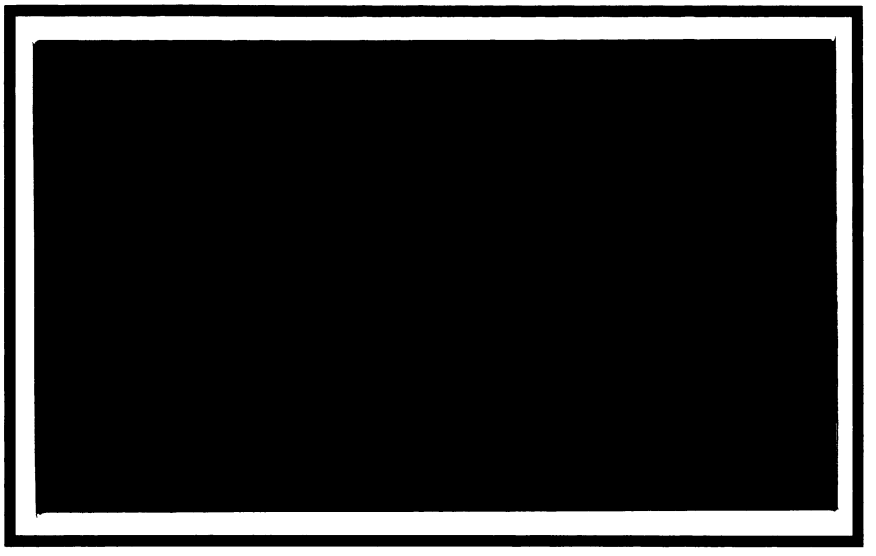
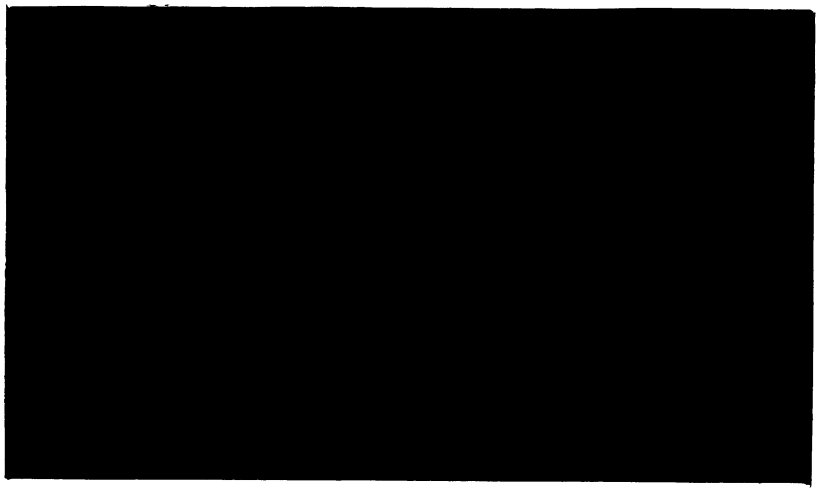
R-114.84

Center for Research on Economic and Social Theory
Research Seminar in Quantitative Economics

Discussion Paper



DEPARTMENT OF ECONOMICS
University of Michigan
Ann Arbor, Michigan 48109



A LACK-OF-FIT TEST WITH AN APPLICATION TO AN EARNINGS FUNCTION FOR WOMEN

H.E. DORAN
University of New England
Armidale, N.S.W. 2351, Australia

J. KMENTA
University of Michigan
Ann Arbor, Michigan 48109, USA

R-114.84

April, 1984

A LACK-OF-FIT TEST WITH AN APPLICATION TO AN EARNINGS FUNCTION FOR WOMEN*

This paper presents a lack-of-fit test of model specification used by experimental statisticians but mostly unknown to econometricians. The test is applicable in situations in which there are replicated observations on the dependent variable. In this paper the test is modified to allow for heteroskedasticity usually encountered when dealing with cross-sectional observations, and applied to an earnings function estimated from a sample survey of Norwegian women.

* A major part of this work was done while the authors were Fellows of the Alexander von Humboldt Foundation at the University of Bonn, Germany. Comments by Michael McAleer, Saul Hymans, Karl Lin, Jerry Thursby, and G.S. Maddala are gratefully acknowledged. Special thanks are due to Helge Brunborg and the Norwegian Central Bureau of Statistics for providing the survey data used in this paper, and to Jeffrey Pliskin who was very helpful with suggestions and who carried out all of the computer work involved.

1. Introduction

The lack-of-fit (LOF) test is a model specification test that can be applied when there are replicated observations on the dependent variable corresponding to observations on the explanatory variables. Basically, the test utilizes the additional information that comes from within-group variation. As the situation of replicated observations is normal in experimental work, the test is well known to experimental statisticians but appears to be almost unknown to econometricians. (A survey of econometric text books has revealed no mention of the test.) This is probably due to the strong emphasis in econometrics on the methodology applicable to time-series data involving a single observation on the dependent variable for each set of observed values of the explanatory variables. By contrast, when cross-sectional data are used, there can be many units (individuals, firms, families) that are characterized by the same values of explanatory variables (e.g., incomes, prices, educational levels). In this situation a lack-of-fit test could often be profitably applied as an aid to appropriate model specification.

The main purpose of the paper is to draw the attention of econometricians to the possibilities offered by the lack-of-fit test (see also Battese [1977]). The paper is organized as follows. Section 2 contains a development and an explanation of the test. In Section 3 the test is generalized to allow for

heteroskedasticity which frequently characterizes relations pertaining to cross-sectional observations. Finally, in Section 4 the test is applied to the standard semilog earnings function due to Mincer [1974], utilizing data from the Norwegian Fertility Survey of 1977.

2. The Lack-of-Fit Test

Suppose that an nM1 random vector Y is normally distributed with mean μ and covariance matrix $\sigma^2 I_n$. The null hypothesis specifies a linear model of the form,

$$\mu = X\beta, \tag{1}$$

where X and β are nMK and KM1, respectively. Let us suppose that there are m (m > K) distinct observations on the explanatory variables X, and that corresponding to the i-th such observation there are n_i observations on the dependent variable Y, where $n_i > 1$ and $\sum_{i=1}^{n} n_i = m$. We will refer to these n_i observations as the "i-th group". The alternative hypothesis is

$$\mu = X\beta + Z\gamma, \tag{2}$$

where Z is a matrix of values of unspecified omitted explanatory variables (including possibly higher powers of X) of dimension nML (L < m-K) such that $Z'X \neq 0$, and $\gamma \neq 0$. Note that we assume that if the model is misspecified, observations on Z within each

group have the same mean.[1]

The error sum of squares (SSE) from the regression of Y on X is given by

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2, \tag{3}$$

where Y_{ij} denotes the j-the observation on the dependent variable in the i-th group. As the i-th group observations are all characterized by the same observation on the explanatory variables, the fitted values $\hat{Y}_{ij}$ $(j = 1,2,\ldots,n_i)$ must all be equal. Therefore we may write

$$\hat{Y}_{ij} = \hat{Y}_i,$$

and it follows that

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} [(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \hat{Y}_i)]^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{Y}_i + \hat{Y}_i)^2$$

where $\bar{Y}_i$ is a sample mean of the i-the group. Thus the error sum

---

[1] If Z is not constant whenever X is constant, the distribution of the test statistic in (11) is unchanged under H_0 and the test is still valid, but the power of the test is adversely affected. In this situation the test is in the class of the Goldfeld-Quandt test when used as specification error test. We are indebted to Jerry Thursby for a lengthy comment on this point.

of squares can be partitioned into two components

$$SSP = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \tag{4}$$

and

$$SSL = \sum_{i=1}^{m} n_i (\bar{Y}_i - \hat{Y}_i)^2. \tag{5}$$

The first of these components represents the 'within-group' variation which, following the experimental literature, we call the 'pure error sum of squares' (SSP). The second component is termed the 'lack-of-fit sum of squares' (SSL). It is the error sum of squares which would be obtained if each group were replaced by its sample mean and these sample means regressed on the same regressor variables with each observation weighted by $\sqrt{n_i}$. Defining

$$s_i^2 = \frac{1}{(n_i-1)} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2, \tag{6}$$

we may write

$$SSP = \sum_{i=1}^{m} (n_i-1)s_i^2 \tag{7}$$

and SSP/(n-m) is seen to be a weighted average of m different estimates s_i^2 of σ^2. This means that we have two different estimators of σ^2, SSE/(n-K) and SSP/(n-m). Under H_0

$$E[SSE/(n-K)] = E[SSP/(n-m)] = \sigma^2$$

whereas under H_A

$$E[SSE/(n-K)] > E[SSP/(n-m)] = \sigma^2$$

since variation around any point other than the mean always exceeds the variation around the mean. This provides the basis for the lack-of-fit test.[2]

Now, under H_0

$$SSE/\sigma^2 \sim \chi^2_{n-K} \tag{8}$$

and

$$SSP/\sigma^2 \sim \chi^2_{n-m}. \tag{9}$$

Further, it can easily be shown that SSL(=SSE-SSP) and SSP are independent so that

$$SSL/\sigma^2 = \chi^2_{m-K}, \tag{10}$$

and, therefore, if the model is correctly specified

---

[2] The paper is directed to situations in which all primary replicated data are available. However, if the data are in the form of aggregates such that for each group we are given (i) the group size (n_i), (ii) the group mean ($\bar{Y}_i$), and (iii) the group variance (s_i^2), then the LOF test can also be applied.

$$F = \frac{SSL/(m-K)}{SSP/(n-m)} \sim F_{m-k,n-m}. \tag{11}$$

A simple way of calculating SSP follows from the fact that SSP is simply the error sum of squares obtained by the application of least squares to the unconstrained regression equation

$$Y_{ij} = \mu_1 D_{1j} + \mu_2 D_{2j} + \ldots + \mu_m D_{mj} + u_{ij} \tag{12}$$

where $D_{ij} = 1$ for all observations in the i-th group,

$= 0$ otherwise.

Thus SSE defined in (3) represents constrained (by the null hypothesis) error sum of squares whereas SSP represents an unconstrained error sum of squares. The F-statistic in (11) can equivalently be written as

$$F = \frac{(SSE - SSP)/(m-K)}{SSP/(n-m)}. \tag{13}$$

Finally, using the well-known result that F_{ν_1,ν_2} approaches $\chi^2_{\nu_i}/\nu_1$ as $\nu_1 \to \infty$, we see that as $n \to \infty$, $n(SSL/SSP)$ is asymptotically distributed as χ^2_{m-k}. In general σ^2 is, of course, unknown. However, if σ^2 were known, a more powerful test would be obtained by using the chi-square result (10) rather than (11) or (13). This remark has relevance to the following section.

3. The Lack-of-Fit Test Under Heteroskedasticity

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 $E[(Y-\mu)(Y-\mu)'] = \sigma^2 I_n$ is unlikely to be realistic. We will generalize it by assuming that

$$E[(Y-\mu)(Y-\mu)'] = \Sigma, \tag{14}$$

where

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

and the σ_i^2 are known. Then, defining

$$D = \text{diag}(\sigma_1 I_{n_1}, \sigma_2 I_{n_2}, \ldots, \sigma_m I_{n_m}),$$

it follows that

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

and the linear specification on the mean under the null hypothesis takes the form

$$D^{-1}\mu = (D^{-1}X)\beta. \tag{15}$$

Thus, provided the analysis is carried out in terms of the weighted observations Y_{ij}/σ_i and X_i'/σ_i ($i = 1,2,\ldots,m$), the test described in the previous section follows through with one difference, namely that σ^2 is known to be equal to unity. Thus by (10), if the model is correctly specified,

$$LSS^* \sim \chi^2_{m-K}, \tag{16}$$

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

In practice the σ_i^2 are rarely known and have to be estimated. Provided the omitted variables are functions of the included variables as specified at the outset, the within group variation in the observations on Y provides consistent estimators of the σ_i^2. These estimators are completely independent of the specification $\mu = X\beta$. Thus the result in (16) is to be regarded as a large sample result; if the observations are weighted inversely by s_i, then as $n_i \to \infty$, the distribution of LSS* is given by (16).

It is commonly assumed in econometrics that the variances σ_i^2 can be related to a single variable w_i say, through a relationship of the form

$$\sigma_i^2 = \alpha w_i^{\delta} \tag{17}$$

where α and δ are unknown parameters. Typically, w_i would be a member of the regressor set, but this need not be the case. If such a model is appropriate and $m > 2$, then there will be gains in asymptotic efficiency if we make use of the information given in (17). A simple method of using (17) proceeds as follows. Defining s_i^2 as in (6), we have

$$v_i = \frac{(n_i-1)s_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1}.$$

Therefore, given that $\sigma_i^2 = aw_i^{\delta}$,

$$\ln s_i^2 = \ln a + \delta \ln w_i + u_i \qquad (18)$$

where

$$u_i = \ln v_i - \ln(n_i-1).$$

Bartlett and Kendall [1946] have shown that for large n_i, $\sqrt{(n_i-2)}u_i$ is approximately normally distributed with mean zero and variance 2. Furthermore, they showed that this approximation is likely to be good for n_i as small as 10.

Thus least squares regression of $\sqrt{(n_i-2)}\ln s_i^2$ on $\sqrt{(n_i-2)}$ and $\sqrt{(n_i-2)}\ln w_i$ will produce consistent and asymptotically efficient estimates of a and δ.

4. An Application of the LOF Test to an Earnings Function for Women

A question of considerable importance to human capital theory concerns the effect of schooling on earnings. Of particular interest is how this effect applies to women whose work history typically differs from that of men. The standard model of earnings proposed by Mincer [1974] and frequently used in applied labor economics -- most recently by Behrman and Birdsall [1983] and by Chiswick [1983] -- is of the form:

$$\ln W_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + u_{ij} \qquad (19)$$

where W_{ij} is the wage-rate of the j-th individual in the i-th group;

X_i is the 'on-job' work experience (common to all individuals in the i-th group);

S_i is the number of years of schooling (common to all individuals in the i-th group);

and u_{ij} is an independently distributed stochastic disturbance. (The subscripts have been adapted to the replicated data case considered by us.)

At least two types of possible misspecifications of equation (19) have been discussed in the literature.[3] First, a study by Mincer and Polachek [1974] provides some evidence that the marginal effect of schooling on the wage-rate of people of equal work experience is not constant for women with families. A similar conclusion with respect to heads of households (of any sex) has been reached by Ryder, Stafford, and Stephan [1976] on the basis of a model of life-cycle decision making with leisure as a choice variable. This suggests the possibility of an incorrect functional form of the equation. Second, it is by no means certain that the model in (19) contains all relevant explanatory variables. In particular, Griliches [1977] contends that models such as that in (19) suffer from the fact that

---

[3] Along with other applied research workers, we do not address the problem of a simultaneous equation bias that may arise from the endogeneity of schooling -- a point discussed at length by Griliches [1977].

individual ability -- likely to be correlated with schooling -- has been left out. The LOF test can certainly be effective in detecting incorrect functional form since this leaves the "omitted" variable Z constant within each group. As for the omitted "ability" variable, for the test to work well it would have to be true that all individuals of given experience and schooling are characterized by nearly the same ability, which may be questionable.[4]

The preceding discussion leads to the conclusion that it may be worthwhile to apply the LOF test not only to equation (19) -- to be labeled Model 1 -- but also to an equation that allows for a nonlinear effect of schooling as in

$$\ln W_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \beta_4 S_i^2 + u_{ij} \qquad (20)$$

which we label Model 2. To carry out the test we used the data on about 2,000 married women taking part in the Fertility Survey 1977 carried out by the Norwegian Central Bureau of Statistics. (The survey involved a random sample of about 4,000 women but this number was reduced to about one-half after eliminating all observations that were incomplete or that corresponded to a small number of replications.) Wages were measured by Kroner per hour; they are represented by actual wages for working women and by potential wages for women who did not work at the time of the

---

[4] See, however, footnote 1 above. Griliches [1977] refers to "ability" as "an unobserved latent variable that both drives people to get relatively more schooling and earn more income, given schooling..." (p. 7).

interview. Experience was measured by the number of years worked since completing the highest education. Finally, education was measured by years of schooling.[5]

The results for both models are presented in Table 1. Part (a) of the table contains the results before correcting for heteroskedasticity whereas Part (b) shows the results after heteroskedasticity has been corrected for. The details of the correction are as follows. By reference to the large-sample procedure described in Section 3, we postulate that heteroskedasticity in the earnings equation takes the form

$$\sigma_i^2 = aS_i^{\delta} \tag{21}$$

and obtain the following estimates:

$$\underset{(0.173)}{\hat{a} = 0.429} \qquad \underset{(0.112)}{\hat{\delta} = -0.610} \qquad (F = 15.52^{**}).$$

Correction for heteroskedasticity was then implemented in accordance with equation (15).

The most important result of Table 1 is that Model 1 is rejected by the LOF test whereas Model 2 passes the test. This holds whether a correction for heteroskedasticity is carried out or not. Thus the evidence of Mincer and Polachek (1974) is confirmed by our results. The linear form of schooling in the

---

[5] When applying the LOF test to Model 2 we also used dummy variables to represent various kinds of education instead of the linear and quadratic number of years of schooling. Since the results were very close for both formulations, we present only the latter.

earnings equation for married women appears to be inappropriate according to our evidence.

We conclude by giving a more detailed consideration to the marginal effect of schooling on the earnings of women. Using the results for Model 2 after correcting for heteroskedasticity, we estimate the marginal effect of schooling as

$$\left[\frac{\partial E(\ln W_{ij})}{\partial S_i}\right]_{X_i \text{ constant}} = -0.104 + 0.016S_i. \qquad (22)$$

This result is consistent with that of Mincer and Polachek [1974] in that the effect of schooling is an increasing function of schooling. In addition, the marginal effect of schooling is positive for $S_i \geq 6.5$ years. As there are at least 8 years of compulsory primary schooling, these data suggest that the marginal effect of schooling (for fixed experience) is positive. This is consistent with the result which has been found for men .

TABLE 1

Regression Results and Lack-of-Fit Computations for Models 1 and 2[a]

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	SSE	SSP	F
(a) Before Correcting for Heteroskedasticity								
MODEL 1 (n=1958, m=154, K=4)	2.23 (0.05)	0.031 (0.006)	-0.00068 (0.00024)	0.078 (0.004)	-----	283.2	255.7	1.29**
MODEL 2 (n=1958, m=154, K=5)	2.95 (0.16)	0.030 (0.006)	-0.00066 (0.00024)	-0.051 (0.028)	0.006 (0.001)	280.1	255.7	1.15
(b) After Correcting for Heteroskedasticity								
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	SSE*	SSP*	χ^2
MODEL 1 (n=1582, m=68, K=4)	2.34 (0.05)	0.042 (0.007)	-0.0013 (0.0004)	0.066 (0.003)	------	1 616.7	1 514	102.7*
MODEL 2 (n=1582, m=68, K=5)	3.26 (0.18)	0.036 (0.007)	-0.0010 (0.0004)	-0.104 (0.032)	0.0080 (0.0014)	1 588.8	1 514	74.8

[a] Standard errors are in parentheses. Single and double asterisks on statistics represent significance at the five and one percent levels, respectively.

References

Bartlett, M.S. and D.G. Kendall, 1946, The statistical analysis of variance heterogeneity and the logarithmic transformation, Journal of the Royal Statistical Society B 8, 128-138.

Battese, G.E., 1977, Agricultural economists, response functions and lack-of-fit tests, Review of Marketing and Agricultural Economics 45, 85-94.

Behrman, J.R. and N. Birdsall, 1983, The quality of schooling: quantity alone is misleading, American Economic Review 73, 928-946.

Central Bureau of Statistics of Norway, 1981, Fertility survey 1977 (Oslo).

Chiswick, C.V., 1983, Analysis of earnings from household enterprises: methodology and application to Thailand, Review of Economics and Statistics 65, 658-662.

Griliches, Z., 1977, Estimating the returns to schooling: some econometric problems, Econometrica 45, 1-22.

Mincer, J., 1974, Schooling, experience and earnings, New York: Columbia University Press.

Mincer, J. and S. Polachek, 1974, Family investments in human capital: earnings of women, Journal of Political Economy 82, 76-108.

Ryder, H.E., F.P. Stafford and P.E. Stephan, 1976, Labor, leisure and training over the life cycle, International Economic Review 17, 651-674.