, we used the fact that $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$. In other words, the trace is unaffected by the cyclical permutation of the product. Of course, these matrices should be conformable and the product should result in a square matrix. Note that $\bar{P}_X = I_n - P_X$ is also a symmetric and idempotent matrix. In this case, $\bar{P}_X y = y - P_X y = y - \widehat{y} = e$ where e denotes the least squares residuals, $y - X\widehat{\beta}_{OLS}$ where $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$, see Chapter 7. Some properties of these projection matrices are the following:

$$P_X X = X, \bar{P}_X X = 0, \bar{P}_X e = e \quad \text{and} \quad P_X e = 0.$$

In fact, $X'e = 0$ means that the matrix X is *orthogonal* to the vector of least squares residuals e. Note that $X'e = 0$ means that $X'(y - X\widehat{\beta}_{OLS}) = 0$ or $X'y = X'X\widehat{\beta}_{OLS}$. These k equations are known as the OLS normal equations and their solution yields the least squares estimates $\widehat{\beta}_{OLS}$. By the definition of $\bar{P}_X$, we have (i) $P_X + \bar{P}_X = I_n$. Also, (ii) P_X and $\bar{P}_X$ are idempotent and (iii) $P_X \bar{P}_X = 0$. In fact, any two of these properties imply the third. The $\text{rank}(\bar{P}_X) = \text{tr}(\bar{P}_X) = \text{tr}(I_n - P_X) = n - k$. Note that P_X and $\bar{P}_X$ are of rank k and $(n - k)$, respectively. Both matrices are not of full column rank. In fact, the only full rank, symmetric idempotent matrix is the identity matrix.

Matrices not of full rank are singular, and their inverse do not exist. However, one can find a *generalized inverse* of a matrix Ω which we will call Ω^- which satisfies the following requirements:

$$\begin{array}{lcl} \text{(i) } \Omega\Omega^-\Omega = \Omega & & \text{(ii) } \Omega^-\Omega\Omega^- = \Omega^- \\ \text{(iii) } \Omega^-\Omega \text{ is symmetric} & \text{and} & \text{(iv) } \Omega\Omega^- \text{ is symmetric.} \end{array}$$

Even if Ω is not square, a unique Ω^- can be found for Ω which satisfies the above four properties. This is called the *Moore-Penrose* generalized inverse.

Note that a symmetric idempotent matrix is its own Moore-Penrose generalized inverse. For example, it is easy to verify that if $\Omega = P_X$, then $\Omega^- = P_X$ and that it satisfies the above four properties. Idempotent matrices have *characteristic roots* that are either zero or one. The number of non-zero characteristic roots is equal to the rank of this matrix. The characteristic roots of Ω^{-1} are the reciprocals of the characteristic roots of Ω, but the characteristic vectors of both matrices are the same.

The determinant of a matrix is non-zero if and only if it has full rank. Therefore, if A is singular, then $|A| = 0$. Also, the determinant of a matrix is equal to the product of its characteristic roots. For two square matrices A and B, the determinant of the product is the product of the determinants $|AB| = |A| \cdot |B|$. Therefore, the determinant of Ω^{-1} is the reciprocal of the determinant of Ω. This follows from the fact that $|\Omega||\Omega^{-1}| = |\Omega\Omega^{-1}| = |I| = 1$. This property is used in writing the likelihood function for *Generalized Least Squares* (GLS) estimation, see Chapter 9. The determinant of a triangular matrix is equal to the product of its diagonal elements. Of course, it immediately follows that the determinant of a diagonal matrix is the product of its diagonal elements.

The constant in the regression corresponds to a vector of ones in the matrix of regressors X. This vector of ones is denoted by ι_n where n is the dimension of this column vector. Note that $\iota_n'\iota_n = n$ and $\iota_n\iota_n' = J_n$ where J_n is a matrix of ones of dimension $n \times n$. Note also that J_n is not idempotent, but $\bar{J}_n = J_n/n$ is idempotent as can be easily verified. The $\text{rank}(\bar{J}_n) = \text{tr}(\bar{J}_n) = 1$. Note also that $I_n - \bar{J}_n$ is idempotent with rank $(n-1)$. $\bar{J}_n y$ has a typical element $\bar{y} = \sum_{i=1}^n y_i/n$ whereas $(I_n - \bar{J}_n)y$ has a typical element $(y_i - \bar{y})$. So that $\bar{J}_n$ is the *averaging* matrix, whereas premultiplying by $(I_n - \bar{J}_n)$ results in deviations from the mean.

For two nonsingular matrices A and B

$$(AB)^{-1} = B^{-1}A^{-1}$$

Also, the transpose of a product of two conformable matrices, $(AB)' = B'A'$. In fact, for the product of three conformable matrices this becomes $(ABC)' = C'B'A'$. The transpose of the inverse is the inverse of the transpose, i.e., $(A^{-1})' = (A')^{-1}$.

The inverse of a partitioned matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is

$$A^{-1} = \begin{bmatrix} E & -EA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}E & A_{22}^{-1} + A_{22}^{-1}A_{21}EA_{12}A_{22}^{-1} \end{bmatrix}$$

where $E = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$. Alternatively, it can be expressed as

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}FA_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}F \\ -FA_{21}A_{11}^{-1} & F \end{bmatrix}$$

where $F = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$. These formulas are used in partitioned regression models, see for example the Frisch-Waugh Lovell Theorem and the computation of the variance-covariance matrix of forecasts from a multiple regression in Chapter 7.

An $n \times n$ symmetric matrix Ω has n distinct characteristic vectors $c_1, \ldots, c_n$. The corresponding n characteristic roots $\lambda_1, \ldots, \lambda_n$ may not be distinct but they are all real numbers. The number of non-zero characteristic roots of Ω is equal to the rank of Ω. The characteristic roots of a positive definite matrix are positive. The characteristic vectors of the symmetric matrix Ω are *orthogonal* to each other, i.e., $c_i'c_j = 0$ for $i \neq j$ and can be made *orthonormal* with $c_i'c_i = 1$ for $i = 1, 2, \ldots, n$. Hence, the matrix of characteristic vectors $C = [c_1, c_2, \ldots, c_n]$ is an *orthogonal* matrix, such that $CC' = C'C = I_n$ with $C' = C^{-1}$. By definition $\Omega c_i = \lambda_i c_i$ or $\Omega C = C\Lambda$ where $\Lambda = \text{diag}[\lambda_i]$. Premultiplying the last equation by C' we get $C'\Omega C = C'C\Lambda = \Lambda$. Therefore, the matrix of characteristic vectors C diagonalizes the symmetric matrix Ω. Alternatively, we can write $\Omega = C\Lambda C' = \sum_{i=1}^n \lambda_i c_i c_i'$ which is the *spectral decomposition* of Ω.

A real symmetric $n \times n$ matrix Ω is *positive semi-definite* if for every $n \times 1$ non-negative vector y, we have $y'\Omega y \geq 0$. If $y'\Omega y$ is strictly positive for any non-zero y then Ω is said to be *positive definite.* A necessary and sufficient condition for Ω to be positive definite is that all the characteristic roots of Ω are positive. One important application is the comparison of efficiency of two unbiased estimators of a vector of parameters β. In this case, we subtract the variance-covariance matrix of the inefficient estimator from the more efficient one and show that the resulting difference yields a positive semi-definite matrix, see the Gauss-Markov Theorem in Chapter 7.

If Ω is a symmetric and positive definite matrix, there exists a nonsingular matrix P such that $\Omega = PP'$. In fact, using the spectral decomposition of Ω given above, one choice for $P = C\Lambda^{1/2}$ so that $\Omega = C\Lambda C' = PP'$. This is a useful result which we use in Chapter 9 to obtain Generalized Least Squares (GLS) as a least squares regression after transforming the original regression model by $P^{-1} = (C\Lambda^{1/2})^{-1} = \Lambda^{-1/2}C'$. In fact, if $u \sim (0, \sigma^2\Omega)$, then $P^{-1}u$ has zero mean and $\text{var}(P^{-1}u) = P^{-1\prime}\text{var}(u)P^{-1\prime} = \sigma^2 P^{-1}\Omega P^{-1\prime} = \sigma^2 P^{-1}PP'P^{-1\prime} = \sigma^2 I_n$.

From Chapter 2, we have seen that if $u \sim N(0, \sigma^2 I_n)$, then $u_i/\sigma \sim N(0,1)$, so that $u_i^2/\sigma^2 \sim \chi_1^2$ and $u'u/\sigma^2 = \sum_{i=1}^n u_i^2/\sigma^2 \sim \chi_n^2$. Therefore, $u'(\sigma^2 I_n)^{-1}u \sim \chi_n^2$. If $u \sim N(0, \sigma^2\Omega)$ where Ω is positive definite, then $u^* = P^{-1}u \sim N(0, \sigma^2 I_n)$ and $u^{*\prime}u^*/\sigma^2 \sim \chi_n^2$. But $u^{*\prime}u^* = u'P^{-1\prime}P^{-1}u = u'\Omega^{-1}u$. Hence, $u'\Omega^{-1}u/\sigma^2 \sim \chi_n^2$. This is used in Chapter 9.

Note that the OLS residuals are denoted by $e = \bar{P}_X u$. If $u \sim N(0, \sigma^2 I_n)$, then e has mean zero and $\text{var}(e) = \sigma^2 \bar{P}_X I_n \bar{P}_X = \sigma^2 \bar{P}_X$ so that $e \sim N(0, \sigma^2 \bar{P}_X)$. Our estimator of σ^2 in Chapter 7 is $s^2 = e'e/(n-k)$ so that $(n-k)s^2/\sigma^2 = e'e/\sigma^2$. The last term can also be written as $u'\bar{P}_X u/\sigma^2$. In order to find the distribution of this quadratic form in Normal variables, we use the following result stated as lemma 1 in Chapter 7.

Lemma 1: For every symmetric idempotent matrix A of rank r, there exists an orthogonal matrix P such that $P'AP = J_r$ where J_r is a diagonal matrix with the first r elements equal to one and the rest equal to zero.

We use this lemma to show that the $e'e/\sigma^2$ is a chi-squared with $(n-k)$ degrees of freedom. To see this note that $e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2$ and that $\bar{P}_X$ is symmetric and idempotent of rank $(n-k)$. Using the lemma there exists a matrix P such that $P'\bar{P}_X P = J_{n-k}$ is a diagonal matrix with the first $(n-k)$ elements on the diagonal equal to 1 and the last k elements equal to zero. An orthogonal matrix P is by definition a matrix whose inverse, is its own transpose, i.e., $P'P = I_n$. Let $v = P'u$ then v has mean zero and $\text{var}(v) = \sigma^2 P'P = \sigma^2 I_n$ so that v is $N(0, \sigma^2 I_n)$ and $u = Pv$. Therefore,

$$e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2 = v'P'\bar{P}_X Pv/\sigma^2 = v'J_{n-k}v/\sigma^2 = \textstyle\sum_{i=1}^{n-k} v_i^2/\sigma^2$$

But, the v's are independent identically distributed $N(0, \sigma^2)$, hence v_i^2/σ^2 is the square of a standardized $N(0,1)$ random variable which is distributed as a χ_1^2. Moreover, the sum of independent χ^2 random variables is a χ^2 random variable with degrees of freedom equal to the sum of the respective degrees of freedom, see Chapter 2. Hence, $e'e/\sigma^2$ is distributed as χ_{n-k}^2.

The beauty of the above result is that it applies to all quadratic forms $u'Au$ where A is symmetric and idempotent. In general, for $u \sim N(0, \sigma^2 I)$, a necessary and sufficient condition for $u'Au/\sigma^2$ to be distributed χ_k^2 is that A is idempotent of rank k, see Theorem 4.6 of Graybill (1961). Another useful theorem on quadratic forms in normal random variables is the following: If $u \sim N(0, \sigma^2\Omega)$, then $u'Au/\sigma^2$ is χ_k^2 if and only if $A\Omega$ is an idempotent matrix of rank k, see Theorem 4.8 of Graybill (1961). If $u \sim N(0, \sigma^2 I)$, the two positive semi-definite quadratic forms in normal random variables say $u'Au$ and $u'Bu$ are independent if and only if $AB = 0$, see Theorem 4.10 of Graybill (1961). A sufficient condition is that $\text{tr}(AB) = 0$, see Theorem 4.15 of Graybill (1961). This is used in Chapter 7 to construct F-statistics to test hypotheses, see for example problem 11. For $u \sim N(0, \sigma^2 I)$, the quadratic form $u'Au$ is independent of the linear form Bu if $BA = 0$, see Theorem 4.17 of Graybill (1961). This is used in Chapter 7 to prove the independence of s^2 and $\widehat{\beta}_{ols}$, see problem 8. In general, if $u \sim N(0, \Sigma)$, then $u'Au$ and $u'Bu$ are independent if and only if $A\Sigma B = 0$, see Theorem 4.21 of Graybill (1961). Many other useful matrix properties can be found. This is only a sample of them that will be implicitly or explicitly used in this book.

The *Kronecker* product of two matrices say $\Sigma \otimes I_n$ where Σ is $m \times m$ and I_n is the identity matrix of dimension n is defined as follows:

$$\Sigma \otimes I_n = \begin{bmatrix} \sigma_{11}I_n & \dots & \sigma_{1m}I_n \\ \vdots & & \vdots \\ \sigma_{m1}I_n & \dots & \sigma_{mm}I_n \end{bmatrix}$$

In other words, we place an I_n next to every element of $\Sigma = [\sigma_{ij}]$. The dimension of the resulting matrix is $mn \times mn$. This is useful when we have a system of equations like Seemingly Unrelated Regressions in Chapter 10. In general, if A is $m \times n$ and B is $p \times q$ then $A \otimes B$ is $mp \times nq$. Some properties of Kronecker products include $(A \otimes B)' = A' \otimes B'$. If both A and B are square matrices of order $m \times m$ and $p \times p$ then $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, $|A \otimes B| = |A|^m |B|^p$ and $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$. Applying this result to $\Sigma \otimes I_n$ we get

$$(\Sigma \otimes I_n)^{-1} = \Sigma^{-1} \otimes I_n \quad \text{and} \quad |\Sigma \otimes I_n| = |\Sigma|^m |I_n|^n = |\Sigma|^m$$

and $\mathrm{tr}(\Sigma \otimes I_n) = \mathrm{tr}(\Sigma)\mathrm{tr}(I_n) = n\ \mathrm{tr}(\Sigma)$.

Some useful properties of matrix differentiation are the following:

$$\frac{\partial x'b}{\partial b} = x \quad \text{where } x' \text{ is } 1 \times k \text{ and } b \text{ is } k \times 1.$$

Also

$$\frac{\partial b'Ab}{\partial b} = (A + A') \quad \text{where } A \text{ is } k \times k.$$

If A is symmetric, then $\partial b'Ab/\partial b = 2Ab$. These two properties will be used in Chapter 7 in deriving the least squares estimator.

References

Bellman, R. (1970), *Introduction to Matrix Analysis* (McGraw Hill: New York).

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics* (John Wiley and Sons: New York).