Derivación de la Prueba Jarque – Bera*

*Prueba estadística utilizada para probar la Normalidad de un conjunto de Observaciones o de Residuos en el Modelo de Regresión. Permite analizar si provienen de la distribución de Gauss (la función de densidad de probabilidad Normal). Por su sencillez y potencia es la más usada y se aplica en inferencia estadística en todos los campos científicos. Los textos provienen de la tesis "The Econometrics of Cross Sections" de Carlos M. Jarque, Universidad Nacional de Australia.

CHAPTER 1

THE ECONOMETRICS OF CROSS-SECTIONS

"As Economics pushes beyond 'statics'
it becomes less like Science and more
like History"

Sir John Hicks
Causality in Economics

1.1 INTRODUCTION

For the statistical analysis of economic models, three types of data are used: cross-sectional, time-series and panel data.

Cross-sectional data consists of a set of observations on a group of entities from a defined population. These entities are typically micro decision-making units in an economy (e.g., consumers). The observations are obtained by 'interviewing', at a particular time, the totality of the given population (i.e., by carrying out a census) or by 'interviewing' only a subgroup of it (i.e., by carrying out a sample survey).

Time-series data comprises a set of observations during a sequence of time periods on a given entity. This is often a macro entity (e.g., a country). The time period can be a year, a quarter, a month or a week. Here the observations are usually successive and equally spaced in time.

Panel data consists of time-series observations on a set of cross-sectional entities (e.g., a time-series of T years of observations on production, capital and labour for a cross-section of N industries).

Various *problems* that arise in econometric analysis are specific to the kind of data we are to use. For example, if we are to use cross-sectional data, then we may have some choice as to the way in which this data is to be obtained (i.e., choice of sample design); also, we may have to deal with a problem of data confidentiality. In addition, for the estimation stage of the model, we usually may assume serially independent disturbances. This is in contrast with the *problems* that arise when using time-series, where the data is typically given and where we would often have to deal with serially correlated disturbances.

In this thesis we concentrate on some problems that arise in *the analysis of econometric models when using cross-sectional data.* The choice of this research topic was motivated by three considerations.

The *first* relates to data availability. In many (primarily developing) countries the data available for econometric studies is cross-sectional data obtained, or to be obtained, through a census or a sample survey.[1] These data-gathering techniques allow us to have a large amount of information on economic variables in a short period of time. The econometric methodology employed in these situations would therefore be that of cross-sections.

---

[1] The non-existence of sufficiently long historical series on micro or macroeconomic variables in these countries, would not allow efficient time-series estimation of economic relationships.

The *second* consideration arises from the recent concern to study more closely the behaviour of micro decision-making units, even when carrying out macroeconomic inferences. The appropriateness of 'macro-studies' (i.e., econometric studies aiming to describe the behaviour of macrovariables through the exclusive use of macrorelations) in both developing and developed countries has been critisized by various authors.[2] Arguments are presented on the grounds that relationships among economic variables should be motivated from economic theory; this theory is often derived for microvariables and it has been noted that results may not necessarily hold for their macro-counterparts [e.g., for a discussion of this point in relation to demand and production function studies see, respectively, Muellbauer (1975) and Fisher (1969); see also Theil (1955) and Green (1964)]. It has also been argued that, even if we were willing to regard the theory as appropriate for the macrovariables, the exclusive use of macrorelations would not give answers to important questions (e.g., see Orcutt (1962, p.231)). As a result, some economists have emphasized the need to develop microanalytic models of economies (e.g., see Orcutt (1962), Goldberger and Lee (1962) and Orcutt, Watts and Edwards (1968)). In this activity we would use extensively, although not exclusively, the econometrics of cross-sections.[3]

---

2 Cases do exist where the use of macrorelations has advantages over the use of microrelations. (e.g., see Grunfeld and Griliches (1960) and Aigner and Goldfeld (1974)).

3 Cross-sectional studies do not provide information on the important question of *dynamics*; for this we have to look at time-series or panel data studies.

The *third* consideration refers to the view expressed by some economists regarding the 'scientific validity' of time-series econometric studies (e.g., see Hicks' (1980) quotation in the beginning of this Chapter). Indeed, to some - in our opinion pessimistic - economists, 'valid econometrics' would need to be limited to the econometrics of cross-sections.[4]

These three considerations led us to value the study of problems that arise in the *econometric analysis of cross-sectional data*. In the next section we give necessary definitions, introduce the model to be considered initially and state some assumptions that are often made in the analysis of this model. In Section 1.3 we give an overview of the thesis.

1.2 THE MODEL AND ITS ASSUMPTIONS

A more precise definition of cross-sectional data is now given and some notation introduced. We say we possess *cross-sectional data* or a *cross-section*, when we have available a set of, say N, observations on the $K+1$ variables $Y, X_1, \ldots, X_K$, obtained at a given point in time and which relate to a fixed date or period. These observations correspond to N units or entities which belong to a defined population. For example, the entities may be households and the defined population may consist of the households in a geographical area. Similarly, we may have employees in a given industry, farms in a region or banks in a particular country. We denote by y_i the observation on variable Y corresponding to the i'th entity and by $x_i' = (X_{i1}, \ldots, X_{iK})$ the

---

[4] Hicks (1980) also questions the 'scientific validity' of cross-sectional econometric studies. We have no intention here of debating Hicks' points (for this see Hendry (1980, p.402) and Sims (1981)).

1 by K vector representing the observation on $X_1,\ldots,X_K$, also corresponding to the i'th entity. □

We assume that - for the population under study - economic theory suggests a relationship between Y and $X_1,\ldots,X_K$, and that this relationship may be written in the form

$$y_i = x_i'\beta + u_i \quad , \tag{1.1}$$

where u_i is the i'th unobservable disturbance and β is a vector of unknown parameters that is to be estimated.

Equation (1.1) is linear in parameters and is sufficiently general to allow the study of a wide range of models. Particular cases are *polynomial* (e.g. see Theil (1971, p.155)), *log-linear* (e.g. see Desai (1976, p.12)), *semi-log* (e.g. see Maddala (1977, p.6)), and *spline* regression models (e.g. see Poirier (1976, p.10)). For instance, in equation (1.1) we could have K = 3 with $y_i = \log Q_i$, $X_{i1} = 1$, $X_{i2} = \log C_i$ and $X_{i3} = \log L_i$, where Q_i, C_i and L_i are - respectively - production, capital and labour for the i'th entity, say, i'th industry. In this example, Q_i, C_i and L_i would be physically observed variables and y_i and x_i' the result of some mathematical transformation applied to the observed variables. For simplicity in terminology, throughout the thesis, y_i and x_i' - and not the originally observed variables - will be referred to as our *observations*. Also, without loss of generality, in our discussion we will refer to the population units or entities as *individuals*. □

To provide a simple framework for our exposition, we now state assumptions that are commonly made for the econometric analysis of the model given by equation (1.1).

ASSUMPTION [1]: The cross-sectional data to be used, consisting of the N vectors $(y_1, x_1'), \ldots, (y_N, x_N')$, is given to the researcher (i.e., there is no choice regarding which N individuals are interviewed).

ASSUMPTION [2]: The cross-sectional data is not in aggregated form (i.e., each observation refers to an individual).

ASSUMPTION [3]: The disturbances u_i are normally distributed.

ASSUMPTION [4]: The disturbances u_i are homoscedastic (i.e., $\sigma_1^2 = \ldots = \sigma_N^2$, where σ_i^2 denotes the variance of u_i).

ASSUMPTION [5]: The dependent variable Y may take any value (i.e., $-\infty < y_i < \infty$).

ASSUMPTION [6]: There is no parameter variation in the population under study (so, in (1.1) the vector β is the same for all $i = 1, \ldots, N$).

ASSUMPTION [7]: The variables $X_1, \ldots, X_K$ are fixed (i.e., there are no current endogenous variables among $X_1, \ldots, X_K$) and no other model exists with disturbance correlated with u_i.

Each of these *seven* assumptions provides a topic for discussion in the thesis. The topics are discussed under *four* maintained hypotheses which we now state.

A *first* maintained hypothesis is that the model is correctly specified, in the sense that there are no omitted deterministic influences. This assumption is commonly made in the development of econometric methodology. Tests for 'model misspecification' - which is an important inferential problem in econometrics - have been suggested by Ramsey (1969) and, more recently, by Hausman (1978) and White (1980a). Additional results are indicated in Chapter 10.

A *second* maintained hypothesis is serial independence of disturbances. In most cross-sectional studies this would be a reasonable assumption and we shall not deal with it here. We note, however, that a specific situation in cross-sectional studies, where autocorrelated disturbances may arise, is in the analysis of geographical data. The problem has come to be known as spatial correlation and a description of this is given by Cliff and Ord (1973).

A *third* maintained hypothesis is that no relation exists between β and the unknown parameters that define the distribution of the disturbances. In particular, we assume that x_i is measured without error.

Lastly, a *fourth* maintained hypothesis is that the rank of X is equal to K, where $X = (x_1,\dots,x_N)'$.

1.3 OVERVIEW OF THE THESIS

Under the assumptions stated in Section 1.2, the model given by equation (1.1) could be 'optimally' estimated by Ordinary Least Squares and inferences about the parameters could be carried out using standard t and F tests. This is - of course - treated in detail in elementary econometrics textbooks. Here we are concerned with the study of econometric models, when one or more of assumptions [1] to [7] have 'questionable validity'. Basically, in each Chapter we consider a particular assumption and note the possible consequences when this is invalidly made; we then discuss the problem of testing its validity and, finally, we comment on the analysis of the model when there is evidence that the assumption does not hold. More specifically, the structure of the thesis - highlighting our main results - is as follows.

In Chapter 2, we concentrate on *assumption* [1]. We discuss the estimation of the regression model when using census and given survey data and study the properties of estimators - taking explicit account of the probabilities of selection of the individuals in the sample. We then discuss the question of sample design to obtain efficient estimators of the regression model, and indicate the use of experimental design results and clustering algorithms in the present setting.

In Chapter 3, we consider *assumption* [2] and discuss the problem of efficient aggregation or grouping of observations in regression analysis. We suggest various grouping criteria that lead to efficient estimators of the model. We also show the usefulness of Ward's and the K-Means clustering algorithms in the computation of an optimal aggregation.

In Chapter 4, we concentrate on *assumption* [3]. A main result here is the suggestion of a test for disturbance normality, which is simple to compute and has maximum asymptotic local power. The finite sample properties of this test are also studied and it is found that it performs with very good power, relative to other existing tests - most of which are considerably more difficult to calculate. Our normality test is obtained by applying the Lagrange Multiplier principle to a family of disturbance distributions (e.g., the Pearson Family). This suggested approach can be used in a very wide range of inferential problems and is applied in various sections of the thesis, illustrating its use in inferential problems of the econometrics of cross-sections.

In Chapter 5, we consider *assumption* [4]. We suggest a fully non-constructive test for homoscedasticity, which may be used when there is weak a-priori knowledge as to the nature or the form of the possible heteroscedasticity. We also apply the procedure suggested in Chapter 4, to obtain a joint test for disturbance normality and homoscedasticity. The power of our tests is studied and found to be good relative to other existing procedures.

In Chapter 6, we discuss *assumption* [5] and consider two limited dependent variable (LDV) models: the Truncated model and the Tobit model. We comment on the consequences of disturbance non-normality and heteroscedasticity in the usual maximum likelihood estimators of these LDV models. Also, we suggest tests for disturbance normality and/or homoscedasticity in LDV situations.

In Chapter 7, we concentrate on *assumption* [6] and present a two stage estimation procedure for models with systematic parameter variation. In the first stage, the individuals would be classified into groups of homogeneous parameter values (regimes) by the use of a clustering criterion suggested; the second stage would consist of the econometric estimation of the regimes.

In Chapter 8, we consider *assumption* [7] and extend some of our inferential results to simultaneous equations models. We suggest tests for multivariate normality, multivariate homoscedasticity and for parameter variation.

In Chapter 9, we present an *illustration* of some of the results obtained in previous Chapters. We estimate the Extended Linear Expenditure System, using data from a 1975 Income-Expenditure Household Survey for México. We apply clustering algorithms to form groups of households of homogeneous demand behaviour. Also, we take into consideration the fact that expenditures are non-negative and therefore use LDV models. Additionally, we carry out a comprehensive statistical analysis of the disturbances. To end, we discuss the patterns of household consumption and saving behaviour that emerge.

In Chapter 10, we comment on the maintained hypothesis of correct specification of the deterministic part of the model, and indicate *extensions* of our work which provide a general specification test. We also highlight the usefulness of our results in time-series studies and present some concluding remarks.

□

Two statistical techniques are largely used throughout; these are *Cluster Analysis* (particularly in Chapters 2, 3 and 7), and the *Lagrange Multiplier Test* (particularly in Chapters 4, 5, 6 and 8). Indeed, the thesis is concerned with the application of these techniques in problems that arise in *the econometrics of cross-sections*.

CHAPTER 4

THE PROBLEM OF NON-NORMAL DISTURBANCES*

"To say that 'errors' must obey the normal law means taking away the right of the free-born to make any 'error' he darn well pleases!"

Sir Arthur Eddington
Cambridge Lecture

4.1 INTRODUCTION

We have presented 'best' estimators of the parameters in the linear regression model for various kinds of cross-sectional data (e.g., *census, survey* and *partially aggregated data*). After estimation, we would typically want to carry out inferences about the model. For this we have to make some assumptions - in addition to the specification of first and second order moments - regarding the *distribution of the disturbances*.

We denote the probability density function (p.d.f.) of the i'th disturbance u_i by $f(u_i)$, and proceed under the maintained hypothesis that - apart from scale differentials - $f(u_i)$ is the same for all $i = 1,\dots,N$. Two additional assumptions frequently made are that disturbances are *homoscedastic* and that $f(u_i)$ is the *normal* p.d.f.. We study the homoscedasticity assumption in Chapter 5. For now, we

---

* Sections 4.3, 4.4 and 4.5 contain results obtained with Anil K. Bera and are based on the paper Jarque and Bera (1981b).

assume its validity, and devote our attention to the normality assumption. The model is given in (1.1) and, apart from the stated maintained hypotheses, we make assumptions [2], [5], [6] and [7] described in Section 1.2. □

Under *disturbance normality*, i.e., under assumption [3], one may justify the use of the OLS estimator for β noting that, by the Rao-Blackwell Theorem, it is efficient (e.g., see Schmidt (1976, p.14)). Also, one may apply the usual t and F-tests of restrictions on β, and one may choose from several tests for homoscedasticity which are derived under normality (e.g., see Goldfeld and Quandt (1965) and Harrison and McCabe (1979)). In addition, one may easily obtain confidence intervals for the dependent variable and arrive at particular conclusions about the economic phenomena being studied; for example, Lillard and Willis (1978) investigate earning mobility, and use disturbance normality to make probability statements about the dependent variable (an individual's earnings) given an observation on the regressors (e.g., job history and education). The assumption also plays an important role in Bayesian procedures (e.g., see Zellner (1971, Chapter 3)). □

The consequences of violation of the *normality assumption* have been studied by various authors. In estimation, for instance, the OLS estimator $b = (X'X)^{-1}X'y$ is known to be very sensitive to long-tailed distributions (e.g., see Hogg (1979)). Regarding inferential procedures, Box and Watson (1962) consider the usual t and F-tests, and demonstrate that sensitivity to non-normality is determined by the numerical values of the regressors. They show that, to obtain the desired significance level, some adjustment in the degrees of freedom of these tests may be required. Similarly, Arnold (1980) studies the asymptotic distribution of $s^2 = (y-Xb)'(y-Xb)/N$ and shows the significance level of the usual χ^2 test of the hypothesis $\sigma^2 = \sigma_o^2$ (or the confidence interval for σ^2) is not asymptotically valid in the presence of non-normality. Also, the significance level and power of several homoscedasticity tests (suggested for normal disturbances) is studied in Chapter 5, and it is found that these tests may result in incorrect conclusions under non-normal disturbances. In all, violation of the normality assumption may lead to

(i) The use of sub-optimal estimators;

(ii) Invalid inferential statements; and to

(iii) Inaccurate conclusions.

These consequences highlight the importance of testing the validity of the assumption. □

In Section 4.2, we present a procedure for the construction of efficient and computationally simple econometric specification tests. This procedure is used in Section 4.3 to obtain a test for the normality of observations, and in Section 4.4 to obtain a test for the normality of (unobserved) regression disturbances. We then present - in Section 4.5 - an extensive simulation study to compare the power of these tests with that of other existing tests. In Section 4.6, we comment on possible estimation methods to follow if the hypothesis of disturbance normality has been rejected. Finally, in Section 4.7, we make some concluding remarks.

4.2 THE LM TEST AND AN INFERENTIAL PROCEDURE

We now present a procedure for the construction of specification tests. This consists of the use of the Lagrange Multiplier, or Rao's score test, on a 'General Family of Distributions'. First, some remarks about the Lagrange Multiplier (LM) test. □

The LM test is fully described elsewhere (e.g., see Rao (1948), Aitchison and Silvey (1960), Breusch (1978) and Engle (1981)). So here we shall only introduce notation, and state required results. Consider a random variable u with probability density function (p.d.f.) $f(u)$. For a given set of N independent observations on u, say $u_1,\ldots,u_N$, denote by $\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta)$ the logarithm of the likelihood function, where $\ell_i(\theta) = \log f(u_i)$, $\theta = (\theta_1',\theta_2')'$ is the vector of parameters (of finite dimension), and θ_2 is of dimension r by 1. Assume we are interested in testing the hypothesis $H_o\colon \theta_2 = 0$.

Define

$$d_j = \partial\ell(\theta)/\partial\theta_j$$

$$= \sum_{i=1}^{N} \partial\ell_i(\theta)/\partial\theta_j$$

and

$$I_{jk} = E[-\partial^2\ell(\theta)/\partial\theta_j\partial\theta_k']$$

$$= E\left[\sum_{i=1}^{N} (\partial\ell_i(\theta)/\partial\theta_j)(\partial\ell_i(\theta)/\partial\theta_k)'\right]$$

for $j = 1,2$ and $k = 1,2$. Let $\hat{d}_j$ and $\hat{I}_{jk}$ denote d_j and I_{jk} evaluated at the restricted (obtained by imposing the restriction

$\theta_2 = 0$) maximum likelihood estimator of θ, say $\hat{\theta}$.

It may be shown that under general conditions, which are satisfied in all the applications we discuss, the statistic defined by

$$LM = \hat{d}_2'(\hat{I}_{22}-\hat{I}_{21}\hat{I}_{11}^{-1}\hat{I}_{12})^{-1}\hat{d}_2 \tag{4.1}$$

is, under H_o: $\theta_2 = 0$, asymptotically distributed as a χ^2 with r degrees of freedom, say $\chi^2_{(r)}$ (e.g., see Breusch and Pagan (1980, p.241)). A test of H_o: $\theta_2 = 0$, based on (4.1), will be referred to as an LM test. Two aspects of this test are worth noting.

Firstly, that it is asymptotically equivalent to the likelihood ratio (LR) test, provided the maximum likelihood estimators under the alternative hypothesis are well defined. We shall assume that the true value of θ, θ^o, is an interior point of Ω, where Ω is the subset of $\mathbb{R}^m$ for which maximum likelihood estimation is well defined and m is the dimension of θ. This implies the LM test has the same asymptotic power characteristics as the LR test, including maximum local asymptotic power, i.e., asymptotic efficiency. This is a most desirable feature since, with large samples, any reasonable test can be expected to have high power for alternatives far away from $\theta_2 = 0$, and it is only for alternatives where θ_2 is near the zero vector that asymptotic power has relevance.[1]

---

1 The LM test is also known to have optimal small sample power properties in some cases, e.g., see King and Hillier (1980).

A *second* aspect of this test is that - to compute it - we only require estimation under the null hypothesis of the parameters in the model. In all the inferential problems studied here estimation under H_o: $\theta_2 = 0$ is easily carried out. This makes the LM test computationally attractive, as compared to other asymptotically equivalent tests (i.e., the LR test and the Wald test).[2]

For these two reasons - *good power properties and computational ease* - we use the LM test, rather than others, in our inferential procedure.

□

The LM test has been recently applied in many econometric inferential problems (e.g., see Byron (1970), Godfrey (1978a,b,c) and Breusch and Pagan (1979,1980)). Our procedure for the construction of specification tests also uses the LM principle, but has its *distinct feature* in the formulation of $\ell(\theta)$. Rather than assuming a 'particular' p.d.f. for u_i (or transformation of u_i), we assume that the true p.d.f. for u_i belongs to a 'General Family' (e.g., the Pearson Family), of which the distribution under H_o is a particular member. We then use the LM principle to test H_o within this 'General Family of Distributions'. [Of course, it may be argued that any application of the LM test (e.g., testing for homoscedasticity) specifies a 'General Family'. Here we use this term to refer to Families of p.d.f.'s in the 'statistical' sense (e.g., see Kendall and

---

2 Cases exist when LM, LR and Wald tests have identical power *for all* sample sizes; and choice of test in these cases is based exclusively on computational ease. This occurs when testing linear restrictions in the single equation model (see Evans and Savin (1980)), and in some simultaneous equations situations, e.g., when testing homogeneity restrictions in demand systems (see Bera, Byron and Jarque (1981)).

Stuart (1969, Chapter 6)). Our later discussion will help make this point clearer]. We also note the tests obtained are known to have optimal large sample power properties for members of the 'General Family' specified, and that this does not imply they will not have good power properties for non-member distributions. Indeed, for the cases studied, we found that the tests performed with extremely good power for distributions not belonging to the 'General Family' used in our derivations.

The suggested approach for the development of specification tests can be applied in a *wide range* of statistical and econometric inferential problems.[3] In this thesis we confine ourselves to those applications of the procedure that are of direct interest to the *econometrics of cross-sections*. The first two applications are presented in the next two sections, where we obtain tests for normality of observations and regression disturbances.

4.3 A TEST FOR NORMALITY OF OBSERVATIONS

Statisticians' interest in fitting curves to data goes a long way back. As noted by Ord (1972, p.1) - although towards the end of the nineteenth century - "not all were convinced of the need for curves other than the normal" (see K. Pearson (1905)), "by the turn of the century most informed opinion had accepted that populations might be non-normal" (some historical accounts may be found in E.S. Pearson (1965)). This naturally led to the development of tests for the

---

[3] For example, as pointed out in Jarque and Bera (1981c), this can be used to test if observations come from a particular truncated distribution, providing an alternative to the *Kolmogorov-Smirnov test* with *fixed truncation point*. Another example is given in Lee (1981), who has used our approach to test distributional assumptions in accelerated failure time models.

normality of observations. Interest in this area is still very much alive, and recent contributions to the literature are the skewness, kurtosis and omnibus tests proposed by D'Agostino and Pearson (1973), Bowman and Shenton (1975) and Pearson, D'Agostino and Bowman (1977). Other approaches include: the Analysis of Variance tests of Shapiro and Wilk (1965), and Shapiro and Francia (1972); LR tests based on specific alternatives such as power transformations; goodness of fit tests such as the χ^2-test and the Kolmogorov-Smirnov test; and graphical methods like normal probability plots. In this section we consider the Pearson Family of distributions, and make use of the LM principle to derive an additional test for the normality of observations. This test is simple to compute and asymptotically efficient.

Before proceeding to our derivations, we note that testing the *normality of observations* has constituted an important and central statistical problem in the Natural Sciences. Furthermore, this is also a relevant problem in the Social Sciences, e.g., Carlson (1975) considers testing the normality of price expectations of a group of economists. Therefore, the results of this section have potential application in many fields of research. □

We now present our derivations. Consider having a set of N independent observations on a random variable ν, say $\nu_1,\ldots,\nu_N$, and assume we are interested in testing the normality of ν. Denote the unknown population mean of ν_i by $\mu = E[\nu_i]$ and, for convenience, write $\nu_i = \mu + u_i$. It follows that $E[u_i] = 0$ and that - apart from location - the p.d.f. of ν_i is equal to the p.d.f. of u_i. Assume the p.d.f. of u_i, $f(u_i)$, has a single mode and "smooth contact with the u_i-axis at the extremities". More specifically, assume $f(u_i)$ is a member of the *Pearson Family*. This is not very restrictive, due to the wide range of distributions that are encompassed in it (e.g., particular members are the Normal, Beta, Gamma, Students' t and F distributions). This means we can write (see Kendall and Stuart (1969, p.148))

$$df(u_i)/du_i = (c_1-u_i)f(u_i)/(c_o-c_1u_i+c_2u_i^2)$$

or

$$f(u_i) = \frac{\exp\left[\int \frac{c_1-u_i}{c_o-c_1u_i+c_2u_i^2}\,du_i\right]}{\int_{-\infty}^{\infty} \exp\left[\int \frac{c_1-u_i}{c_o-c_1u_i+c_2u_i^2}\,du_i\right]du_i} , \qquad (4.2)$$

with $-\infty < u_i < \infty$, and where the denominator in equation (4.2) makes $f(u_i)$ a proper p.d.f. .

It follows that the logarithm of the likelihood function (or log-likelihood) of our N observations $\nu_1,\ldots,\nu_N$ may be written as

$$\ell(\mu,c_o,c_1,c_2) = -N\log\left[\int_{-\infty}^{\infty} \exp\left[\int \frac{c_1-u_i}{c_o-c_1u_i+c_2u_i^2}\,du_i\right]du_i\right]$$

$$+ \sum_{i=1}^{N} \left[\int \frac{c_1-u_i}{c_o-c_1u_i+c_2u_i^2}\,du_i\right] . \qquad (4.3)$$

Although our procedure is more general, our interest here is to test the hypothesis of normality, which means, from our expression for $f(u_i)$, that we want to test H_o: $c_1 = c_2 = 0$. Let $\theta_1 = (\mu,c_o)'$, $\theta_2 = (c_1,c_2)'$ and $\theta = (\theta_1',\theta_2')'$. Using these, and the definitions of Section 4.2, we can show that - for this problem - the LM test statistic is given by (see Proposition 1 in Appendix A in page 274)

$$LM = N[(\sqrt{b_1})^2/6 + (b_2-3)^2/24] \ . \qquad (4.4)$$

where $\sqrt{b_1} = \mu_3/\mu_2^{3/2}$; $b_2 = \mu_4/\mu_2^2$; $\mu_j = \sum_{i=1}^{N} (\nu_i-\bar{\nu})^j/N$ and $\bar{\nu} = \sum_{i=1}^{N} \nu_i/N$. (Note $\sqrt{b_1}$ and b_2 are, respectively, the skewness of kurtosis sample coefficients). From the results stated in Section 4.2 we know that, under H_o: $c_1 = c_2 = 0$, LM is asymptotically distributed as $\chi^2_{(2)}$, and that a test based on (4.4) is asymptotically locally most powerful. H_o is rejected, for large samples, if the computed value of (4.4) is greater than the appropriate significance point of a $\chi^2_{(2)}$. □

Several tests for normality of observations are available. For example, there are tests based on either of the quantities $\sqrt{b_1}$ or b_2. These have optimal properties, for large samples, if the departure from normality is due to either skewness or kurtosis (see Geary (1947)).

We have shown expression (4.4) is an LM test statistic. Therefore, we have uncovered a principle that proves its *asymptotic efficiency*. This finding encourages the study of its finite sample properties. For finite N, the distributions of $\sqrt{b_1}$ and b_2, under H_o, are *still unknown*. The problem has engaged statisticians for a number of years, and only approximations to the true distributions are available (e.g., for $\sqrt{b_1}$ see D'Agostino and Tietjen

(1973); and for b_2 see D'Agostino and Pearson (1973)). This highlights, together with the fact that $\sqrt{b_1}$ and b_2 are not independent (e.g., see Pearson, D'Agostino and Bowman (1977, p.233)), the difficulty of analytically obtaining the finite sample distribution of (4.4) under H_o.

An alternative is to resort to *computer simulation*. We see that LM is invariant to the scale parameter, i.e., that the value of LM is the same if computed with ν_i/σ rather than ν_i (for all finite $\sigma > 0$). Therefore, we may assume $V[\nu_i^2] = 1$, and generate n sets of N pseudo-random variates from a $N(0,1)$. Then, for each of these n sets, LM would be computed, giving n values of LM under H_o. By choosing n large enough, we may obtain as good an approximation as desired to the distribution of LM and, so, determine the critical point of the test for a given significance level α, or the probability of a Type I error for the computed value of LM from a particular set of observations. Computer simulation is used in Subsection 4.5.1. There, we present a study comparing the finite sample power of LM with that of other existing tests for normality; and a Table of significance points for $\alpha = .10$ and $.05$. □

To finalize this section, we note the procedure utilized here may be applied in a similar way to other families of distributions. We have used the *Gram-Charlier (type A) Family* (e.g., see Kendall and Stuart (1969, p.156) or Cramer (1946, p.229)), and derived the LM normality test, obtaining the same expression as for the *Pearson Family*, i.e., equation (4.4). Our approach may also be used to test the hypothesis that $f(u)$ is any particular member of, say, the *Pearson Family*. This may be done by forming H_o with the appropriate values of c_o, c_1 and c_2 that define the desired distribution, e.g., to test if $f(u)$ is a Gamma distribution we would test H_o: $c_1 = 0$ (see Kendall and Stuart (1969, p.152)). In some cases this may involve testing non-linear inequalities in c_o, c_1 and c_2. For example, to test if $f(u)$ is a Pearson Type IV we would test H_o: $c_1^2-4c_oc_2 < 0$. This requires the development of the Lagrange multiplier procedure to test non-linear inequalities, and should be an important area for further research.

4.4 A TEST FOR NORMALITY OF DISTURBANCES

Now we consider the regression model given by equation (1.1). We note that - by our maintained hypotheses - the regression disturbances $u_1,\ldots,u_N$ are assumed to be independent and identically distributed with population mean equal zero. In addition, we now assume the p.d.f. of u_i, $f(u_i)$, is a member of the *Pearson Family* (the same result is obtained if we use the *Gram-Charlier (type A) Family*). This means we can write $f(u_i)$ as in (4.2) and the log-likelihood of our N observations $y_1,\ldots,y_N$ as in (4.3), where now the parameters, i.e. the arguments in $\ell(\cdot)$, are β, c_o, c_1 and c_2, and $u_i = y_i-x_i'\beta$.

We define $\theta_1 = (\beta',c_o)'$ and $\theta_2 = (c_1,c_2)'$, and note that we want to test the *normality of the disturbances*. This is equivalent to testing H_o: $\theta_2 = 0$. It is shown in Proposition 2 in Appendix A (see page 275), that - in this case - the LM test statistic becomes

$$LM_N = N[\hat{\mu}_3^2/(6\hat{\mu}_2^3) + ((\hat{\mu}_4/\hat{\mu}_2^2)-3)^2/24] + N[3\hat{\mu}_1^2/(2\hat{\mu}_2) - \hat{\mu}_3\hat{\mu}_1/\hat{\mu}_2^2] , \qquad (4.5)$$

where $\hat{\mu}_j = \sum_{i=1}^{N} \hat{u}_i^j/N$, and the $\hat{u}_i$ are the OLS residuals, i.e.,

$\hat{u}_i = y_i - x_i'b$. We have written the resulting test statistic with a suffix N to indicate this refers to a *disturbance normality* test. Using the results of Section 4.2, we know LM_N is, under H_o, asymptotically distributed as $\chi^2_{(2)}$, and that it is asymptotically efficient. Obtaining the finite sample distribution of LM_N by analytical procedures appears to be intractable. For a given matrix X, we may resort to computer simulation, generating u_i from a N(0,1) (e.g., see Section 4.3 and note LM_N is invariant to the scale parameter σ^2). In Subsection 4.5.2 we use computer simulation to study the finite sample power of LM_N.

To finalize, we recall that - in linear models with a constant term - OLS residuals satisfy the condition $\hat{u}_1 + \ldots + \hat{u}_N = 0$. In these cases we have $\hat{\mu}_1 = 0$ and, therefore, (4.5) would reduce to

$$LM_N = N[(\sqrt{\hat{b}_1})^2/6 + (\hat{b}_2 - 3)^2/24] \quad , \qquad (4.6)$$

where $\hat{b}_1 = \hat{\mu}_3^2/\hat{\mu}_2^3$ and $\hat{b}_2 = \hat{\mu}_4/\hat{\mu}_2^2$.

4.5 POWER OF NORMALITY TESTS

In this section we present results of a *Monte Carlo study*. This was done to compare the power of various tests for normality of observations and regression disturbances.

Not all cross-sectional studies have large samples. There are situations where the sample may be small due to splitting of the original data set (e.g., see Chapter 7), or because of a small cross-section to start with (e.g., observations on a group of countries). Keeping this in mind, we carried out simulations for small and moderate *sample sizes*. More specifically, we used N = 20, 35, 50, 100, 200 and 300.

We consider four distributions members of the Pearsón Family: the Normal, Gamma (2,1), Beta (3,2) and Students t with 5 degrees of freedom; and one distribution which is a non-member of the Pearson Family: the Lognormal. These *distributions* were chosen because they cover a wide range of values of third and fourth standardized moments (see Shapiro, Wilk and Chen (1968, p.1346)). To generate pseudo-random variates u_i, from these and other distributions considered throughout the study, we used the subroutines described in Naylor et al. (1966) on a UNIVAC 1100/42. Each of the five variates mentioned above was *standardized* so as to have zero mean.

4.5.1 Testing for Normality of Observations

We first note that since $\mu = 0$, we have $\nu_i = u_i$ (see Section 4.3 for notation). The *tests* we consider for the normality of the observations u_i are the following:

1. Skewness measure test [with this we would reject normality, i.e. H_o, if $\sqrt{b_1}$ is outside the interval $(\sqrt{b_{1L}},\sqrt{b_{1U}})$. For the definition of $\sqrt{b_{1L}},\ldots$ etc. see below];
2. Kurtosis measure test [reject H_o if b_2 is outside (b_{2L},b_{2U})];
3. D'Agostino (1971) D^* test [reject H_o if $D^* = [\Sigma(i/N^2-(N+1)/(2N^2))e_i^o/\mu_2^{\frac{1}{2}}-(2\sqrt{\pi})^{-1}]N^{\frac{1}{2}}/.02998598$ is outside (D_L^*,D_U^*), where e_i^o is the i'th order statistic of $u_1,\ldots,u_N$];
4. Pearson, D'Agostino and Bowman (1977) R test [reject H_o if either $\sqrt{b_1}$ is outside (R_{1L},R_{1U}) or b_2 is outside (R_{2L},R_{2U})];

5. Shapiro and Wilk (1965) W test [reject H_o if $W = (\Sigma a_{iN} e_i^o)^2/(N\mu_2)$ is less than W_L, where the a_{iN} are coefficients tabulated in Pearson and Hartley (1972, p.218)];

6. Shapiro and Francia (1972) W' test [reject H_o if $W' = (\Sigma a'_{iN} e_i^o)^2/(N\mu_2)$ is less than W'_L, where the a'_{iN} are coefficients that may be computed using the tables in Harter (1961)]; and

7. LM test [reject H_o if $LM > LM_U$].

We did not include *distance tests* (such as the Kolmogorov-Smirnov test, Cramer-Von Mises test, weighted Cramer-Von Mises test and the Durbin test) because it has previously been reported that, for a wide range of alternative distributions, the W test - considered here - was superior to these (see Shapiro, Wilk and Chen (1968)). □

The values $\sqrt{b}_{1L}$, $\sqrt{b}_{1U}$, b_{2L}, b_{2U}, D^*_L, D^*_U, R_{1L}, R_{1U}, R_{2L}, R_{2U}, W_L, W'_L and LM_U are appropriate *significance points*. We considered a 10 per cent significance level, i.e., we set $\alpha = .10$. All the points we used are summarized in Table 4.1. For $N = 20, 35, 50$ and 100 and tests $\sqrt{b}_1$, b_2, D^* and R, the points are as given in White and MacDonald (1980, p.20). For $N = 200$ and 300, significance points for $\sqrt{b}_1$, b_2 and D^* were obtained respectively from Pearson and Hartley (1962, p.183), Pearson and Hartley (1962, p.184) and D'Agostino (1971, p.343); and for the R test we extrapolated the points for $N \leq 100$. For W, W' and LM we computed the significance points by simulation using 250 replications so that the empirical α, say $\hat{\alpha}$, was equal to .10. For example, for a given N, we set $W_L = W(25)$, where $W(25)$ was the 25'th largest of the values of W in the 250

TABLE 4.1

Significance points for normality tests (α = .10)

N	20	35	50	100	200	300
$\sqrt{b}_{1L}$	-.769	-.621	-.534	-.389	-.280	-.230
$\sqrt{b}_{1U}$	.769	.621	.534	.389	.280	.230
b_{2L}	1.82	2.03	2.15	2.35	2.51	2.59
b_{2U}	4.17	4.10	3.99	3.77	3.57	3.47
D^*_L	-2.440	-2.295	-2.210	-2.070	-1.960	-1.906
D^*_U	.565	.805	.937	1.140	1.290	1.357
R_{1L}	-.891	-.722	-.624	-.457	-.332	-.285
R_{1U}	.891	.722	.624	.457	.332	.285
R_{2L}	1.762	1.973	2.078	2.302	2.450	2.500
R_{2U}	4.530	4.415	4.230	3.955	3.650	3.500
W_L	.925	.945	.957			
W'_L	.933	.946	.967	.980	.989	.991
LM_U	2.18	2.56	2.63	3.36	3.71	4.29

replications under Normal observations. Similarly for $\dot{W}'$. For LM we set $LM_U = LM(225)$. [Initially we used, for W, the points from Shapiro and Wilk (1965, p.605); for W' from Weisberg (1974, p.645) and Shapiro and Francia (1972, p.216); and for LM from the values of Table 4.3. With $\hat{\alpha} = .10$, easier power comparisons among the one-sided tests W, W' and LM can be made. Note that $\sqrt{b_1}$, b_2 and D^* are two-sided tests and that R is a four-sided test, and hence, for these it is troublesome to adjust the significance points so that $\hat{\alpha} = .10$.] □

Every *experiment* in this simulation study consists of generating N pseudo random variates from a given distribution; computing the values of $\sqrt{b_1}$, b_2, D^*, W, W' and LM and seeing whether H_o is rejected by each individual test. We carried out 250 replications. The *estimated power* of each test (obtained by dividing the number of times H_o was rejected by 250) for each of the 5 distributions and 6 sample sizes considered are given in Table 4.2, except for W which cannot be computed for N > 50 because of the unavailability of the coefficients a_{iN}. The power for the Lognormal and N = 50, 100, 200 and 300 is not reported; this was equal to 1 for all tests. In the Table, the highest power is underlined for each distribution and sample size, except when three or more tests have this power.

If we have large samples, and we are considering members of the Pearson Family, the theoretical results of Section 4.3 justify the use of the LM test. For finite sample performance we resort to Table 4.2. For N = 20, the preferred test would probably be the W test, followed by LM and W'. For N = 35, tests W and LM may be considered best, and we find these are followed by W'. For N = 50, perhaps LM would be preferred, followed by the W and W' tests. LM has highest power

TABLE 4.2

Normality of Observations

Estimated *power* with 250 replications ($\alpha = .10$)

	$\sqrt{b_1}$	b_2	D^*	R	W	W'	LM
N = 20							
Normal	.068	.080	.060	.084	.100	.100	.100
Beta	.072	.128	.124	.120	.208	.132	.116
Students t	.272	.212	.240	.252	.280	.300	.340
Gamma	.796	.476	.604	.772	.920	.884	.872
Lognormal	.996	.916	.988	.996	.996	.996	.996
N = 35							
Normal	.100	.108	.132	.128	.100	.100	.100
Beta	.108	.208	.164	.200	.276	.120	.116
Students t	.332	.396	.384	.372	.316	.428	.444
Gamma	.968	.600	.804	.940	.992	.980	.992
Lognormal	1.000	.980	1.000	1.000	1.000	1.000	1.000
N = 50							
Normal	.064	.072	.080	.064	.100	.100	.100
Beta	.192	.232	.204	.292	.480	.360	.412
Students t	.372	.404	.404	.420	.332	.496	.508
Gamma	1.000	.768	.920	1.000	1.000	1.000	1.000
N = 100							
Normal	.084	.100	.104	.084		.100	.100
Beta	.276	.488	.372	.568		.652	.684
Students t	.484	.680	.680	.672		.736	.744
Gamma	1.000	.948	.988	1.000		1.000	1.000
N = 200							
Normal	.092	.144	.156	.100		.100	.100
Beta	.540	.836	.628	.916		.944	.964
Students t	.520	.844	.848	.844		.848	.856
Gamma	1.000	.996	1.000	1.000		1.000	1.000
N = 300							
Normal	.124	.132	.132	.100		.100	.100
Beta	.776	.940	.804	.972		.996	1.000
Students t	.560	.984	.988	.980		.964	.992
Gamma	1.000	1.000	1.000	1.000		1.000	1.000

for all distributions and <u>N = 100, 200 and 300</u>, but the differences in power compared with the W' test are small. We should also note that LM may have good relative power even when the distribution is not a member of the Pearson Family (e.g., see power for Lognormal in Table 4.2). Overall, LM *is preferred*, followed by W and W', which in turn dominate the other four tests. [This uniformly good relative performance of W and W', is in contrast with the findings of White and MacDonald (1980, p.22). The differences in the results may be due to our use of a one-sided rejection region and their use, apparently, as pointed out by Weisberg (1980, p.30), of a two-sided rejection region for the one-sided tests W and W'].

Apart from power considerations, LM has an *advantage* over W (and W') in that, for its computation, one requires neither ordered observations (which may be expensive to obtain for large N) <u>nor</u> expectations and variances and covariances of standard normal order statistics (which may <u>not</u> be available for a particular N, e.g., as noted previously W cannot be computed for $N > 50$ because of the unavailability of a_{iN}).[4]

> *These results - together with its asymptotic properties - suggest the* LM *test may be the preferred test in many situations.* Therefore, it appeared worthwhile to carry out extensive simulations to obtain - under normality - finite sample significance points for LM. Using expression (4.4) we carried out 10 000 replications and present, in Table 4.3, significance points for $\alpha = .10$ and .05

---

4 LM has an additional advantage in providing a convenient framework in which simultaneous specification tests may be derived (e.g., see Chapters 5, 6 and 10).

for a range of sample sizes. The Table suggests that, for large samples, a test of approximately the desired level may be carried out using the asymptotic distribution of LM, i.e. a $\chi^2_{(2)}$, in the choice of the significance point.

TABLE 4.3

Normality of Observations
Significance points for LM normality test
(10000 replications)

N	α = .10	α = .05
20	2.13	3.26
30	2.49	3.71
40	2.70	3.99
50	2.90	4.26
75	3.09	4.27
100	3.14	4.29
125	3.31	4.34
150	3.43	4.39
200	3.48	4.43
250	3.54	4.51
300	3.68	4.60
400	3.76	4.74
500	3.91	4.82
800	4.32	5.46
∞	4.61	5.99

□

4.5.2 Testing for Normality of Regression Disturbances

In the present subsection, we study the power of *tests for normality* of (unobserved) *regression disturbances*. The tests we consider are the same as those described in Subsection 4.5.1, but we computed them with estimated regression residuals rather than the true disturbances u_i. We denote these by $\sqrt{\hat{b}_1}$, $\hat{b}_2$, $\hat{D}^*$, $\hat{R}$, $\hat{W}$, $\hat{W}'$ and LM_N. The first six are the modified large-sample tests discussed in White and MacDonald (1980). The seventh test is the LM test suggested in

Section 4.4. The modified Shapiro-Wilk test, $\hat{W}$, has been reported to be superior to modified distance tests, so these were excluded (see Huang and Bolch (1974, p.334)). □

We consider a linear model with a constant term and three additional regressors, i.e. with $K = 4$, and utilize the OLS residuals $\hat{u}_i$ to compute the modified tests. (Huang and Bolch (1974) and Ramsey (1974, p.36) have found that the power of modified normality tests, computed using OLS residuals, is higher than when using Theil's (1971, p.202) BLUS residuals.) To obtain $\hat{u}_i$ we use the same u_i's as those generated in Subsection 4.5.1. For comparison purposes, our regressors $X_1,\ldots,X_K$ are defined as in White and MacDonald (1980, p.20), i.e., we set $X_{i1} = 1$ $(i = 1,\ldots,N)$ and generate X_2, X_3 and X_4 from a Uniform distribution. [The specific values of the means and variances of these regressors have no effect on the simulation results. This invariance property follows from the fact that, for a linear model with regressor matrix $X = (x_1,\ldots,x_N)'$, the OLS residuals are the same as those of a linear model with regressor matrix XR, where R is any K by K non-singular matrix of constants (see Weisberg (1980, p.29))]. For $N = 20$ we use the first 20 of the 300 (generated) observations x_i. Similarly for $N = 35$, 50, 100 and 200. □

For this part of the study we utilize the same significance points as those of Subsection 4.5.1, except for $\hat{W}$, $\hat{W}'$ and LM_N, for which we use the points corresponding to $\hat{\alpha} = .10$ (e.g., as significance point of $\hat{W}$ we use $\hat{W}(25)$, where $\hat{W}(25)$ is the 25'th largest of the values of $\hat{W}$ in the 250 replications under Normal disturbances). The *estimated power* of each test is given in Table 4.4

TABLE 4.4

Normality of Disturbances

Estimated *power* with 250 replications ($\alpha = .10$)

(K=4) (Regressors: $X_1 = 1$; $X_2, X_3, X_4 \sim$ uniform) (N varies)

	$\sqrt{\hat{b}_1}$	$\hat{b}_2$	$\hat{D}^*$	$\hat{R}$	$\hat{W}$	$\hat{W}'$	LM_N
N = 20							
Normal	.084	.100	.140	.100	.100	.100	.100
Beta	.068	.108	.096	.100	.124	.072	.084
Students t	.224	.192	.188	.204	.168	.192	.256
Gamma	.640	.356	.416	.572	.644	.600	.644
Lognormal	.920	.844	.904	.912	.924	.932	.944
N = 35							
Normal	.108	.092	.120	.120	.100	.100	.100
Beta	.128	.164	.124	.184	.216	.128	.116
Students t	.292	.340	.332	.324	.236	.340	.360
Gamma	.804	.544	.708	.856	.872	.872	.892
Lognormal	1.000	.968	.988	1.000	.996	1.000	1.000
N = 50							
Normal	.092	.084	.088	.084	.100	.100	.100
Beta	.160	.180	.148	.212	.344	.172	.188
Students t	.360	.388	.400	.412	.300	.456	.464
Gamma	.984	.724	.856	.976	.988	.988	.988
N = 100							
Normal	.100	.096	.108	.108		.100	.100
Beta	.244	.416	.296	.512		.496	.536
Students t	.444	.648	.664	.628		.676	.724
Gamma	1.000	.940	.988	1.000		1.000	1.000
N = 200							
Normal	.088	.128	.132	.112		.100	.100
Beta	.520	.788	.592	.872		.924	.928
Students t	.532	.824	.820	.808		.820	.828
Gamma	1.000	.996	1.000	1.000		1.000	1.000
N = 300							
Normal	.108	.116	.124	.088		.100	.100
Beta	.740	.932	.780	.964		.992	.992
Students t	.540	.984	.984	.980		.972	.988
Gamma	1.000	1.000	1.000	1.000		1.000	1.000

For N = 20 we find that probably the best tests are LM_N and $\hat{W}$, followed by $\hat{W}'$ and $\sqrt{\hat{b}_1}$. For N = 35 and 50 we obtain that LM_N and $\hat{W}$ are the best, and that these are followed by $\hat{W}'$. For N = 100, 200 and 300, LM_N has highest power for all distributions and we see that the $\hat{W}'$ test performs quite well also. Our results agree with those of White and MacDonald (1980) in that - in almost all the cases - the modified tests give, correspondingly, lower powers than those using the original disturbances; these power differences diminish as N increases (compare Tables 4.2 and 4.4). We also find that, for a given N and a given distribution, the ranking of the tests in Table 4.2 is approximately the same as that of Table 4.4. To obtain a measure of closeness between the true and modified statistics we computed their correlation. The numerical results are given in Table 4.5. Our findings agree with those of White and MacDonald (1980, p.22): $\sqrt{\hat{b}_1}$ appears to be closer to $\sqrt{b_1}$; and $\hat{b}_2$ appears to be closer to b_2, than the other modified statistics. In our study, these would be followed by $(D^*, \hat{D}^*)$ and then by (LM, LM_N). We would then have $(W', \hat{W}')$ and, lastly, $(W, \hat{W})$. □

A *further comment* is required. It is clear that the OLS residuals $\hat{u} = (I-Q_X)u$ are a linear transformation (defined by Q_X) of the unobserved disturbances u, where $\hat{u} = (\hat{u}_1,\ldots,\hat{u}_N)'$, $u = (u_1,\ldots,u_N)'$ and Q_X is an N by N matrix defined by $Q_X = X(X'X)^{-1}X'$. As noted by White and MacDonald (1980) and Weisberg (1980), simulation results studying the relative power of tests for the normality of u - computed using $\hat{u}$ - depend on the particular form of Q_X. If one is to carry out a Monte Carlo study then, to have a less restrictive result, one should consider various forms of Q_X. Different forms may arise due to changes in N; due to variations in the way the regressors $X_1,\ldots,X_K$

TABLE 4.5

Estimated *correlations* between true and modified statistics

(K=4) (Regressors: $X_1 = 1$; $X_2, X_3, X_4 \sim$ uniform) (N varies)

	$(\sqrt{b_1}, \sqrt{\hat{b}_1})$	$(b_2, \hat{b}_2)$	$(D^*, \hat{D}^*)$	$(W, \hat{W})$	$(W', \hat{W}')$	(LM, LM_N)
N = 20						
Normal	.754	.718	.669	.532	.594	.726
Beta	.679	.580	.556	.283	.326	.321
Students t	.874	.829	.794	.740	.778	.821
Gamma	.803	.862	.796	.674	.724	.787
Lognormal	.816	.894	.764	.688	.728	.816
N = 35						
Normal	.893	.821	.833	.704	.768	.786
Beta	.839	.819	.804	.630	.664	.782
Students t	.950	.944	.925	.904	.921	.953
Gamma	.925	.959	.916	.845	.877	.946
Lognormal	.957	.974	.870	.834	.859	.961
N = 50						
Normal	.902	.837	.844	.747	.790	.881
Beta	.898	.829	.842	.726	.782	.790
Students t	.969	.974	.958	.952	.961	.994
Gamma	.942	.968	.932	.862	.892	.969
Lognormal	.980	.987	.922	.892	.910	.982
N = 100						
Normal	.956	.929	.932		.844	.907
Beta	.944	.924	.927		.841	.864
Students t	.989	.989	.979		.984	.983
Gamma	.980	.990	.962		.936	.991
Lognormal	.995	.997	.954		.951	.995
N = 200						
Normal	.976	.972	.972		.926	.935
Beta	.968	.954	.955		.933	.949
Students t	.996	.997	.992		.995	.997
Gamma	.991	.996	.981		.961	.998
Lognormal	.998	.999	.971		.973	.999
N = 300						
Normal	.980	.973	.975		.926	.941
Beta	.968	.957	.964		.954	.958
Students t	.997	.998	.995		.996	.999
Gamma	.994	.997	.987		.966	.998
Lognormal	.999	.999	.978		.980	.999

are generated; and/or due to changes in the number of regressors K.

So far we have studied the power of the tests for *different values* of N, using $K = 4$ and generating the regressors as in White and MacDonald (1980). This was done to compare our results with theirs. In addition, we have repeated our experiments but *generating the regressors in a different way*. We set $X_{i1} = 1$ $(i = 1,\ldots,N)$ and generated X_2 from a Normal, X_3 from a Uniform and X_4 from a χ^2_{10}. These regressor-distributions are of interest because they are commonly found in cross-sectional studies (we shall also use this regressor set in future parts of the thesis). The numerical results are presented in Tables 4.6 and 4.7 (Tables 4.6 to 4.13 are in Appendix B, page 280). Our findings do not vary substantially from those stated for the White and MacDonald regressor set. LM_N is a preferred test (together with $\hat{W}$ and $\hat{W}'$) for $N \leq 50$ and is preferable to all tests for $N \geq 100$ (see Table 4.6). The conclusions from the analysis of the correlations between the true and modified statistics are also the same (see Table 4.7).

□

As a *final exercise*, we carried out our experiment fixing $N = 20$ and using the three regressor Data Sets reported in Weisberg (1980, p.29). Following Weisberg, we *varied* K, for each Data Set, using $K = 4, 6, 8$ and 10. The numerical results are summarized in Tables 4.8-4.13. Weisberg found that the power of the $\hat{W}'$ test may vary as K and/or the regressors are changed. We find this to be the case for all the tests considered. For example, for Data Set 1 (see Table 4.8), we obtain that for the Lognormal the power of $\sqrt{\hat{b}_1}$, say $P(\sqrt{\hat{b}_1})$, is equal to .636, for $K = 10$, and .956 for $K = 4$, i.e., $.636 \leq P(\sqrt{\hat{b}_1}) \leq .956$. Similarly we obtain that $.572 \leq P(\hat{b}_2) \leq .812$; $.608 \leq P(\hat{D}^*) \leq .888$; $.632 \leq P(\hat{R}) \leq .940$; $.592 \leq P(\hat{W}) \leq .928$;

$.636 \le P(\hat{W}') \le .944$ and $.616 \le P(LM_N) \le .952$. We also find that for Data Sets 1 and 2 (see Tables 4.8 and 4.9), the empirical significance level $\hat{\alpha}$ is close to .10 for all statistics and all K. For Data Set 3, however, $\hat{\alpha}$ increased considerably as K increased (e.g., see Table 4.10 and note that for $\sqrt{\hat{b}_1}$, $\hat{\alpha}$ = .088, .104, .200 and .216 for K = 4, 6, 8 and 10 respectively). This shows that the *power and the level* of a test may depend on the specific form of Q_X. Nevertheless, when comparing the *relative power*, it is interesting to note that, for all K and all three regressor Data Sets, LM_N, $\hat{W}$, $\hat{W}'$ and $\sqrt{\hat{b}_1}$ are the preferred tests (as it was found in our earlier 2 sets of experiments with N = 20). Regarding the correlations between the true and modified statistics, we observe that (for each Data Set) as K increases, all correlations decrease (e.g., see Table 4.11 and note that for Data Set 1, the correlation between D^* and $\hat{D}^*$, for the Normal was equal to .640 when K = 4 and to .329 when K = 10). However, the ranking among the tests remains the same, i.e., from high to low correlations the order remains being $(\sqrt{b_1},\sqrt{\hat{b}_1})$, $(b_2,\hat{b}_2)$, $(D^*,\hat{D}^*)$, (LM,LM_N), $(W',\hat{W}')$ and $(W,\hat{W})$ (see Tables 4.11, 4.12 and 4.13). □

Our main result of Subsection 4.5.2 is that, for all the forms of matrices Q_X that we have studied, LM_N *performed with good relative power*. This was true for both small N (e.g., N = 20) and large N (e.g., N = 300). The above findings encourage the use of LM_N in testing for the normality of u_i. The statistic LM_N is simple to compute and, in any regression problem, we may easily obtain an approximation to its finite sample distribution, under H_o, by computer simulation. This should not represent a serious problem, particularly with the fast speed and increased availability of modern computers.

4.6 ANALYSIS OF MODELS WITH NON-NORMAL DISTURBANCES

In Section 4.4 we suggested an asymptotically efficient test for the normality of disturbances, and in Section 4.5 we noted its good power - relative to other existing tests - even for very small sample sizes. If the hypothesis that disturbances are normally distributed (H_o) is accepted, *classical econometric analysis* may be carried out.[5] In this section, we make a very brief comment on possible procedures to follow if H_o is rejected.

Some econometricians consider that disturbances represent the sum of effects of omitted variables, and justify the normality assumption by making an appeal to a central limit theorem (e.g., see Johnston (1972, p.11)). Others regard normality as an assumption made for computational convenience, arguing it gives a quasi-likelihood function. Whatever the motivations underlying the assumption, typically, if H_o is rejected, no alternative disturbance distribution would exist in an econometricians' mind.[6] Under these circumstances one may consider - as an *alternative* to Least Squares estimation - the use of

(1) Robust estimation on the linear model,

(2) Estimation within a Family of transformations, or

(3) Estimation within a Family of distributions.

We shall - very briefly - describe these.

---

5 In particular, we could regard OLS estimators as MLE. It is difficult to determine the properties of estimators after a disturbance specification test has been carried out. Apparently this problem has not been dealt with in the econometric and statistical literature. As it is common practice (e.g., see Malinvaud (1980, footnote in pg.292)) we do not pursue this point, and proceed with our presentation disregarding the effect of the pretest.

6 If one had a specific alternative p.d.f. for u_i, one could write down the log-likelihood, obtain MLE's of β, and use the general theory of MLE to obtain approximate significance tests. (See Subsection 4.6.3).

4.6.1 Robust Estimation

Consider a situation where, after rejection of H_o, we decided to proceed analysing the linear model, but regarding the disturbances as non-normal. If u_i is non-normal, but its variance is finite, the OLS estimator b would still be BLUE. Furthermore, under fairly general conditions, we could apply - for large samples - the usual t and F tests (e.g., see Arnold (1980)). Then it would appear that - particularly for large samples - there is strong reason for the use of OLS estimates even when disturbance normality is rejected. Yet, some statisticians question the 'appropriateness' of the OLS estimator by arguing the class of linear estimators is too restrictive, and the unbiasedness property as being of doubtful value (e.g., see Hampel (1973, p.90)). Others question the use of the OLS estimator because of the finding that it may be quite sensitive to outliers and long-tailed distributions (e.g., see Hogg (1979)). These considerations - among others - have contributed to the development of alternative estimation methods.

An important class of these alternative methods goes under the rubric of *robust regression*. As stated by Hill and Holland (1977, p.828),

> "the emphasis in robust regression is on methods which are not sensitive to deviations from normal distributions and to the effects of outliers in the data".

Work in this area is extensive and the literature is voluminous. This is evident from the five-part article of H. Leon Harter entitled '*The method of Least Squares and some alternatives*', which includes some historical accounts and a survey of recent developments (see Harter (1974,1975)). A good, brief illustration of some of these techniques is found in the textbooks of Maddala (1977, p.308) and Malinvaud (1980, p.318).

4.6.2 Estimation within a Family of Transformation

Analysis of linear models with non-normal disturbances is complicated. This has led people to consider the use of *transformations* applied to the measured variables y_i so that - in particular - disturbance normality is better suited.[7] More formally, if u_i^o denotes the i'th disturbance from a transformed model; then we would want the transformation to be such that u_i^o is normally distributed and, at the same time, to have (as was assumed for u_i) $E[u_i^o|x_i] = 0$, and $E[u_i^{o2}] = \sigma_o^2$ for all $i = 1,\dots,N$.

Several transformations exist that aim to achieve this, and a popular one is the Box and Cox (1964) transformation. This defines $y_i^{(\lambda)} = (y_i^\lambda - 1)/\lambda$ for $\lambda \neq 0$; and $y_i^{(\lambda)} = \log y_i$ for $\lambda \to 0$, where λ is an unknown parameter. If the Box-Cox transformation is applied, the transformed model would be

$$y^{(\lambda)} = X\beta^o + u^o \quad ,$$

where $y^{(\lambda)} = (y_1^{(\lambda)},\dots,y_N^{(\lambda)})$ is the vector of the N transformed observations; β^o is a K by 1 vector of unknown parameters associated with $y^{(\lambda)}$, and u^o is a vector that contains the N disturbances u_i^o.

---

7 Data transformation is considered here as a statistical device to achieve disturbance normality, so one may consider applying transformations to the dependent variable only. Transformations may also be applied to the variables x_i (e.g., see Box and Tidwell (1962)), and are a tool used in the choice of regression functional form (e.g., see Zarembka (1968,1974)). In addition, transformations provide a convenient framework for the derivation of specification tests (see Savin and White (1978), Godfrey and Wickens (1981b), and Bera and Jarque (1981b)).

A fundamental assumption of the Box-Cox transformation is that, for some λ, the disturbances $u_1^o,\ldots,u_N^o$ *'can be treated as'* N normally distributed variables with mean zero and finite variance σ_o^2.[8] This allows writing the log-likelihood, from which MLE's for λ, β^o and σ_o^2, and corresponding standard errors can be obtained. Before concluding we note that - in econometrics - our usual interest is to explain Y (the 'measured variable'), and not some function of Y. In this case, as stated by Box and Cox (1964, p.214)

> "we either analyse linearly the untransformed data or, if we do apply a transformation in order to make a more efficient and valid analysis, we convert the conclusions back to the original scale".

In particular, interest could reside in computing elasticities, and for this we may proceed as in Savin and White (1978, p.3).

4.6.3 Estimation within a Family of Distributions

A final alternative to OLS estimation, that we shall briefly comment on, is MLE under a particular *family of distributions*. For example, an attempt may be made to obtain the MLE's of β, c_o, c_1 and c_2 for the *Pearson Family* - unfortunately this is difficult. Indeed, no results are presently available (even for the non-regression case), and this is an area that requires further study (Pearson distributions are usually fitted by the method of moments, e.g., see Kendall and Stuart (1969, p.152)). However, other computationally more manageable

---

[8] The Box-Cox transformation is well defined for positive values. This implies limits on $y_i^{(\lambda)}$, which means - strictly speaking - that u_i^o cannot be normal (see Poirier (1978)). In the various applications of the transformation, this *'truncation problem'* has been typically neglected. This approach is not entirely incorrect when truncation is not severe.

families may be considered. For instance, Goldfeld and Quandt (1980) derive maximum likelihood estimates under the *Sargan Family* of distributions (see also Zeckhauser and Thompson (1970) and Anscombe (1967)). An advantage of this approach (as the one described in Subsection 4.6.2) is that the general theory of MLE could be used to obtain approximate significance tests.

4.7 CONCLUDING REMARKS

Throughout this Chapter we have discussed normality under the assumption that the variance of the disturbances was constant. In Chapter 5 we study the *problem of heteroscedasticity*, firstly, under normal disturbances, and then we consider both problems jointly. In particular, we extend the procedure of Section 4.4 to derive a joint test for disturbance normality and homoscedasticity.

We end by noting there are cases in applied econometrics where we would not carry out a disturbance normality test because, by the nature of the model, normality would not hold. This is the case, for example, when estimating a frontier production function (e.g., see Maddala (1977, p.317)). Another example arises when y_i is restricted, say, to non-negative values, i.e., $y_i = x_i'\beta + u_i \geq 0$. Then the range of u_i would be restricted to $u_i \geq -x_i'\beta$ and, therefore, disturbance normality could not hold. These types of *limited dependent variable models* are studied in Chapter 6.

Appendix A

Derivations of Sections 4.3 and 4.4

PROPOSITION 1: The LM test statistic for testing the *normality of observations* is given by equation (4.4).

Proof:

The problem of testing the normality of the observations $\nu_1,\ldots,\nu_N$, is the same as testing the normality of the 'regression disturbances' $u_1,\ldots,u_N$ in the regression model $y_i = x_i'\beta + u_i$, with K (the dimension of x_i') equal to one, $x_i = 1$, $\beta = \mu$ and $y_i = \nu_i$. So, the proof of this proposition is a particular case of the proof given for Proposition 2 (see below). Note that, in the present case, the MLE of β (i.e., μ) under H_o: $c_1 = c_2 = 0$ is $\hat{\beta} = \hat{\mu} = (\nu_1 + \ldots + \nu_N)/N = \bar{\nu}$, which implies that $\hat{u}_i = \nu_i - \bar{\nu}$, and that $\hat{\mu}_1 = (\hat{u}_1 + \ldots + \hat{u}_N)/N = 0$.

□

PROPOSITION 2: The LM test statistic for testing the *normality of regression disturbances*, LM_N, is given by equation (4.5).

Proof:

Define

$$\phi(\theta,u_i) = \int \frac{c_1-u_i}{c_o-c_1u_i+c_2u_i^2}\,du_i ,$$

$v_i = c_o - c_1u_i + c_2u_i^2$, $\theta_1 = (\beta',c_o)'$, $\theta_2 = (c_1,c_2)'$ and $\theta = (\theta_1',\theta_2')'$. Then the log-likelihood for the i'th observation can be written as

$$\ell_i(\theta) = -\log\left[\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)]du_i\right] + \phi(\theta,u_i) .$$

We can show that

$$\frac{\partial\ell_i(\theta)}{\partial\beta} = -\frac{x_i\int_{-\infty}^{\infty}\exp[\phi(\theta,u_i)]\left[\int \frac{v_i-(c_1-u_i)(c_1-2c_2u_i)}{v_i^2}du_i\right]du_i}{\int_{-\infty}^{\infty}\exp[\phi(\theta,u_i)]du_i}$$

$$+\; x_i\int \frac{v_i-(c_1-u_i)(c_1-2c_2u_i)}{v_i^2}\,du_i$$

$$\frac{\partial\ell_i(\theta)}{\partial c_o} = -\frac{\int_{-\infty}^{\infty}\exp[\phi(\theta,u_i)]\left[\int -\frac{(c_1-u_i)}{v_i^2}\,du_i\right]du_i}{\int_{-\infty}^{\infty}\exp[\phi(\theta,u_i)]du_i} + \int -\frac{(c_1-u_i)}{v_i^2}\,du_i$$

$$\frac{\partial \ell_i(\theta)}{\partial c_1} = - \frac{\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)] \left[\int \frac{v_i - (c_1 - u_i)(-u_i)}{v_i^2} du_i \right] du_i}{\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)] du_i}$$

$$+ \int \frac{v_i - (c_1 - u_i)(-u_i)}{v_i^2} du_i$$

$$\frac{\partial \ell_i(\theta)}{\partial c_2} = - \frac{\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)] \left[\int \frac{-(c_1 - u_i)u_i^2}{v_i^2} du_i \right] du_i}{\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)] du_i} + \int \frac{-(c_1 - u_i)u_i^2}{v_i^2} du_i \ .$$

Setting $c_1 = c_2 = 0$ in the above expressions, and noting that *under normality* $E[u_i] = E[u_i^3] = 0$ and $E[u_i^4] = 3c_o^2$, we obtain

$$\frac{\partial \ell_i(\theta)}{\partial \beta} = \frac{x_i u_i}{c_o} \quad ,$$

$$\frac{\partial \ell_i(\theta)}{\partial c_o} = - \frac{1}{2c_o} + \frac{u_i^2}{2c_o^2} \quad ,$$

$$\frac{\partial \ell_i(\theta)}{\partial c_1} = \frac{u_i}{c_o} - \frac{u_i^3}{3c_o^2} \quad \text{and} \tag{A.1}$$

$$\frac{\partial \ell_i(\theta)}{\partial c_2} = - \frac{3}{4} + \frac{u_i^4}{4c_o^2} \quad .$$

(In particular, note that when $c_1 = c_2 = 0$, we have $\int_{-\infty}^{\infty} \exp[\phi(\theta,u_i)] du_i = \sqrt{(2\pi c_o)}$).

Adding (A.1) from $i = 1$ to N, and evaluating the resulting quantities at the MLE of θ under H_o: $c_1 = c_2 = 0$ (i.e. setting $\beta = \hat{\beta} = (X'X)^{-1}X'y$ and $c_o = \hat{\sigma}^2$) we obtain

$$\hat{d}_2 = \left[\sum_{i=1}^{N} \frac{\partial \hat{\ell}_i(\theta)}{\partial c_1}, \sum_{i=1}^{N} \frac{\partial \hat{\ell}_i(\theta)}{\partial c_2}\right]'$$

$$= N\left[\frac{\hat{\mu}_1}{\hat{\mu}_2} - \frac{\hat{\mu}_3}{3\hat{\mu}_2^2}, -\frac{3}{4} + \frac{\hat{\mu}_4}{4\hat{\mu}_2^2}\right]' , \qquad (A.2)$$

where $\hat{\mu}_j = \sum_{i=1}^{N} \hat{u}_i^j/N$ and $\hat{u}_i = y_i - x_i'\hat{\beta}$.

Now we use (A.1) to compute $\psi(\theta) = \sum_{i=1}^{N} (\partial\ell_i(\theta)/\partial\theta)(\partial\ell_i(\theta)/\partial\theta)'$, obtaining

$$\psi(\theta) = N \left[\begin{array}{cc|cc} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} \\ \psi_{12}' & \psi_{22} & \psi_{23} & \psi_{24} \\ \hline \psi_{13}' & \psi_{23} & \psi_{33} & \psi_{34} \\ \psi_{14}' & \psi_{24} & \psi_{34} & \psi_{44} \end{array}\right] ,$$

with

$$\psi_{11} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i x_i' u_i^2}{c_o^2}$$

$$\psi_{12} = \frac{1}{N} \sum_{i=1}^{N} x_i \left(-\frac{u_i}{2c_o^2} + \frac{u_i^3}{2c_o^3}\right)$$

$$\psi_{13} = \frac{1}{N} \sum_{i=1}^{N} x_i \left(\frac{u_i^2}{c_o^2} - \frac{u_i^4}{3c_o^3}\right)$$

$$\psi_{14} = \frac{1}{N} \sum_{i=1}^{N} x_i \left(-\frac{3u_i}{4c_o} + \frac{u_i^5}{4c_o^3}\right)$$

$$\psi_{22} = \frac{1}{4c_o^2} + \frac{\mu_4}{4c_o^4} - \frac{\mu_2}{2c_o^3}$$

$$\psi_{23} = -\frac{\mu_1}{2c_o^2} + \frac{2\mu_3}{3c_o^3} - \frac{\mu_5}{6c_o^4}$$

$$\psi_{24} = \frac{3}{8c_o} - \frac{\mu_4}{8c_o^3} - \frac{3\mu_2}{8c_o^2} + \frac{\mu_6}{8c_o^4}$$

$$\psi_{33} = \frac{\mu_2}{c_o^2} - \frac{2\mu_4}{3c_o^3} + \frac{\mu_6}{9c_o^4}$$

$$\psi_{34} = \frac{\mu_5}{4c_o^3} + \frac{\mu_3}{4c_o^2} - \frac{\mu_7}{12c_o^4} - \frac{3\mu_1}{4c_o}$$

$$\psi_{44} = \frac{9}{16} + \frac{\mu_8}{16c_o^4} - \frac{6\mu_4}{16c_o^2}$$

where

$$\mu_j = \sum_{i=1}^{N} u_i^j/N .$$

Taking the expectation of $\psi(\theta)$ (noting that - *under normality* - $E[u_i] = E[u_i^3] = E[u_i^5] = E[\mu_1] = E[\mu_3] = E[\mu_5] = E[\mu_7] = 0$, $E[u_i^2] = E[\mu_2] = c_o$, $E[u_i^4] = E[\mu_4] = 3c_o^2$, $E[\mu_6] = 15c_o^3$ and $E[\mu_8] = 105c_o^4$), we obtain

$$I = E[\psi(\theta)] = N \left[\begin{array}{cc|cc} X'X/c_o & 0 & 0 & 0 \\ 0 & 1/(2c_o^2) & 0 & 3/(2c_o) \\ \hline 0 & 0 & 2/(3c_o) & 0 \\ 0 & 3/(2c_o) & 0 & 6 \end{array}\right] .$$

After evaluating I at $\theta = \hat{\theta}$ (i.e. setting $c_o = \hat{\sigma}^2 = \hat{\mu}_2$) we obtain

$$(\hat{I}_{22}-\hat{I}_{21}\hat{I}_{11}^{-1}\hat{I}_{12})^{-1} = \frac{1}{N}\begin{bmatrix} 3\hat{\mu}_2/2 & 0 \\ 0 & 2/3 \end{bmatrix} . \qquad \text{(A.3)}$$

Using (A.3) and (A.2) in (4.1) we obtain our result. □