Elementary Regression Theory

Richard E. Quandt
Princeton University

Theorem 1. *If the $k \times 1$ vector Y is distributed as $N(\mu, V)$, and if B is an $m \times k$ matrix ($m \leqq k$) of rank m, then $Z = BY$ is distributed as $N(B\mu, BVB')$.*

Proof. The moment generating function for the multivariate normal distribution of y is

$$E(e^{\theta' Y}) = e^{\theta'\mu + \theta' V\theta/2}.$$

Hence, the moment generating function for Z is $E(e^{\theta' BY}) = e^{\theta' B\mu + \theta' BVB'\theta/2}$, which is the moment generating function for a multivariate normal distribution with mean vector $B\mu$ and covariance matrix BVB'. ∎

Theorem 2. *If B is a $q \times n$ matrix, A an $n \times n$ matrix, and if $BA = 0$, and Y is distributed as $N(\mu, \sigma^2 I)$, then the linear form BY and the quadratic form $Y'AY$ are independently distributed.*

Proof. Without loss of generality, A can be taken to be symmetric. Then there exists an orthogonal matrix P such that

$$P'AP = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix},$$

where D_1 and D_2 are diagonal, and writing the matrix in partitioned forms is needed in the proof below. Let $P'Y = Z$. Then Z is distributed as $N(P'\mu, \sigma^2 I)$, since $E(P'Y) = P'\mu$ and $E[(Z - P'\mu)(Z - P'\mu)'] = P'E[(Y - \mu)(Y - \mu)']P = \sigma^2 I$ by the orthogonality of P.

Now let

$$BP = C. \tag{1}$$

By the hypothesis of the theorem,

$$0 = BA = BAP = BPP'AP = CD \tag{2}$$

and partitioning C and D conformably, CD can be written as

$$CD = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix},$$

where either D_1 or D_2 must be zero. If neither D_1 nor D_2 is zero, Eq.(2) would imply that $C_{11} = C_{12} = C_{21} = C_{22} = 0$; hence $C = 0$, and from Eq.(1), $B = 0$, which is a trivial case. So assume that D_2 is zero. But if $D_2 = 0$, Eq.(2) implies only $C_{11} = C_{21} = 0$. Then

$$BY = BPP'Y = CZ = [\,0 \quad C_2\,]\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = C_2 Z_2$$

where C_2 denotes $[\,C'_{12} \quad C'_{22}\,]'$, and

$$Y'AY = Y'PP'APP'Y = Z'DZ = Z'_1 D_1 Z_1.$$

Since the elements of Z are independent and BY and $Y'AY$ share no element of Z in common, they are independent. ∎

Corollary 1. *It follows immediately that if the elements of the vector $x' = (x_1, \ldots, x_n)$ are independent drawings from a normal distribution, the sample mean, $\overline{x}$ and (n times) the sample variance, $\sum_i (x_i - \overline{x})^2$, are independently distributed.*

This follows because, defining i' as the vector of n 1's, $(1, \ldots, 1)$, we can write

$$\overline{x} = \frac{i'}{n} x$$

and

$$\sum_i (x_i - \overline{x})^2 = x'\big(I - \frac{ii'}{n}\big)\big(I - \frac{ii'}{n}\big)x = x'\big(I - \frac{ii'}{n}\big)x$$

and we can verify that the matrices $\big(I - \frac{ii'}{n}\big)$ and $\big(\frac{i'}{n}\big)$ have a zero product.

Theorem 3. *If A and B are both symmetric and of the same order, it is necessary and sufficient for the existence of an orthogonal transformation $P'AP$ =diag, $P'BP$ =diag, that $AB = BA$.*

Proof. (1) Necessity. Let $P'AP = D_1$ and $P'BP = D_2$, where D_1 and D_2 are diagonal matrices. Then $P'APP'BP = D_1 D_2 = D$ and $P'BPP'AP = D_2 D_1 = D$. But then it follows that $AB = BA$.

(2) Sufficiency. Assume that $AB = BA$ and let $\lambda_1, \ldots, \lambda_r$ denote the distinct eigenvalues of A, with multiplicities $m_1, \ldots, m_r$. There exists an orthogonal Q_1 such that

$$Q'_1 A Q_1 = D_1 = \begin{bmatrix} \lambda_1 I_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 I_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_r I_r \end{bmatrix},$$

where I_j denotes an identity matrix of order m_j. Then D_1 commutes with $Q'_1 B Q_1$, since

$$\begin{aligned} D_1 Q'_1 B Q_1 = Q'_1 A Q_1 Q'_1 B Q_1 &= Q'_1 A B Q_1 = Q'_1 B A Q_1 \\ &= Q'_1 B Q_1 Q'_1 A Q_1 = Q'_1 B Q_1 D_1. \end{aligned}$$

It follows that the matrix $Q_1'BQ_1$ is a block-diagonal matrix with symmetric submatrices $(Q_1'BQ_1)_i$ of dimension m_i $(i = 1, \ldots, r)$ along the diagonal. Then we can find orthogonal matrices P_i such that $P_i'(Q_1'BQ_1)_iP_i$ are diagonal. Now form the matrix

$$Q_2 = \begin{bmatrix} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & P_r \end{bmatrix}.$$

Q_2 is obviously orthogonal. Then the matrix Q defined as Q_1Q_2 is orthogonal and diagonalizes both A and B, since

$$Q'AQ = Q_2'D_1Q_2 = Q_2'Q_2D_1 = D_1$$

(because the blockdiagonality of Q_2 causes it to commute with D_1), and

$$Q'BQ = Q_2'Q_1'BQ_1Q_2$$

is a diagonal matrix by construction. ∎

Theorem 4. *If Y is distributed as $N(\mu, I)$, the positive semidefinite forms $Y'AY$ and $Y'BY$ are independently distributed if and only if $AB = 0$.*

Proof. (1) Sufficiency. Let $AB = 0$. Then $B'A' = BA = 0$ and $AB = BA$. Then, by Theorem 3, there exists an orthogonal P such that $P'AP = D_1$ and $P'BP = D_2$. It follows that $D_1D_2 = 0$, since

$$D_1D_2 = P'APP'BP = P'ABP$$

and $AB = 0$ by hypothesis. Then D_1 and D_2 must be of the forms

$$D_1 = \begin{bmatrix} \Delta_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3}$$

and

$$D_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Delta_2 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where Δ_1 and Δ_2 are diagonal submatrices and where D_1 and D_2 are partitioned conformably.

Now let $Z = P'Y$. Then Z is distributed as $N(P'\mu, I)$, and

$$Y'AY = Z'P'APZ = Z_1'\Delta_1Z_1$$

and

$$Y'BY = Z'P'BPZ = Z_2'\Delta_2Z_2.$$

Since the elements of Z are independent and the two quadratic forms share no element of Z in common, they are independent.

(2) Necessity. If $Y'AY$ and $Y'BY$ are independent, and $Y = PZ$, where P is orthogonal and $P'AP = D_1$, then $Z'D_1Z$ and $Z'P'BPZ$ cannot have elements of Z in common. Thus, if D_1 has the form of Eq.(3), then $P'BP$ must be of the form

$$D_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Delta_{22} & \Delta_{23} \\ 0 & \Delta_{32} & \Delta_{33} \end{bmatrix}.$$

Then $P'APP'BP = 0$, from which it follows that $AB = 0$. ∎

Definition 1. *A square matrix A is said to be idempotent if $A^2 = A$.*

In what follows, we consider only symmetric idempotent matrices.

Theorem 5. *The eigenvalues of a symmetric matrix A are all either 0 or 1 if and only if A is idempotent.*

Proof. (1) Sufficiency. If λ is an eigenvalue of the matrix A, and x is the corresponding eigenvector, then

$$Ax = \lambda x. \tag{4}$$

Multiplying (4) by A yields

$$A^2x = \lambda AX = \lambda^2 x,$$

but since A is idempotent, we also have

$$A^2x = Ax = \lambda x.$$

From the last two equations it follows that

$$(\lambda^2 - \lambda)x = 0,$$

but since x is an eigenvector and hence not equal to the zero vector, it follows that λ is equal to either 0 or 1. It also follows that if A has rank r, there exists an orthogonal matrix P such that

$$P'AP = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \tag{5}$$

where I_r is an identity matrix of order r.

(2) Necessity. Let the eigenvalues of A all be 0 or 1. Then there exists orthogonal P such that $P'AP = E_r$, where E_r is the matrix on the right hand side of Eq.(5). Then $A = PE_rP'$, and $A^2 = PE_rP'PE_rP' = PE_rP' = A$, and A is idempotent.

Theorem 6. *If n-vector Y is distributed as $N(0, I)$, then $Y'AY$ is distributed as $\chi^2(k)$ if and only if A is idempotent of rank k.*

Proof. (1) Sufficiency. Let P be the orthogonal matrix that diagonalizes A, and define $Z = P'Y$. Then

$$Y'AY = Z'P'APZ = \sum_{i=1}^{k} z_i^2$$

by Theorem 5. But the right hand side is the sum of squares of k normally and independently distributed variables with mean zero and variance 1; hence it is distributed as $\chi^2(k)$.

(2) Necessity. Assume that $Y'AY$ is distributed as $\chi^2(k)$. Since A is symmetric, there exists orthogonal P such that $P'AP = D$, where D is diagonal. Let $Z = P'Y$; then

$$Y'AY = Z'P'APZ = Z'DZ = \sum_{i=1}^{n} d_i z_i^2,$$

and we define the right hand side as ϕ. Since the z_i are $N(0, 1)$ and independent, the moment generating function of ϕ is

$$\prod_{i=1}^{n} (1 - 2\theta d_i)^{-1/2}.$$

The moment generating function for $Y'AY$ (since it is distributed as $\chi^2(k)$) is $(1 - 2\theta)^{-k/2}$. These two moment generating functions must obviously equal one another, which is possible only if k of the d_i are equal to 1 and the rest are equal to 0; but this implies that A is idempotent of rank k. ∎

In the next theorem we introduce matrices A_r and A_u; the subscripts that identify the matrices refer to the context of hypothesis testing in the regression model and indicate the model restricted by the hypothesis or the unrestricted model.

Theorem 7. *Let u be an n-vector distributed as $N(0, \sigma^2 I)$, and let A_r and A_u be two idempotent matrices, with $A_r \neq A_u$ and $A_r A_u = A_u$. Then, letting $u_r = A_r u$ and $u_u = A_u u$, the quantity $F = \frac{(u_r'u_r - u_u'u_u)/(\mathrm{tr}(A_r) - \mathrm{tr}(A_u))}{u_u'u_u/\mathrm{tr}(A_u)}$ has the F distribution with $\mathrm{tr}(A_r) - \mathrm{tr}(A_u)$ and $\mathrm{tr}(A_u)$ degrees of freedom.*

Proof. Dividing both numerator and denominator by σ^2, we find from Theorem 6 that the denominator has $\chi^2(\mathrm{tr}(A_u))$ distribution. From the numerator we have $u_r'u_r - u_u'u_u = u'A_r'A_r u - u'A_u'A_u u = u'(A_r - A_u)u$. But $(A_r - A_u)$ is idempotent, because

$$(A_r - A_u)^2 = A_r^2 - A_r A_u - A_u A_r + A_u^2 = A_r - A_u - A_u + A_u = A_r - A_u.$$

Hence, the numerator divided by σ^2 has $\chi^2(\mathrm{tr}(A_r) - \mathrm{tr}(A_u))$ distribution. But the numerator and the denominator are independent, because the matrices of the respective quadratic forms, $A_r - A_u$ and A_u, have a zero product. ∎

Theorem 8. *If Y is distributed as $N(\mu, \sigma^2 I)$, then $Y'AY/\sigma^2$ is distributed as noncentral $\chi^2(k, \lambda)$, where $\lambda = \mu' A\mu/2\sigma^2$ if and only if A is idempotent of rank k.*

The proof is omitted.

Theorem 9. *The trace of a product of square matrices is invariant under cyclic permutations of the matrices.*

Proof. Consider the product BC, where B and C are two matrices of order n. Then $\mathrm{tr}(BC) = \sum_i \sum_j b_{ij} c_{ji}$ and $\mathrm{tr}(CB) = \sum_i \sum_j c_{ij} b_{ji}$. But these two double sums are obviously equal to one another. ∎

Theorem 10. *All symmetric, idempotent matrices not of full rank are positive semidefinite.*

Proof. For symmetric and idempotent matrices, we have $A = AA = A'A$. Pre- and postmultiplying by x' and x respectively, $x'A'Ax{=}y'y$, which is $\geqq 0$. ∎

Theorem 11. *If A is idempotent of rank r, then $\mathrm{tr}(A) = r$.*

Proof. It follows from Eq.(4) in Theorem 5 that $P'AP = E_r$. Hence, from Theorem 9 it follows that $\mathrm{tr}(A) = \mathrm{tr}(P'AP) = \mathrm{tr}(APP') = \mathrm{tr}(E_r) = r$. ∎

Theorem 12. *If Y is distributed with mean vector $\mu = 0$ and covariance matrix $\sigma^2 I$, then $E(Y'AY) = \sigma^2 \mathrm{tr}(A)$.*

Proof.

$$E(Y'AY) = E\{\sum_i \sum_j a_{ij} y_i y_j\} = E\{\sum_i a_{ii} y_i^2\} + E\{\sum_{\substack{i \\ i \neq j}} \sum_j a_{ij} y_i y_j\}$$

. But when $i \neq j$, $E(y_i y_j) = E(y_i)E(y_j) = 0$; hence $E(Y'AY) = \sigma^2 \mathrm{tr}(A)$. ∎

We now specify the regression model as

$$Y = X\beta + u \tag{6}$$

where Y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times k$ matrix of observations on the independent variables, and u is an $n \times 1$ vector of unobservable error terms. We make the following assumptions:

Assumption 1. *The elements of X are nonstochastic (and, hence, may be taken to be identical in repated samples).*

Assumption 2. *The elements of the vector u are independently distributed as $N(0, \sigma^2)$.*

It follows from Assumption 2 that the joint density of the elements of u is

$$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-u'u/2\sigma^2}.$$

The loglikelihood is

$$\log L = -\left(\frac{n}{2}\right)\log(2\pi) - \left(\frac{n}{2}\right)\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta),$$

and its partial derivatives are

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\big[X'(Y - X\beta)\big]$$

$$\frac{\partial \log L}{\sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta).$$

Setting these equal to zero yields the maximum likelihood estimates, which are

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{7}$$

and

$$\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/n. \tag{8}$$

Remark. It is obvious from the form of the likelihood function that the estimate $\hat{\beta}$ is also the least squares estimate of β, i.e., the estimate that minimizes $(Y - X\beta)'(Y - X\beta)$, without any assumptions about the distribution of the error term vector u.

Theorem 13. *$\hat{\beta}$ is unbiased.*

Proof. $E\big[(X'X)^{-1}X'Y\big] = (X'X)^{-1}X'E(X\beta + u) = \beta + E(u) = \beta.$ ∎

Theorem 14. $E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma^2$.

Proof. Substituting into Eq.(6) for $\hat{\beta}$ yields

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n}E\big[(Y - X(X'X)^{-1}X'Y)'(Y - X(X'X)^{-1}X'Y)\big] \\ &= \frac{1}{n}E\big[Y'(I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X')Y\big] \\ &= \frac{1}{n}E\big[Y'(I - X(X'X)^{-1}X')Y\big], \end{aligned} \tag{9}$$

since $I - X(X'X)^{-1}X'$ is idempotent, as may be easily verified by multiplying it by itself. Substituting $X\beta + u$ for Y in Eq.(9) yields

$$E(\hat{\sigma}^2) = \frac{1}{n}E\big\{\big[u'(I - X(X'X)^{-1}X')u\big]\big\},$$

and applying Theorem 12,

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n}\sigma^2\mathrm{tr}(I - X(X'X)^{-1}X') = \frac{1}{n}\sigma^2\big[\mathrm{tr}(I) - \mathrm{tr}\big(X(X'X)^{-1}X'\big)\big] \\ &= \frac{1}{n}\sigma^2\big[n - \mathrm{tr}\big((X'X)(X'X)^{-1}\big)\big] = \frac{n-k}{n}\sigma^2. \end{aligned}$$

∎

It follows from Theorem 14 that an unbiased estimator of σ^2 can be defined as $\hat{\sigma}^2_{ub} = (n/(n-k))\hat{\sigma}^2$.

Theorem 15. *$\hat\beta$ has the k-variate normal distribution with mean vector β and covariance matrix $\sigma^2(X'X)^{-1}$.*

Proof. Normality of the distribution follows from Theorem 1. The fact that the mean of this distribution is β follows from Theorem 13. The covariance matrix of $\hat\beta$ is

$$\begin{aligned} E\big[(\hat\beta-\beta)(\hat\beta-\beta)'\big] &= E(\hat\beta\hat\beta') - \beta\beta' = E\{(X'X)^{-1}X'(X\beta+u)(X\beta+u)'X(X'X)^{-1}\} - \beta\beta' \\ &= E\big[(X'X)^{-1}X'\sigma^2 IX(X'X)^{-1}\big] = \sigma^2(X'X)^{-1}. \end{aligned}$$

∎

Theorem 16. *$(n-k)\hat\sigma^2_{ub}/\sigma^2$ is distributed as $\chi^2(n-k)$.*

Proof. $(n-k)\hat\sigma^2_{ub}$ can be written as $\frac{u'}{\sigma}\big[I-X(X'X)^{-1}X\big]\frac{u}{\sigma}$. Since $I-X(X'X)^{-1}X'$ is idempotent of rank $n-k$, Theorem 6 applies. ∎

Theorem 17. *$\hat\beta$ and $\hat\sigma^2_{ub}$ are independently distributed.*

Proof. Multiplying together the matrix of the linear form, $(X'X)^{-1}X'$, and of the quadratic form, $I-X(X'X)^{-1}X'$, we obtain

$$(X'X)^{-1}X'\big(I-X(X'X)^{-1}X'\big) = 0,$$

which proves the theorem. ∎

Theorem 18. *(Markov) Given $Y = X\beta+u$, $E(u)=0$, $E(uu')=\sigma^2 I$, and $E(u'X)=0$, the best, linear, unbiased estimate of β is given by $\hat\beta = (X'X)^{-1}X'Y$.*

Proof. First note that Assumption 2 is not used in this theorem. Let C be any constant matrix of order $k\times n$ and define the linear estimator

$$\tilde\beta = CY$$

. Without loss of generality, let $C = (X'X)^{-1}X' + B$, so that

$$\tilde\beta = \big[(X'X)^{-1}X' + B\big]Y. \tag{10}$$

Then unbiasedness requires that

$$E(\tilde\beta) = E\big\{\big[(X'X)^{-1}X'+B\big]Y\big\} = \beta + BX\beta = \beta.$$

For the last equality to hold, we require that $BX=0$.

The covariance matrix of $\tilde\beta$ is

$$\begin{aligned} E(\tilde\beta-\beta)(\tilde\beta-\beta)' &= E(\tilde\beta\tilde\beta') - \beta\beta' \\ &= E\big\{\big[(X'X)^{-1}X'+B\big]\big[X\beta+u\big]\big[\beta'X'+u'\big]\big[X(X'X)^{-1}+B'\big]\big\} - \beta\beta', \end{aligned}$$

where we have substituted for $\tilde{\beta}$ from Eq.(10) and for Y we have substituted $X\beta + u$. Noting that any expression in the product with BX in it is zero due to the requirement of unbiasedness, and any expression with a single u becomes zero when we take expectations, the covariance matrix simplifies to $\sigma^2\big[(X'X)^{-1} + BB'\big]$. But BB' is a positive semidefinite matrix; hence the covariance matrix of $\tilde{\beta}$ exceeds that of $\hat{\beta}$ by a positive semidefinite matrix. Hence the least squares estimator is "best."

Theorem 19. *Denoting the* i^{th} *diagonal element of* $(X'X)^{-1}$ *by* $(X'X)_i^{-1}$*, the quantity* $\big[(\hat{\beta}_i - \beta_i)/\big((X'X)_i^{-1}\sigma^2\big)^{\frac{1}{2}}\big]/\big[\sigma^2_{ub}/\sigma^2\big]^{\frac{1}{2}}$ *has the* t *distribution with* $n-k$ *degrees of freedom.*

Proof. $(\hat{\beta}_i - \beta_i)/\big[(X'X)_i^{-1}\sigma^2\big]^{\frac{1}{2}}$ is normally distributed as $N(0,1)$ by Theorem 15. $\sigma^2_{ub}(n-k)/\sigma^2$ is distributed as $\chi^2(n-k)$ by Theorem 16. The two are independently distributed by Theorem 17. But the ratio of an $N(0,1)$ variable to the squareroot of an independent χ^2 variate, which has been first divided by its degrees of freedom, has $t(n-k)$ distribution. ∎

We next consider a test on a subset of the regression parameters. For this purpose we partition the regression model as

$$Y = X_1\beta_1 + X_2\beta_2 + u, \tag{11}$$

where X_1 is $n \times (k-q)$ and X_2 is $n \times q$, and where β_1 and β_2 are $(k-q)$- and q-vectors respectively. We shall test a hypothesis about the vector β_2, leaving β_1 unrestricted by the hypothesis.

Define $S(\hat{\beta}_1, \hat{\beta}_2)$ as the sum of the squares of the regression deviations when the least-squares estimates for both β_1 and β_2 are used, and denote the sum of the squares of deviations when β_2 is assigned a fixed value, and a least-squares estimate is obtained for β_1 as a function of the chosen β_2 by $S(\hat{\beta}_1(\beta_2), \beta_2)$.

Theorem 20. *The quantity* $\big[\big(S(\hat{\beta}_1(\beta_2), \beta_2) - S(\hat{\beta}_1, \hat{\beta}_2)\big)/q\big]/\big[S(\hat{\beta}_1, \hat{\beta}_2)/(n-k)\big]$ *is distributed as* $F(q, n-k)$.

Proof. The least squares estimate for β_1, as a function of β_2, is

$$\hat{\beta}_1(\beta_2) = (X_1'X_1)^{-1}X_1'(Y - X_2\beta_2)$$

and the restricted sum of squares is

$$\begin{aligned} S(\hat{\beta}_1(\beta_2), \beta_2) &= \big[Y - X_2\beta_2 - X_1(X_1'X_1)^{-1}X_1'(Y - X_2\beta_2)\big]'\big[Y - X_2\beta_2 - X_1(X_1'X_1)^{-1}X_1'(Y - X_2\beta_2)\big] \\ &= (Y' - \beta_2'X_2')A_r(Y - X_2\beta_2) = (\beta_1'X_1' + u')A_r(X_1\beta_1 + u) \\ &= u'A_r u, \end{aligned}$$

where $A_r = I - X_1(X_1'X_1)^{-1}X_1'$, and where we have used Eq.(11) to replace Y. Since $S(\hat{\beta}_1, \hat{\beta}_2)$ is $u'A_u u$, where $A_u = I - X(X'X)^{-1}X'$, the ratio in the theorem can be written as

$$F = \frac{u'(A_r - A_u)u/q}{u'A_u u/(n-k)}. \tag{12}$$

$A_r - A_u$ is idempotent, for

$$(A_r - A_u)^2 = A_r - A_u - A_r A_u - A_u A_r,$$

but

$$\begin{aligned} A_u A_r &= \left[I - X(X'X)^{-1}X'\right]\left[I - X_1(X_1'X_1)^{-1}X_1'\right] \\ &= I - X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1' + X(X'X)^{-1}X'X_1(X_1'X_1)^{-1}X_1', \end{aligned}$$

where the last matrix on the right can be written as

$$X\begin{pmatrix} I \\ 0 \end{pmatrix}(X_1'X_1)^{-1}X_1' = X_1(X_1'X_1)^{-1}X_1'.$$

Hence, $A_u A_r = A_u$ and $A_r - A_u$ is idempotent. It further follows immediately that $A_u(A_r - A_u) = 0$; hence the numerator and denominator are independently distributed. But the ratio of two independent χ^2 variates, when divided by their respective degrees of freedom, has the indicated F-distribution. ∎

Corollary 2. *If β_1 is the scalar parameter representing the constant term, and β_2 the vector of the remaining $k-1$ parameters, the F-statistic for testing the null hypothesis $H_0 : \beta_2 = 0$ is $\left[R^2/(k-1)\right]/\left[(1-R^2)/(n-k)\right]$, where R^2 is the coefficient of determination.*

Proof. Define $y = Y - \overline{Y}$, $x = X_2 - \overline{X}_2$, where $\overline{Y}$ is the vector containing for each element the sample mean of the elements of Y and $\overline{X}_2$ is the matrix the j^{th} column of which contains for each element the sample mean of the corresponding column of X_2. Then write the regression model in deviations from sample mean form as

$$y = x\beta_2 + v$$

where v represents the deviations of the error terms from their sample mean. The numerator of the test statistic in Theorem 20 can then be written as

$$\left[S(\beta_2) - S(\hat{\beta}_2)\right]/(k-1). \tag{13}$$

We have

$$S(\beta_2) = y'y + \beta_2'x'x\beta_2 - 2\beta_2'x'y \tag{14}$$

and

$$S(\hat{\beta}_2) = (y - x\hat{\beta}_2)'(y - x\hat{\beta}_2) = y'y - \hat{\beta}_2'x'y' - (y'x - \hat{\beta}_2'x'x)\hat{\beta}_2 \tag{15}$$

Note that the parenthesized expression in the last equation is zero by the definition of $\hat{\beta}_2$ as the least squares estimate. Then

$$S(\beta_2) - S(\hat{\beta}_2) = \beta_2'x'x\beta_2 - 2\beta_2'x'y + \hat{\beta}_2'x'y$$

$$
\begin{aligned}
&= \beta_2' x' x \beta_2 - 2\beta_2' x' y + \hat{\beta}_2' x' y + \hat{\beta}_2' x' x \hat{\beta}_2 - \hat{\beta}_2' x' x \hat{\beta}_2 \\
&\qquad \text{(by adding and subtracting } \hat{\beta}_2' x' x \hat{\beta}_2\text{)} \\
&= \beta_2' x' x \beta_2 - 2\beta_2' x' y + \hat{\beta}_2' x' x \hat{\beta}_2 \\
&\qquad \text{(by noting that the third and fifth terms cancel)} \\
&= \beta_2' x' x \beta_2 + \hat{\beta}_2' x' x \hat{\beta}_2 - 2\beta_2' x' x \hat{\beta}_2 \\
&\qquad \text{(by replacing } x'y \text{ by } x' x \hat{\beta}_2\text{)} \\
&= (\beta_2 - \hat{\beta}_2)' x' x (\beta_2 - \hat{\beta}_2). \qquad (16)
\end{aligned}
$$

We also define

$$R^2 = 1 - \frac{S(\hat{\beta}_2)}{y'y}. \qquad (17)$$

Under H_0, $S(\beta_2) - S(\hat{\beta}_2) = \hat{\beta}_2 x' y = y' y R^2$ from the first line of Eq.(16) by the definition of R^2 in Eq.(17). Combining the definition of $S(\hat{\beta}_2)$ with that of R^2 yields for the denominator $y'y - R^2 y'y$. Substituting these expressions in the definition of F and cancelling out $y'y$ from the numerator and the denominator yields the result. ∎

Theorem 21. *Let X_r be a $n \times k$ and X an $n \times p$ matrix, and assume that there exists a matrix C of order $p \times k$ such that $X_r = XC$. Then the matrices $A_r = I - X_r(X_r' X_r)^{-1} X_r'$ and $A_u = I - X(X'X)^{-1}X'$ satisfy the conditions of Theorem 7.*

Proof. A_r and A_u are obviously idempotent. To show that $A_r A_u = A_u$, multiply them to obtain

$$
\begin{aligned}
A_r A_u &= I - X_r(X_r' X_r)^{-1} X_r' - X(X'X)^{-1}X' + X_r(X_r' X_r)^{-1} X_r' X (X'X)^{-1} X' \\
&= I - X_r(X_r' X_r)^{-1} X_r' - X(X'X)^{-1}X' + X_r(X_r' X_r)^{-1} C' X' \\
&= I - X_r(X_r' X_r)^{-1} X_r' = A_u,
\end{aligned}
$$

where we have replaced X_r' in the first line by $C'X'$, cancelled the term $(X'X)(X'X)^{-1}$, and finally replaced $C'X'$ by X_r'.∎

The precise form of the test statistics in performing tests on subsets of regression coefficients depends on whether there are enough observations ("degrees of freedom") to obtain least squares regression coefficients under the alternative hypothesis. We first consider the case of sufficient degrees of freedom.

Sufficient Degrees of Freedom.

Case 1: Test on a Subset of Coefficients in a Regression. Write the model as

$$Y = X_1 \beta_1 + X_2 \beta_2 + u = X\beta + u, \qquad (18)$$

where $X_1$is $n \times k_1$ and X_2 is $n \times k_2$, and where we wish to test $H_0 : \beta_2 = 0$. Under H_0, the model is

$$Y = X_1\beta_1 + u, \tag{19}$$

and we can write

$$X_1 = X \begin{pmatrix} I \\ 0 \end{pmatrix},$$

where the matrix on the right in parentheses is of order $(k_1 + k_2) \times k_1$. Hence the conditions of Theorem 7 are satisfied. Denote the residuals from Eq.(18) and (19) respectively by $\hat{u}_u$ and $\hat{u}_r$. Then the F-test of Theorem 20 can also be written as

$$\frac{(\hat{u}_r'\hat{u}_r - \hat{u}_u'\hat{u}_u)/k_2}{\hat{u}_u\hat{u}_u/(n - k_1 - k_2)}, \tag{20}$$

which is distributed as $F(k_2, n - k_1 - k_2)$.

Since $\hat{u}_u = (I - X(X'X)^{-1}X')u$ and $\hat{u}_r = (I - X_1(X_1'X_1)^{-1}X_1')u$, the numerator is $u'(A_r - A_u)u/\mathrm{tr}(A_r - A_u)$, where the trace in question is $n - k_1 - (n - k_1 - k_2)) = k_2$. By the same token, the denominator is $u'A_uu/(n - k_1 - k_2)$. But this is the same as the statistic (12).

Case 2: Equality of Regression Coefficients in Two Regressions. Let the model be given by

$$Y_i = X_i\beta_i + u_i \qquad i = 1, 2, \tag{21}$$

where X_1 is $n_1 \times k$ and X_2 is $n_2 \times k$, and where $k < \min(k_1, k_2)$. We test $H_0 : \beta_1 = \beta_2$. The unrestricted model is then written as

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{22}$$

and the model restricted by the hypothesis can be written as

$$Y = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta_1 + u. \tag{23}$$

We obviously have

$$X_r = X \begin{bmatrix} I \\ I \end{bmatrix},$$

where the the matrix in brackets on the right is of order $2k \times k$ and the conditions of Theorem 7 are satisfied. The traces of the restricted and unrestricted A-matrices are

$$\mathrm{tr}(A_r) = \mathrm{tr}\big(I - X_r(X_r'X_r)^{-1}X_r'\big) = \mathrm{tr}\Big\{I_{n_1+n_2} - \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} (X_1'X_1 + X_2'X_2)^{-1}(X_1' \;\; X_2')\Big\} = n_1 + n_2 - k$$

and

$$\mathrm{tr}(A_u) = \mathrm{tr}\Big\{I - \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1' & 0 \\ 0 & X_2' \end{bmatrix}\Big\} = n_1 + n_2 - 2k.$$

Letting $\hat{u}_u$ and $\hat{u}_r$ denote, as before, the unrestricted and restricted residuals respectively, the test statistic becomes

$$\frac{(\hat{u}_r'\hat{u}_r - \hat{u}_u'\hat{u}_u)/k}{\hat{u}_u'\hat{u}_u/(n_1+n_2-2k)},$$

which is distributed as $F(k, n_1+n_2-2k)$.

Case 3: Equality of Subsets of Coefficients in Two Regressions. Now write the model as

$$Y_i = X_i\beta_i + Z_i\gamma_i + u_i \qquad i = 1, 2,$$

where X_i is of order $n_i \times k_1$ and Z_i is of order $n_i \times k_2$, with $k_1 + k_2 < \min(n_1, n_2)$. The hypothesis to be tested is $H_0 : \beta_1 = \beta_2$.

The unrestricted and restricted models are, respectively,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 & Z_1 & 0 \\ 0 & X_2 & 0 & Z_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = X\delta + u \tag{24}$$

and

$$Y = \begin{bmatrix} X_1 & Z_1 & 0 \\ X_2 & 0 & Z_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + u = X_r\delta_r + u. \tag{25}$$

Clearly,

$$X_r = X \begin{bmatrix} I & 0 & 0 \\ I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

and hence the conditions of Theorem 7 are satisfied. Hence we can form the F-ratio as in Eq.(12) or (20), where the numerator degrees of freedom are $\mathrm{tr}(A_r - A_u) = n_1 + n_2 - k_1 - 2k_2 - (n_1 + n_2 - 2k_1 - 2k_2) = k_1$ and the denominator degrees of freedom are $\mathrm{tr}(A_u) = n_1 + n_2 - 2k_1 - 2k_2$.

Insufficient Degrees of Freedom.

Case 4: Equality of Regression Coefficients in Two Regressions. This case is the same as Case 2, except we now assume that $n_2 \leqq k$. Denote by $\hat{u}_r$ be the residuals from the restricted regression using the full set of $n_1 + n_2$ observations and let $\hat{u}_u$ denote the residuals from the regression using only the first n_1 observations. Then $\hat{u}_r = A_r u$, $\hat{u}_u = A_1 u_1$, where $A_1 = I - X_1(X_1'X_1)^{-1}X_1'$. We can then also write $\hat{u}_u = [A_1 \quad 0]u$, and $\hat{u}_u'\hat{u}_u = u'A_u u$, where

$$A_u = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} \tag{26}$$

is an $(n_1 + n_2) \times (n_1 + n_2)$ matrix. Since $X_1'A_1 = 0$, we have

$$X'A_u = [X_1' \ \ X_2'] \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

Hence $A_r A_u = A_u$, and the conditions of Theorem 7 are satisfied. The relevant traces are $\mathrm{tr}(A_r) = n_1 + n_2 - k$ and $\mathrm{tr}(A_u) = n_1 - k$.

Case 5: Equality of Subsets of Coefficients in Two Regressions. This is the same case as Case 3, except that we now assume that $k_2 < n_2 \leqq k_1 + k_2$; hence there are not enough observations in the second part of the dataset to estimate a separate regression equation. As before, let $\hat{u}_r$ be the residuals from the restricted model, and $\hat{u}_u$ be the residuals from the regression on the first n_1 observations. Denote by W_1 the matrix $[X_1 \quad Z_1]$. Again as before, $\hat{u}_r = A_r u$, $\hat{u}_u = A_1 u_1 = [A_1 \quad 0]u$, where $A_1 = I - W_1(W_1'W_1)^{-1}W_1']$. Defining A_u as in Eq.(26), we again obtain $A_r A_u = A_u$ and the conditions of Theorem 7 are satisfied. The relevant traces then are $\mathrm{tr}(A_r) = n_1 + n_2 - k_1 - 2k_2$ and $\mathrm{tr}(A_u) = n_1 - k_1 - k - 2$. Notice that the requirement that $tr(A_r) - \mathrm{tr}(A_u)$ be positive is fulfilled if and only if $n_2 > k_2$, as we assumed.

Irrelevant Variables Included.

Consider the case in which

$$Y = X_1\beta_1 + u$$

is the "true" model, but in which the investigator mistakenly estimates the model

$$Y = X_1\beta_1 + X_2\beta_2 + u.$$

The estimated coefficient vector $\hat{\beta}' = (\hat{\beta}_1' \quad \hat{\beta}_2')$ becomes

$$\begin{aligned} \hat{\beta} &= \begin{bmatrix} (X_1'X_1) & (X_1'X_2) \\ (X_2'X_1) & (X_2'X_2) \end{bmatrix}^{-1} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} [X_1\beta_1 + u] \\ &= \begin{bmatrix} I \\ 0 \end{bmatrix} \beta_1 + \begin{bmatrix} (X_1'X_1) & (X_1'X_2) \\ (X_2'X_1) & (X_2'X_2) \end{bmatrix}^{-1} \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} u \end{aligned},$$

from which it follows that

$$E\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix},$$

and hence the presence of irrelevant variables does not affect the unbiasedness of the regression parameter estimates. We next prove two lemmas needed for Theorem 22.

Lemma 1. *If A is $n \times k$, $n \geqq k$, with rank k, then $A'A$ is nonsingular.*

Proof. We can write

$$A'A = [A_1' \quad A_2'] \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = A_1'A_1 + A_2'A_2,$$

where $A_1'A_1$ is of order and rank k. We can then write $C'A_1 = A_2$, where C' is an $(n-k) \times k$ matrix. Then $A'A = A_1'[I + CC']A_1$. The matrix CC' is obviously positive semidefinite; but then $I + CC'$ is positive definite (because its eigenvalues exceed the corresponding eigenvalues of CC' by unity). But then $A'A$ is the product of three nonsingular matrices. ∎

Lemma 2. *Let X_1 be $n \times k_1$, X_2 be $n \times k_2$, let the columns of $X = [X_1 \quad X_2]$ be linearly independent, and define $M = I - X_1(X_1'X_1)^{-1}X_1'$. Then the matrix $(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)$ is nonsingular.*

Proof. Write $(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)$ as $X_2'MX_2$. M obviously has rank $\rho(M) = n - k_1$. Since $MX_1 = 0$, the columns of X_1 span the null-space of M. It follows that $\rho(MX_2) = k_2$, for if the rank of MX_2 were smaller than k_2, there would exist a vector $c_2 \neq 0$ such that $MX_2c_2 = 0$, and the vector X_2c_2 would lie in the null-space of M, and would therefore be spanned by the columns of X_1. But then we could write $X_2c_2 + X_1c_1 = 0$, for a vector $c_1 \neq 0$, which contradicts the assumption that the columns of X are linearly independent.

But then it follows that $X_2'MX_2 = X_2'M'MX_2$ has rank k_2 by Lemma 1. ∎

Theorem 22. *The covariance matrix for $\hat{\beta}_1$ with irrelevant variables included exceeds the covariance matrix for the correctly specified model by a positive semidefinite matrix.*

Proof. The covariance matrix for $\hat{\beta}_1$ for the incorrectly specified model is obviously the upper left-hand block of of

$$\sigma^2 \begin{bmatrix} (X_1'X_1) & (X_1'X_2) \\ (X_2'X_1) & (X_2'X_2) \end{bmatrix}^{-1}$$

whereas the covariance matrix in the correctly specified model is

$$\sigma_2(X_1'X_1)^{-1}.$$

Since the inverse of a partitioned matrix can be written as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}\big[I + B(D - CA^{-1}B\big)^{-1}CA^{-1}\big] & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix},$$

the required upper left-hand block of the covariance matrix in the misspecified model is

$$\begin{aligned} &\sigma^2(X_1'X_1)^{-1}\big[I + (X_1'X_2)[(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)]^{-1}(X_2'X_1)(X_1'X_1)^{-1}\big] \\ &= \sigma^2\Big\{(X_1'X_1)^{-1} + (X_1'X_1)^{-1}(X_1'X_2)\big[X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2\big]^{-1}(X_2'X_1)(X_1'X_1)^{-1}\Big\} \end{aligned}$$

Subtracting $\sigma^2(X_1'X_1)^{-1}$, we obtain the difference between the two covariance matrices as

$$\sigma^2\Big\{(X_1'X_1)^{-1}(X_1'X_2)\big[X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2\big]^{-1}(X_2'X_1)(X_1'X_1)^{-1}\Big\}$$

Since the matrix in square brackets is positive definite, its inverse exists and the matrix in $\{\}$ is positive semidefinite. ∎

Relevant Variables Omitted.

Consider the case in which the true relation is

$$Y = X_1\beta_1 + X_2\beta_2 + u, \tag{27}$$

but in which the relation

$$Y = X_1\beta_1 + u \tag{28}$$

is estimated. Then we have

Theorem 23. *For the least squares estimator $\hat{\beta}_1$ we have $E(\hat{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2$.*

Proof. We have $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1Y = (X_1'X_1)^{-1}X_1'[X_1\beta_1 + X_2\beta_2 + u]$, and taking expectations leads to the result. ∎

Estimation with Linear Restrictions.

Now consider the model

$$Y = X\beta + u, \tag{29}$$

$$r = R\beta, \tag{30}$$

where r is $p \times 1$, R is $p \times k$, $p < k$, and where the rank of R is $\rho(R) = p$. We assume that the elements of r and R are known numbers; if the rank of R were less than p, then some restrictions could be expressed as linear combinations of other restrictions and may be omitted. Minimizing the sum of the squares of the residuals subject to (30) requires forming the Lagrangian

$$V = (Y - X\beta)'(Y - X\beta) - \lambda'(R\beta - r), \tag{31}$$

where λ is a vector of Lagrange multipliers. Now denote by $\tilde{\beta}$ and $\tilde{\lambda}$ the estimates obtained by setting the partial derivatives of Eq.(29) equal to zero, and let $\hat{\beta}$ denote the least squares estimates without imposing the restrictions (30). Then we have

Theorem 24. *$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - R\hat{\beta})$ and $\tilde{\lambda} = 2(R(X'X)^{-1}R')^{-1}(r - R\hat{\beta})$.*

Proof. Setting the partial derivatives of Eq.(31) equal to zero yields

$$\frac{\partial V}{\partial \beta} = -2X'Y + 2(X'X)\tilde{\beta} - R'\tilde{\lambda} = 0 \tag{32}$$

$$\frac{\partial V}{\partial \lambda} = R\tilde{\beta} - r = 0. \tag{33}$$

Multiplying Eq.(32) on the left by $R(X'X)^{-1}$ yields

$$-2R(X'X)^{-1}X'Y + 2R\tilde{\beta} - R(X'X)^{-1}R'\tilde{\lambda} = 0$$

or

$$-2R\hat{\beta} + 2R\tilde{\beta} - R(X'X)^{-1}R'\tilde{\lambda} = 0. \tag{34}$$

Since $\tilde{\beta}$ satisfies the constraints by definition, Eq.(34) yields

$$\tilde{\lambda} = 2\big[R(X'X)^{-1}R'\big]^{-1}(r - R\hat{\beta}).$$

Substituting this result in Eq.(32) and solving for $\tilde{\beta}$ yields

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}(r - R\hat{\beta}). \tag{34a}$$

$\blacksquare$

Corollary 3. *If* $r - R\hat{\beta} = 0$, *then* $E(\tilde{\beta}) = \beta$ *and* $\tilde{\lambda} = 0$.

Proof. Substituting $X\beta + u$ for Y and $(X'X)^{-1}X'Y$ for $\hat{\beta}$ in the expression for $\tilde{\beta}$ in Eq.(34a) and taking expectations yields the result for $E(\tilde{\beta})$. Using this by replacing $\tilde{\beta}$ by β in the formula for $\tilde{\lambda}$ yields the result for $E(\tilde{\beta})$.

Define

$$A = I - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R. \tag{35}$$

We then have

Theorem 25. *The covariance matrix of* $\tilde{\beta}$ *is* $\sigma^2 A(X'X)^{-1}$.

Proof. Substituting $(X'X)^{-1}X'Y$ for $\hat{\beta}$ and $X\beta + u$ for Y in $\tilde{\beta}$ (Eq.(34a)), we can write

$$\begin{aligned}\tilde{\beta} - E(\tilde{\beta}) &= (X'X)^{-1}\big[X' - R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1}X'\big]u \\ &= \big[I - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R\big](X'X)^{-1}X'u = A(X'X)^{-1}X'u.\end{aligned} \tag{36}$$

Multiplying Eq.(36) by its transpose and taking expectations, yields

$$\begin{aligned}\mathrm{Cov}(\tilde{\beta}) &= \sigma^2 A(X'X)^{-1}A' \\ &= \sigma^2\big[I - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R\big](X'X)^{-1}\big[I - R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1}\big] \\ &= \sigma^2\big[(X'X)^{-1} - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1} \\ &\qquad - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1} \\ &\qquad + (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1}\big] \\ &= \sigma^2\big[(X'X)^{-1} - (X'X)^{-1}R'\big[R(X'X)^{-1}R'\big]^{-1}R(X'X)^{-1}\big] \\ &= \sigma^2 A(X'X)^{-1}.\end{aligned}$$

$\blacksquare$

We now consider the test of the null hypothesis $H_0 : R\beta = r$. For this purpose we construct an F-statistic as in Theorem 20 (see also Eq.(12)).

The minimum sum of squares subject to the restriction can be written as

$$\begin{aligned} S_r &= \{Y - X[\hat\beta + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta)]\}' \\ &\quad \times \{Y - X[\hat\beta + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta)]\} \\ &= (Y - X\hat\beta)'(Y - X\hat\beta) - [X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta)]'(Y - X\hat\beta) \\ &\quad - (Y - X\hat\beta)'[X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta)] \\ &\quad + (r - R\hat\beta)'[R(X'X)^{-1}R']^{-1}(X'X)^{-1}(X'X)(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta) \\ &= S_u + (r - R\hat\beta)'[R(X'X)^{-1}R']^{-1}(r - R\hat\beta), \end{aligned} \tag{37}$$

where S_u denotes the unrestricted minimal sum of squares, and where the disappearance of the second and third terms in the third and fourth lines of the equation is due to the fact that that $X'(Y - X\hat\beta) = 0$ by the definition of $\hat\beta$. Substituting the least squares estimate for $\hat\beta$ in (37), we obtain

$$\begin{aligned} S_r - S_u &= [r - R(X'X)^{-1}X'Y]'[R(X'X)^{-1}R']^{-1}[r - R(X'X)^{-1}X'Y] \\ &= [r - R\beta - R(X'X)^{-1}X'u]'[R(X'X)^{-1}R']^{-1}[r - R\beta - R(X'X)^{-1}X'u] \\ &= u'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'u = u'B_1u, \end{aligned} \tag{38}$$

since under H_0, $r - R\beta = 0$. The matrix B_1 is idempotent and of rank p because

$$\begin{aligned} &X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}(X'X)(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \\ &\qquad = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \end{aligned}$$

and

$$\begin{aligned} \mathrm{tr}(B_1) &= \mathrm{tr}(X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X') \\ &= \mathrm{tr}(R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}(X'X)(X'X)^{-1}) \\ &= \mathrm{tr}([R(X'X)^{-1}R']^{-1}R(X'X)^{-1}R') = \mathrm{tr}(I_p) = p. \end{aligned}$$

The matrix of the quadratic form S_u is clearly $B_2 = I - X(X'X)^{-1}X'$ which is idempotent and or rank $n - k$. Moreover, $B_1B_2 = 0$, since

$$X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'(I - X(X'X)^{-1}X') = 0.$$

Hence

$$\frac{(S_r - S_u)/p}{S_u/(n-k)}$$

is distributed as $F(p, n - k)$.

We now turn to the case in which the covariance matrix of u is Ω and we wish to test the hypothesis $H_0 : R\beta = r$. We first assume that Ω is known. We first have

Theorem 26. *If u is distributed as $N(0, \Omega)$, and if Ω is known, then the Lagrange Multiplier, Wald, and likelihood ratio test statistics are identical.*

Proof. The loglikelihood function is

$$\log L(\beta) = (2\pi)^{-n/2} + \frac{1}{2}\log|\Omega^{-1}| - \frac{1}{2}(Y - X\beta)'\Omega^{-1}(Y - X\beta),$$

where $|\Omega^{-1}|$ denotes the determinant of Ω^{-1}, and the score vector is

$$\frac{\partial \log L}{\partial \beta} = X'\Omega^{-1}(Y - X\beta).$$

By further differentiation, the Fischer Information matrix is

$$I(\beta) = X'\Omega^{-1}X.$$

The unrestricted maximum likelihood estimator for β is obtained by setting the score vector equal to zero and solving, which yields

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

Letting $\hat{u}$ denote the residuals $Y - X\hat{\beta}$, the loglikelihood can be written as

$$\log L = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Omega^{-1}| - \frac{1}{2}\hat{u}'\Omega^{-1}\hat{u}.$$

To obtain the estimates restricted by the linear relations $R\beta = r$, we form the Lagrangian

$$L(\beta, \lambda) = \log L(\beta) + \lambda'(R\beta - r)$$

and set its partial derivatives equal to zero, which yields

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= X'\Omega^{-1}(Y - X\beta) + R'\lambda = 0 \\ \frac{\partial \log L}{\partial \lambda} &= R\beta - r = 0. \end{aligned} \tag{39}$$

Multiply the first equation in (39) by $(X'\Omega^{-1}X)^{-1}$, which yields

$$\tilde{\beta} = \hat{\beta} + (X'\Omega^{-1}X)^{-1}R'\tilde{\lambda}.$$

Multiplying this further by R, and noting that $R\tilde{\beta} = r$, we obtain

$$\tilde{\lambda} = -\big[R(X'\Omega^{-1}X)^{-1}R'\big]^{-1}(R\hat{\beta} - r) \tag{40}$$

$$\tilde{\beta} = \hat{\beta} - (X'\Omega^{-1}X)^{-1}R'\big[R(X'\Omega^{-1}X)^{-1}R'\big]^{-1}(R\hat{\beta} - r). \tag{41}$$

The loglikelihood, evaluated at $\tilde\beta$ is

$$\log L(\tilde\beta) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Omega^{-1}| - \frac{1}{2}\tilde u'\Omega^{-1}\tilde u.$$

We now construct the test statistics. The Lagrange multiplier statistic is

$$\begin{aligned} LM &= \left[\frac{\partial \log L}{\partial\tilde\beta}\right]' I(\tilde\beta)^{-1}\left[\frac{\partial \log L}{\partial\tilde\beta}\right] = \tilde u'\Omega^{-1}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\tilde u \\ &= \tilde\lambda' R(X'\Omega^{-1}X)^{-1}R'\tilde\lambda. \end{aligned} \tag{42}$$

The Wald statistic is

$$W = (R\tilde\beta - r)'\big[R(X'\Omega^{-1}X)^{-1}R'\big]^{-1}(R\tilde\beta - r), \tag{43}$$

and since the covariance matrix of $(R\tilde\beta - r)$ is $R(X'\Omega^{-1}X)^{-1}R'$, W can be written as

$$\begin{aligned} W &= (R\tilde\beta - r)'\big[R(X'\Omega^{-1}X)^{-1}R'\big]^{-1}\big[R(X'\Omega^{-1}X)^{-1}R'\big]\big[R(X'\Omega^{-1}X)^{-1}R'\big]^{-1}(R\tilde\beta - r) \\ &= \tilde\lambda'\big[R(X'\Omega^{-1}X)^{-1}R'\big]\tilde\lambda = LM, \end{aligned}$$

where we have used the definition of $\tilde\lambda$ in (40). The likelihood ratio test statistic is

$$LR = -2\big[\log L(\tilde\beta) - \log L(\hat\beta)\big] = \tilde u'\Omega^{-1}\tilde u - \hat u'\Omega^{-1}\hat u. \tag{44}$$

Since $\Omega^{-1/2}\tilde u = \Omega^{-1/2}(Y - X\tilde\beta)$, and substituting in this for $\tilde\beta$ from its definition in (41), we obtain

$$\Omega^{-1/2}\tilde u = \Omega^{-1/2}\big[Y - X\hat\beta - X(X'\Omega^{-1}X)^{-1}R'\tilde\lambda\big]. \tag{45}$$

We multiply Eq.(45) by its transpose and note that terms with $(Y - X\hat\beta)\Omega^{-1}X$ vanish; hence we obtain

$$\tilde u'\Omega^{-1}\tilde u = \hat u'\Omega^{-1}\hat u + \tilde\lambda' R(X'\Omega^{-1}X)^{-1}R'\tilde\lambda.$$

But the last term is the Lagrange multiplier test statistic from (42); hence comparing this with (44) yields $LR = LM$. ∎

We now consider the case when Ω is unknown, but is a smooth function of a p-element vector α, and denoted by $\Omega(\alpha)$. We then have

Theorem 27. *If u is normally distributed as $N(0, \Omega(\alpha))$, then $W \geqq LR \geqq LM$.*

Proof. Denote by θ' the vector (β', α'). The loglikelihood is

$$\log L(\theta) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Omega^{-1}(\alpha)| + \frac{1}{2}(Y - X\beta)'\Omega^{-1}(\alpha)(Y - X\beta).$$

Denoting the unrestricted estimates by $\hat{\theta}$ and the restricted estimates by $\tilde{\theta}$, as before, and in particular, denoting by $\hat{\Omega}$ the matrix $\Omega(\hat{\alpha})$ and by $\tilde{\Omega}$ the matrix $\Omega(\tilde{\alpha})$, the three test statistics can be written, in analogy with Eqs.(42) to (44), as

$$\begin{aligned} LM &= \tilde{u}'\tilde{\Omega}^{-1}X(X'\tilde{\Omega}^{-1}X)^{-1}X'\tilde{\Omega}^{-1}\hat{u} \\ W &= (R\hat{\beta}-r)'\big[R(X'\hat{\Omega}^{-1}X)^{-1}R'\big]^{-1}(R\hat{\beta}-r) \\ LR &= -2\big(\log L(\tilde{\alpha},\tilde{\beta}) - \log L(\hat{\alpha},\hat{\beta})\big). \end{aligned}$$

Now define

$$LR(\tilde{\alpha}) = -2\big(\log L(\tilde{\alpha},\tilde{\beta}) - \log L(\tilde{\alpha},\tilde{\beta}_u)\big), \tag{46}$$

where $\tilde{\beta}_u$ is the unrestricted maximizer of $\log L(\tilde{\alpha},\beta)$ and

$$LR(\hat{\alpha}) = -2\big(\log L(\hat{\alpha},\hat{\beta}_r) - \log L(\hat{\alpha},\hat{\beta})\big), \tag{47}$$

where $\hat{\beta}_r$ is the maximizer of $\log L(\hat{\alpha},\beta)$ subject to the restriction $R\beta - r = 0$. $LR(\tilde{\alpha})$ employs the same Ω matrix as the LM statistic; hence by the argument in Theorem 26,

$$LR(\tilde{\alpha}) = LM.$$

It follows that

$$LR - LM = LR - LR(\tilde{\alpha}) = 2\big(\log L(\hat{\alpha},\hat{\beta}) - \log L(\tilde{\alpha},\tilde{\beta}_u)\big] \geqq 0,$$

since the $\hat{\alpha}$ and $\hat{\beta}$ estimates are unrestricted. We also note that W and $LR(\hat{\alpha})$ use the same Ω, hence they are equal by Theorem 26. Then

$$W - LR = LR(\hat{\alpha}) - LR = 2\big(\log L(\tilde{\alpha},\tilde{\beta}) - \log L(\hat{\alpha},\hat{\beta}_r)\big) \geqq 0, \tag{48}$$

since $\hat{\beta}_r$ is a restricted estimate and the highest value of the likelihood with the restriction that can be achieved is $\log L(\tilde{\alpha},\tilde{\beta})$. Hence $W \geqq LR \geqq LM$. ▮

We now prove a matrix theorem that will be needed subsequently.

Theorem 28. *If Σ is symmetric and positive definite of order p, and if H is of order $p \times q$, with $q \leqq p$, and if the rank of H is q, then*

$$\begin{bmatrix} \Sigma & H \\ H' & 0 \end{bmatrix}$$

is nonsingular.

Proof. First find a matrix, conformable with the first,

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix}$$

such that

$$\begin{bmatrix} \Sigma & H \\ H' & 0 \end{bmatrix} \begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Performing the multiplication and equating the two sides, we obtain

$$\Sigma P + HQ' = I \tag{49}$$

$$\Sigma Q + HR = 0 \tag{50}$$

$$H'P = 0 \tag{51}$$

$$H'Q = I \tag{52}$$

From (49) we have

$$P + \Sigma^{-1}HQ' = \Sigma^{-1}. \tag{53}$$

Multiplying Eq.(53) on the left by H', and noting from Eq.(51) that $H'P = 0$, we have

$$H'\Sigma^{-1}HQ' = H'\Sigma^{-1}. \tag{54}$$

Since H is of full rank, $H'\Sigma^{-1}H$ is nonsingular by a straightforward extension of Lemma 1. Then

$$Q' = (H'\Sigma^{-1}H)^{-1}H'\Sigma^{-1}, \tag{55}$$

which gives us the value of Q. Substituting (55) in Eq.(53) gives

$$P = \Sigma^{-1} - \Sigma^{-1}H(H'\Sigma^{-1}H)^{-1}H'\Sigma^{-1}. \tag{56}$$

From Eq.(50) we have

$$\Sigma^{-1}HR = -Q,$$

and multiplying this by H' and using Eq.(52) yields

$$H'\Sigma^{-1}HR = -I$$

and

$$R = -(H'\Sigma^{-1}H)^{-1}, \tag{57}$$

which determines the value of R. Since the matrix

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix}$$

is obviously the inverse of the matrix in the theorem, the proof is complete. ∎

We now consider the regression mode $Y = X\beta + u$, where u is distributed as $N(0, \sigma^2 I)$, subject to the restrictions $R\beta = 0$; hence this is the same model as considered before with $r = 0$. Minimize the sum of squares subject to the restrictions by forming the Lagrangian

$$L = (Y - X\beta)'(Y - X\beta) + \lambda' R\beta. \tag{58}$$

The first order conditions can be written as

$$\begin{bmatrix} (X'X)^{-1} & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} X'Y \\ 0 \end{bmatrix}. \tag{59}$$

Denote the matrix on the left hand side of (59) by A, and write its inverse as

$$A^{-1} = \begin{bmatrix} P & Q \\ Q' & S \end{bmatrix}. \tag{60}$$

We can then write the estimates as

$$\begin{bmatrix} \tilde{\beta} \\ \tilde{\lambda} \end{bmatrix} = \begin{bmatrix} PX'Y \\ Q'X'Y \end{bmatrix}, \tag{61}$$

and taking expectations, we have

$$E \begin{bmatrix} \tilde{\beta} \\ \tilde{\lambda} \end{bmatrix} = \begin{bmatrix} PX'X\beta \\ Q'X'X\beta \end{bmatrix}. \tag{62}$$

From multiplying out $A^{-1}A$ we obtain

$$PX'X + QR = I \tag{63}$$

$$Q'X'X + SR = 0 \tag{64}$$

$$PR' = 0 \tag{65}$$

$$Q'R' = I \tag{66}$$

Hence we can rewrite Eq.(62) as

$$E \begin{bmatrix} \tilde{\beta} \\ \tilde{\lambda} \end{bmatrix} = \begin{bmatrix} (I - QR)\beta \\ -SR\beta \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \end{bmatrix}, \tag{67}$$

since $R\beta = 0$ by definition. This, so far, reproduces Corollary 3.

Theorem 29. *Given the definition in Eq.(61), the covariance matrix of $(\tilde{\beta}, \tilde{\lambda})$ is*

$$\sigma^2 \begin{bmatrix} P & 0 \\ 0 & -S \end{bmatrix}.$$

Proof. It is straightforward to note that

$$\operatorname{cov}(\tilde{\beta}, \tilde{\lambda}) = E \left[\begin{pmatrix} \tilde{\beta} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} \beta \\ 0 \end{pmatrix} \right] \left[\begin{pmatrix} \tilde{\beta} \\ \tilde{\lambda} \end{pmatrix} - \begin{pmatrix} \beta \\ 0 \end{pmatrix} \right]' = \sigma^2 \begin{bmatrix} PX'XP & PX'XQ \\ QX'XP & Q'X'XQ \end{bmatrix}. \tag{68}$$

From (65) and (66), multiplying the second row of A into the first column of A^{-1} gives

$$RP = 0,$$

and multiplying it into the second column gives

$$RQ = I.$$

Hence, multiplying Eq.(63) on the right by P gives

$$PX'XP + QRP = P$$

or, since $RP = 0$,

$$PX'XP = P.$$

Multiplying Eq.(63) by Q on the right gives

$$PX'XQ + QRQ = Q,$$

or, since $RQ = I$,

$$PX'XQ = 0.$$

Finally, multiplying (64) by Q on the right gives

$$Q'X'XQ + SRQ = 0,$$

which implies that

$$Q'X'XQ = -S.$$

∎

We now do large-sample estimation for the general unconstrained and constrained cases. We wish to estimate the parameters θ of the density function $f(x, \theta)$, where x is a random variable and θ is a parameter vector with k elements. In what follows, we denote the true value of θ by θ_0. The loglikelihood is

$$\log L(x, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta). \tag{69}$$

Let $\hat{\theta}$ be the maximum likelihood estimate and let D_θ be the differential operator. Also define $I_1(\theta)$ as $\mathrm{var}(D_\theta \log f(x, \theta))$. It is immediately obvious that $\mathrm{var}(D_\theta \log L(x, \theta)) = nI_1(\theta)$. Expanding in Taylor Series about θ_0, we have

$$0 = D_\theta \log L(x, \hat{\theta}) = D_\theta \log L(x, \theta_0) + (\hat{\theta} - \theta_0) D_\theta^2 \log L(x, \theta_0) + R(x, \theta_0, \hat{\theta}) \tag{70}$$

Theorem 30. *If $\hat{\theta}$ is a consistent estimator and the third derivative of the loglikelihood function is bounded, then $\sqrt{n}(\hat{\theta} - \theta_0)$ is distributed as $N(0, I_1(\theta_0)^{-1})$.*

Proof. From Eq.(70) we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{n^{-1/2} D_\theta \log L(x, \theta_0) + n^{-1/2} R(x, \theta_0, \hat{\theta})}{n^{-1} D_\theta^2 \log L(x, \theta_0)} \tag{71}$$

where R is a remainder term of the form $(\hat{\theta} - \theta_0)^3 D_\theta^3 (\log L(x, \overline{\theta})/2$, $\overline{\theta}$ being between $\hat{\theta}$ and θ_0. The quantity $n^{-1/2} D_\theta \log L(x, \theta_0)$ is a sum of n terms, each of which has expectation 0 and variance $I_1(\theta_0)$; hence by the Central Limit Theorem, $n^{-1/2} D_\theta \log L(x, \theta_0)$ is asymptotically normally distributed with mean zero and variance equal to $(1/n) n I_1(\theta_0) = I_1(\theta_0)$. The remainder term converges in probability to zero. The denominator is $1/n$ times the sum of n terms, each of which has expectation equal to $-I_1(\theta_0)$; hence the entire denominator has the same expectation and by the Weak Law of Large Numbers the denominator converges to this expectation. Hence $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a random variable which is $I_1(\theta_0)^{-1}$ times an $N(0, I_1(\theta_0))$ variable and hence is asymptotically distributed as $N(0, I_1(\theta_0)^{-1})$.

We now consider the case when there are p restrictions given by $h(\theta)' = (h_1(\theta), \ldots, h_p(\theta)) = 0$. Estimation subject to the restrictions requires forming the Lagrangian

$$G = \log L(x, \theta) - \lambda' h(\theta)$$

and setting its first partial derivatives equal to zero:

$$\begin{aligned} D_\theta \log L(x, \tilde{\theta}) - H_\theta \tilde{\lambda} &= 0 \\ h(\tilde{\theta}) \qquad\qquad &= 0 \end{aligned} \tag{72}$$

where H_θ is the $k \times p$ matrix of the derivatives of $h(\theta)$ with respect to θ. Expanding in Taylor Series and neglecting the remainder term, yields asymptotically

$$\begin{aligned} D_\theta \log L(x, \theta_0) + D_\theta^2 \log L(x, \theta_0)(\tilde{\theta} - \theta_0) - H_\theta \tilde{\lambda} &= 0 \\ H_\theta'(\tilde{\theta} - \theta_0) \qquad\qquad &= 0 \end{aligned} \tag{73}$$

The matrix H_θ should be evaluated at $\tilde{\theta}$; however, writing $H_\theta(\tilde{\theta})\tilde{\lambda} = H_\theta(\theta_0)\tilde{\lambda} + H_\theta'(\theta_0)(\tilde{\theta} - \theta_0)$ and noting that if the restrictions hold, $\tilde{\theta}$ will be near θ_0 and $\tilde{\lambda}$ will be small, we may take H_θ to be evaluated at θ_0.

Theorem 31. *The vector*

$$\begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta_0) \\ \frac{1}{\sqrt{n}} \tilde{\lambda} \end{bmatrix}$$

is asymptotically normally distributed with mean zero and covariance matrix

$$\begin{bmatrix} P & 0 \\ 0 & S \end{bmatrix}$$

where

$$\begin{bmatrix} I_1(\theta_0) & H_\theta \\ H'_\theta & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P & Q \\ Q' & S \end{bmatrix}.$$

Proof. Dividing the first line of (73) by $\sqrt{n}$ and multiplying the second line by $\sqrt{n}$, we can write

$$\begin{bmatrix} -\frac{1}{n}D^2_\theta \log L(x,\theta_0) & H_\theta \\ H'_\theta & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \frac{1}{n}\tilde{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}}D_\theta \log L(x,\theta_0) \\ 0 \end{bmatrix}. \tag{74}$$

The upper left-hand element in the left-hand matrix converges in probability to $I_1(\theta_0)$ and the top element on the right hand side converges in distribution to $N(0, I_1(\theta_0))$. Thus, (74) can be written as

$$\begin{bmatrix} I_1(\theta_0) & H_\theta \\ H'_\theta & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \frac{1}{n}\tilde{\lambda} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}}D_\theta \log L(x,\theta_0) \\ 0 \end{bmatrix}. \tag{75}$$

Eq.(75) is formally the same as Eq.(59); hence by Theorem 29,

$$\begin{bmatrix} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \frac{1}{\sqrt{n}}\tilde{\lambda} \end{bmatrix}$$

is asymptotically normally distributed with mean zero and covariance matrix

$$\begin{bmatrix} P & 0 \\ 0 & S \end{bmatrix}$$

where

$$\begin{bmatrix} I_1(\theta_0) & H_\theta \\ H'_\theta & 0 \end{bmatrix}^{-1} = \begin{bmatrix} P & Q \\ Q' & S \end{bmatrix}. \tag{76}$$

∎

We now turn to the derivation of the asymptotic distribution of the likelihood ratio test statistic. As before, $\hat{\theta}$ denotes the unrestricted, and $\tilde{\theta}$ the restricted estimator.

Theorem 32. *Under the assumptions that guarantee that both the restricted and unrestricted estimators ($\tilde{\theta}$ and $\hat{\theta}$ respectively) are asymptotically normally distributed with mean zero and covariance matrices $I_1(\theta_0)$ and P respectively, and if the null hypothesis $H_0 : h(\theta) = 0$ is true, the* *likelihood ratio test statistic,* $2\log\mu = 2(\log L(x,\hat\theta) - \log L(x,\tilde\theta))$ *is asymptotically distributed as* $\chi^2(p)$.

Proof. Expand $\log L(x,\tilde\theta)$ in Taylor Series about $\hat\theta$, which yields to an approximation

$$\log L(x,\tilde\theta) = \log L(x,\hat\theta) + D_\theta \log L(x,\hat\theta)(\hat\theta-\tilde\theta) + \frac{1}{2}(\hat\theta-\tilde\theta)'\big[D_\theta^2(\log L(x,\hat\theta)\big](\hat\theta-\tilde\theta). \tag{77}$$

Since the second term on the right hand side is zero by definition, the likelihood ratio test statistic becomes

$$2\log\mu = (\hat\theta-\tilde\theta)'\big[-D_\theta^2 \log L(x,\hat\theta)\big](\hat\theta-\tilde\theta). \tag{78}$$

Let v be a k-vector distributed as $N(0, I_1(\theta_0))$. Then we can write

$$\begin{aligned}\sqrt{n}(\hat\theta-\theta_0) &= I_1(\theta_0)^{-1}v\\ \sqrt{n}(\tilde\theta-\theta_0) &= Pv\end{aligned} \tag{79}$$

where P is the same P as in Eq.(76). Then, to an approximation,

$$\begin{aligned}2\log\mu &= v'(I_1(\theta_0)^{-1}-P)'I_1(\theta_0)(I_1(\theta_0)^{-1}-P)v\\ &= v'(I_1(\theta_0)^{-1}-P-P+PI_1(\theta_0)P)v\end{aligned}. \tag{80}$$

We next show that $P = PI_1(\theta_0)P$. From Eq.(56) we can write

$$P = I_1(\theta_0)^{-1} - I_1(\theta_0)^{-1}H(H'I_1(\theta_0)^{-1}H)^{-1}H'I_1(\theta_0)^{-1}. \tag{80}$$

Multiplying this on the left by $I_1(\theta_0)$ yields

$$I_1(\theta_0)P = I - H(H'I_1(\theta_0)^{-1}H)^{-1}H'I_1(\theta_0)^{-1},$$

and multiplying this on the left by P (using the right-hand side of (81)), yields

$$\begin{aligned}PI_1(\theta_0)P = I_1(\theta_0)^{-1} &- I_1(\theta_0)^{-1}H[H'I_1(\theta_0)^{-1}H]^{-1}H'I_1(\theta_0)^{-1}\\ &- I_1(\theta_0)^{-1}H[H'I_1(\theta_0)^{-1}H]^{-1}H'I_1(\theta_0)^{-1}\\ &+ I_1(\theta_0)^{-1}H[H'I_1(\theta_0)^{-1}H]^{-1}H'I_1(\theta_0)^{-1}H[H'I_1(\theta_0)^{-1}H]^{-1}H'I_1(\theta_0)^{-1}\\ = P\end{aligned} \tag{82}$$

Hence,

$$2\log\mu = v'(I_1(\theta_0)^{-1}-P)v. \tag{83}$$

Since $I_1(\theta_0)$ is symmetric and nonsingular, it can always be written as $I_1(\theta_0) = AA'$, where A is a nonsingular matrix. Then, if z is a k-vector distributed as $N(0, I)$, we can write

$$v = Az$$

and $E(v) = 0$ and $\mathrm{cov}(v) = AA' = I_1(\theta_0)$ as required. Then

$$\begin{aligned} 2\log\mu &= z'A'(I_1(\theta_0)^{-1} - P)Az \\ &= z'A'I_1(\theta_0)^{-1}Az - z'A'PAz \\ &= z'A'(A')^{-1}A^{-1}Az - z'A'PAz\,. \\ &= z'z - z'A'PAz \\ &= z'(I - A'PA)z \end{aligned} \tag{84}$$

Now $(A'PA)^2 = A'PAA'PA = A'PI_1(\theta_0)PA$, but from Eq.(82), $P = PI_1(\theta_0)P$; hence $A'PA$ is idempotent, and its rank is clearly the rank of P. But since the k restricted estimates must satisfy p independent restrictions, the rank of P is $k-p$. Hence the rank of $I - A'PA$ is $k-(k-p) = p$. ∎

We next turn to the Wald Statistic. Expanding $h(\hat\theta)$ in Taylor Series about θ_0 gives asymptotically

$$h(\hat\theta) = h(\theta_0) + H'_\theta(\hat\theta - \theta_0)$$

and under the null hypothesis

$$h(\hat\theta) = H'_\theta(\hat\theta - \theta_0). \tag{85}$$

Since $\sqrt{n}(\hat\theta - \theta_0)$ is asymptotically distributed as $N(0, I_1(\theta_0)^{-1})$, $\sqrt{n}h(\hat\theta)$, which is asymptotically the same as $H'_\theta\sqrt{n}(\hat\theta - \theta_0)$, is asymptotically distributed as $N(0, H'_\theta I_1(\theta_0)^{-1}H_\theta)$. Then the Wald Statistic, $h(\hat\theta)'[\mathrm{cov}(h(\hat\theta))]^{-1}h(\hat\theta)$ becomes

$$W = nh(\hat\theta)'[H'_\theta I_1(\theta_0)^{-1}H_\theta]^{-1}h(\hat\theta). \tag{86}$$

Theorem 33. *Under $H_0 : h(\theta) = 0$, and if H_θ is of full rank r, W is asymptotically distributed as $\chi^2(p)$.*

Proof. Let z be distributed as $N(0, I)$ and let $I_1(\theta_0)^{-1} = AA'$, where A is nonsingular. Then AZ is distributed as $N(0, I_1(\theta_0)^{-1})$, which is the asymptotic distribution of $\sqrt{n}(\hat\theta - \theta_0)$. Thus, when $h(\theta) = 0$,

$$\sqrt{n}h(\hat\theta) = H'_\theta\sqrt{n}(\hat\theta - \theta_0)$$

is asymptotically distributed as $H'_\theta Az$. The Wald Statistic can be written as

$$W = z'A'H_\theta[H'_\theta I_1(\theta_0)^{-1}H_\theta]^{-1}H'_\theta Az, \tag{87}$$

which we obtain by substituting in Eq.(86) the asymptotic equivalent of $\sqrt{n}h(\hat\theta)$. But the matrix in Eq.(87) is idempotent of rank p, since

$$\begin{aligned} A'H_\theta[H'_\theta I_1(\theta_0)^{-1}H_\theta]^{-1}H'_\theta AA'H_\theta[H'_\theta I_1(\theta_0)^{-1}H_\theta]^{-1}H'_\theta A &= \\ A'H_\theta[H'_\theta I_1(\theta_0)^{-1}H_\theta]^{-1}H'_\theta A & \end{aligned}$$

where we have substituted $I_1(\theta_0)^{-1}$ for AA', $I_1(\theta_0)^{-1}$ is of rank k, H_θ is of full rank p, and A is nonsingular. ▮

We next turn to the Lagrange Multiplier test. If the null hypothesis that $h(\theta) = 0$ is true, then the gradient of the loglikelihood function is likely to be small, where the appropriate metric is the inverse covariance matrix for $D_\theta \log L(x, \theta)$. Hence the Lagrange Multiplier statistic is written generally as

$$LM = [D_\theta \log L(x, \tilde{\theta})]' \big[\operatorname{cov}(D_\theta \log L(x, \tilde{\theta}))\big]^{-1} [D_\theta \log L(x, \tilde{\theta})]. \tag{88}$$

Theorem 34. *Under the null hypothesis, LM is distributed as $\chi^2(p)$.*

Proof. Expanding $D_\theta \log L(x, \tilde{\theta})$ in Taylor Series, we have asymptotically

$$D_\theta \log L(x, \tilde{\theta}) = D_\theta \log L(x, \hat{\theta}) + D_\theta^2 \log L(x, \hat{\theta})(\tilde{\theta} - \hat{\theta}). \tag{89}$$

$D_\theta^2 \log L(x, \hat{\theta})$ converges in probability to $-nI_1(\theta_0)$, $D_\theta \log L(x, \hat{\theta})$ converges in probability to zero, and $D_\theta \log L(x, \tilde{\theta})$ converges in probability to $-nI_1(\theta_0)(\tilde{\theta} - \hat{\theta})$. But asymptotically $\hat{\theta} = \tilde{\theta}$ under the null; hence $n^{-1/2} D_\theta \log L(x, \tilde{\theta})$ is asymptotically distributed as $N(0, I_1(\theta_0))$. Hence the appropriate test is

$$LM = n^{-1} [D_\theta \log L(x, \tilde{\theta})] I_1(\theta_0)^{-1} [D_\theta \log L(x, \tilde{\theta})]$$

which by (88) is asymptotically

$$LM = n^{-1} \big[n(\tilde{\theta} - \hat{\theta})' I_1(\theta_0) I_1(\theta_0)^{-1} I_1(\theta_0)(\tilde{\theta} - \hat{\theta}) n\big] = n(\tilde{\theta} - \hat{\theta})' I_1(\theta_0)(\tilde{\theta} - \hat{\theta}). \tag{90}$$

But this is the same as Eq.(78), the likelihood ratio statistic, since the term $-D_\theta^2 \log L(x, \tilde{\theta})$ in Eq.(78) is $nI_1(\theta_0)$. Since the likelihood ratio statistic has asymptotic $\chi^2(p)$ distribution, so does the LM statistic. ▮

We now illustrate the relationship among W, LM, and LR and provide arguments for their asymptotic distributions in a slightly different way than before with a regression model $Y = X\beta + u$, with u distributed as $N(0, \sigma^2 I)$, and the restrictions $R\beta = r$.

In that case the three basic statistics are

$$\begin{aligned} W &= (R\hat{\beta} - r)' \big[R(X'X)^{-1}R'\big]^{-1} (R\hat{\beta} - r)/\hat{\sigma}^2 \\ LM &= (R\hat{\beta} - r)' \big[R(X'X)^{-1}R'\big]^{-1} (R\hat{\beta} - r)/\tilde{\sigma}^2 \\ LR &= -\frac{n}{2}(\log \hat{\sigma}^2 - \log \tilde{\sigma}^2) \end{aligned} \tag{91}$$

where W is immediate from Eq.(45) when $\hat{\Omega}$ is set equal to $\hat{\sigma}^2 I$, LM follows by substituting (40) in to (42) and setting $\tilde{\Omega} = \tilde{\sigma}^2$, and $LR = -2 \log \mu$ follows by substituting $\hat{\beta}$, respectively $\tilde{\beta}$ in the likelihood function and computing $-2\log\mu$. The likelihood ratio μ itself can be written as

$$\begin{aligned}\mu &= \left(\frac{\hat{u}'\hat{u}/n}{\tilde{u}'\tilde{u}/n}\right)^{n/2} \\ &= \left[\frac{1}{1+\frac{1}{n\hat{\sigma}^2}(R\hat{\beta}-r)'(R\hat{\beta}-r)}\right]^{n/2}\end{aligned} \tag{92}$$

where we have utilized Eq.(37) by dividing both sides by S_u and taking the reciprocal. We can also rewrite the F-statistic

$$\frac{(S_r - S_u)/p}{S_u/(n-k)}$$

as

$$F = \frac{(R\hat{\beta}-r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-r)/p}{S_u/(n-k)}. \tag{93}$$

Comparing (92) and (93) yields immediately

$$\mu = \left(\frac{1}{1+\frac{p}{n-k}F}\right)^{n/2} \tag{94}$$

and comparing W in (91) with (92) yields

$$\mu = \left(\frac{1}{1+W/n}\right)^{n/2}. \tag{95}$$

Equating (94) and (95) yields

$$\frac{W}{n} = \frac{p}{n-k}F$$

or

$$W = p\left(1+\frac{k}{n-k}\right)F. \tag{96}$$

Although the left-hand side is asymptotically distributed as $\chi^2(p)$ and F has the distribution of $F(p, n-k)$, the right hand side also has asymptotic distribution $\chi^2(p)$, since the quantity $pF(p, n-k)$ converges in distribution to that chi^2 distribution.

Comparing the definitions of LM and W in (91) yields

$$LM = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}W\right) \tag{97}$$

and from Eq.(37) we have

$$\tilde{\sigma}^2 = \hat{\sigma}^2(1+W/n). \tag{98}$$

Hence, from (97) and (98) we deduce

$$LM = \frac{W}{1+W/n}, \tag{99}$$

and using (96) we obtain

$$LM = \frac{p\left(\frac{n}{n-k}\right)F}{1+\frac{p}{n}\frac{n}{n-k}F} = \frac{npF}{n-k+pF}, \tag{100}$$

which converges in distribution as $n \to \infty$ to $\chi^2(p)$. From (95) we obtain

$$-2\log\mu = LR = n\log\left(1+\frac{W}{n}\right). \tag{101}$$

Since for positive z, $e^z > 1+z$, it follows that

$$\frac{LR}{n} = \log\left(1+\frac{W}{n}\right) < \frac{W}{n}$$

and hence $W > LR$.

We next note that for $z \geqq 0$, $\log(1+z) \geqq z/(1+z)$, since (a) at the origin the left and right hand sides are equal, and (b) at all other values of z the derivative of the left-hand side, $1/(1+z)$ is greater than the slope of the right-hand side, $1/(1+z)^2$. It follows that

$$\log\left(1+\frac{W}{n}\right) \geqq \frac{W/n}{1+W/n}.$$

Using (99) and (101), this shows that $LR \geqq LM$.

Recursive Residuals.

Since least squares residuals are correlated, even when the true errors u are not, it is inappropriate to use the least squares residuals for tests of the hypothesis that the true errors are uncorrelated. It may therefore be useful to be able to construct residuals that are uncorrelated when the true errors are. In order to develop the theory of uncorrelated residuals, we first prove a matrix theorem.

Theorem 35 (Barlett's). *If A is a nonsingular $n \times n$ matrix, if u and v are n-vectors, and if $B = A + uv'$, then*

$$B^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}.$$

Proof. To show this, we verify that pre- or postmultiplying the above by b yields an identity matrix. Thus, postmultiplying yields

$$\begin{aligned} B^{-1}B = I &= \left(A^{-1} - \frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}\right)(A+uv') \\ &= I - \frac{A^{-1}uv'}{1+v'A^{-1}u} + A^{-1}uv' - \frac{A^{-1}uv'A^{-1}uv'}{1+v'A^{-1}u} \\ &= I + \frac{-A^{-1}uv' + A^{-1}uv' + A^{-1}uv'(v'A^{-1}u) - A^{-1}u(v'A^{-1}u)v'}{1+v'A^{-1}u} \\ &= I \end{aligned} \tag{102}$$

We consider the standard regression model $Y = X\beta + u$, where u is distributed as $N(0, \sigma^2 I)$ and where X is $n \times k$ of rank k. Define X_j to represent the first j rows of the X-matrix, Y_j the first j rows of the Y-vector, x'_j the j^{th} row of X, and y_j the j^{th} element of Y. It follows from the definitions, for example, that

$$X_j = \begin{bmatrix} X_{j-1} \\ x'_j \end{bmatrix} \qquad \text{and} \qquad Y_j = \begin{bmatrix} Y_{j-1} \\ y_j \end{bmatrix}.$$

Define the regression coefficient estimate based on the first j observations as

$$\hat{\beta}_j = (X'_j X_j)^{-1} X'_j Y_j. \tag{103}$$

We then have the following

Theorem 36.

$$\hat{\beta}_j = \hat{\beta}_{j-1} + \frac{(X'_{j-1}X_{j-1})^{-1} x_j (y_j - x'_j \hat{\beta}_{j-1})}{1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j}.$$

Proof. By Theorem 35,

$$(X'_j X_j)^{-1} = (X'_{j-1}X_{j-1})^{-1} - \frac{(X'_{j-1}X_{j-1})^{-1} x_j x'_j (X'_{j-1}X_{j-1})^{-1}}{1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j}.$$

We also have by definition that

$$X_j Y_j = X_{j-1} Y_{j-1} + x_j y_j.$$

Substituting this in Eq.(103) gives

$$\begin{aligned}
\hat{\beta}_j &= \left[(X'_{j-1}X_{j-1})^{-1} - \frac{(X'_{j-1}X_{j-1})^{-1} x_j x'_j (X'_{j-1}X_{j-1})^{-1}}{1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j} \right] (X_{j-1} Y_{j-1} + x_j y_j) \\
&= \hat{\beta}_{j-1} + (X'_{j-1}X_{j-1})^{-1} x_j y_j \\
&\quad \frac{-(X'_{j-1}X_{j-1})^{-1} x_j x'_j [(X'_{j-1}X_{j-1})^{-1} X'_{j-1} Y_{j-1}] + (X'_{j-1}X_{j-1})^{-1} x_j x'_j (X'_{j-1}X_{j-1})^{-1} x_j y_j}{1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j} \\
&= \hat{\beta}_{j-1} + \frac{(X'_{j-1}X_{j-1})^{-1} x_j (y_j - x'_j \hat{\beta}_{j-1})}{1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j},
\end{aligned}$$

where, in the second line, we bring the second and third terms on a common denominator and also note that the bracketed expression in the numerator is $\hat{\beta}_{j-1}$ by definition. ∎

First define

$$d_j = \left[1 + x'_j (X'_{j-1}X_{j-1})^{-1} x_j \right]^{1/2} \tag{104}$$

and also define the recursive residuals $\tilde{u}_j$ as

$$\tilde{u} = \frac{y_j - x_j'\hat{\beta}_{j-1}}{d_j} \qquad j = k+1, \ldots, n \tag{105}$$

Hence, recursive residuals are defined only when the β can be estimated from at least k observations, since for j less than $k+1$, $(X_{j-1}'X_{j-1})^{-1}$ would not be nonsingular. Hence the vector $\tilde{u}$ can be written as

$$\tilde{u} = CY, \tag{106}$$

where

$$C = \begin{bmatrix} \frac{-x_{k+1}'(X_k'X_k)^{-1}X_k'}{d_{k+1}} & \frac{1}{d_{k+1}} & 0 & 0 & \ldots & 0 \\ \frac{-x_{k+2}'(X_{k+1}'X_{k+1})^{-1})X_{k+1}'}{d_{k+2}} & & \frac{1}{d_{k+2}} & 0 & \ldots & 0 \\ \vdots & & & & \ddots & \vdots \\ \frac{-x_n'(X_{n-1}'X_{n-1})^{-1})X_{n-1}'}{d_n} & & & & & \frac{1}{d_n} \end{bmatrix}. \tag{107}$$

Since the matrix X_j' has j columns, the fractions that appear in the first column of C are rows with increasingly more columns; hence the term denoted generally by $1/d_j$ occurs in columns of the C matrix further and further to the right. Thus, the element $1/d_{k+1}$ is in column $k+1$, $1/d_{k+2}$ in column $k+2$, and so on. It is also clear that C is an $(n-k) \times n$ matrix. We then have

, **Theorem 37.** *(1) $\tilde{u}$ is linear in Y; (2) $E(\tilde{u}) = 0$; (3) The covariance matrix of the $\tilde{u}$ is scalar, i.e., $CC' = I_{n-k}$; (4) For all linear, unbiased estimators with a scalar covariance matrix, $\sum_{i=k+1}^{n} \tilde{u}_i^2 = \sum_{i=1}^{n} \hat{u}_i^2$, where $\hat{u}$ is the vector of ordinary least squares residuals.*

Proof. (1) The linearity of $\tilde{u}$ in Y is obvious from Eq.(106).

(2) It is easy to show that $CX = 0$ by multiplying Eq.(107) by X on the right. Multiplying, for example, the $(p-k)^{\text{th}}$ row of C, $(p = k+1, \ldots, n)$, into X, we obtain

$$\frac{-x_p'(X_{p-1}'X_{p-1})^{-1}X_{p-1}'}{d_p}X_{p-1} + \frac{1}{d_p}x_p' = 0.$$

It then follows that $E(\tilde{u}) = E(CY) = E(C(X\beta + u)) = E(u) = 0.$∎

(3) Multiplying the $(p-k)^{\text{th}}$ row of C into the $(p-k)^{\text{th}}$ column of C', we obtain

$$\frac{1}{d_p^2} + \frac{x_p'(X_{p-1}'X_{p-1})^{-1}X_{p-1}'X_{p-1}(X_{p-1}'X_{p-1})^{-1}x_p}{d_p^2} = 1$$

by definition. Multiplying the $(p-k)^{\text{th}}$ row of C into the $(s-k)^{\text{th}}$ column of C', $(s > p)$, yields

$$\begin{bmatrix} \frac{-x_p'(X_{p-1}'X_{p-1})^{-1}X_{p-1}'}{d_p} & \frac{1}{d_p} & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} X_{p-1} \\ x_p' \\ x_{p+1}' \\ \vdots \end{bmatrix} \frac{(X_{s-1}'X_{s-1})^{-1}}{d_s}.$$

$$= \left[\frac{-x_p'}{d_pd_s} + \frac{x_p'}{d_pd_s}\right](X_{s-1}'X_{s-1})^{-1}x_s = 0$$

∎

(4) We first prove that $C'C = I - X(X'X)^{-1}X'$. Since $CX = 0$ by (2) of the theorem, so is $X'C'$. Define $M = I - X(X'X)^{-1}X'$; then

$$MC' = (I - X(X'X)^{-1}X')C' = C'. \tag{108}$$

Hence,

$$MC' - C' = (M - I)C' = 0. \tag{109}$$

But for any square matrix A and any eigenvalues λ of A, if $(A - \lambda I)w = 0$, then w is an eigenvector of A. Since M is idempotent, and by Theorem 5 the eigenvalues of M are all zero or 1, the columns of C' are the eigenvectors of M corresponding to the unit roots (which are $n - k$ in number, becase the trace of M is $n - k$).

Now let G' be the $n \times k$ matrix which contains the eigenvectors of M corresponding to the zero roots. Then, since M is symmetric, the matrix of all the eigenvectors of M is orthogonal and

$$[\,C' \quad G'\,]\begin{bmatrix} C \\ G \end{bmatrix} = I.$$

Let Λ denote the diagonal matrix of eigenvalues for some matrix A and let W be the matrix of its eigenvectors. Then $AW = W\Lambda$; applying this to the present case yields

$$M[\,C' \quad G'\,] = [\,C' \quad G'\,]\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = [\,C' \quad 0\,].$$

Hence

$$M = MI = M[\,C' \quad G'\,]\begin{bmatrix} C \\ G \end{bmatrix} = [\,C' \quad 0\,]\begin{bmatrix} C \\ G \end{bmatrix} = C'C.$$

But

$$\sum_{i=k+1}^{n} \tilde{u}^2 = Y'C'CY = Y'[I - X(X'X)^{-1}X']Y = \sum_{i=1}^{n} \hat{u}^2.$$

∎

Now define S_j by $S_j = (Y_j - X_j\hat{\beta}_j)'(Y_j - X_j\hat{\beta}_j)$; thus S_j is the sum of the squares of the least squares residuals based on the first j observations. We then have

Theorem 38. $S_j = S_{j-1} + \tilde{u}^2$.

Proof. We can write

$$\begin{aligned} S_j &= (Y_j - X_j\hat{\beta}_j)'(Y_j - X_j\hat{\beta}_j) = Y_j'(I - X_j(X_j'X_j)^{-1}X_j)Y_j \\ &= Y_j'Y_j - Y_j'X_j(X_j'X_j)^{-1}X_j'X_j(X_j'X_j)^{-1}X_j'Y_j \\ &\quad \text{(where we have multiplied by } X_j'X_j(X_j'X_j)^{-1}) \\ &= Y_j'Y_j - \hat{\beta}_j'X_j'X_j\hat{\beta}_j + 2\hat{\beta}_{j-1}(-X_j'Y_j + X_j'X_j\hat{\beta}_j) \\ &\quad \text{(where we replaced } (X_j'X_j)^{-1}X_j'Y_j \text{ by } \hat{\beta}_j \text{ and where the third} \\ &\quad \text{term has value equal to zero)} \\ &= Y_j'Y_j - \hat{\beta}_j'X_j'X_j\hat{\beta}_j + 2\hat{\beta}_{j-1}X_j'Y_j + 2\hat{\beta}_{j-1}X_j'X_j\hat{\beta}_j \\ &\quad + \hat{\beta}_{j-1}'X_j'X_j\hat{\beta}_{j-1} - \hat{\beta}_{j-1}'X_j'X_j\hat{\beta}_{j-1} \\ &\quad \text{(where we have added and subtracted the last term)} \\ &= (Y_j - X_j\hat{\beta}_{j-1})'(Y_j - X_j\hat{\beta}_{j-1}) - (\hat{\beta}_j - \hat{\beta}_{j-1})'X_j'X_j(\hat{\beta}_j - \hat{\beta}_{j-1}) \end{aligned} \quad . \tag{110}$$

Using the definition of X_j and Y_j and the definition of regression coefficient estimates, we can also write

$$\begin{aligned} X_j'X_j\hat{\beta}_j &= X_j'Y_j = X_{j-1}'Y_{j-1} + x_jy_j = X_{j-1}'X_{j-1}\hat{\beta}_{j-1} + x_jy + j \\ &= (X_j'X_j - x_jx_j')\hat{\beta}_{j-1} + x_jy_j \\ &= X_j'X_j\hat{\beta}_{j-1} + x_j(y_j - x_j'\hat{\beta}_{j-1}) \end{aligned} ,$$

and multiplying through by $(X_j'X_j)^{-1}$,

$$\hat{\beta}_j = \hat{\beta}_{j-1} + (X_j'X_j)^{-1}x_j(y_j - x_j'\hat{\beta}_{j-1}). \tag{111}$$

Substituting from Eq.(111) for $\hat{\beta}_j - \hat{\beta}_{j-1}$ in Eq.(110), we obtain

$$S_j = S_{j-1} + (y_j - x_j'\hat{\beta}_{j-1})^2 - x_j'(X_j'X_j)^{-1}x_j(y_j - x_j'\hat{\beta}_{j-1})^2. \tag{112}$$

Finally, we substitute for $(X_j'X_j)^{-1}$ in Eq.(112) from Bartlett's Identity (Theorem 35), yielding

$$\begin{aligned} S_j &= S_{j-1} + (y_j - x_j'\hat{\beta}_{j-1})^2 \times \\ &\times \left[\frac{1 + x_j'(X_{j-1}'X_{j-1})^{-1}x_j - x_j'(X_{j-1}'X_{j-1})^{-1}x_j - (x_j'(X_{j-1}'X_{j-1})^{-1}x_j)^2 + (x_j'(X_{j-1}'X_{j-1})^{-1}x_j)^2}{1 + x_j'(X_{j-1}'X_{j-1})^{-1}x_j} \right] \end{aligned}$$

from which the Theorem follows immediately, since $\tilde{u}$ is defined as

$$(y_j - x_j'\hat{\beta}_{j-1})/(1 + x_j'(X_{j-1}'X_{j-1})^{-1}x_j).$$

∎

We now briefly return to the case of testing the equality of regression coefficients in two regression in the case of insufficient degrees of freedom (i.e., the Chow Test). As in Case 4, on p. 13, the number of observations in the two data sets is n_1 and n_2 respectively. Denoting the sum of squares from the regression on the first n_1 observations by $\hat{u}_u'\hat{u}_u$ and the sum of squares using all n_1+n_2 observations by $\hat{u}_r'\hat{u}_r$, where the $\hat{u}$s are the ordinary (not recursive) least squares residuals, the test statistic can be written as

$$\frac{(\hat{u}_r'\hat{u}_r - \hat{u}_u'\hat{u}_u)/n_2}{\hat{u}_u'\hat{u}_u/(n_1-k)}.$$

By Theorem 37, this can be written as

$$\frac{(\sum_{i=k+1}^{n_1+n_2}\tilde{u}_i^2 - \sum_{i=k+1}^{n_1}\tilde{u}_i^2)/n_2}{\sum_{i=k+1}^{n_1}\tilde{u}_i^2/(n_1-k)} = \frac{\sum_{i=n_1+1}^{n_1+n_2}\tilde{u}_i^2/n_2}{\sum_{i=k+1}^{n_1}\tilde{u}_i^2/(n_1-k)}.$$

It may be noted that the numerator and denominator share no value of $\tilde{u}_i$; since the $\tilde{u}$s are independent, the numerator and denominator are independently distributed. Moreover, each $\tilde{u}_i$ has zero mean, is normally distributed and is independent of every other $\tilde{u}_j$, and has variance σ^2, since

$$\begin{aligned} E(\tilde{u}_i^2) &= E\left[\frac{(x_i'\beta - u_i - x_i'\hat{\beta}_{i-1})^2}{1+x_i'(X_{i-1}'X_{i-1})^{-1}x_i}\right] \\ &= \frac{x_i'E\big[(\beta-\hat{\beta}_{i-1})(\beta-\hat{\beta}_{i-1})'\big]x_i + E(u_i^2)}{1+x_i'(X_{i-1}'X_{i-1})^{-1}x_i} = \sigma^2. \end{aligned}$$

Hence, the ratio has an F distribution, as argued earlier.

Cusum of Squares Test. We consider a test of the hypothesis that a change in the true values of the regression coefficients occured at some observation in a series of observations. For this purpose we define

$$Q_i = \frac{\sum_{j=k+1}^{i}\tilde{u}_j^2}{\sum_{j=k+1}^{n}\tilde{u}_j^2}, \tag{113}$$

where $\tilde{u}_j$ represent the recursive residuals.

We now have

Theorem 39. *On the hypothesis that the values of the regression coefficients do not change, the random variable* $1-Q_i$ *has Beta distribution, and* $E(Q_i) = (i-k)/(n-k)$.

Proof. From Eq.(113), we can write

$$Q_i^{-1} - 1 = \frac{\sum_{j=i+1}^{n}\tilde{u}_j^2}{\sum_{j=k+1}^{i}\tilde{u}_j^2}. \tag{114}$$

Since the numerator and denominator of Eq.(114) are sums of iid normal variables with zero mean and constant variance, and since the numerator and denominator share no common $\tilde{u}_j$, the quantity

$$z = (Q_i^{-1} - 1)\frac{i-k}{n-k}$$

is distributed as $F(n-i, i-k)$. Consider the distribution of the random variable w, where W is defined by

$$(n-i)z/(i-k) = w/(1-w)$$

. Then the density of w is the Beta density

$$\frac{\Gamma(\alpha+\beta-1)}{\Gamma(\alpha)\Gamma(\beta)}\alpha^{\alpha}(1-w)^{\beta},$$

with $\alpha = -1 + (n-i)/2$ and $\beta = -1 + (i-k)/2$. It follows that

$$E(1-Q_i) = \frac{\alpha+1}{\alpha+\beta+2} = \frac{n-i}{n-k},$$

and

$$E(Q_i) = \frac{i-k}{n-k}. \tag{115}$$

Durbin (*Biometrika*, 1969, pp.1-15) provides tables for constructing confidence bands for Q_i of the form $E(Q_i) \pm c_0$.