Notes and Comments

A Note on Heteroscedasticity in Cross-Section Regressions Estimated from Irish County Data*

J. PETER NEARY

Nuffield College, Oxford

It is well known that, even when they are expressed in log or *per capita* form, many statistical series on Irish counties (by which is meant here the twenty-six counties of the Republic) are characterised by a value for Dublin which is considerably greater than the values for all other counties. The fact that the Dublin observation is an "outlier" does not of itself pose a problem for an econometrician who wishes to estimate cross-section regressions from Irish county data. However, a problem does arise when the inclusion of Dublin makes a significant difference to estimated parameters. The purpose of the present note is to show that, in at least one case where this common phenomenon occurs, the frequently adopted procedure of accepting the equation which excludes Dublin may be given a statistical justification. Specifically, parameter estimates obtained in this way are shown to be statistically indistinguishable from those estimated by applying a heteroscedastic correction procedure to the equation estimated from observations on all twenty-six counties.

The case study to be considered is introduced in Table 1, which presents alternative estimates of a cross-section demand function for Irish postal services in 1965, where the (natural) logarithm of total mail per head in each county is regressed on the (natural) logarithm of personal income per head. (Further details of the specification adopted and the data used may be found in Neary (1975).) Equations (1) and (2) show that the estimated coefficients are very sensitive to the inclusion of Dublin: when it is omitted the estimated income elasticity is nearly halved, and the significance of this coefficient is somewhat reduced. Ideally, one would wish to choose between the two equations on *a*

*This note is a revision of a section from my thesis 'Some Problems in the Specification and Evaluation of Econometric Models', submitted in 1974 in partial fulfilment of the requirements for the Travelling Studentship of the National University of Ireland. I am grateful to B.M. Walsh and an anonymous referee for helpful comments.

Table 1: *Tests for heteroscedasticity in cross section mail demand function, 1965*

Description of equation	*Equation*	*Estimated equations (t—values in parentheses)*	*R*	*R′₁*	*R′₂*
Demand function, estimated by OLS (26 observations)	1	$\log Y = -\,7{\cdot}174 + 1{\cdot}459 \log X$ (5·84) (6·13)	·781**	·840	·635
Demand function, estimated by OLS (25 observations: Dublin omitted)	2	$\log Y = -\,3{\cdot}692 + {\cdot}780 \log X$ (3·30) (3·59)	·599*	·802	·628
Tests for heteroscedasticity in residuals from equation (1)	3	$\lvert e_1 \rvert = -\,1{\cdot}481 + {\cdot}311 \log X$ (2·22) (2·41)	·441*		
	4	$\lvert e_1 \rvert = 41{\cdot}007 - 15{\cdot}93 \log X + 1{\cdot}551 (\log X)^2$ (2·74) (2·79) (2·85)	·636**		
	5	$\lvert e_1 \rvert = -\,{\cdot}267 \log X + {\cdot}056 (\log X)^2$ (2·15) (2·34)	·458**		
	6	$\lvert e_1 \rvert = {\cdot}195 - {\cdot}01803 (1/\sqrt{N})$ (3·04) (1·14)	·227		
Tests for heteroscedasticity in residuals from equation (2)	7	$\lvert e_2 \rvert = -\,{\cdot}125 + {\cdot}041 \log X$ (·18) (·31)	·065		
	8	$\lvert e_2 \rvert = -\,48{\cdot}08 + 18{\cdot}65 \log X - 1{\cdot}804 (\log X)^3$ (1·31) (1·31) (1·31)	·276		
	9	$\lvert e_2 \rvert = -\,{\cdot}00391 \log X + {\cdot}00407 (\log X)^2$ (.15) (·03)	·062		
	10	$\lvert e_2 \rvert = {\cdot}05806 + {\cdot}00749 (1/\sqrt{N})$ (1·08) (·57)	·119		
Demand functions estimated by GLS (26 observations); (i.e., equation (1) re-estimated after deflating all variables by predicted values of $\lvert e_1 \rvert$ from equation (4)	11	$\frac{1}{W} \log Y = 2{\cdot}677 - 3{\cdot}105 \frac{1}{W} + 6{\cdot}07 \frac{1}{W} \log X$ (2·36) (2·41) (2·34)	·461	·966	·655
	12	$\frac{1}{W} \log Y = -\,4{\cdot}019 \frac{1}{W} + {\cdot}842 \frac{1}{W} \log X$ (3·23) (3·00)	·148**	·812	·644
Non-linear demand function, estimated by OLS (26 observations)	13	$\log Y = 124{\cdot}237 - 45{\cdot}614 \log X + 4{\cdot}195 (\log X)^2$ (5·06) (5·21) (5·37)	·906**	·965	·638

Key· R = Multiple correlation coefficient (unadjusted):
*indicates that the associated F-statistic is significant at the 5 per cent level;
**indicates that it is significant at the 1 per cent level.

R'_1= Correlation coefficient between the actual values of Y and the values predicted by the equation in question: 26 observations.

R'_2= as R'_1, but recalculated with Dublin omitted (i.e., 25 observations).

Variables:

Y = Total mail per head posted in each county, 1965. X = County personal income per head, 1965. $\lvert e_1 \rvert$ = Absolute value of residuals from equation (1). $\lvert e_2 \rvert$ = Absolute value of residuals from equation (2). $\sqrt{N}$ = Square root of county population, 1965. W = GLS Weights, calculated from equation (4) and used to re-estimate equation (1) (i.e., equals predicted values of dependent variable, $\lvert e_1 \rvert$, from equation (4)).

priori grounds, but in this case there are no strong arguments either way.[1] Alternatively, one might have recourse to extraneous information: in this context it may be noted that the estimated income elasticity from equation (2) is more compatible than that from equation (1) with estimates from time series models in Section 2.3 of Neary (op. cit.). However, such a comparison is inconclusive, since there are many reasons why time series and cross-section estimates of the same equation need not yield similar results. So without further analysis there is no way of choosing between equations (1) and (2).

One plausible explanation of the difference between the two equations is that the assumption of homoscedasticity—constant variance of the disturbance term—does not hold in the present sample. If it does not, then parameter estimates obtained by ordinary least squares (OLS) are inefficient, though unbiased, while their estimated standard errors are biased and hence likely to lead to incorrect inferences. Previous studies have recognised that this is "a potentially serious problem in any regression involving units of such different size as the Irish counties" (Walsh 1970–71, Appendix 2; see also Geary 1966). Hence it seems desirable to test the residuals from equations (1) and (2) for heteroscedasticity, and such testing is carried out in the next eight equations in Table 1.

The tests employed were suggested by Glejser (1969)[2], and involve assuming a structure for the error term variance of the general form: $E(u_i^2) = \sigma^2 f(X_i)$ and then testing various simple forms of $f(X_i)$, by regressing the absolute values of the OLS residuals on them. In addition to trying three forms of $f(X_i)$, the absolute values of the residuals were also regressed on the reciprocal square root of the county population $(1/\sqrt{N_i})$. This attempts to test the hypothesis put forward by Walsh (op. cit.), that aggregation over county units of data and equations pertaining to individuals leads to a structure for the error term variance of the form: $E(u_i^2) = \sigma^2/N_i$.

The outstanding feature of these eight equations is that for three of the four specifications tried, the residuals from equation (1) show significant evidence of heteroscedasticity, whereas none of the equations fitted to the residuals from equation (2) show any evidence of departure from homoscedasticity. As for the alternative specifications of the heteroscedastic structure for the residuals from equation (1), equation (4), a non-homogeneous quadratic in log X, gives the most satisfactory results. It may be noted that the coefficient of $(1/\sqrt{N_i})$ in equation (6) is not significant; this suggests that, despite its theoretical appeal, the aggregation model used by Walsh does not provide an adequate explanation for the presence of heteroscedasticity in equation (1).

[1] This situation may be contrasted with that in Neary (op. cit), Sections 3.4 and 3.5, where, in fitting production functions to post office operations, the very different technological and spatial conditions prevailing in Dublin were taken to justify its exclusion from the regression.

[2] Glejser's approach was anticipated by Geary (1966) for the case of a linear $f(X_i)$. It may be remarked that Glejser's tests performed extremely well in the Monte-Carlo studies reported in Goldfeld and Quandt (1972), Chapter 5.

Having established that the residuals from equation (1) support the hypothesis of heteroscedasticity, while those from equation (2), do not, all that remains is to re-estimate the former equation using the standard generalised least squares (GLS) weighting procedure. The weights used are the predicted values of the dependent variable from equation (4), which appears to provide the best available estimate of the heteroscedastic structure, and the resulting equations are given in the next two lines of the table. Of these, equation (12), in which the intercept has been suppressed, is the finally corrected GLS estimate of the cross-section demand function.[3] It is evident that the coefficients of this equation are totally different from those of equation (1); however, they are very similar to those of equation (2). In other words, the major conclusion to be drawn from the table is that the uncorrected regression using twenty-five counties and the GLS regression using all twenty-six counties yield estimates of the demand curve parameters which are statistically indistinguishable from one another.

Finally, the fact that the preferred specification for the heteroscedastic structure of the error term in equation (1) is a quadratic in log X might be thought to suggest that the true specification of the demand function is not a linear equation with a heteroscedastic error term, but a non-linear equation with homoscedastic errors.[4] Fortunately, it transpires that this alternative specification of the demand function does not affect the conclusion that the OLS estimate of the income elasticity based on twenty-five counties is preferable to that based on all twenty-six counties including Dublin. This may be seen from equation (13), which is an OLS estimate of a non-linear (in logs) specification of the demand function. The explanatory power of this equation is comparable with, and, if anything, marginally inferior to, that of equations (11) and (12). More importantly, the estimated income elasticity implied by it is found to be 0·825, with a t-value of 4·17.[5] As with the estimate from equation (12), this is well within sampling error of the value obtained from equation (2). This suggests that an estimated value for the income elasticity of about 0·8 is extremely robust with respect to alternative specifications of the demand function.

Obviously the conclusions of this note cannot be assumed to apply to other data sets without further research. Nevertheless they suggest, first, that OLS parameter estimates obtained from cross-section equations which include Dublin should be treated with caution; and secondly, that in cases where an unweighted regression applied to all twenty-six Irish counties would yield inefficient estimates, re-estimating the equation with Dublin omitted is in practice equivalent to applying a full correction for heteroscedasticity.

[3] The only reason for including equation (11) in the table is to show that its intercept is significant. Strictly speaking, this is evidence of mis-specification of the original equation, (1); however, since the reduction in explanatory power (as judged by R'_1 and R'_2) in passing from (11) to (12) does not appear to be large, it seems reasonable to ignore the mis-specification and to take equation (12) as the final estimate.

[4] This possibility was suggested by a referee.

[5] This is the value obtained when the income elasticity is evaluated at the mean of the twenty-six counties including Dublin. When it is evaluated instead at the mean of the twenty-five non-Dublin counties, the value obtained is 0·682, with a t-value of 3·19.

REFERENCES

GEARY, R. C., 1966. "A note on residual heterovariance and estimation efficiency in regression", *American Statistician,* Vol. 20, 30–31.

GLEJSER, H., 1969. "A new test for heteroscedasticity", *Journal of the American Statistical Association,* Vol. 64, 316–323.

GOLDFELD, S. M., and R. E. QUANDT, 1972. *Nonlinear Methods in Econometrics,* Amsterdam: North-Holland.

NEARY, P., 1975. *An Econometric Study of the Irish Postal Services,* Dublin: The Economic and Social Research Institute. Paper No. 80.

WALSH, B. M., 1970–71. "Aspects of labour supply and demand with special reference to the employment of women in Ireland", *Journal of the Statistical and Social Inquiry Society of Ireland,* Vol. 22, Part 3, 88–118.