Syllabus

M.Sc (Applied Statistics) Semester III

STAS3 – III: Elective I (B) – Econometric Models (EM)

Unit–I

Meaning and scope of econometrics. Concepts of dummy variables and proxy variable.

Problems and methods of estimation in single equation regression Models
Multicollinearity: Consequences of multicollinearity, tests to detect its presence and solutions to the problem of multicollinearity.
Generalised Least Squares: Estimates of regression parameters – Properties of these estimates.

Unit–II

Heteroscedasticity: Consequences of heteroscedastic disturbances – test to detect its presence and solutions to the problem of heteroscedasticity.

Auto Correlation: Consequences of autocorrelated disturbances, Durbin – Watson test – Estimation of autocorrelation coefficient (for a first order autoregressive scheme).

Unit–III

Distributed lag models: study of simple finite lag distribution models – Estimation of the coefficients of Kayak geometric lag model.

Instrumental Variable: Definition – derivation of instrument variable estimates and their properties.

Unit–IV

Errors in variables: Problem of errors in variables simple solutions using instrumental variables technique.

Simulation equation models and methods of estimation: distinction between structure and Model–Exogenous and Endogenous variables – Reduced form of a model.

Problem of identification – Rank and order conditions and their application.
Methods of estimation: Indirect least squares. Two stages least squares, three stages least squares. A study of merits and demerits of these methods.

References:

1) Johnston – Econometrics Methods (2nd Edition) :
Chapter 1, Chapter 7: Section 7-1,7-3, Chapter 9 : Section 9-3, 9-4, Chapter 12 : Section 12-2,12-3, Chapter 13, Section 13-2,13-6

Meaning of Econometrics ;

Dummy Variables : the variables

Proxy Variables : A proxy variable is a variable that is used to measure an unobservable quantity of interest.

For Example a researcher wishes to find the living standards of the respondent , but the respondent is not willing , then one can use proxy variable.

Viz., affordability index --- ; Income --- ;

Problems and methods of estimation in single equation regression Models ,

Let X be the independent variable and Y be the dependent variables

The conditional distribution be $f_{y/x}$

$\mu_{y/x}$ be the Conditional mean , is denoted by μ_x

Similarly $v_{Y/x}$ be the conditional variance is denoted by $v_{\dot{x}}$

.. μ_x is the average value of y for given values of x.

$$\mu_x = g(X, \theta) \tag{10.1}$$

where θ is a vector of parameters.

A particularly simple form for $g(X, \theta)$ is the linear form

$$g(X, \theta) = \alpha + \beta X \tag{10.2}$$

which, combined with Equation (10.1) leads to

$$\mu_x = \alpha + \beta X \tag{10.3}$$

where $\theta = (\alpha, \beta)$ is the vector of unknown parameters to be estimated from the data.

Let Y_x denote the earnings of a worker with level of schooling X. Then Y_x may not necessary equal μ_x, but may be expressed as:

$$Y_x = \mu_x + \varepsilon_x \tag{10.4}$$

where ε_x may be referred to as a disturbance or an error.

Since $E(Y_x) = \mu_x$ by definition, from Equation (10.4)

$$E(\varepsilon_x) = 0 \tag{10.5}$$

Combining Equations (10.3) and (10.4), we may write our model as:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (i = 1, \cdots, n) \tag{10.6}$$

where Y is referred to as the dependent variable and X as the independent variable or regressor. The model (10.6) is called the two-variable linear regression model.

Model (10.6) can be generalized in at least two ways.

First, instead of making Y a function of a single variable X, we could have k dependent variables $X_1 \cdots X_k$ and write the counterpart of Model (10.6) (dropping the sample suffixes), as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta k X_k + \varepsilon \tag{10.7}$$

where X_1 (the first regressor) is a constant term. Model (10.7) is the so-called k-variable linear multiple regression model.

Second, instead of letting μ_x be a linear function of X and θ as in Equation (10.3), we could make it non-linear in either X, or θ, or both. This gives us the two-variable non-linear regression model.

$$Y = g(X, \theta) + \varepsilon \tag{10.8}$$

If $g(X, \theta)$ is non-linear only in the variable X, we may rewrite Model (10.8) as, say,

$$Y = \alpha + \beta X + \gamma X^2 + \delta X^3 + \varepsilon \tag{10.9}$$

(where we have assumed g to be a cubic in X). We now define $Z = X_2$ and $W = X^3$, and Equation (10.9) becomes:

$$Y = \alpha + \beta X + \gamma Z + \delta W + \varepsilon \tag{10.10}$$

which is of the form (10.7). Thus, non-linearity in variables is not much of a complication. We can still retain the linear structure by adding suitably defined new variables.

Non-linearity in parameters is, however, a different matter altogether. Thus, when we refer to non-linear models, further in the text, we essentially mean non-linearity in parameters (since the case of non-linear variables can be typically reduced to the linear case).

A non-linear model in several variables is simply

$$Y = g(X_1, X_2, \cdots, X_k, \theta) + \varepsilon \tag{10.11}$$

Definition 10.1 (Population regression function): The Model (10.11) is in general referred to as a population regression function (PRF). Note that in (Model 10.11), the dimensions of the parameter vector θ need not be equal to k (the number of regressors).

Model (10.11) admits an additional generalization. Suppose, instead of making the conditional mean μ_x dependent on X in Equation (10.1), we had made the conditional variance V_x dependent on X, that is,

$$V_x = g(X, \theta) \tag{10.12}$$

We would then obtain the so-called scedastic regression model, which is the basis of the ARCH models (considered in a later chapter). Similarly, by making the conditional

Two Variable Linear Model

Assumptions

Let us now take a closer look at Model (10.6), in particular, the error term ε_i. From Equation (10.3), we note that α incorporates the effects of all factors, apart from X, which are likely to affect μ_x, the conditional mean of the dependent variable Y. The term ε_i includes the following factors:

(i) non-linear terms in X, which may have been omitted as a simplification;
(ii) the influences of factors, other than X (say W, Z etc.), on μ_x;
(iii) the influence of all factors affecting the conditional variance and other higher order conditional moments of Y (the influence on the conditional mean having been accounted for already in Equation (10.3);
(iv) measurement errors in X and Y; and
(v) intrinsic randomness in economic data, which arises because human and organizational behaviour can, by its very nature, never be captured fully in any mathematical model, however comprehensive.

Thus, effectively, the error term ε_i includes a large number of factors which are likely to be independent (or nearly so). Hence the ε_i term is likely to be (approximately) normally distributed (by appeal to the Central Limit Theorem of Chapter 7).

Standard regression analysis commences by making the following ideal assumptions.

(a.1) $E(\varepsilon_i) = 0 \ \forall \ i$ (no specification error);
(a.2) $E(\varepsilon_i \varepsilon_j) = 0 \ i \neq j$ (absence of autocorrelation);
(a.3) Var $(\varepsilon_i) = \sigma^2$ $(\forall i)$ (homoscedasticity);
(a.4) ε_i is normally distributed for each i.

Actually all the assumptions above can be combined into one single assumption

$$\varepsilon_i \text{ are } iid \quad N(0, \sigma^2) \tag{10.13}$$

However, in the interest of conceptual clarity, it is best to retain all the assumption (a.1) to (a.4).

Our final assumption is:

(a.5) Either the X_i $(i = 1 \cdots n)$ are fixed in repeated samples or if the X_i are stochastic, then Cov $(X_i, \varepsilon_i) = 0$, $i = 1 \cdots n$.

The significance of these assumptions will emerge when they are discussed in detail in Chapter 11.

Note: The independent variable X in the illustration referred to years of schooling, which could typically take the values 0, 1, 2, ..., 12. These values would remain the same, even though the dependent variable Y (that is, earnings) could vary across different samples. Here X is non-stochastic and Assumption (a.5) holds. Suppose, however, that we wanted to examine how dividends (Y) declared by a firm depend on its net worth (X), then it would be unrealistic to assume that X is fixed in repeated samples. For the standard results to hold, we then need to assume that Cov $(X_i, \varepsilon_i) = 0$. Violations of Assumption (a.5) are discussed in Chapter 11.

Least Squares Principle

We now turn to the problem of deriving estimates of α and β in Model (10.6) from n pairs of sample observations (X_i, Y_i) $i = 1 \cdots n$. We assume that Assumptions (a.1) to (a.5) hold.

Plotting the observations in the $X - Y$ plane, yields a 'scatter diagram'. It seems sensible to choose α and β in such a way that the straight line $Y = \alpha + \beta X$, lies 'as close as possible' to the various points in the scatter diagram.

In Figure 10.2 we have plotted the scatter of observations, and also a tentative regression line. If A is a typical observation, the vertical distance AP is an indication of

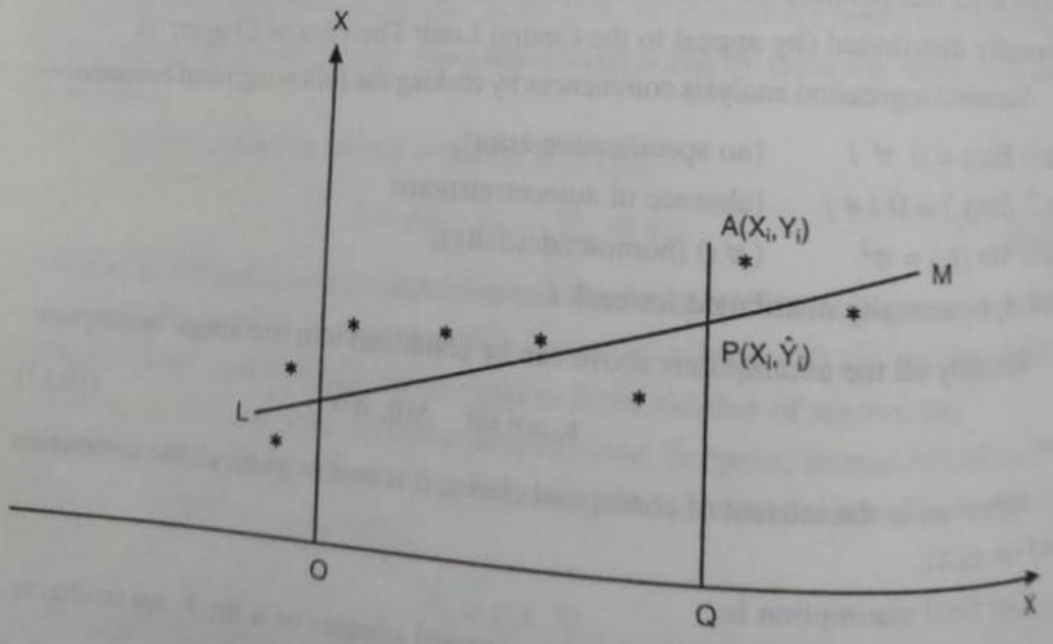


Figure 10.2

the proximity of A to the regression line. We denote the height of the regression line at Q by $\hat{Y}_i (= QP)$.

$$AP = AQ - QP = (Y_i - \hat{Y}_i) = (Y_i - \alpha - \beta X_i) = e_i \text{ (say)},$$

Then, e_i may be termed the ith residual associated with the regression line LM. When deciding which line fits the data best, the e_is have to be made as small as possible, in the aggregate. However, a simple sum $\Sigma_{i=1}^{n} e_i$ is not appropriate since large positive residuals will be set off against large negative residuals, yielding the 'wrong' line of best fit. To avoid this problem, we may minimize

$$\sum_{1}^{n} e_i^2 \tag{10.14}$$

(Expression 10.14) is referred to as residual sum of squares (RSS).

Thus, we have to select the line (that is, the parameters α and β) such that RSS is a minimum.

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \alpha - \beta X_i\right)^2 \tag{10.15}$$

The first order conditions for a minimum are:

$$\frac{\partial(RSS)}{\partial\alpha} = \frac{\partial(RSS)}{\partial\beta} = 0 \tag{10.16}$$

which easily lead to the 'normal equations' (see Exercise 10.1)

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} X_i \tag{10.17}$$

$$\sum_{i=1}^{n} X_i Y_i = \alpha \sum_{i=1}^{n} X_i + \beta \sum_{i=1}^{n} X_i^2 \tag{10.18}$$

Their solution leads to the following estimates of α and β

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \tag{10.19}$$

$$\hat{\beta} = \left\{\sum_{i=1}^{n} x_i y_i\right\} \Big/ \sum_{i=1}^{n} x_i^2 \tag{10.20}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means

$$\overline{X} = \frac{1}{n}\left[\sum_{i=1}^{n} X_i\right]; \overline{Y} = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i\right]$$

and

$$x_i = X_i - \overline{X}; y_i = Y_i - \overline{Y}$$

$\hat{\alpha}$ and $\hat{\beta}$ are called the ordinary least squares (OLS) estimators (the word ordinary being used to distinguish these estimators from other least squares estimators that are discussed later).

The line of 'best fit' is then given by

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X \tag{10.21}$$

For a given X_i, the best fit line yields the 'predicted' value of Y, which we have denoted as $\hat{Y}_i$, that is,

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \tag{10.22}$$

The actual observation may of course depart from the Value (10.22) predicted by the line of best fit,

$$Y_i = \hat{Y}_i + e_i \tag{10.23}$$

Thus, e_i is the error in fitting the line to the observation.

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i, \qquad (i = 1 \cdots n) \tag{10.24}$$

Definition 10.2 (Sample regression function): The relation (10.24) is termed the sample regression function (SRF).

Note: The PRF is the hypothesized relation between Y and X, whereas the SRF is an estimate of this relation, based on a specific sample. Two investigators using the same PRF may get different SRFs, simply because their samples are different. There is a corresponding difference between ε_i (the 'true' errors) and e_i (the sample residuals). By their very nature, ε_i are unobservable, but we can try to estimate their characteristics from the observable residuals e_i.

Note: Berztiss (1964) considers the general mathematical problem of minimising the E_p (norm of errors) defined as:

$$\| E \|_p = \left\{ \sum_{i=1}^{n} | \in_i |^p \right\}^{1/p} \tag{10.25}$$

The case $p = 2$ is, of course, the least squares problem already considered. The case $p = 1$ is called the 'Gershgorin norm' and leads to the minimum sum of absolute deviations considered by Boscovich (1757), whereas the case $P = \infty$ leads to the Chebyshev (1854) criterion of minimizing the maximum absolute value of the error. Berztiss claims some superiority for the Chebyshev norm on theoretical grounds, though from a computational point of view, it is universally agreed that the least squares method is the most advantageous.

Moving Further ,

GLM – Generalized Linear

Date 3/7/20.

→ Problems & Methods of Estimation in Simple Regression Models ←

(I)

Intro – Let us find a situation where you have age of Customers and Income., let them be x and y.

one is interested in $f_{y|x}$ (or) $f_{x|y}$

the conditional distribution of $f_{y|x}$.

This is part of the distribution

$f(x,y)$, $f(x,y) = f(x)\, f(y|x)$

Concentrating on the Conditional distribution

① – $E[y|x] = g(x)$; $g(x)$ is an Increasing function.

If $E[y|x]$ is linear then

$E[y|x]$ will assume a linear form

$E[y|x] = \alpha + \beta x$ – ②

For Example the income of the Customer, whose age is x_i; $E[y_i|x_i] = \alpha + \beta x_i$.

The Error in the Estimation of y_i is

$y_i - \alpha - \beta x_i$: let it be ϵ_i

$\epsilon_i = y_i - \alpha - \beta x_i \quad \forall i = 1, 2, \ldots, n$

the assumptions of ϵ_i

$E(\epsilon_i) = 0$ for all i

$\text{Var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$ for all i

$\text{Cov}(\epsilon_i, \epsilon_j) = E(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$.

These assumptions are embodied in the simple statement

The ϵ_i's are iid $(0, \sigma^2)$

Estimates of parameters...

Let $y = a + bx$ for

$y_i = a + bx_i$ for $i = 1, 2, \ldots, n$

$e_i = y_i - \hat{y}_i$

$= y_i - a - bx_i$

$RSS = \sum e_i^2 = f(a, b)$ {Residual sum of squares}

$= \sum (y_i - a - bx_i)^2$

From principle of Maxima and Minima

$\frac{\partial (\sum e_i^2)}{\partial a} = 2 \sum (y_i - a - bx_i)(-1)$

$= -2 \sum e_i = 0 \quad -(1)$

$\frac{\partial (\sum e_i^2)}{\partial b} = \sum 2 (y_i - a - bx_i)(-x_i) = 0$

$= -2 \sum xe = 0 \quad -(2)$

$\sum (y_i - a - bx_i) = 0 \Rightarrow$

(1) — $\sum y = na + b \sum x$

(2) — $\sum xy = a \sum x + b \sum x^2$

Normal Equations

2020/12/3 07:49

dividing both sides by "n"

$$\frac{\Sigma y}{n} = \frac{na}{n} + b\frac{\Sigma x}{n}$$

$$\bar{y} = a + b\bar{x} \quad -③$$

⇒ The Regression Equation passes through the Point $(\bar{x}, \bar{y})$

To find the parameters,

let $\mu = Cov(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\cdot\bar{y}$

$$\frac{1}{n}\Sigma xy = \mu + \bar{x}\bar{y}$$

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \bar{x}^2$$

$$\frac{1}{n}\Sigma x^2 = \sigma_x^2 + \bar{x}^2$$

dividing Eqn ② by 'n'

$$\frac{\Sigma xy}{n} = a\frac{\Sigma x}{n} + b\frac{\Sigma x^2}{n}$$

$$\mu + \bar{x}\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2)$$

Mul ③ by $(\bar{x})$

$$\bar{x}\bar{y} = a\bar{x} + b\bar{x}^2$$

$$\mu = b\cdot\sigma_x^2$$

$$b = \frac{\mu}{\sigma_x^2} = \frac{Cov(x,y)}{\sigma_x^2}$$

$$b = r\cdot\frac{\sigma_y}{\sigma_x} \quad -④$$

Substituting "b" value in ② we shall get the value ⓐ

$$r = \frac{Cov(x,y)}{\sigma_x\cdot\sigma_y}$$

$$= \frac{\sigma_x}{\sigma_y}\cdot\frac{Cov(x,y)}{\sigma_x^2}$$

$$r\cdot\frac{\sigma_y}{\sigma_x} = \frac{Cov(x,y)}{\sigma_x^2}$$

2020/12/3 07:49

b is the slope of the Regression line $y = a + bx$ and which passes through $(\bar{x}, \bar{y})$. The Regression Equation is

$$(y - \bar{y}) = b(x - \bar{x})$$

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Regression Equations with "k" variables.

$$(y-\bar{y}) = b(x-\bar{x})$$

$$(y-\bar{y}) = r\frac{s_y}{s_x}(x-\bar{x})$$

(II) Regression Equations with "K" Variables.

But before that let us check the decomposition of Sum of Squares.

Error | Residual | Disturbance.

$$\epsilon_i = y_i - \hat{y}_i = y_i - a - bx_i \quad ; i = 1, 2, \ldots$$

The first Normal Equation is

$$a = \bar{y} - b\bar{x}$$

$$\Rightarrow \quad \epsilon_i = (y_i - \bar{y}) - b(x_i - \bar{x})$$

$$\epsilon_i = y_i - bx_i$$

Squaring on both sides

$$\sum \epsilon_i^2 = \sum y_i^2 - 2b\sum x_i y_i + b^2 \sum x_i^2$$

$\Rightarrow$ The Residual Sum of Squares is a quadratic function of b.

2020/12/3 20:52

$$\sum \epsilon_i^2 = \sum y_i^2 - 2b \sum x_i y_i + b^2 \sum x_i^2$$

$$\sum \epsilon_i^2 = \sum y_i^2 - 2b \sum x_i y_i + b \sum x_i y_i$$

$$\sum \epsilon_i^2 = \sum y_i^2 - b \sum x_i y_i$$

$$\Rightarrow \sum y^2 = b \sum xy + \sum \epsilon^2 \quad -(I)$$

$$\Rightarrow \sum y^2 = r^2 \sum y^2 + \sum \epsilon^2$$

$-1 \le r \le +1$

$0 \le r^2 \le 1$

r^2 is a proper fraction

$$r^2 \sum y^2 = r^2 \sum (y - \bar{y})^2 \quad \text{— portion of deviation}$$

$(I) \Rightarrow$ TSS = ESS + RSS

Total sum of squares — TSS

Explained sum of squares — ESS

Residual sum of squares — RSS.

Earlier derivation

$$b = \frac{\sum xy}{\sum x^2}$$

$$r^2 = \frac{(\sum xy)^2}{\sum x^2 \sum y^2}$$

$$r^2 = \frac{(\sum xy)(\sum xy)}{\sum x^2 \sum y^2}$$

$$r^2 \sum y^2 = b \sum xy$$

2020/12/3 20:52

now the General Model will be

$$Y_i = \beta_1 X_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon$$

ϵ_i is the i^{th} disturbance (Error) term

β_j ; $j = 1, 2 \dots k$ are unknown parameters to be Estimated

This can be represented as $Y = X\beta + \varepsilon$

$$Y_{(n\times 1)} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \; ; \; \beta_{(k\times 1)} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \; ; \; \varepsilon_{(n\times 1)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$X_{(n\times k)} = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{k1} \\ X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & & \vdots \\ X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}$$

The Variance Co-Variance matrix of the Error ~~and~~ disturbances

$$COV = \begin{bmatrix} Cov(\varepsilon_1^2) & cov(\varepsilon_1\varepsilon_2) & \dots & cov(\varepsilon_1\varepsilon_n) \\ cov(\varepsilon_2\varepsilon_1) & cov(\varepsilon_2^2) & \dots & cov(\varepsilon_2\varepsilon_n) \\ \dots & \dots & \dots & \dots \\ cov(\varepsilon_n\varepsilon_1) & cov(\varepsilon_n\varepsilon_2) & \dots & cov(\varepsilon_n^2) \end{bmatrix} = \begin{bmatrix} E(\varepsilon_i^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \dots & E(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & & \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \dots & E(\varepsilon_n^2) \end{bmatrix}$$

Assumptions of our Model

Page No.

The general Regression Equation $Y = X\beta + \epsilon \Rightarrow \hat{Y} = X\hat{\beta}$.

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_n - \hat{Y}_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = Y - \hat{Y}$$

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon$$

$$= (Y - \hat{Y})^T (Y - \hat{Y})$$

$$= (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

$$= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta})$$

$$= Y^T Y - Y^T X\hat{\beta} - \hat{\beta} X^T Y + \hat{\beta}^T X^T X \hat{\beta} \quad -(1)$$

partially diff with $\hat{\beta}$ & equate to 0

$$\frac{\partial(RSS)}{\partial \hat{\beta}} = \frac{\partial(Y^T Y - Y^T X\hat{\beta} - \hat{\beta} X^T Y + \hat{\beta}^T X^T X \hat{\beta})}{\partial \hat{\beta}} = 0$$

$$\Rightarrow \frac{\partial(Y^T Y)}{\partial \hat{\beta}} - \frac{\partial(Y^T X\hat{\beta})}{\partial \hat{\beta}} - \frac{\partial(\hat{\beta} X^T Y)}{\partial \hat{\beta}} + \frac{\partial(\hat{\beta}^T X^T X \hat{\beta})}{\partial \hat{\beta}} = 0$$

$$\Rightarrow 0 - Y^T X - (X^T Y)^T + 2\hat{\beta}^T X^T X = 0$$

$$\Rightarrow 0 - Y^T X - Y^T X + 2\hat{\beta}^T X^T X = 0$$

$$\Rightarrow \hat{\beta}^T (X^T X) = Y^T X$$

$$\hat{\beta}^T = Y^T X (X^T X)^{-1}$$

$$\boxed{\hat{\beta} = (X^T X)^{-1} Y X^T}$$

[$X^T X$ is Symmetric so it doesn't change its Value]

$X = m \times l$
$A = n \times m$

$Y = A \rightarrow \frac{\partial y}{\partial x} = 0$

$Y = AX = \frac{\partial y}{\partial x} = A$

$Y = xA = \frac{\partial y}{\partial x} = A^T$

$Y = x^T A x = \frac{\partial y}{\partial x} = 2x^T A$

Partial derivatives

Goodness of Fit

Even though by its very construction, the line (10.24) is the best fit (among all straight lines) to the observations, it is still a moot question as to how good this best fit really is; that is, how much of the variation in the data is explained by this line of best fit.

We may write (Equation 10.24) as

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i) = (Y_i - \bar{Y}) - \hat{\beta}\,(X_i - \bar{X})$$

[on using Equation (10.19)]

Expressing variables in mean deviation form,

$$e_i = y_i - \hat{\beta} x_i \qquad (i = 1 \cdots n) \tag{10.26}$$

From Equation (10.26), on squaring and summing (using 10.20) we get

$$\sum e_i^2 = \sum y_i^2 - \hat{\beta}^2 \sum x_i^2$$

or

$$\sum y_i^2 = \hat{\beta}^2 \sum x_i^2 + \sum e_i^2 \tag{10.27}$$

Henceforth we do not always mention the summation index explicitly. It is understood that the summation ranges over the entire sample, i.e. from 1 to n.

In Equation (10.27), the Left hand side (LHS) is the total variation in the dependent variable—of this the portion explained by the regression line is given by the first term on the right hand side (RHS), the second term is the residual sum of squares. Equation (10.27) may thus be written as:

Total sum of squares (TSS) = explained sum of squares (ESS) + residual sum of squares (RSS)

The 'goodness of fit' of the regression line may be judged by the quantity R^2 which is called the coefficient of determination.

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} \tag{10.28}$$

Note that R^2 is simply the square of the correlation coefficient between Y and X (see Exercise 10.2). Unfortunately, R^2 does not follow a standard distribution and hence it cannot be used directly to make statistical inferences about the goodness-of-fit of the regression line. However, it will be seen later that a suitable transform of R^2 follows the F-distribution.

Apart from α and β, the Model (10.6) contains an additional parameter σ^2 [the variance of the disturbance term ε_i, assumed to be constant via assumption (a.3)]

$$\hat{\sigma}^2 = \frac{S^2}{(n-2)} \tag{10.29}$$

$$S^2 = \sum e_i^2$$

where

The reasons for choosing $(n - 2)$ rather than n, as the denominator in Equation (10.29) will be discussed in the next section.

Several questions of interest may be posed in the context of Model (10.6) and the OLS parameter estimates.

(1) How good are these OLS estimates in terms of the desirable properties of estimators considered in Chapter 8 such as unbiasedness, consistency, etc?

(2) Can we construct confidence intervals for the true values of parameters based on these estimates?

(3) Can we test hypotheses involving one or several parameters?

(4) How are the OLS estimates related to estimates derived from other methods such as maximum likelihood?

(5) What are the consequences of relaxing the assumptions (a.1) to (a.5)?

We try to confront these issues in the more general context of the k-variable linear model.

Coeff of Correlation -

$\frac{ESS}{TSS}(R^2)$ - Coefficient of determination.

- Coefficient of Square Correlation.

$1 - \frac{ESS}{TSS} = \frac{RSS}{TSS}$ } which is unexplained.

GLM.

we have seen the Regression line with the two variables $y = \alpha + \beta x$

Consider a Regression line with "K" variables

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

y - Dependent | x_1, x_2 are Independent

Assume that there are 'K' outlets of a Retail Chain

x_1 — Sales in outlet 1

$\cdots$ x_n - Sales in outlet n.

y - average sales.

Further consider weekly sales for "n" weeks 1, 2, ..., n weeks

2020/12/7 08:02

y_1 - Sales in 1st week

... y_n - Sales in nth week.

x_{11} - Sales in 1st week at outlet 1

... x_{1i} - Sales in 1st week at outlet i

... x_{in} - Sales in nth week at outlet n.

Now the general model will be

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + \epsilon_i$$

$i = 1, 2, \ldots, n; K < n$

ϵ_i is the ith disturbance term.

β_J, $J = 1, \ldots, K$ are the unknown parameters to be estimated.

2020/12/7 08:02

General Linear Model (GLM)

Matrix Specification of GLM

We now generalize Model (10.6) to the case of several independent variables (regressors). Suppose we have a sample of n observations on the dependent variable Y and k independent variables $X_1, X_2, \cdots, X_k$. To allow for an intercept term in the model, the first variable X_1 is assumed to be constant at unity. The ith observation on Y is denoted as Y_i and the ith observation on the variable X_p is X_{pi} $(p = i \cdots k,\ i = 1 \cdots n)$. The k-variable linear regression model is written as

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (i = 1 \cdots n),\ k < n \tag{10.30}$$

where ε_i is the ith disturbance term, and $\beta_j, j = 1 \cdots k$ are the unknown parameters to be estimated.

Note that $X_{1i} = 1,\ i = 1 \cdots n$.

The Model (10.30) is called the general linear model (GLM), and may be written in matrix form as:

$$Y = X\beta + \varepsilon \tag{10.31}$$

where

$$\underset{(n\times 1)}{Y} = \begin{bmatrix} Y \\ Y_2^1 \\ \cdots \\ Y_n \end{bmatrix}$$

$$\underset{(k\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix}$$

$$\underset{(n\times 1)}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

.....

$$\underset{(n\times k)}{X} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}$$

X is often called the data matrix. Each column of X contains all the observations on a particular regressor.

Next, the so-called variance–covariance matrix V of the disturbances is defined. The (i, j)th element v_{ij} of V is simply the covariance between ε_i and ε_j.

Thus,

$$V = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \cdots & \varepsilon_2\varepsilon_n \\ \cdots & \cdots & \cdots & \cdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \cdots & \cdots & \cdots & \cdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{bmatrix}$$

We make the following assumptions on our model.

(A.1) $E(\varepsilon) = 0$ (no specification error).

(A.2) $V = \sigma^2 I_n$, where σ^2 is a constant and I_n is the (nxn) identity matrix (absence of autocorrelation and heteroscedasticity).

(A.3) ε is multivariate normal with mean 0 and variance–covariance matrix V.

(A.4) X is a non–stochastic matrix.

(A.5) X has rank k.

Assumptions (A.1) to (A.4) are simply the k-variable analogues of the assumptions (a.1) to (a.5) in the 2-variable case. Note that (A.2) is a compact way of combining the assumptions of constant variance and zero autocorrelation. However, (A.5) is a new assumption.

OLS Estimation

As in the 2-variables case, the OLS estimator is achieved by minimizing the residual sum of squares Σe_i^2,

where $e_i = Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \cdots - \hat{\beta}_k X_{ki}, \ (i = 1 \cdots n)$

$(\hat{\beta}_j, j = 1 \cdots k$ are the OLS estimates) (10.32)

Let e be the column vector of residuals

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

$$e = (Y - X\hat{\beta}).$$

so $\hat{\beta}_{\hat{\beta}_j (j=1 \cdots k)}$ is obtained from β by replacing each β_j by its estimate

Then,

$$RSS = \sum e_i^2 = e'e = (Y - X\hat{\beta})' (Y - X\hat{\beta})$$
$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y - \hat{\beta}'X'X\hat{\beta}$$

Now $(\hat{\beta}'X'Y)$ is a scalar and hence equals its transpose $Y'X\hat{\beta}$.

Thus, $$RSS = e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \quad (10.33)$$

The first order conditions for minimizing RSS are

$$\frac{\partial(e'e)}{\partial\hat{\beta}} = 0 \quad (10.34)$$

Applying the rules of matrix differentiation (see Appendix A10.6) to Equation (10.34) we get

$$-X'Y + X'X\ \hat{\beta} = 0 \quad (10.35)$$

Thus, the OLS estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (X'X)^{-1} X'Y \quad (10.36)$$

provided $(X'X)^{-1}$ exists which is guaranteed by (A.5). The equation of the best fitting hyperplane is simply

$$Y = X\hat{\beta} \quad (10.37)$$

The vector $\hat{Y} = X\hat{\beta}$ represents the value of Y as predicted by the best fitting hyperplane (10.37). By analogy with the 2-variable case, we may define

$$TSS = \sum Y_i^2 - n\bar{Y}^2 = Y'Y - n\bar{Y}^2$$
$$ESS = \sum \hat{Y}_i^2 - n\bar{\hat{Y}}^2$$

(where $\bar{\hat{Y}} = \left(\frac{1}{n}\right)\Sigma\hat{Y}_i$)

But when one of the regressors is a constant $\bar{\hat{Y}} = \bar{Y}$ (see Exercise 10.5).

Hence
$$ESS = \sum \hat{Y}_i^2 - n\bar{Y}^2 = \hat{Y}'\hat{Y} - n\bar{Y}^2$$
$$RSS = e'e$$

Exercise 10.3 asks the reader to show that

$$TSS = RSS + ESS$$

just as in the 2-variable case.

We now define the coefficient of determination as

$$R^2 = \frac{ESS}{TSS} = \frac{\hat{Y}'\hat{Y} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} \tag{10.38}$$

There is one problem, however, in using R^2 to test the overall goodness-of-fit of the Model (10.31), viz., the possibility of inflating R^2 artificially by simply increasing k (the total number of regressors) including redundant and irrelevant variables (see Exercise 10.4). Hence, a supplementary measure $\bar{R}^2$ (or adjusted R^2) is defined as

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-k)}(1 - R^2) \tag{10.39}$$

Note that $\bar{R}^2$ cannot be increased by simply increasing k. Further, $0 \leq R^2 \leq 1$, whereas $\bar{R}^2$ can take negative values.

The following proposition lists some of the basic properties of OLS estimates. (Their proofs are left for the reader in Exercise 10.5).

Theorem 10.1:

(i) $X'e = 0$ where e is the vector of OLS residuals.

(ii) $\Sigma e_i = 0$ provided one of the regressors is a constant.

(iii) $\text{Corr}(Y, \hat{Y}) = \sqrt{R^2}$ provided one of the regressors is a constant.

We now come to a fundamental result for the GLM, which asserts a certain desirable property of the OLS estimator $\hat{\beta}$.

Theorem 10.2 (Gauss–Markov): The OLS estimator $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β.

Proof: A linear estimate of β means an estimator which is a linear function of the observation vector Y of the dependent variable. Thus, an estimator β^* is a linear estimator, if $\beta^* = CY$ where C is a constant matrix.

Best here means most efficient. In the scalar case, as seen in Chapter 8, efficiency is simply a matter of comparing variances. In the vector case, we have to compare the variance–covariance matrices of the vector of parameters. An estimator $\tilde{\beta}$ is more efficient than another estimator β^+ if

$$\text{Var}(\beta^+) - \text{Var}(\tilde{\beta}) = Q$$

where Q is a positive semi-definite matrix and Var(.) stands for the variance-covariance matrix.

Thus, the theorem asserts that:

1. OLS estimator $\hat{\beta}$ is linear and unbiased; and
2. If β^* is another linear unbiased estimator, then $\hat{\beta}$ is more efficient than β^*.

Now, $\hat{\beta} = (X'X)^{-1}X'Y = PY$, where $P = (X'X)^{-1}X'$ is a constant matrix by assumption (A.4).

Further,

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) \quad \text{(using 10.31))}$$
$$= \beta + (X'X)^{-1}X'\varepsilon$$
$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(\varepsilon) = \beta \quad \text{(by A.1))}$$

Thus, $\hat{\beta}$ is unbiased. This proves Assertion (1).

Let β^* be any other linear unbiased estimator of β. Suppose $\beta^* = CY$, where C is a constant matrix. Define another matrix D:

$$D = C - (X'X)^{-1}X'$$
$$E(\beta^*) = E[CY] = E\,[(X'X)^{-1}X' + D]\,[X\beta + \varepsilon]$$
$$= E[\beta + DX\beta + (X'X)^{-1}X'\varepsilon + D\,\varepsilon\,] \quad (10.40)$$
$$= \beta + DX\beta$$

since β^* is unbiased, we must have from Expression (10.40),

$$DX = 0 \quad (10.41)$$

Now,

$$\text{Var}(\hat{\beta}) = E(\hat{\beta} - \beta)\,(\hat{\beta} - \beta)' = E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}]$$
$$= (X'X)^{-1}X'\,E(\varepsilon\varepsilon')\,X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\,(\sigma^2 I_n)\,X\,(X'X)^{-1}$$

so that

on using (A.2)

$$\text{Var}(\hat{\beta}) = \sigma^2\,(X'X)^{-1}X'X\,(X'X)^{-1} = \sigma^2\,(X'X)^{-1} \quad (10.42)$$

Further, $\beta^* = CY$ and on using the relevant expressions for C and Y

$$\beta^* = [(X'X)^{-1}X' + D]\,[(X\beta + \varepsilon\,]$$
$$\beta^* = \beta + DX\beta + (X'X)^{-1}X'\varepsilon + D\varepsilon$$

so that

$$= \beta + (X'X)^{-1}X'\varepsilon + D\varepsilon$$

on using (10.41)

$$\beta^* - \beta = (X'X)^{-1}X'\varepsilon + D\varepsilon$$

Then
$$\begin{aligned}\text{Var}(\beta^*) &= E\{(\beta^* - \beta)(\beta^* - \beta)'\} \\ &= E[(X'X)^{-1}X'\varepsilon + D\varepsilon][\varepsilon'X(X'X)^{-1} + \varepsilon'D'] \\ &= \sigma^2[(X'X)^{-1} + DD'] \qquad \text{(on using (10.41)}\end{aligned}$$

Thus,
$$\text{Var}(\beta^*) = \text{Var}(\hat{\beta}) + \sigma^2(DD') \tag{10.43}$$

We now use the mathematical result that DD' is a positive semi-definite matrix for any D. Thus, Expression (10.43) implies that $\hat{\beta}$, the OLS estimator, is more efficient than any other linear unbiased estimator.

This proves the theorem.

Estimation of σ^2

There is one additional unknown parameter in Model (10.31), viz., σ^2, (the common variance of the ε_i) which needs to be estimated.

$$\begin{aligned}e = Y - X\hat{\beta} &= Y - X(X'X)^{-1}X'Y \\ &= [I - X(X'X)^{-1}X']Y \qquad (10.44)\end{aligned}$$

Let
$$M = [I - X(X'X)^{-1}X'] \tag{10.45}$$

M can be shown to have the following properties (see Exercise 10.6)

(i) M is symmetric and idempotent;

(ii) $e = M\varepsilon$; and

(iii) $MX = 0$.

Using (ii), we find

$$RSS = e'e = \varepsilon'M'M\varepsilon = \varepsilon'M\varepsilon \qquad \text{[using (i)]} \tag{10.46}$$

Thus,
$$E(RSS) = E[\varepsilon'M\varepsilon] \tag{10.47}$$

But $\varepsilon'M\varepsilon$ is a scalar, and hence is the same as its trace. Thus,

$$\begin{aligned}E(RSS) &= E[Tr(\varepsilon'M\varepsilon)] = E[Tr(M\varepsilon\varepsilon')] \\ &= Tr[M\,\sigma^2 I_n] = \sigma^2\,Tr(M)\end{aligned} \tag{10.48}$$

(on going (A10.12) and (A10.13))

Further,
$$\begin{aligned}\text{Tr}(M) &= Tr[I_n - X(X'X)^{-1}X'] \\ &= Tr(I_n) - Tr\{X(X'X)^{-1}X'\} \\ &= n - Tr\{X'X)^{-1}X'X\} \\ &= n - Tr(I_k) = n - k\end{aligned} \tag{10.49}$$

Combining Equations (10.48) and (10.49),

$$E(RSS) = \sigma^2(n-k) \tag{10.50}$$

Let

$$S^2 = \frac{RSS}{(n-k)} \tag{10.51}$$

Then Equation (10.50) implies that

$$E(S^2) = \sigma^2 \tag{10.52}$$

Thus, S^2 as defined by Expression (10.51), is an unbiased estimator of σ^2. S^2 is called the adjusted residual sum of squares.

Statistical Properties

Some important statistical properties of $\hat{\beta}$ and S^2 are listed below. [Proofs may be found in Theil (1971), Schmidt (1976) and in the original work of Anscombe and Tukey (1963).]

Theorem 10.3: $\hat{\beta}$ and S^2 are unbiased and consistent estimators of β and σ^2 respectively.

Theorem 10.4:

(i) $\hat{\beta}$ has a multivariate normal distribution with mean β and variance–covariance matrix $\sigma^2 (X'X)^{-1}$.

(ii) $\left[(n-k)\frac{S^2}{\sigma^2}\right]$ has a χ^2 distribution with $(n-k)$ d.f.

(iii) $\hat{\beta}$ and S^2 are independently distributed.

Theorem 10.5: $\hat{\beta}$ and S^2 are jointly sufficient for β and σ^2. Further, $\hat{\beta}$ is a minimum variance bound (MVB) estimator, while S^2 is not MVB but it is still MVUE (see Chapter 8 for definitions of joint sufficiency, MVB, and MVUE).

Note: The validity of Theorem 10.1 to 10.5 depends upon all the assumptions (A.1) to (A.5) being fulfilled.

These propositions indicate that the OLS estimators possess several desirable statistical properties. The following result demonstrates the equivalence of OLS and MLE estimators under ideal assumptions.

Theorem 10.6: Under assumptions (A.1) to (A.5), the MLE and OLS estimators of β in Model (10.31) coincide.

Proof: Suppose there is a sample of n observations $Y_1 \cdots Y_n$ on the dependent variable Y. Because ε follows a multivariate normal distribution, and Y is linearly related to ε, the conditional pdf of Y, for any given value of X, is also multivariate normal. Let this conditional pdf be denoted by $h(Y \mid X)$. Then it is easily seen that the conditional mean of Y is

$$E[Y \mid X] = X\beta \tag{10.53}$$

and the conditional variance–covariance matrix of Y is

$$\text{Var}[Y \mid X] = \text{Var}[\in] = V = \sigma^2 I_n \tag{10.54}$$

using (A.2) and (A.3)

$$h(Y \mid X) \text{ is } N[X\beta, \sigma^2 I_n]$$

Thus,

The conditional likelihood of our sample is then

$$L(Y_1 \cdots Y_n \mid X) = \prod_{i=1}^{n} h(Y_i \mid X)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\} \tag{10.55}$$

and the log-likelihood is

$$l = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \tag{10.56}$$

The first-order condition for maximizing l is

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \sigma^2} = 0$$

$$\frac{\partial l}{\partial \beta} = 0 \Rightarrow X'(Y - X\beta) = 0 \tag{10.57}$$

$$\frac{\partial l}{\partial \sigma^2} \Longrightarrow -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}(Y - X\beta)'(Y - X\beta) = 0 \tag{10.58}$$

Let $\tilde{\beta}$ and $\tilde{\sigma}^2$ denote the MLE estimators.

Condition (10.57) $\rightarrow X'Y = X'X\tilde{\beta}$ or $\tilde{\beta} = (X'X)^{-1}X'Y = \hat{\beta}$

so that the MLE $\tilde{\beta}$ coincides with the OLS estimator $\hat{\beta}$.

Condition (10.58) now implies that

$$\tilde{\sigma}^2 = \frac{RSS}{n} \tag{10.59}$$

where RSS are the (OLS) residual sum of squares.

Thus, while the MLE and OLS estimators of β coincide, the MLE $\tilde{\sigma}^2$ is related to the OLS estimator S^2 via

$$\tilde{\sigma}^2 = \frac{(n-k)}{n} S^2 \tag{10.60}$$

using Expression (10.51).

Model in Deviation Form

It is quite often convenient to express each variable that figures in Model (10.30) in deviation form.

Let $\bar{X}_j, j = 1 \cdots k$ denote the sample means of the k independent variables $X_1 \cdots X_k$ and $\bar{Y}$ the sample mean of the dependent variable Y. Further, we introduce the mean deviation variables (denoted by lowercase letters)

$$y_i = Y_i - \bar{Y}; i = 1 \cdots n$$
$$x_{ji} = X_{ji} - \bar{X}_j; i = 1 \cdots n; j = 1 \cdots k$$

Note: $x_{1i} = 0$, since $X_{1i} = 1$ $(i = 1 \cdots n)$

Let us now introduce a square matrix A.

$$A = I_n - \left(\frac{1}{n}\right) jj' \tag{10.61}$$

where I_n is the $(n \times n)$ identity matrix and j is the n-dimensional column vector with each element unity.

Exercise 10.7 asks the reader to check the following properties of A

(i) A is symmetric and idempotent.

(ii) $AZ = z$, where

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \cdots \\ Z_n \end{bmatrix} \qquad z = \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{bmatrix}$$

and $z_i = Z_i - \bar{Z}$ where $\bar{Z} = \left(\frac{1}{n}\right)\left(\Sigma_1^n Z_i\right)$.

(ii) implies that the matrix A, on premultiplication, converts a vector of observations into mean deviation form.

We have

$$Y = X\hat{\beta} + e \tag{10.62}$$

where $\hat{\beta}$ are the OLS estimators and e is the vector of residuals.

Premultiplying both sides of Expression (10.62) by A

$$AY = AX\hat{\beta} + Ae \tag{10.63}$$

Since A converts each variable into mean deviation form, we have

$$y = x\hat{\beta} + e \tag{10.64}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad x = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

Note: We have used Theorem 10.1 (ii) to note that $Ae = e$, in deriving (10.64) from Equation (10.63).

The first column of x is 0 as noted above, so x may be partitioned as

$$x = [0, x^*] \tag{10.65}$$

where x^* is the $n \times (k-1)$ submatrix of all columns of x, excluding the first. Thus, the Equation (10.64) may be written as

$$y = [0 : x^*] \begin{bmatrix} \hat{\beta}_1 \\ \cdots \\ \hat{\beta}^* \end{bmatrix} + e = x^* \hat{\beta}^* + e \tag{10.66}$$

where $\hat{\beta}^*$ is the $(k-1)$ subvector of all elements of $\hat{\beta}$ excluding $\hat{\beta}_1$

Thus,

$$RSS = e'e = (y - x^* \hat{\beta}^*)' (y - x^* \hat{\beta}^*) \tag{10.67}$$

Repeating the derivations (10.33) to (10.36) yields

$$\hat{\beta}^* = (x^{*\prime} x^*)^{-1} x^{*\prime} y \tag{10.68}$$

Equation (10.68) furnishes an alternative way of deriving the OLS estimates $\hat{\beta}_j$ $(j = 2 \cdots k)$ of β_j, using variables in mean deviation form. These estimates, of course, coincide with the corresponding estimates derived from Equation (10.36) except that the estimate for the intercept term β_1 is missing from (10.68). This is easily obtained by noting that

$$0 = \bar{e} = \frac{\sum e_i}{n} = \frac{1}{n}\left\{\sum Y_i - \hat{\beta}_1 \sum X_{1i} \cdots - \hat{\beta}_k \sum X_{ki}\right\}$$

i.e.

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 \cdots - \hat{\beta}_k \bar{X}_k \tag{10.69}$$

$\left(\text{since } \frac{1}{n}(\Sigma X_{1i}) = 1\right)$

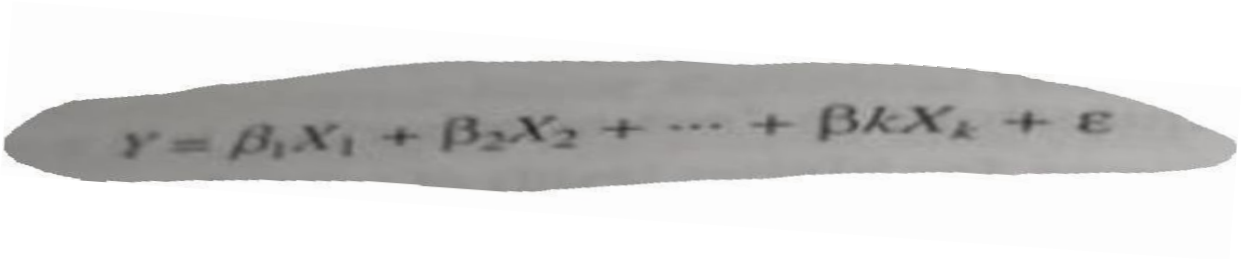
so that once $\hat{\beta}_j$ $(j = 2 \cdots k)$ are obtained from Equation (10.68), $\hat{\beta}_1$ follows from (10.69).

Expression (10.28) is R^2 for the 2-variable case. The generalization of (10.28) to the k-variable case is simply

$$R^2 = \frac{\hat{\beta}^{*\prime} x^{*\prime} x^* \hat{\beta}^*}{y' y} = \frac{ESS}{TSS} \tag{10.70}$$

Multicollinearity : Consequences of multicollinearity, tests to detect its presence and solutions to the problem of multicollinearity.

Multicollinearity is a situation where there exists a linear relationship between the independent variables , Let



The regressors in the regression are "**k**", the Multi collinearity could be as follows,

$X_3 = a_0 + a_1 X_1 + a_2 X_2$,the variables X_1 , X_2 and X_3 have a linear relationship.
Common sense indicates that all the three variables need not be taken to build the regression equation , Only one can be taken in place of the three.

A **matrix** which has its determinant close to zero, and whose inverse is unreliable, is called **near- singular matrix** or ill - conditioned **matrix**.

12 General Linear Model— Relaxation of Assumptions (Part II)

Introduction

In the previous chapter, assumptions (A.1) and (A.2) of the General Linear Model (GLM) were relaxed. This chapter examines the consequences of relaxing the remaining GLM assumptions.

(A.3) refers to the assumption of normality of the error term of the GLM and violation of (A.4) leads to the so-called problem of 'stochastic regressors'. The extensively discussed problem of 'multicollinearity' stems from the breakdown of (A.5).

We commence this chapter with a discussion of multicollinearity, and then move on to the general problem of specification errors. The problem of stochastic regressors and non-normal errors are discussed within the general rubric of specification errors.

Multicollinearity

Definition

The GLM (of section 'General Linear Model', Chapter 10) may be written in either of the two equivalent forms:

$$Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots \beta_k X_{kt} + \varepsilon_t, \; t = 1 \cdots n \tag{12.1a}$$

or

$$Y = X\beta + \varepsilon = [X_1 \; X_2 \cdots X_k]\, \beta + \varepsilon \tag{12.1b}$$

where X_1 is an $(n \times 1)$ vector of unities, and each of $X_2, \cdots, X_k$ is an $(n \times 1)$ vector of observations on the other variables. β is a $(k \times 1)$ vector of regression parameters, and ε is an $(n \times 1)$ vector of errors. We have k regressors (including the intercept) and n observations.

We adhere to this notation throughout the chapter. The OLS estimator of β in model (12.1b), as we have seen, is given by

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{12.2}$$

We will reserve the notation $\hat{\beta}_j$ for the OLS estimate of β_j $(j = 1 \cdots k)$

Definition 12.1 (Multicollinearity): If the matrix $(X'X)$ is singular, we have a case of 'perfect collinearity' and if $(X'X)$ is near-singular we get the case of 'quasi-collinearity' or 'multicollinearity'.

If (A.5) is violated, that is, Rank $(X) < k$, then it can be shown that $(X'X)$ will be singular (perfect collinearity). This is rather rare in economics. It would occur if, for example, an exact linear relationship prevailed between two or more regressors. Near-singularity of $(X'X)$ can, however, occur quite frequently in applied economics. This is due to the fact that economic variables are typically highly interdependent (correlated), so that it is quite common to encounter a pronounced degree of association among regressors in an economic model. To take a common example, let a demand for money function be formulated as:

$$M_t = \alpha + \beta_1 Y_t + \beta_2 Y_{t-1} + \beta_3 (TB)_t + \beta_4 (GS)_t + \varepsilon_t \qquad (12.3)$$

where M_t is current demand for nominal money balances, Y_t and Y_{t-1} are current and previous period (nominal incomes), $(TB)_t$ is the current treasury bill yield (91 days), and $(GS)_t$ the current yield on 1-year government securities.

Multicollinearity is very likely to arise in (12.3) both because of high correlation between Y_t and Y_{t-1} as also between $(TB)_t$ and $(GS)_t$.

Note: Because multicollinearity is a consequence of observed correlation among a set of regressors, it is a data-related problem, rather than one due to the intrinsic structure of the model. Model (12.3) may exhibit multicollinearity if applied to a particular data set (say India over 1970–2000) but may not exhibit multicollinearity over some other data set.

Consequences

The consequences of multicollinearity have been well documented in the literature and may be summarized as follows.

(1) The OLS estimator $\hat{\beta}$ of (12.2), continues to be unbiased but the variance–covariance (var–cov.) matrix var$(\hat{\beta})$ is likely to have large elements (recall that var$(\hat{\beta}) = \sigma^2(X'X)^{-1}$ from (10.42), as $(X'X)$ is near singular). This mens that the variances of the individual regression coefficients are likely to be unduly large that is, the OLS estimates are 'imprecise'.

(2) Following from (1), we get two practical consequences. First, low t-values for some (or even all) regression coefficients may co-exist with a high R^2. Second, the parameter estimates could be very sensitive to small changes in the data

* If the matrix $(X'X)$ is singular - i.e have a case of Perfect collinearity

The Assumption (A.5) is violated, that is Rank $(X) < K$, then one can check that $X'X$ will be singular.

This indicates that there is a linear Relationship among the Independent variables.

* Near singularity of $(X'X)$ — mean the matrix whose determi have close to zero, and whose Inverse is unreliable, is called near - singular matrix (or) Ill conditioned matrix.

When $X'X$ is near singular then we get a case of Quasi collinearity (or) Multicollinearity.

Multicollinearity is quite common in ECONOMETRICS.

Multi collinearity is Data - related Problem.

What are the Problems of Multi collinearity.

Consequences of Multi collinearity.

→ High variance among the β coefficients.

$\beta_1, \beta_2, \ldots, \beta_K$

β_2 — if variance $(\beta_2$ is $\uparrow$

then the Estimation of y is imprecise.

Thus β_i's have less variance.

→ To detect the Multicollinearity the following methods are used.

(I)

→ Correlation Matrix of Regressors / Explanatory Variables

→ There could be several problems with this method.

(a) two variables have High Correlation but they may not have a linear relationship, Converse may not be true

(II) Variance Inflation Factor [VIF]

$$VIF(\hat{\beta}_J) = \frac{1}{1 - R_J^2}$$

where R_J^2 is coefficient of determination of X_J, with respect to the Explanatory Variables

$$X_J = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_J X_{J+1} + \cdots + a_n X_n$$

If R_J's is Large then X_J can be deleted from the Regression Equation.

Common Sense-ically

$VIF(\hat{\beta}_J) = \frac{1}{1 - R_J^2}$ is Large then X_J can be deleted from Regression Equations. —①

From theil [1971]

$$Var(\hat{\beta}_J) = \frac{\sigma^2}{(X_J' X_J)} VIF(\hat{\beta}_J)$$

High VIF indicates, high $Var(\hat{\beta}_J)$ hence can be deleted from Reg.

Detection

Since economic data frequently exhibit multicollinearity, it is important to have powerful tests to detect the phenomenon. The first generation tests focused on an examination of R_x, the matrix of correlation coefficients among the $(k-1)$ explanatory variables (excluding the constant). There are several problems with this approach (see Belsley 1991). First, whereas a high correlation between a pair of variables does indicate collinearity, the converse need not be true. It is possible for a near-collinear relationship to exist between a set of variables, without any two of them being significantly correlated. Second, an examination of R_x by itself, will not enable us to assess the number of collinear relations in the data. Farrar and Glauber (1967) suggested a test (which at one time was quite popular) based on the fact that a transformation of $\det(R_x)$ was approximately distributed as a chi-square. This test has been criticized, among others, by Kumar (1975) and Belsley et al. (1980), on the ground that it relies on the distributional properties of the regressors, when in fact, multicollinearity is a 'data problem' (also see Kennedy 1998).

The next generation of tests employs the concept of the variance inflation factor (VIF) defined below.

Definition 12.2 (VIF): The VIF of $\hat{\beta}_j$ in (12.1a) is defined as:

$$\text{VIF}\left(\hat{\beta}_j\right) = \frac{1}{\left(1 - R_j^2\right)} \tag{12.4}$$

where R_j^2 is the coefficient of determination of the regression of X_j on all the other explanatory variables (including the intercept).

As shown by Theil (1971, p. 166),

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left(X_j' X_j\right)} \text{VIF}\left(\hat{\beta}_j\right) \tag{12.5}$$

where $\sigma^2 = \text{var}(\varepsilon)$.

In view of (12.5), a high value of VIF $(\hat{\beta}_j)$ is associated with a high value of $\text{var}(\hat{\beta}_j)$ except when $(X_j' X_j)$ is high.

The VIF has been suggested as a multicollinearity diagnostic by Theil (1971), Berk (1977), and Simon and Lesage (1988).

Unfortunately, VIFs suffer from many of the same weaknesses as correlation coefficients, adversely affecting their utility as diagnostics. Thus, high VIFs are sufficient but not necessary, for multicollinearity and they are unable to indicate the number of collinear relationships. Finally, the statistical distribution of VIFs remains unknown, so that no obvious thresholds exist to discriminate between high and low values (in empirical work, this threshold is often arbitrarily fixed at 10).

Multicollinearity tests of recent vintage, correctly emphasize the properties of the data matrix X of the GLM. They focus not only on the detection of multicollinearity but also on isolating the set of variables responsible for the problem. These tests make use of several key concepts from matrix algebra.

Definition 12.3 (Condition Number (CN)): Let P be a square positive semi-definite matrix, (Def. A10.15) such that each of its columns has unit Euclidean length. Let μ_{max} and μ_{min} denote the maximum and minimum eigenvalues of P. The CN of P is defined as:

$$K(P) = \{\mu_{max}/\mu_{min}\}^{1/2} \tag{12.6}$$

Note: As P is positive semi-definite, all its eigenvalues are non-negative, so that $K(P)$ is well-defined, except when $\mu_{min} = 0$. The scaling is accomplished by dividing a typical element p_{ij} of P by $\sqrt{\sum_i p_{ij}^2}$. This removes the ambiguity in the CN arising from the fact that the eigenvalues are scale-dependent.

Definition 12.4 (Spectral Decomposition): Any $(k \times k)$ matrix P may be expressed as:

$$P = \sum_{s=1}^{k} \mu_s q_s q_s' \tag{12.7}$$

where μ_s is the s^{th} eigenvalue and q_s the associated (column) eigenvector. We refer to (12.7) as the spectral decomposition of P.

Definition 5 (Singular value decomposition (SVD)): Let Q be any $(n \times k)$ matrix. Then it can be decomposed as:

$$Q = U D V' \tag{12.8}$$

(U is $(n \times k)$ and V and D are both $(k \times k)$); additionally

(a) $U'U = V'V = I_k$ ($k \times k$ identity matrix)

(b) D is a diagonal matrix with non-negative diagonal elements $s_1 \cdots s_k$, which are called the singular values of Q.

For more details on the mathematical properties of some of the above concepts, see Laub and Klema (1980) or Golub and Loan (1996). We now discuss two tests based on some of the above concepts. Both these tests presume that the data matrix X of (12.1b) is scaled in the manner described in Definition 12.3.

Variance Decomposition Test

The Belsley, Kuh and Welsch (1980) (BKW) test is derived from the mathematical result that near dependencies among the columns of Q get reflected in low values of some of the singular values s_j (see (12.8))—while exact relationships lead to some zero singular values.

Definition 12.7 (Variance decomposition proportions): The quantities

$$\pi_{rj} = (\phi_{rj}/\phi_j) \tag{12.16}$$

are referred to as the variance decomposition proportions. Each π_{rj} indicated the proportion of var (β_j) attributable to the rth singular value s_r of X.

Judge et al. (1985) and Peracchi (2001) indicate how the CIs $\eta(j)$ and the π_{rj} can be used in conjunction, both for detection of multicollinearity as well as for identifying the variables responsible for the problem. A suggested threshold value for π_{rj} is 0.5 (Belsley 1991) and for η_j such a value is 30 (as discussed already).

Thus suppose in a specific GLM, the number of regressors $k = 6$ and we find that $\eta(3) = 38$. This implies that the third singular value s_3 of X is low (relative to the maximum singular value). A low singular value could (from (12.11)) inflate any of the variances var $(\hat{\beta}_j)$.

If π_{32} is high, it implies that the proportion of var $(\hat{\beta}_2)$ attributable to s_3 is significant. Suppose π_{32}, π_{33}, and π_{35} all exceed the threshold of 0.5, then we may be led to suspect an interrelationship among X_2, X_3, and X_5 which is leading to multicollinearity, since s_3 is inflating the variances of these three regressors. Additionally suppose that $\eta(4) > 30$ and π_{44} and π_{46} are both greater than 0.5, then there is an additional linear association between X_4 and X_6.

Notes:

1. If a high value of $\eta(j)$ is associated with a single high value, say, π_{jl} then that cannot be interpreted as evidence favouring multicollinearity (see Exercise 12.3).
2. Throughout the above discussion, the matrix X is assumed to be scaled.
3. The BKW test requires the computation of the singular values of X. A computer programme (in C-language) for the purpose is given in Vetterling and Press (1988).

Belsley (1991) introduces an important distinction between 'co-existing' and 'simultaneous' relationships among the regressors of a GLM. Co-existing relationships are two or more relationships in which no common variates figure and in this case the BKW test works fairly well. In the case of 'simultaneous' relationships, common variables are involved and the procedure needs to be used with much greater care. The illustration just discussed yields two co-existing relationships.

Signal to Noise Ratio Test

While the BKW test is useful in detecting mulcolinearity, it does not indicate how harmful (or otherwise) the multicollinearity in a particular problem is. Large values of var $(\hat{\beta}_j)$ may not necessarily be a cause for worry, if β_j is large too (see Smith and Campbell 1980, p. 77). Belsley (1982) suggests a 'signal to noise ratio' (SNR) test to assess whether the multicollinearity is, in fact, harmful. A related advantage of this procedure is that it

The critical values for the asymptotic distribution of L_c are also available in Hansen (1992).

Note: A great advantage of the Hansen procedure is that it also applies if OLS estimators are replaced by GLM estimators. In that case, all the above quantities (such as e_t, $\hat{\sigma}^2$ and h_{it}) have to be replaced by their GLS counterparts.

Omitted Variables and Selection of Regressors

One of the major causes of specification error is the incorrect specification of the variables that figure in the model. We often omit relevant explanatory variables from a model (underfitting) either out of ignorance about their possible influence (on the dependent variable) or because data on these variables are not available. There is also the converse problem that irrelevant variables may be included as regressors in the model (overfitting). Both possibilities raise their own problems but as we discuss here, problems related to underfitting can be more serious than those stemming from overfitting.

Underfitting—Consequences and Detection

Suppose that our true model is (12.1a) with k explanatory variables but we fit a model including only the first k_1 explanatory variables. In matrix form, the original model (12.1b) can be written as:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{12.164}$$

where X_i is $(n \times k_i)$, $i = 1,2$ and $\beta_i = (k_i \times 1)$, $i = 1, 2$ with $k_1 + k_2 = k$. Y and ε are both $(n \times 1)$ vectors.

The researcher fits a partial model:

$$Y = X_1\beta_1 + \varepsilon \tag{12.165}$$

[illegible] model (12.164) and $\hat{b}_1$ the

11 General Linear Model—Relaxation of Assumptions (Part I)

Introduction

In Chapter 10, in the section entitled 'General Linear Model' Matrix Specification of GLM we made several assumptions on the general linear model (GLM), which formed the basis for deriving several critical properties of its estimators. Recall that the GLM specified that

$$Y = X\beta + \varepsilon \tag{11.1}$$

Throughout this chapter, we assume that we have a sample of n observations on the dependent variable and k independent variables (the first of which is the intercept). The matrices Y, X, β, and ε are exactly as in Section 'General Linear Model' of Chapter 10.

For convenience we reproduce assumptions (A.1) to (A.5).

(A.1) $E(\varepsilon) = 0$, where E is the expectations operator.

(A.2) $V = \sigma^2 I_n$, where V is the variance–covariance (var–cov.) matrix of the error term, I_n is the $(n \times n)$ identity matrix and σ^2 is a constant.

(A.3) ε is a multivariate normal vector with mean 0 and var–cov. matrix V.

(A.4) X is a non-stochastic matrix, or if X is stochastic, each regressor is uncorrelated with the error term ε in the limit, that is, plim $\left\{\frac{1}{n}(X'\varepsilon)\right\} = 0$

(A.5) X has rank k.

It may be noted that we have re-stated (A.4) in a more general form, than we had done in Chapter 10.

The above assumptions are ideal and not necessarily realized in practice. In this chapter and the next, we examine the consequences of the relaxation of these assumptions. However, in order to keep the analysis manageable, we relax the assumptions one at a time, keeping the other assumptions intact.

Our primary focus is on what happens to ordinary least squares (OLS) estimators under such relaxations and whether new estimators may be found, when there is pronounced deterioration in the properties of OLS estimators.

Non-zero Mean of Errors

We begin by examining what happens when (A.1) is violated. In this situation, we may distinguish between two situations. If:

$$E(\varepsilon_i) = \mu \qquad (i = 1, \cdots, n) \tag{11.2}$$

with μ a constant, then as shown by Schmidt (1976), $\hat{\beta}$ and S^2 (see (10.51) are unbiased and consistent estimators of β and σ^2, respectively.

The situation, however, becomes complicated if:

$$E(\varepsilon_i) = \mu_i \qquad (i = 1, \cdots, n) \tag{11.3}$$

for then $\hat{\beta}$ and S^2 are biased and inconsistent in general (Maddala 1987). The phenomenon (11.3) most commonly manifests itself when a relevant variable has been omitted from the regression. Such omissions constitute 'specification errors' which we shall discuss in Chapter 12.

Heteroscedasticity

Definition

Assumption (A.2) really encompasses two distinct parts viz.,

(a) $\text{Var}(\varepsilon_i) = \sigma^2$ (a constant) and

(b) $\text{Cov}(\varepsilon_r, \varepsilon_s) = 0, r \neq s$

The first assumption is one of constant variances (of the errors) and is referred to as homoscedasticity. Violation of this assumption is then referred to as the problem of heteroscedasticity.

The second part of the assumption says that distinct error terms are uncorrelated. A failure of this assumption leads to the problem of serial correlation or autocorrelation.

In this section, we discuss the issues related to heteroscedasticity.

Consequences

We now assume that heteroscedasticity is present (with all the other ideal assumptions, including that of the absence of autocorrelation, continuing to hold). The var–cov. matrix of the error terms of (11.1) may now be written as:

$$E(\varepsilon\varepsilon') = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & 0 \\ 0 & 0 & & \sigma_n^2 \end{bmatrix} = \sigma^2 V \tag{11.4}$$

where we impose the restriction that:

$$Tr \cdot (V) = n \tag{11.5}$$

With this restriction

$$\sum_{j=1}^{n} \sigma_j^2 = \sigma^2 \; Tr(V) = n\,\sigma^2$$

that is,

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n} \sigma_j^2 \tag{11.6}$$

σ^2 is thus, so to speak, the average variance of the disturbances.

Suppose we apply OLS to the model (11.1) when the disturbances are characterized by the structure (11.4). Then as shown by Amemiya (1985), we have the following result.

Theorem 11.1: The OLS estimators $\hat{\beta}$ of β in (11.1) when (11.4) prevails, continue to satisfy the following properties (under fairly general conditions): unbiasedness, consistency, and asymptotic normality. Additionally S^2, as defined in (10.51) is a consistent estimator of σ^2.

However, there are two major problems with the OLS estimators in this context.

1. First, the standard procedures for hypothesis testing can be misleading because the var–cov matrix of the OLS estimator $\hat{\beta}$ is not given by the usual formula, $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ but by

$$\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}(X'VX)\,(X'X)^{-1} \tag{11.7}$$

(see Taylor,1977 and Greene 1997.)

As (11.7) involves V, we cannot get a reasonable estimate of $\text{Var}(\hat{\beta})$ unless we estimate V in a consistent manner (this is discussed a little later).

2. Second, OLS estimators are 'inefficient' in the sense that they fail the 'minimum variance property' (of Theorem 10.2). In particular, a new class of estimators known as generalized least squares (GLS) estimators may be defined; these possess lower variance than the OLS estimators.

GLS Estimators

The concept of GLS estimators is applicable in a wide variety of situations, including heteroscedasticity and autocorrelation, and hence we introduce them in a rather general context here.

Suppose that the var–cov. matrix of the error term ε is $(\sigma^2 V)$ with σ^2 a constant and V not necessarily an identity matrix. V will of course be positive definite and symmetric (see Appendix, Chapter 10). Such matrices possess the property (see Johnston and DiNardo 1997, p. 484) that they can be factorized in a specific way.

Hence,

$$V = Q \Lambda Q' \tag{11.8}$$

Where the columns of Q are the eigenvectors of V and the eigenvalues of V are arranged in the diagonal matrix Λ (all these concepts are defined in Appendix, Chapter 10). Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the n eigenvalues of V (which are all positive because V is positive definite and we also assume that the λ_i are all distinct). Then

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{11.9}$$

We may define

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} \tag{11.10}$$

with $\Lambda^{-1/2}$ as its inverse.

Let $P = Q \Lambda^{1/2}$ and $R = \Lambda^{-1/2} Q'$

then

$$P P' = Q \Lambda Q' = V \tag{11.11}$$

$$R'R = Q \Lambda^{-1} Q' = V^{-1} \tag{11.12}$$

Note: (11.12) follows from the well-known result in matrix algebra that $Q' = Q^{-1}$ because in view of V being symmetric, the eigenvectors are pairwise orthogonal (see Johnston and DiNardo 1997, p. 479).

By premultiplying (11.1) by R, we get the transformed model

$$Y^* = X^* \beta + \varepsilon^* \tag{11.13}$$

where $Y^* = RY$, $X^* = RX$, and $\varepsilon^* = R\varepsilon$.

It can be shown that (11.13) satisfies the ideal assumptions (see Exercise 11.1) and in particular:

$$E(\varepsilon^* \varepsilon^{*\prime}) = \sigma^2 I \tag{11.14}$$

The OLS estimators of β in model (11.13) would thus be BLUE and we denote them as $\hat{\beta}_{GLS}$, the GLS estimators of β.

It is easy to see that (see Exercise 11.2):

$$\hat{\beta}_{GLS} = (X'V^{-1}X)^{-1} X'V^{-1}Y \tag{11.15}$$

The GLS estimator of β is thus simply the OLS estimator in the transformed model (11.13). Apart from being BLUE, $\hat{\beta}_{GLS}$ is also consistent and asymptotically normal under fairly general conditions.

The interpretation of the coefficient of determination R^2, however, becomes problematic in the GLS context (Buse 1973 discusses this problem in detail).

The GLS method is quite general and can be applied (as we shall see) in the context of heteroscedasticity as well as autocorrelation. We take up the case of heteroscedasticity first. We may distinguish three situations:

1. The matrix V of (11.4) is completely known.
2. The matrix V is partially known, for example, we may suspect that σ_j^2 $(j = 1, \cdots, n)$ is some known function $f\left(X_{s_1}, X_{s_2}, \cdots, X_{s_m}\right)$ of a subset of the regressors.
3. V is completely unknown.

Each of these cases is discussed below.

Case 1 (V completely known)

Consider now the rather unrealistic case, viz., V fully known but σ^2 unknown, which is a useful starting point. This means that the disturbance variances are known up to a multiplicative constant. Let

$$V = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & v_n \end{bmatrix} \tag{11.16}$$

where the v_i are known.

Now V^{-1} is a diagonal matrix with ith diagonal entry $(1/v_i)$. Further, using the definitions of eigenvalues and eigenvectors (given in Appendix, Chapter 10) we note that the ith eigenvalue of V is simply v_i and the corresponding eigenvector is the unit vector $(0, 0, \cdots, 1, \cdots 0)$ with unity in the ith place.

Hence if we put $\Lambda = V$, in terms of the discussion in the previous section, Q is the identity matrix I and

$$R = V^{-1/2} = \begin{bmatrix} \left(\frac{1}{\sqrt{v_1}}\right) & 0 & \cdots & 0 \\ 0 & \left(\frac{1}{\sqrt{v_2}}\right) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \left(\frac{1}{\sqrt{v_n}}\right) \end{bmatrix}$$

Thus in model (11.13) we have:

$$Y^* = \begin{bmatrix} \left(\frac{Y_1}{\sqrt{v_1}}\right) \\ \left(\frac{Y_2}{\sqrt{v_2}}\right) \\ \cdots\cdots \\ \left(\frac{Y_n}{\sqrt{v_n}}\right) \end{bmatrix}$$

and

$$X^* = \begin{bmatrix} \left(\frac{X_{11}}{\sqrt{v_1}}\right) & \left(\frac{X_{21}}{\sqrt{v_1}}\right) & \cdots & \left(\frac{X_{k1}}{\sqrt{v_1}}\right) \\ \left(\frac{X_{12}}{\sqrt{v_2}}\right) & \left(\frac{X_{22}}{\sqrt{v_2}}\right) & \cdots & \left(\frac{X_{k2}}{\sqrt{v_2}}\right) \\ \cdots & \cdots & \cdots & \cdots \\ \left(\frac{X_{1n}}{\sqrt{v_n}}\right) & \left(\frac{X_{2n}}{\sqrt{v_n}}\right) & \cdots & \left(\frac{X_{kn}}{\sqrt{v_n}}\right) \end{bmatrix}$$

It is clear that model (11.13) now represents a regression in which the ith observation $(Y_i, X_{1i}, \cdots, X_{ki})$ is divided by $\sqrt{v_i}$.

Thus the GLS estimator assumes a particularly simple form in this case. It is simply the OLS estimator from the regression in which the jth observation is weighted by the reciprocal of $\sqrt{v_j}$ (which is the standard deviation of the jth disturbance term). Hence the GLS estimator in this case is often referred to as a weighted least squares (WLS) estimator.

Case 2 (V known partially)

As mentioned earlier, it is quite unrealistic to assume that V is fully known. In reality, either V is completely unknown or we have some a priori idea about its structure. In this section we discuss the latter case.

We now assume (as in Case 1) that σ^2 is unknown but we have some a priori idea about the behaviour of σ_i^2. This information could take any one of the several following forms.

1. $\sigma_j^2 = \sigma^2 X_{kj}^p, (j = 1, 2 \cdots n)$, where p is a known constant (11.17)
   that is, the variances are proportional to a known exponent of one of the regressor variables X_k.

2. $\sigma_j^2 = \sigma^2 \left[\sum_{i=1}^{k} \gamma_i X_{ij} \right]^2, j = 1, \cdots, n$ (11.18)

with γ_i being unknown constants. Here the variances are proportional to the square of an (unknown) linear function of the regressors.

3. $$\sigma_j^2 = \sigma^2\left[\sum_{i=1}^{s}\gamma_i Z_{ij}\right]^2 = \sigma^2\left[Z_j'\gamma\right]^2 \qquad (j = 1, \cdots, n) \tag{11.19}$$

(with γ_i unknown)

$$Z_j = \begin{bmatrix} Z_{1j} \\ Z_{2j} \\ \vdots \\ Z_{sj} \end{bmatrix} \quad \text{and} \quad \gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_s \end{bmatrix}$$

Here the variances are proportional to the square of a linear combination of s variables (which are not necessarily the regressor variables).

This case is more general than (2) and reduces to (2), if $s = k$ and $z_{ij} = X_{ij}$ $(i = 1, \cdots k;$ $j = 1 \cdots n)$

4. $$\sigma_j^2 = \sigma^2 \exp\left[\sum_{i=1}^{s}\gamma_i Z_{ij}\right], \qquad j = 1, \cdots, n \tag{11.20}$$

This case is referred to as exponential heteroscedasticity.

Amemiya (1985) suggests a two-step method to tackle these problems.

Let e_i denote the OLS residuals of model (11.1). To obtain consistent estimates of β, we proceed as follows.

Step 1: Depending on the form of heteroscedasticity suspected, we run an appropriate OLS regression involving the residuals e_i. Suppose, for example, it is felt that heteroscedasticity is of the form (11.19). Then we run the regression:

$$|e_j| = \sum_{i=1}^{s}\gamma_i Z_{ij} + w_j \quad (j = 1, \cdots, n) \tag{11.21}$$

with the constant term absent.

Amemiya has shown that the OLS estimates $\hat{\gamma}_i$ from (11.21) are consistent, and can be used to give a consistent estimator of σ_j^2 viz.,

$$\hat{\sigma}_j^2 = \left(\sum_{i=1}^{s}\hat{\gamma}_i Z_{ij}\right)^2 \tag{11.22}$$

(For the appropriate regression when heteroscedasticity is of type (11.20), see Exercise 11.3.)

Step 2: Let $\hat{V}$ be obtained from V [see (11.16)] by replacing V_j by

$$\hat{v}_j = \left(\hat{\sigma}_j^2/\hat{\sigma}^2\right) \text{ where } \hat{\sigma}^2 = \left(\frac{1}{n}\right)\left\{\sum_{j=1}^{n}\hat{\sigma}_j^2\right\}.$$

Amemiya now proposes the following estimator, feasible generalized least squares estimator (FGLS) denoted by $\hat{\beta}_{FGLS}$.

$$\hat{\beta}_{FGLS} = \left(X'\hat{V}^{-1}X\right)^{-1} X'\hat{V}^{-1}Y \tag{11.23}$$

While the FGLS estimator is asymptotically efficient, its small sample properties are unclear.

An FGLS estimator of β based on maximum likelihood estimates of σ_j^2 has been suggested by Oberhofer and Kmenta (1974), but it is computationally burdensome.

Case 3 (V unknown)

Here since we have absolutely no a priori idea about V at all, efficient estimation via FGLS is not possible. We then adopt the following strategy. We obtain the OLS estimator $\hat{\beta}$ and try to devise a way of estimating its var–cov. matrix var($\hat{\beta}$) in a consistent manner.

White (1980) defines the matrix:

$$W = \sum_{i=1}^{n} e_i^2 x_i' x_i \tag{11.24}$$

where x_i is the i^{th} row of the matrix X and the e_i ($i = 1, \cdots, n$) are OLS residuals. He demonstrates how, based on W, one may construct a consistent estimator of var($\tilde{\beta}$), the estimated asymptotic matrix

$$\text{Est. Asy. Var}(\hat{\beta}) = (X'X)^{-1}\ W(X'X)^{-1} \tag{11.25}$$

The importance of estimator (11.25) stems from the fact that it enables us to get asymptotically valid inferences about $\hat{\beta}$.

Detection of Heteroscedasticity I (Tests based on OLS residuals)

Several tests have been suggested in the literature to detect the presence of heteroscedasticity. We discuss some of the more important ones here. These tests may broadly be divided into two categories viz., (i) tests based on OLS residuals, and (ii) tests based on best linear unbiased scalar covariance matrix (BLUS) and recursive residuals (these types of residuals are defined later).

Here we discuss the first group of tests.

Goldfeld–Quandt (1965) Test

This is possibly one of the oldest tests suggested in the literature. This test proceeds as follows.

Let $\hat{Y}_i$ denote the ith element of the vector $\hat{Y} = X\hat{\beta}$. We now arrange the n observations in ascending order of the values $|\hat{Y}_i|$. The ordered observations are divided into three groups.

1. Group I consisting of the lowest l observations;
2. Group II consisting of the highest l observations; and
3. Group III consisting of the middle $(n - 2l)$ observations. (l is a suitably chosen number.)

The central group of observations (that is, group III) are discarded, and model (11.1) is estimated separately over group I and group II observations, with e_1 and e_2 denoting the vectors of residuals from these two regressions, and S_1, S_2 the corresponding residual sums of squares (RSS). Then Golfeld and Quandt (1965) define the statistic:

$$GQ = (S_2/S_1) = \left\{\frac{e_2'e_2}{e_1'e_1}\right\} \tag{11.26}$$

and show that GQ follows the $F(m, m)$ distribution with $m = l - k$ (recall that k is the number of regressors in (11.1)), under the null hypothesis of no heteroscedasticity. Significant values of GQ point to a rejection of the null in favour of the alternative of the error variances increasing with either one of the independent variables or with a linear combination of a subset of the regressors.

Note:

1. The residuals e_1 and e_2 are obtained via separate regressions on the group I and group II observations, and hence the test is apt to have low power unless n is fairly large.
2. We have assumed that l, the number of observations in group I and group II, is the same. This was originally suggested by Goldfeld and Quandt (1965) who further recommended setting $l = n/3$.

Glejser Test

Glejser (1969) suggests an auxiliary regression with the OLS residuals as the dependent variable. The form of the regression depends upon the type of heteroscedasticity suspected. Thus, for example, if the suspected heteroscedasticity is of the form (11.19) we execute the OLS regression (11.21) and test the null hypothesis (of no heteroscedasticity) via the usual F-test, testing:

$$\gamma_1 = \gamma_2 = \cdots = \gamma_s = 0 \tag{11.27}$$

Rejection of this null (via a significant value of F) then points to heteroscedasticity of the form (11.19).

Unlike the Goldfeld–Quandt test which simply indicates the presence or absence of heteroscedasticity without any information about its likely form the Glejser test is explicit on this point.

An improved version of the Glejser test is presented in Harvey (1976) and Fomby et al. (1984).

White Test

White (1980) proposes a very general test with null hypothesis, H_0: $\sigma_i^2 = \sigma^2$ (homoscedasticity), and alternative hypothesis H_1: $\sigma_i^2 \neq \sigma_j^2$ (for at least one pair of distinct i, j)

He suggests running the auxiliary regression:

$$e_j^2 = \alpha_0 + \sum_{r=1}^{k}\sum_{s=1}^{k} X_{rj}X_{sj} + \text{error term} \quad (j = 1, \cdots, n) \tag{11.28}$$

where e_j are the OLS residuals from (11.1).

Let R^2 denote the coefficient of determination of (11.28), then the statistic:

$$WH = nR^2 \tag{11.29}$$

is shown by White (1980) to be asymptotically distributed as a χ^2 with $(p-1)$ degrees of freedom (d.f.) where p is the total number of regressors in (11.28) including the constant.

The White test also suffers from the problem of being non-informative about the type of heteroscedasticity indicated in the event of rejection of H_0.

Breusch–Pagan Test

A general formulation of heteroscedasticity is

$$\sigma_j^2 = \sigma^2 g\left(\sum_{i=1}^{s}\gamma_i Z_{ij}\right) \qquad j = 1, \cdots, n \tag{11.30}$$

where $(Z_{1j}, Z_{2j}, \cdots, Z_{sj})$ is the jth observation on s variables $Z_1, Z_2, \cdots, Z_s$ (of which Z_1 is a constant) and $\gamma_1, \gamma_2, \cdots, \gamma_s$ are unknown parameters. The function g is assumed to be continuously differentiable.

Note that (11.30) includes the various types of heteroscedasticity listed in (11.17) to (11.20).

The null hypothesis of no heteroscedasticity corresponds to

$$H_0 : \gamma_2 = \cdots = \gamma_s = 0 \tag{11.31}$$

with the alternative corresponding to

$$H_1 : \gamma_j \neq 0 \text{ for at least one } j, \ (2 \leq j \leq s) \tag{11.32}$$

Let e_j denote the OLS residuals from (11.1) and define

$$\hat{\sigma}^2 = \left(\sum_{j=1}^{n} e_j^2 \Big/ n\right) \tag{11.33}$$

and

$$\eta_j = \left(e_j^2 / \hat{\sigma}^2\right) \tag{11.34}$$

regress n_j on the s variables $Z_{1j}, Z_{2j}, \cdots, Z_{sj}$ $(j = 1, \cdots, n)$ and let R^2 denote the coefficient of determination from this regression. The Breusch–Pagan statistic is defined as:

$$BP = \left(\frac{R^2}{2}\right) \tag{11.35}$$

which is asymptotically distributed as a χ^2 with $(s-1)$ d.f. under H_0.

The Breusch–Pagan test assumes that the disturbances in (11.1) are normally distributed. Koenker and Bassett (1982) have shown that the test is quite sensitive to the normality assumption, and suggest replacing $\hat{\sigma}^2$ in (11.34) with:

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left[e_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} e_j^2\right)\right]^2 \tag{11.36}$$

with the rest of the procedure intact.

This modified statistic is asymptotically equivalent to the BP statistic but yields a more powerful test in finite samples.

Harvey Test

Harvey (1976) presents a likelihood ratio test for a particular type of heteroscedasticity viz., the exponential type described in (11.20). As before, the first variable Z_1 is assumed to be a constant. The null of homoscedasticity corresponds to (11.31).

Let $\hat{\gamma}$ be the maximum likelihood estimate of $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_s)$, then Harvey's statistic is given by:

$$HLR = n \ln\left(\sum_{j=1}^{n} e_j^2 \Big/ n\right) - \sum_{j=1}^{n}\left(Z_j'\hat{\gamma}\right) \tag{11.37}$$

where Z_j is defined immediately after (11.19).

Asymptotically, HLR has a χ^2 distribution with $(s-1)$ d.f. and significant values of HLR point to the presence of heteroscedasticity of the exponential type.

Detection of Heteroscedasticity II (Tests based on BLUS and Recursive Residuals)

BLUS Residuals

In two fundamental papers, Theil (1965, 1968) introduced the concept of BLUS residuals.

Definition 11.1 (BLUS residuals): In model (11.1), a linear residual vector u is said to be BLUS if:

(a) u is expressible as $u = CY$ where C is an $(n \times n)$ matrix of constants,

(b) $E(u) = 0$, and

(c) $E(uu') = \sigma^2 I$.

The basic motivation for BLUS residuals is to find for model (11.1) a vector of residuals u which is a linear function of Y, whose expectation is zero and whose var-cov. matrix is a scalar matrix.

Theil (1971) shows that in the model (11.1), only $(n-k)$ BLUS residuals can be defined (provided $n > k$), obtained by partitioning the set of n observations into two sets of k and $(n-k)$ observations, respectively.

Definition 11.2 (Base of BLUS residuals): The set of k observations used in obtaining the BLUS residuals is called a base.

Note: There is some ambiguity in the choice of the base, since the BLUS residuals will be specific to the base chosen. As a convention, the base is chosen to consist of the first k observations.

Selecting the first k observations as the base, model (11.1) may be rewritten as:

$$\begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix} = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \end{bmatrix} \tag{11.38}$$

with Y_0 and ε_0 having dimensions $(k \times 1)$ and Y_1 and ε_1 having dimensions $((n-k) \times 1)$. Further, X_0 is a square matrix of order $(k \times k)$ and is assumed to be non-singular. X_1 is of order $((n-k) \times k)$.

Let e_0 and e_1 be the OLS residuals corresponding to ε_0 and ε_1, respectively. Define:

$$P = X_0 (X'X)^{-1} X_0' \tag{11.39}$$

The $(k \times k)$ matrix P can be shown to be symmetric and positive definite and Theil (1971, p. 209) shows that all its k eigenvalues $\lambda_1, \lambda_2, \cdots \lambda_k$ satisfy $0 \leq \lambda_i \leq 1$ $(i = 1, \cdots, k)$.

Let $\lambda_1, \lambda_2, \cdots, \lambda_H$ be the H eigenvalues of P which are all strictly less than 1, and q_1, $q_2, \cdots q_H$ are the corresponding eigenvectors.

Then the $(n-k)$ BLUS residuals (corresponding to the chosen base) are given by the vector u defined as:

$$u = e_1 - X_1 X_0^{-1} \left[\sum_{i=1}^{H} \left(\frac{\sqrt{\lambda_i}}{1+\sqrt{\lambda_i}} \right) q_i q_i' \right] e_0 \tag{11.40}$$

Recursive Residuals

Recursive residuals were introduced in the econometrics literature by Phillips and Harvey (1974) and Hedayat and Robson (1970).

In the method developed by Phillips and Harvey (1974), we select the first l observations (unlike for the BLUS residuals, l need not be equal to k, the number of regressors) and compute the OLS estimates of β in (11.1) based on these l observations

We denote the estimate based on the first l observations by $\hat{\beta}^{(l)}$. We now add observations one at a time and repeat the procedure. We thus get a series of estimates

$$\hat{\beta}^{(l)}, \hat{\beta}^{(l+1)}, \cdots, \hat{\beta}^{(n)}$$

(where $\hat{\beta}^{(s)}$ is the OLS estimate of β based on the first s observations).

Let $X_{[j]}$ denote the submatrix of X formed by taking its first j rows and let x_j denote the jth row of X. Then the recursive estimates may be easily obtained by the updating formula

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + \frac{\left[X'_{[r-1]}X_{[r-1]}\right]^{-1} x'_r \left(y_r - x_n \hat{\beta}^{(r-1)}\right)}{\left\{1 + x_r \left(X'_{[r-1]}X_{[r-1]}\right)^{-1} x'_r\right\}} \tag{11.41}$$

$$(r = l+1, \cdots, n)$$

The recursive residuals are then defined as:

$$u_r = \frac{\left(Y_r - x_r \hat{\beta}^{(r-1)}\right)}{\left\{1 + x_r \left(X'_{[r-1]}X_{[r-1]}\right)^{-1} x'_r\right\}^{1/2}} \tag{11.42}$$

where: $r = l+1, \cdots, n$.

Theil Test

Theil (1971) employs the Goldfeld–Quandt strategy in combination with the BLUS residuals. The middle k observations are selected as the base for the BLUS residuals, and the total observations are divided into there groups (after arranging them in the order suggested for the original Goldfeld–Quandt test).

1. group I consisting of the first $\frac{(n-k)}{2}$ observations,
2. group II consisting of the last $\frac{(n-k)}{2}$ observations,
3. group III consisting of the middle k observations.

The Theil statistic is identical to the GQ statistic (11.26) except for two features:

1. S_1 and S_2 are computed from BLUS residuals rather than OLS residuals.
2. Instead of running 2 separate regressions on group I and group II observations (as in the Goldfeld–Quandt test), a single regression is executed on all the observations (using the middle k observations as the base).

The Theil statistic is shown to have the $F(m, m)$ distribution with $m = (n-k)/2$, under the null hypothesis of no heteroscedasticity.

Harvey–Phillips Test

Harvey and Phillips (1974) propose a test similar to the Goldfeld–Quandt test, but based on recursive residuals. We once again divide the total observations into three

groups as in the Theil test, except that group III now contains $p \geq k$ observations and groups I and II contain $(n-p)/2$ observations each.

Selecting the middle p observations to start with, we build up recursive residuals for the set of observations in group I by adding one observation at a time backwards. The recursive residuals for group II are built up by successive observations added forwards.

The Harvey–Phillips statistic has the same form as the GQ statistic except that S_1 and S_2 are computed from the recursive residuals. Under the null of homoscedasticity, the Harvey–Phillips statistic follows an $F(m, m)$ distribution with $m = (n-p)/2$.

Ramsey Test

The Ramsey (1969) test also follows the Godlfeld–Quandt approach of arranging the variables in a specific order (see the sub-section 'Detection of Heteroscedasticity') and then dividing the observations into three groups. However, in the Ramsey approach, the number of observations in groups I and II need not be equal. Let n_1 and n_2 denote the number of observations in group I and group II, respectively while n_3' is the number of observations in group III with $n_3' > k$. Choose the middle k observations (of the entire set of observations) which will be located in the middle group (that is, group III) and use them as the base for obtaining BLUS residuals. Let $n_3 = (n_3' - k)$. Then, in effect, we will get n_1, n_2, and n_3 BLUS residuals from groups I, II, and III, respectively. Let S_1, S_2, and S_3 denote the corresponding *RSS* for the three groups. The Ramsey statistic is defined as:

$$RMS = \sum_{i=1}^{3} n_i \left[\ln\left(S_i/n_i\right)\right] - n \ln\left[\frac{\left(S_1 + S_2 + S_3\right)}{n}\right] \tag{11.43}$$

Asymptotically, RMS follows a χ^2 distribution with 2 d.f.

We have discussed a number of tests suggested in the literature for detecting heteroscedasticity. An exhaustive comparison of the powers of these various tests may be found in Ali and Giaccotto (1984) and Harvey and Phillips (1974). Certain non-parametric tests for heteroscedasticity have also been suggested in the literature, which we do not discuss here. The interested reader may refer to Bickel (1978) and Hajek and Sidak (1967).

Illustrative E...