A LACK-OF-FIT TEST IN THE PRESENCE OF HETEROSEDASTICITY

Howard E. Doran and Jan Kmenta *

No. 12 - April 1981

* Kmenta is in the Department of Economics at the University of Michigan. Most of this work was done while the authors were Fellows of the Alexander von Humboldt Foundation at the University of Bonn.

ISSN 0157-0188
ISBN 0 95834 378 9
1. Introduction

The lack-of-fit test is a model specification test which can be applied when there are replicated observations on the dependent variable. Basically, the test utilises the information that comes from within-group variation. As this situation is normal in experimental work, the test is well-known to experimental statisticians, but appears to be almost unknown to econometricians. A survey by these authors of econometric text books has revealed no mention of the test. This is almost certainly due to the strong emphasis in econometrics on the methodology applicable to time-series data, in which each observation on the regressor variables corresponds to a single observation on the dependent variable. By contrast, when cross-sectional data are analysed, there are typically many units (individuals, firms, families) which are characterised by the same values of explanatory variables (e.g. incomes, prices, rainfall, etc.). In this situation a lack-of-fit test could often profitably be applied as an aid to appropriate model specification.

The classical lack-of-fit test, however, assumes homoscedastic observations, whereas heteroscedasticity is an almost universal characteristic of cross-sectional data.

This article is written for two purposes. Firstly, to draw the attention of econometricians to the possibilities which the test offers, (see also Battese (1977)) and secondly, to generalise the test to cater for heteroscedasticity.

In section 2 the classical lack-of-fit test is developed, this is generalised in section 3 to a large sample test appropriate for heteroscedastic data, and some concluding remarks are made in section 4.

2. The Lack-of-Fit Test

Let us suppose that an Nxl random vector y is normally distributed with mean $\mu$ and variance $\sigma^2 I$, and that a linear model is specified which implies that
\[ \mu = X\beta, \quad (2.1) \]

where \( X \) and \( \beta \) are \( NxK \) and \( Kx1 \), respectively. Thus, if this specification is correct, it follows that

\[ M\mu = 0, \quad (2.2) \]

where \( M \) is the idempotent projection matrix given by

\[ M = I_N - X(X'X)^{-1}X'. \quad (2.3) \]

For the purposes of this paper we will consider a \textit{misspecification} to have occurred when

\[ M\mu \neq 0. \quad (2.4) \]

That is, the model is misspecified when the vector \( \mu \) does not lie completely in the linear space spanned by \( X \).

We quote at this point a fundamental result concerning quadratic forms, which will be used in what follows. Let \( A \) be any \( NxN \) matrix. Then

\[ E(y' Ay) = \sigma^2 \text{tr} A + \mu'A\mu. \quad (2.5) \]

Furthermore, if \( A \) is idempotent,

\[ \mu'A\mu = (A\mu)'(A\mu) \geq 0, \]

and it follows that

\[ E(y' Ay) \geq \sigma^2 \text{tr} A, \quad (2.6) \]

with the equality holding if, and only if, \( A\mu = 0 \).
In particular, the expectation of the residual sum of squares (RSS) obtained by regressing \( y \) on \( X \) is given by

\[
E(\text{RSS}) = E(y'My) \geq \sigma^2 \text{tr } M
\]

with the equality occurring if and only if \( M\mu = 0 \). Thus, the usual estimator of \( \sigma^2 \), denoted by \( \hat{\sigma}^2 \) and defined by

\[
\hat{\sigma}^2 = \frac{\text{RSS} \text{tr } M = \text{RSS}/(N-K)}
\]

is positively biased if the model is misspecified.

If there is only one observation on \( y \) corresponding to each observation on the regressor variables, this property of \( \hat{\sigma}^2 \) has no practical use. On the other hand, if there are replications, it is possible to utilize this property to detect model misspecification.

Let us assume now that there are \( t \) distinct observations on the regressor variables, and that corresponding to the \( i \)-th observation \( x_i \) there are \( n_i \geq 1 \) observations on the dependent variable. Clearly \( \sum_{i=1}^{t} n_i = N \). We will denote by \( y_{ij} \) the \( j \)-th observation (\( j = 1, 2, \ldots, n_i \)) from the \( i \)-th group (\( i = 1, 2, \ldots, t \)) and assume the observations are ordered so that those from the first group (\( i = 1 \)) are followed by those from the second (\( i = 2 \)), and so on. We define the obvious partitioning

\[
y = (y_1', y_2', \ldots, y_t')', \quad \text{where } y_i \text{ is the } n_i \times 1 \text{ vector of dependent variable observations from the } i\text{-th group, and denote by } 1_i \text{ the column of ones of length } n_i.
\]

Then defining the \( N \times t \) block diagonal matrix \( V \) by

\[
V = \text{diag } (1_1, 1_2, \ldots, 1_t),
\]

the matrix \( X \) may be written in the form
\[ X = VX^*, \quad (2.9) \]

where \( X^* = (x_1, x_2, \ldots, x_t)' \) is the \( t \times K \) matrix of distinct observations on the regressor variables. We note here two useful properties of \( V \):

(i) \[ V'V = \text{diag} (n_1, n_2, \ldots, n_t) = W'W, \quad (2.10) \]

where \( W = \text{diag} (n_1^2, n_2^2, \ldots, n_t^2) \).

(ii) \[ V'y = (\Sigma y_1, \Sigma y_2, \ldots, \Sigma y_t)' = (V'V)\vec{y}, \quad (2.11) \]

where \( \vec{y} = (\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_t)' \) is the vector of group sample means.

Consider now the following partitioning of the residual sum of squares:

\[ \text{RSS} = y'My = y'Hy + y'(M-H)y, \quad (2.12) \]

where

\[ H = \text{diag} (H_1, H_2, \ldots, H_t), \quad (2.13) \]

and

\[ H_i = [I_{n_i} - \frac{1}{n_i} \frac{1}{n_i}] \quad (i = 1, 2, \ldots, t). \quad (2.14) \]

It is easily seen that \( H \) is idempotent and \( HV = 0 \), which implies by (2.9) that

\[ HX = 0. \quad (2.15) \]

Also,

\[ (I-H)y = V\vec{y}. \quad (2.16) \]

---

(1) If \( n_i = 1 \), then \( H_i = 0 \). Thus, if there are no replications \( (n_i = 1 \) for all \( i \) \), \( H = 0 \) and the partitioning (2.12) cannot be made.
The partitioning (2.12) can be interpreted as follows:

(i) \[ y'Hy = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \]

This factor gives the total within group variation, and is called the pure error sum of squares (PSS).

(ii) \[ y'(M-H)y = y'(I-H)y - y'(I-M)y \]
\[ = \bar{y}'(Y'Y)\bar{y} - \bar{y}'(Y'Y)X^\delta(X'^\delta Y'YWX^\delta)^{-1}X'^\delta(Y'Y)\bar{y} \]

from (2.3), (2.16), (2.9) and (2.11).

Thus, by (2.10)

\[ y'(M-H)y = (W\bar{y})'[I-WX^\delta(X'^\delta WX^\delta)^{-1}(WX^\delta)'](W\bar{y}), \]

which is seen to be the residual sum of squares which would be obtained if a weighted regression of \( \bar{y} \) on \( X^\delta \) were carried out with each observation weighted by the square root of the corresponding group size. This factor is known as the lack-of-fit sum of squares (LSS).

The partitioning (2.12) may thus be written in the form

\[ \text{RSS} = \text{PSS} + \text{LSS}. \quad (2.17) \]

We have already defined \( \hat{\sigma}^2 = \text{RSS}/\text{tr} \ M \). Thus, on dividing (2.17) by \( \text{tr} \ M \), we obtain

\[ \hat{\sigma}^2 = \alpha \hat{\sigma}_1^2 + (1-\alpha) \hat{\sigma}_2^2, \quad (2.18) \]

where

\[ \hat{\sigma}_1^2 = \text{PSS}/\text{tr} \ H = \text{PSS}/(N-t), \quad (2.19) \]

\[ \hat{\sigma}_2^2 = \text{LSS}/\text{tr}(M-H) = \text{LSS}/(t-K), \quad (2.20) \]

and \( \alpha = \text{tr} \ H/\text{tr} \ M = (N-t)(N-K) \).
Provided \( t < K < N \), \( 0 < \alpha < 1 \) and (2.18) expresses \( \hat{\sigma}^2 \) as a weighted average of \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \). The estimator \( \hat{\sigma}^2 \) will be unbiased if both \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \) are unbiased estimators of \( \sigma^2 \).

We now assume that even if the model is misspecified, observations from within a group have the same mean. It follows then that

\[
H_\mu = 0. \tag{2.21}
\]

The implication of this assumption is the following:

Suppose the truth is

\[
\mu = X\beta + Z\gamma, \quad (MZ \neq 0),
\]

then the within-group means will only be the same if the observations \( Z \) are a function of \( X \).

Now from (2.5) and (2.21)

\[
E(\hat{\sigma}_1^2) = E(y'Hy)\text{tr }H = \sigma^2 + \mu'H\mu/\text{tr }H = \sigma^2,
\]

and so even under a misspecification, \( \hat{\sigma}_1^2 \) is an unbiased estimator of \( \sigma^2 \).

Again, by (2.5)

\[
E(\hat{\sigma}_2^2) = E(y'(M-H)y)/\text{tr}(M-H)
\]

\[
= \sigma^2 + [(M-H)\mu]'[(M-H)\mu]/\text{tr}(M-H)
\]

\[
= \sigma^2 + (M\mu)'(M\mu)/\text{tr}(M-H).
\]

Thus, \( \hat{\sigma}_2^2 \) will be an unbiased estimator only if the model is correctly specified \( (M\mu = 0) \), and under a misspecification will tend to overestimate \( \sigma^2 \).

The basis for the lack-of-fit test is now clear: \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \) are compared, and if \( \hat{\sigma}_2^2 \) is 'large relative to \( \hat{\sigma}_1^2 \)', misspecification is suspected.

The lack-of-fit test may be formally derived using well-known results on the distribution of quadratic forms (see for example, Graybill (1961, pp.83-85)) as follows:

(2)

From now on we will assume this restriction holds.
PSS/$\sigma^2 = y'Hy/\sigma^2 \sim \chi^2(k_1, \lambda_1)$

where

$$k_1 = \text{tr} \ H = N-t,$$

$$\lambda_1 = \mu'H\mu/2\sigma^2 = 0.$$

Thus,

$$\text{PSS}/\sigma^2 \sim \chi^2(N-t). \quad (2.22)$$

Also,

$$\text{LSS}/\sigma^2 = y'(M-H)y/\sigma^2 \sim \chi^2(k_2, \lambda_2),$$

where

$$k_2 = \text{tr}(M-H) = t-K,$$

$$\lambda_2 = \mu'(M-H)\mu/2\sigma^2 = \mu'M\mu/2\sigma^2. \quad (2.23)$$

Thus,

$$\text{LSS}/\sigma^2 \sim \chi^2(t-K, \lambda_2), \quad (2.24)$$

and if the model is correctly specified ($M\mu = 0$), then $\lambda_2 = 0$ and

$$\text{LSS}/\sigma^2 \sim \chi^2(t-K). \quad (2.25)$$

Furthermore, it is easily shown that

$$H(M-H) = 0,$$

and hence, the quadratic forms $\text{PSS}/\sigma^2$ and $\text{LSS}/\sigma^2$ are independent. It follows immediately that if the model is correctly specified,

$$F = \frac{\text{LSS}/(t-K)}{\text{PSS}/(N-t)} = \frac{\hat{\lambda}_2^2}{\hat{\lambda}_1^2} \sim F_{t-K,N-t}. \quad (2.26)$$

The null hypothesis of correct specification is rejected at the $100\alpha\%$ level of significance if $\hat{\lambda}_2^2/\hat{\lambda}_1^2$ exceeds the $\alpha$ critical point of the $F_{t-K,N-t}$ distribution.
In general $\sigma^2$ is, of course, unknown. However if $\sigma^2$ were known, then a more powerful test would be obtained by using the chi-square result (2.25) rather than the $F$-distribution as in (2.26). This remark has relevance to the following section.

3. The Lack-of-Fit Test Under Heteroscedasticity

As mentioned in the introduction, the main area of usefulness (in econometrics) of the lack-of-fit test is in the analysis of cross-sectional data. In this context the underlying assumption that $E[(y-B)(y-B)'] = \sigma^2 I_N$ is unlikely to be realistic, and we will generalise it by assuming that

$$E[(y-\mu)(y-\mu)'] = \Sigma,$$  \(\text{(3.1)}\)

where

$$\Sigma = \text{diag}(\sigma_1^2 I_{n_1}, \sigma_2^2 I_{n_2}, \ldots, \sigma_t^2 I_{n_t})$$

and the $\sigma_i^2$ are known. Then, defining

$$D = \text{diag}(\sigma_1^2 I_{n_1}, \sigma_2^2 I_{n_2}, \ldots, \sigma_t^2 I_{n_t})$$

it follows that

$$D^{-1}y \sim N(D^{-1}\mu, I_N)$$

and the linear specification on the mean takes the form

$$D^{-1}\mu = (D^{-1}X)\beta.$$  \(\text{(3.2)}\)

Thus, provided the analysis is carried out in terms of the weighted observations $y_{ij}/\sigma_i$ and $x_{ij}/\sigma_i \ (i = 1, 2, \ldots, t)$, the test described in the previous section follows through, with one difference - $\sigma^2$ is now known, and equal to unity. Thus, by (2.25), if the model is correctly specified,
LSS* \sim \chi^2(t-K), \quad (3.3)

where we have used LSS* to emphasise that weighted observations are used.

In practice, the \( \sigma_i^2 \) would never be known and would have to be estimated. The within group variation in the observations on \( y \) provides consistent estimates of the \( \sigma_i^2 \), completely independently of the specification \( u = X\beta \). Thus, the result (3.3) is to be regarded as a large sample result; if the observations are weighted inversely by \( \hat{\sigma}_i \), then as \( n_i \to \infty \), the distribution of LSS* is given by (3.3).

It is commonly assumed in econometrics that the variances \( \sigma_i^2 \) can be related to a single variable \( s_i \) say, through a relationship of the form

\[
\sigma_i^2 = k s_i^\delta, \quad (3.4)
\]

where \( k \) and \( \delta \) are unknown parameters. Typically, \( s_i \) would be a member of the regressor vector \( x_i' \), but this need not be the case. If such a model is appropriate, and \( t > 2 \), then there will be gains in efficiency if \( k \) and \( \delta \) are estimated rather than just each \( \sigma_i^2 \) individually. We present below two methods for the estimation of \( k \) and \( \delta \).

(i) **Maximum Likelihood Estimation**

The log-likelihood of the sample is easily seen to be given by

\[
\ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{t} n_i \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^{t} \sum_{j=1}^{n_i} \left[ \frac{(y_{ij} - \mu_i)^2}{2\sigma_i^2} \right].
\]

Thus, assuming \( \sigma_i^2 = k s_i^\delta \)
\[ \ln L = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{t} n_i [\ln k + \delta \ln s_i] - \frac{1}{2} \sum_{i=1}^{t} (ks_i^\delta)^{-1} \sum_{j=1}^{n_i} (y_{ij} - u_i)^2. \]

Setting the partial differentials \( \partial \ln L / \partial u_i \) (i = 1, 2, ..., t), \( \partial \ln L / \partial k \), \( \partial \ln L / \partial \delta \) equal to zero and solving for the maximum likelihood estimates \( \tilde{u}_i \), \( \tilde{k} \) and \( \tilde{\delta} \), we obtain

\[ \tilde{k} = n^{-1} \sum_{i=1}^{t} (n_i - 1) \tilde{s}_i \tilde{\delta}_i, \]

(3.5)

and

\[ \sum_{i=1}^{t} s_i \tilde{\delta}_i(n_i - 1) \tilde{s}_i^2 [n_i s_i - \sum_{j=1}^{t} n_i s_{i,j}] = 0, \]

(3.6)

where

\[ \tilde{s}_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - \tilde{y}_i)^2. \]

(3.7)

The non-linear equation (3.6) must be solved for \( \tilde{\delta} \), and \( \tilde{k} \) is then obtained by (3.5).

(ii) **Large Sample Regression Procedure**

Because of the non-linear nature of (3.6) necessitating an iterative procedure to obtain \( \tilde{\delta} \), a regression approach to the estimation of \( k \) and \( \delta \) may be more convenient.

If we define \( \tilde{s}_i^2 \) as in (3.7) above, then

\[ v_i = (n_i - 1) \tilde{s}_i^2 / s_i^2 \sim \chi^2(n_i - 1). \]

Thus, if \( s_i^2 = ks_i^\delta \), we have

\[ \ln \tilde{s}_i^2 = \ln k + \delta \ln s_i + w_i \]

(3.8)

where \( w_i = \ln v_i - \ln(n_i - 1). \)
Bartlett and Kendall (1946) have shown that for large \( n \), \( \sqrt{n-2} w_i \) is approximately normally distributed, with mean zero and variance 2. Furthermore, they showed that this approximation is likely to be good for \( n \) as small as 10.

Thus, regression of \( \sqrt{n-2} \ln \sigma_i^2 \) on \( \sqrt{n-2} \ln \sigma_i \) and \( \sqrt{n-2} \ln s_i \) will produce consistent, asymptotically efficient estimates of \( k \) and \( \delta \).

Estimation of \( k \) and \( \delta \) by either of the above methods enables the observations to be adjusted through (3.4).

4. Summary and Concluding Remarks

The lack-of-fit test, which can be applied when there are replicated observations on the dependent variable of a regression model, has been expounded. As mentioned in the Introduction, its main area of application in non-experimental work is to the analysis of cross-sectional data. As such data is almost always characterised by heteroscedasticity, the test has been extended to take account of this. This modification has been accomplished in a completely standard way, and results in a large sample test. It is important to note that what is necessary is that the number of replications should be large, but not the number of distinct observations on the regressor variables. As this is almost invariably the case when cross-sectional data are analysed, we believe that this large sample test provides a tool which could often be profitably used as an aid to model specification.
REFERENCES

