Review of Classical Least Squares

James L. Powell
Department of Economics
University of California, Berkeley

The Classical Linear Model

The object of least squares regression methods is to model and estimate the relationship between a scalar dependent variable Y and a vector $\mathbf{x}$ of explanatory variables. In the classical model, the dependent and explanatory variables are treated differently, with Y assumed to be a random variable, with nondegenerate distribution that depends upon $\mathbf{x}$, which is viewed as nonrandom and under the control (in principle) of the researcher. A sample of N observations on Y, $\{Y_1, ..., Y_N\}$, with corresponding values $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ of the explanatory variables (termed "regressors"), can be written in the more compact matrix form

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ ... \\ Y_N \end{pmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ ... \\ \mathbf{x}_N' \end{bmatrix},$$

with $\mathbf{y}$ an N-dimensional vector and $\mathbf{X}$ an $(N \times K)$ matrix, where K is the number of regressors, i.e., the number of components of $\mathbf{x}_i$.

The classical linear regression model imposes strong assumptions on the nature of the relationship of $\mathbf{y}$ to $\mathbf{X}$; these assumptions, which are typically quite unrealistic in empirical economics, nonetheless provide an essential starting point for econometric practice. These conditions, referred henceforth as the "standard assumptions" are:

1. (Linear Expectation) $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, for some K-dimensional vector of unknown "regression coefficients" $\boldsymbol{\beta}$.

2. (Scalar Covariance Matrix) $\mathbf{V}(\mathbf{y}) \equiv \mathbf{E}\left((\mathbf{y} - \mathbf{E}(\mathbf{y}))(\mathbf{y} - \mathbf{E}(\mathbf{y}))'\right) = \sigma^2\mathbf{I}$, for some nonnegative "variance parameter" σ^2, with $\mathbf{I}$ being an $(N \times N)$ identity matrix.

3. (Nonstochastic Regressors) The $(N \times K)$ matrix $\mathbf{X}$ is nonrandom.

4. (Full Rank Regressors) The rank of the matrix $\mathbf{X}$ is K, or, equivalently, the $(K \times K)$ matrix $(\mathbf{X}'\mathbf{X})$ is invertible.

While these assumptions restrict only the first and second moments of the joint distribution of $\mathbf{y}$ (and the invertibility of $\mathbf{X}'\mathbf{X}$, which can be directly checked), those restrictions reduce the problem of determining the "relationship" between $\mathbf{y}$ and $\mathbf{X}$ to estimation of the vector of unknown coefficients $\boldsymbol{\beta}$, and determination of the strength of that relationship reduces to estimation of σ^2. Typically, the first component of $\mathbf{x}_i$ is assumed to be identically equal to one, so that the corresponding component $\boldsymbol{\beta}_1$ of $\boldsymbol{\beta}$ is interpreted as an intercept term in the linear relation of the mean of the dependent variable in terms of the regressors.

Often the standard assumptions are stated, not in terms of the vector of dependent variables $\mathbf{y}$, but in terms of a vector of "error terms" $\boldsymbol{\varepsilon}$, defined as

$$\boldsymbol{\varepsilon} \equiv \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

In this notation, the standard assumptions are

1. (Linear Expectation) $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. (Scalar Covariance Matrix) $\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.
3. (Nonstochastic Regressors) The $(N \times K)$ matrix $\mathbf{X}$ is nonrandom.
4. (Full Rank Regressors) The rank of the matrix $\mathbf{X}$ is K, or, equivalently, the $(K \times K)$ matrix $(\mathbf{X}'\mathbf{X})$ is invertible.

We will treat these assumptions as interchangeable; sometimes it is convenient to investigate their plausibility for the observable variable $\mathbf{y}$, and other times for the unobservable error vector $\boldsymbol{\varepsilon}$, which represents the effect of "left out regressors" in the determination of $\mathbf{y}$.

Classical Least Squares

The classical least squares (LS) estimator $\hat{\boldsymbol{\beta}}$ of the unknown parameter $\boldsymbol{\beta}$ is defined in terms of a minimization problem:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\equiv \arg\min_{\mathbf{c}\in\mathbf{R}^{\mathbf{K}}} (\mathbf{y} - \mathbf{X}\mathbf{c})'(\mathbf{y} - \mathbf{X}\mathbf{c}) \\ &= \arg\min_{\mathbf{c}\in\mathbf{R}^{\mathbf{K}}} \sum_{i=1}^{N} (Y_i - \mathbf{x}_i'\mathbf{c})^2, \end{aligned} \tag{1}$$

which, under standard assumption 4, has the algebraic form

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i Y_i\right).\end{aligned}$$

In the latter representation, the estimator $\hat{\boldsymbol{\beta}}$ is evidently a function of the sample second moments of the dependent and explanatory variables, and in either form the estimator is clearly a linear function of $\mathbf{y}$, with coefficients a function of the nonrandom matrix of regressors, $\mathbf{X}$. The estimator $\hat{\boldsymbol{\beta}}$ is the solution to the "normal equations"

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

Accompanying the LS estimator of $\boldsymbol{\beta}$ is an estimator of the unknown variance parameter $\boldsymbol{\sigma}^2$,

$$s^2 \equiv \frac{1}{N-K}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right),$$

which is a quadratic form in the vector of residuals

$$\begin{aligned}\hat{\mathbf{e}} &\equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\ &\equiv \mathbf{N} \\ &\quad \mathbf{M}_{\mathbf{x}}\mathbf{y}.\end{aligned}$$

The unusual normalization (dividing by $N-K$ rather than N) yields an estimator that is mean-unbiased under the standard assumptions.

If one of the columns of $\mathbf{X}$ (usually the first) is a column vector "ι" identically equal to one, so that the corresponding component of $\boldsymbol{\beta}$ is interpreted as an intercept term, a summary measure of "goodness of fit" of the fitted values

$$\begin{aligned}\hat{\mathbf{y}} &\equiv \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ &\equiv \mathbf{P}_{\mathbf{x}}\mathbf{y}\end{aligned}$$

to the original dependent variable $\mathbf{y}$ is the squared multiple correlation coefficient,

$$\begin{aligned} R^2 &\equiv 1 - \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{(\mathbf{y} - \bar{y}\boldsymbol{\iota})' (\mathbf{y} - \bar{y}\boldsymbol{\iota})} \\ &= 1 - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{(\mathbf{y} - \bar{y}\boldsymbol{\iota})' (\mathbf{y} - \bar{y}\boldsymbol{\iota})}, \end{aligned}$$

where $\bar{y}$ is the (scalar) sample average of the dependent variable,

$$\bar{y} \equiv \frac{1}{N} \sum_i Y_i,$$

and $\boldsymbol{\iota}$ is the N-dimensional vector of ones, i.e., $\boldsymbol{\iota} \equiv (\mathbf{1}, \mathbf{1}, ..., \mathbf{1})'$. The last term in the definition of R^2 never exceeds one, since the denominator can be viewed as a constrained minimizer of the least squares criterion in (1), subject to the constraint that all components of $\mathbf{c}$ equal zero except for the coefficient on the column of $\mathbf{X}$ that equals $\boldsymbol{\iota}$. (When no column of $\mathbf{X}$ equals $\boldsymbol{\iota}$, it is customary instead to use the "no-constant-adjusted" R^2 measure, which substitutes 0 for $\bar{y}$ in the usual formula for R^2.)

Regression Algebra

Often it is useful to know the algebraic relations between the LS regression coefficients using the entire $\mathbf{X}$ matrix and those which use only an $N \times K_1$ submatrix $\mathbf{X}_1$ of the original matrix of regressors – that is, the relation between the unconstrained LS estimator and the LS estimator which constrains the coefficients on the remaining submatrix $\mathbf{X}_2$ of $\mathbf{X}$ to be zero. Partitioning the matrix of regressors as

$$\mathbf{X} \equiv [\mathbf{X}_1, \mathbf{X}_2],$$

with a compatible partitioning of the unconstrained LS estimator

$$\hat{\boldsymbol{\beta}} \equiv \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix},$$

we can derive a relationship between the subvector $\hat{\boldsymbol{\beta}}_1$ of the "long regression" coefficients using all of $\mathbf{X}$ to the "short regression" coefficients

$$\hat{\boldsymbol{\beta}}_1^* \equiv (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

which use only the submatrix $\mathbf{X}_1$ of regressors. That relationship is

$$\hat{\boldsymbol{\beta}}_1^* = \hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2,$$

i.e., the difference $\hat{\boldsymbol{\beta}}_1^* - \hat{\boldsymbol{\beta}}_1$ between the short and long regression coefficients is the product of the matrix of regression coefficients of the omitted regressors $\mathbf{X}_2$ on the included regressors $\mathbf{X}_1$ and the subvector $\hat{\boldsymbol{\beta}}_2$ of the long regression coefficients for the omitted regressors. (Say that three times, quickly!)

Another useful algebraic relationship is the "residual regression" representation of a subvector, say $\hat{\boldsymbol{\beta}}_1$, of the long regression coefficients $\hat{\boldsymbol{\beta}}$. Defining the (idempotent) projection matrix

$$\mathbf{P}_2 \equiv \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',$$

which projects vectors into the linear subspace spanned by the columns of $\mathbf{X}_2$, the long regression coefficients can be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_2)\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_2)\mathbf{X}_1 \\
&= (\mathbf{X}_1^{*\prime}\mathbf{X}_1^*)^{-1}\mathbf{X}_1^{*\prime}\mathbf{y} \\
&= (\mathbf{X}_1^{*\prime}\mathbf{X}_1^*)^{-1}\mathbf{X}_1^{*\prime}\mathbf{y}^*,
\end{aligned}$$

where

$$\mathbf{X}_1^* \equiv (\mathbf{I}-\mathbf{P}_2)\mathbf{X}_1$$

and

$$\mathbf{y}^* \equiv (\mathbf{I}-\mathbf{P}_2)\mathbf{y}$$

are the residuals of the regression of $\mathbf{X}_1$ and $\mathbf{y}$ on $\mathbf{X}_2$. In (many) words, the long regression coefficients can be obtained by first getting the residuals of a regression of $\mathbf{X}_1$ and $\mathbf{y}$ on $\mathbf{X}_2$, and then regressing the residuals for $\mathbf{y}$ on the residuals for $\mathbf{X}_2$. When $\mathbf{X}_2$ is a column vector of ones, i.e., $\mathbf{X}_2 = \boldsymbol{\iota}$, application of the residual regression formula yields an alternative representation for the squared multiple correlation coefficient,

$$R^2 \equiv \frac{(\hat{\mathbf{y}}-\bar{y}\boldsymbol{\iota})'(\hat{\mathbf{y}}-\bar{y}\boldsymbol{\iota})}{(\mathbf{y}-\bar{y}\boldsymbol{\iota})'(\mathbf{y}-\bar{y}\boldsymbol{\iota})},$$

so that R^2 measures the variation (squared length) of the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ around their mean values relative to the corresponding variation in the vector of dependent variables $\mathbf{y}$.

Moments of the LS Estimator

The rules for calculation of the mean vector and variance covariance matrix of a linear function $\mathbf{Ay}$ of

a random vector $\mathbf{y}$ (with $\mathbf{A}$ nonrandom) are:

$$\begin{aligned} E[\mathbf{Ay}] &= \mathbf{A}E[\mathbf{y}], \\ \mathbf{V}[\mathbf{Ay}] &= \mathbf{A}\mathbf{V}[\mathbf{y}]\mathbf{A}'. \end{aligned}$$

Applying these rules to the LS estimator $\hat{\boldsymbol{\beta}}$ (with $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$) yields, under the standard assumptions,

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

so the LS estimator is mean-unbiased, and

$$\begin{aligned} \mathbf{V}[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}[\mathbf{y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}[\sigma^2\mathbf{I}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Similar, but more complicated, calculations (involving interchange of the trace and expectations operators) can be used to show that s^2 is a mean-unbiased estimator of σ^2,

$$E[s^2] = \sigma^2,$$

but calculation of its variance would require more restrictions than imposed in the standard assumptions.

The renowned efficiency result for the classical least squares estimator, known as the *Gauss-Markov Theorem*, states that, under the standard conditions, the LS estimator $\hat{\boldsymbol{\beta}}$ defined above is the "best linear unbiased estimator", or BLUE, where "best" is defined in terms of smallest variance-covariance matrix. More precisely, if $\tilde{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$ that is linear in $\mathbf{y}$, i.e.,

$$\tilde{\boldsymbol{\beta}} = \mathbf{Ay}$$

for some $K \times N$ nonrandom matrix $\mathbf{A}$, and if $\tilde{\boldsymbol{\beta}}$ is mean unbiased, meaning

$$E[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

for all possible $\boldsymbol{\beta} \in R^K$, then under the standard assumptions the covariance matrix of $\tilde{\boldsymbol{\beta}}$ is at least as large as that for $\hat{\boldsymbol{\beta}}$, in the sense that $\mathbf{V}[\tilde{\boldsymbol{\beta}}] - \mathbf{V}[\hat{\boldsymbol{\beta}}]$ is positive semi-definite. This result is obtained by decomposing

the alternative estimator $\tilde{\boldsymbol{\beta}}$ as the sum of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$, and showing that these two components have zero covariance (using the unbiasedness restriction $\mathbf{AX} = \mathbf{I}$), so that

$$\mathbf{V}[\tilde{\boldsymbol{\beta}}] = \mathbf{V}[\hat{\boldsymbol{\beta}}] + \mathbf{V}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}],$$

from which the result immediately follows. As a consequence, the best linear unbiased estimator of any linear combination $\theta \equiv \mathbf{a}'\boldsymbol{\beta}$ of the unknown regression coefficients (with $\mathbf{a}$ a fixed, nonrandom K-dimensional vector) is $\hat{\theta} = \mathbf{a}'\hat{\boldsymbol{\beta}}$, since its variance exceeds that of the alternative linear unbiased estimator $\tilde{\theta} = \mathbf{a}'\tilde{\boldsymbol{\beta}}$ by the quantity $\mathbf{a}'\mathbf{V}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}]\mathbf{a}$, which is nonnegative.

The mean-variance rules can also be used to calculate the expectation and variance of the "short regression coefficients" $\hat{\boldsymbol{\beta}}_1^*$ defined above:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}_1^*] &= E[\hat{\boldsymbol{\beta}}_1] + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 E[\hat{\boldsymbol{\beta}}_2] \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2, \end{aligned}$$

which equals the true value $\boldsymbol{\beta}_1$ only if the regression coefficients of $\mathbf{X}_2\beta_2$ on $\mathbf{X}_1$ are all zero. (This result is known as the "omitted variable bias" formula.) Since, using the variance calculation rules,

$$\mathbf{V}[\hat{\boldsymbol{\beta}}_1^*] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$$

and

$$\mathbf{V}[\hat{\boldsymbol{\beta}}_1] = \sigma^2(\mathbf{X}_1^{*\prime}\mathbf{X}_1^*)^{-1},$$

it follows that the variance-covariance matrix for the short regression coefficients $\hat{\boldsymbol{\beta}}_1^*$ is no larger (in a positive definite sense) than that for $\hat{\boldsymbol{\beta}}_1$, since

$$\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1^{*\prime}\mathbf{X}_1^* = \mathbf{X}_1'\mathbf{P}_2\mathbf{X}_1$$

is nonnegative definite.

The Classical Normal Regression Model

To obtain the finite-sample distributions of the LS estimator $\hat{\beta}$ and the variance estimator s^2, stronger assumptions are needed than those imposed by the standard assumptions, which only restrict the first and second moments of the random vector $\mathbf{y}$. A very convenient distribution for $\mathbf{y}$ is the multivariate normal distribution. To obtain the *normal linear regression model*, we append to standard assumptions 1 through 4 the additional assumption

5. (Normality) The vector $\mathbf{y}$ (or, equivalently, $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$) has a multivariate normal distribution.

Under assumptions 1 through 5, the joint distribution of the vector $\mathbf{y}$ is

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

with

$$\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \sim N(0, \sigma^2\mathbf{I}).$$

Since $\hat{\boldsymbol{\beta}}$ is a linear function of the vector $\mathbf{y}$, and since linear functions of multinormals are themselves multinormal, it follows that $\hat{\beta}$ is itself multinormal under the normal linear model,

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right),$$

and thus that a linear function $\hat{\boldsymbol{\theta}} = \mathbf{R}\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$ (with $\mathbf{R}$ a nonrandom $r \times K$ matrix with full row rank K) is also multivariate normal

$$\hat{\boldsymbol{\theta}} \equiv \mathbf{R}\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\theta}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right),$$

with $\boldsymbol{\theta} \equiv \mathbf{R}\boldsymbol{\beta}$. A more delicate derivation, which uses the fact that s^2 is proportional to a quadratic form in $\mathbf{y}$, yields the result that s^2 is proportional to a chi-squared random variable with $N - K$ degrees of freedom,

$$\frac{(N-K)\,s^2}{\sigma^2} \sim \chi^2_{N-K};$$

furthermore, it can be shown that s^2 and $\hat{\boldsymbol{\beta}}$ are statistically independent under assumptions 1. through 5.

These distributional results provide the foundation for statistical inference regarding the unknown regression coefficient vector $\boldsymbol{\beta}$ (as well as the variance parameter σ^2). For example, if $\mathbf{R}$ is a row vector (i.e., a $1 \times K$ matrix), then it follows that a standardized version of $\hat{\theta} = \mathbf{R}\hat{\boldsymbol{\beta}}$, replacing the unknown σ^2 with the estimator σ^2, will have a Student's t distribution with $N - K$ degrees of freedom,

$$\frac{\hat{\theta} - \theta}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}} \sim t_{N-K},$$

which can be used to construct confidence intervals for the unknown $\theta = \mathbf{R}\boldsymbol{\beta}$, or to test null hypotheses like $H_0 : \mathbf{R}\boldsymbol{\beta} \leq \theta_0$ (against a one-sided alternative) or $H_0' : \mathbf{R}\boldsymbol{\beta} = \theta_0$ (with a two-sided alternative), in the usual way. And, if $\mathbf{R}$ has more than one row, then a quadratic form in the estimated vector $\hat{\boldsymbol{\theta}} = \mathbf{R}\hat{\boldsymbol{\beta}}$ around the

inverse of its estimated covariance matrix will have Snedecor's F distribution with r and $N-K$ degrees of freedom,

$$\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)'\left[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)/r \sim F_{r,N-K}.$$

This result can be used to test the linear hypothesis $H_0 : \mathbf{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0$ by replacing the unknown $\boldsymbol{\theta}$ with its hypothesized value $\boldsymbol{\theta}_0$ in this formulae and comparing the result to a critical value from an F table. Furthermore, the set of possible values of θ_0 for which the F-test fails to reject the null hypothesis $H_0 : \mathbf{R}\boldsymbol{\beta} = \boldsymbol{\theta}_0$ at size α forms a (random) $1-\alpha$ confidence region for the unknown value of θ.

For the special case that the matrix of regressors $\mathbf{X}$ can be partitioned as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, with $\mathbf{X}_1$ being a column vector of ones ($\mathbf{X}_1 = \boldsymbol{\iota}$) and with the corresponding partition of $\boldsymbol{\beta}$, the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ can be tested using the R^2 statistic. Specifically, under assumptions 1. through 5. and the null hypothesis,

$$\frac{(N-K)}{(K-1)}\frac{R^2}{1-R^2} \sim F_{K-1,N-K},$$

so under the normal linear regression model the squared multiple correlation coefficient R^2 is monotonically related to an F-statistic for testing the null hypothesis that all slope (i.e., non-intercept) regression coefficients are zero. Variations of this approach can be used to test the null hypothesis that a particular subvector of $\boldsymbol{\beta}$ is equal to zero, by constructing a test statistic using the difference in R^2 statistics when the null hypothesis is or is not imposed.

Departures from the Standard Assumptions

Given the strong assumptions of the normal linear regression model, the next part of a traditional econometrics course (affectionately known as "Part II", although it is in Parts III and IV in Ruud's text) is investigation of the consequences of relaxing those assumptions, and appropriate adjustment of the statistical procedures should they fail to hold. Working in reverse order, here is an outline of the issues that arise when each assumption fails, along with the jargon that accompanies each problem.

5. (*Nonnormality*) If $\mathbf{y}$ is not multinormally distributed, then the exact distributional results for the LS estimators (normality of $\hat{\boldsymbol{\beta}}$ and a chi-squared distribution for s^2) no longer apply. Fortunately, *asymptotic theory* can be applied to show that $\hat{\boldsymbol{\beta}}$ is approximately normally distributed, and the approximation error shrinks to zero as the sample size increases. Asymptotic normal theory combines two different types of approximations. First, there are classical *limit theorems*, which give general conditions under which the distributions of weighted sums of random variables are approximately multinormal, with shrinking

variances of the approximating distributions; second, there are *Slutsky theorems*, which show how smooth functions of sample averages (like $\hat{\boldsymbol{\beta}}$ and s^2) are approximately weighted sums of random variables, to which the limit theorems can be applied. If the standard assumptions are strengthened to make the limit and Slutsky theorems applicable, the result is that the LS estimator is approximately normal,

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right),$$

where "$\overset{A}{\sim}$" means "is approximately distributed as," in a sense to be made more precise later. Results like these imply that the inference procedures developed for the normal linear regression model can be approximately valid when the dependent variable is not assumed to be normally distributed.

4. (*Multicollinearity*) If the $\mathbf{X}$ matrix is not of full column rank, then $\mathbf{X}'\mathbf{X}$ is noninvertible and the true parameter vector $\boldsymbol{\beta}$ is not identified from the observed data, though some linear combinations (e.g., the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ into the space spanned by the columns of $\mathbf{X}$) can be uniquely determined. If the goal is to obtain point estimates of $\boldsymbol{\beta}$ for some policy exercise (e.g., manipulation of one of the regressors), then such perfect multicollinearity is a killer. Enough said.

3. (*Stochastic Regressors*) If $\mathbf{X}$ is random, but the standard assumptions (and the extra normality assumption) hold conditional on $\mathbf{X}$ (that is, $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{V}(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}$, etc.), then little change in the inference procedures is required; LS is still the BLU estimator *conditional on the observed* $\mathbf{X}$, and the normal and chi-squared distributions of $\hat{\boldsymbol{\beta}}$ and s^2 hold conditionally on $\mathbf{X}$. In fact, since the conditional distribution of s^2 does not depend on $\mathbf{X}$, it is proportional to a chi-squared random variable regardless of whether $\mathbf{X}$ is fixed or random. And since the F- and t-statistics also have distributions under the null that do not depend upon $\mathbf{X}$, they too have the same null distributions whether $\mathbf{X}$ is viewed as stochastic or not.

A more complex setting for random regressors is a *dynamic model,* in which some lagged values of $\mathbf{y}$ are used as regressors for time series data. Here the standard assumptions cannot hold conditional on $\mathbf{X}$, because of the overlap between $\mathbf{y}$ and some components of $\mathbf{X}$. The classical LS estimator $\hat{\boldsymbol{\beta}}$ is necessarily a nonlinear function of $\mathbf{y}$ in this context. Nevertheless, asymptotic theory can usually be used (with appropriate limit theorems for dependent data) to show that the LS estimator is still approximately normal, since it is a smooth function of (approximately normal) weighted sums of the components of $\mathbf{y}$.

2. (*Nonscalar Covariance Matrix*) When the covariance matrix of $\mathbf{y}$ (or $\boldsymbol{\varepsilon}$) is not proportional to an identity matrix – $\mathbf{V}(\mathbf{y}) \equiv \boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}$ for any σ^2 – then the classical LS estimator, while linear and unbiased,

is no longer "best" in that class. If $\mathbf{\Sigma}$ is known up to a constant of proportionality – $\mathbf{\Sigma} = \boldsymbol{\sigma}^2\mathbf{\Omega}$, with $\mathbf{\Omega}$ known – then the original $\mathbf{y}$ and $\mathbf{X}$ data can be transformed (by premultiplying by a matrix square root of the inverse of $\mathbf{\Omega}$) to yield a BLU estimator of $\boldsymbol{\beta}$ by applying LS to the transformed data. This approach yields *Aitken's Generalized Least Squares* (GLS) estimator,

$$\hat{\boldsymbol{\beta}}_{GLS} \equiv \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y},$$

which is BLU for a given (nonsingular) $\mathbf{\Omega}$, and includes classical LS as a special case (when $\mathbf{\Omega} = \mathbf{I}$). If $\mathbf{y}$ is multinormal and $\mathbf{\Omega}$ is known (not estimated), then the multinormality of the GLS estimator follows in the same way as for LS. If $\mathbf{\Omega}$ involves unknown parameters, which must be estimated using the dependent variable $\mathbf{y}$, then the "feasible" version of GLS which uses the estimated $\hat{\mathbf{\Omega}}$ in place of $\mathbf{\Omega}$ will be a nonlinear function of $\mathbf{y}$, and the distribution of the GLS estimator will not be exactly normal. Again, asymptotic theory can be used to show that the feasible GLS estimator has approximately the same normal distribution as its exact counterpart, provided $\hat{\mathbf{\Omega}}$ is a suitably close approximation for $\mathbf{\Omega}$ as the sample size increases.

Depending upon the particular application, $\mathbf{\Omega}$ can depart from an identity matrix in a number of different ways; each sort of departure has its own nomenclature. *Heteroskedastic* models have $\mathbf{\Omega}$ being a diagonal matrix with non-constant diagonal elements, so that the different components of $\mathbf{y}$ have different variances (but are mutually uncorrelated). Models with *serial correlation* have $\mathbf{\Omega}$ being a band-diagonal matrix, with some nonzero components off the main diagonal. Models which have both nonconstant variances and nonzero covariances among the components of $\mathbf{y}$ include Zellner's *seemingly unrelated regressions* model and *panel data* (i.e., pooled cross-section and time series) models.

1. (*Endogenous Regressors*) Failure of the assumption that the expectation of $\mathbf{y}$ (given $\mathbf{X}$) is not a linear combination of $\mathbf{X}$ is the most serious complication for the classical LS procedure. While $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ may fail because the true conditional mean is nonlinear in the regressors, a typical problem in empirical economics is nonzero correlation between the error terms $\boldsymbol{\varepsilon} \equiv \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ and some columns of $\mathbf{X}$, termed *endogenous regressors*. Such endogeneity can arise for a number of reasons, including measurement error in the observed regressors, simultaneity, sample selectivity, omitted regressors, and other empirical problems (to be defined and described in more detail later). The standard econometric approach to estimation of $\boldsymbol{\beta}$ with endogenous regressors involves collection of data on additional variables, termed *instrumental variables* and often denoted by an $L \times K$ matrix $\mathbf{Z}$ (with number of rows L at least as large as the corresponding number of rows K in $\mathbf{X}$), which are posited to be uncorrelated with the error vector $\boldsymbol{\varepsilon}$ but correlated (in an

appropriate sense) with the columns of the matrix $\mathbf{X}$. Variants of such instrumental variables estimation methods, including *two-stage least squares* and *generalized method of moment* estimators, are arguably the most original contribution of econometrics to statistical methodology, and will be covered in detail henceforth.