On Testing for Uniformity of Fit in Regression: An Econometric Case Study

JOHN L. PRATSCHKE*

IT is frequently necessary to test regression results for uniformity of fit—i.e. to test the randomness of the distribution of the error terms. Such a test is necessary because an indication of a systematic trend in the pattern of residuals would suggest that the regression equation fitted to the data did not properly reflect the curvature of the true relationship between the variables regressed. If, for instance, the residuals after regression fell into three groups of positive residuals, followed by a group of negative residuals, followed again by positive residuals, then the implication would clearly be that the estimated regression function misrepresented the true relationship. Indeed, econometricians frequently regard such bunching of residuals as *prima facie* evidence of serial correlation—see, for example, Leser [9, p. 17]. The problem is, perhaps, most familiar in the analysis of serial correlation errors in time-series analysis. The problem also arises in regressions based on cross-section data, provided that the regression estimates of the residuals after regression are first ordered in ascending order of the independent variable. In a two-variable case, the analogy with time-series studies is then complete, because time is there the independent variable and the residuals are automatically ordered in ascending order of the time variable. In cross-section studies, the independent variable (if there is but one) is the logical, and indeed only, possible choice with which to order the regression residuals. Where there is more than one independent variable, the more important of the two appears the logical choice, though the possibility of ordering the residuals by any other independent variable does exist; for example, one might use the principal component of the independent variables. (Again, the analogy with time-series exists in the case of multiple regression studies: time is generally assumed to be the relevant independent variable when the Durbin-Watson *d*-test is applied [3, 4]). When applying such tests to cross-section data, the term "uniformity of fit" is used instead of "serial correlation of errors."

It is the purpose of this study to describe and compare a number of possible tests for uniformity of fit. It will be obvious from the foregoing that the best known test is a modified version of the Durbin-Watson *d*-test (*op. cit.*), where the

*The author wishes to thank his colleagues for comments on earlier drafts, and in particular Dr. R. C. Geary (Dublin). The author is solely responsible for any errors remaining.

residuals after regression are reordered in ascending order of the major independent variable, and the *d*-test then applied in the ordinary way as

$$d = \sum_{j=2}^{n} (\hat{e}_j - \hat{e}_{j-1})^2 / \sum_{j=1}^{n} \hat{e}_j^2$$

where $\hat{e}_j$ is the *jth* regression residual.

Griliches *et alia* [8] have reported that, in their view, the *d*-test may be unduly influenced by even one aberrant observation. They point out that one large $\hat{e}_j$—which might be attributable to inaccuracies in statistical data—may have such a large effect on the *d* statistic as to give a significant statistic, though a regression analysis based on all but that one observation which led to the large $\hat{e}_j$ may be free of serial correlation problems. For this reason, they recommend that alternative non-parametric tests be developed, and the results compared with those yielded from the application of the *d*-test.

Among the first to propose such a non-parametric test were Stevens [14], Wald and Wolfowitz [16] and Swed and Eisenhart [15]. Their test, when applied to the examination of regression residuals—as has recently been discussed by Draper and Smith [2]—examines the possibility of sequences of similar sign (+ or −) being drawn at random from a normal population. The test, which may be referred to as the "Runs Test" is, essentially, a development of the well-known statistical problem of whether or not two independent samples are derived from one continuous distribution. It may be shown (see Mood [11]) that the probability distribution of *u*, where *u* is the number of runs, or sequences of one or more residuals of the same sign, tends to normality, with mean and variance defined as

$$E(u) = 2n_1n_2/(n_1+n_2)+1$$

$$\text{and } \sigma^2 u = 2n_1n_2/(2n_1n_2-n_1n-_2)/(n_1+n_2)^2(n_1+n_2-1)$$

where n_1 is the number of residuals with positive signs, n_2 is the number of residuals with negative signs, and $N(=n_1+n_2)$ is the total number of residuals. This tendency toward normality is confirmed by Geary [7] when $N = 40$.

Indeed, Geary (*op. cit.*) has recently proposed a simpler test, in which account is taken only of the number of changes in sign (+ or −) between successive residuals. Clearly, the number of sign-changes τ is one less than the number of runs, i.e. $\tau = u-1$. He shows that the probability of τ sign changes is the point binomial, slightly modified, i.e.

$$P(\tau = t) = \frac{(N-1)!}{t!(N-1-t)!}/(2^1N+1)$$

where τ is the number of sign changes. Obviously the probability of τ or less sign changes is defined by

$$(\tau \leqslant t) = \sum_{\tau=0}^{t} \frac{(N-1)!}{t!(N-1-t)!}$$

The probability distributions of τ and of u are quite alike. The τ-test is, of course, easier to use.

Another non-parametric test suggested by Griliches *et alia* (*op. cit.*) compares the sign of the ith residual with that of the $i+1$th and the frequencies of the observed combination of signs of successive residuals are arranged in a 2×2 contingency table of the form

		Sign of the i *th* Residual		
		+	−	Total
Sign of the $i+1$ th Residual	+	a_{11}	a_{12}	$a_{11}+a_{12}$
	−	a_{21}	a_{22}	$a_{21}+a_{22}$
	Total	$a_{11}+a_{21}$	$a_{12}+a_{22}$	$N-1 = \Sigma_i \Sigma_j a_{ij}$

and chi-squared (χ^2) is defined as

$$\chi^2 = \frac{(N-1)(|a_{11}a_{22}-a_{21}a_{22}|-(N-1)/2)^2}{(a_{11}+a_{12})(a_{21}+a_{22})(a_{12}+a_{22})(a_{11}+a_{21})}$$

when Yates' correction is used.

If there is positive first order autocorrelation of residuals, one would expect a_{11} and a_{22} to be significantly larger than a_{21} and a_{12}. The null hypothesis of no autocorrelation (i.e. of uniformity of fit) is tested by applying the ordinary chi-squared test (with one degree of freedom) to the table.

These four tests—the modified Durbin-Watson d-test, the Runs Test, the Geary sign-change test, and the chi-squared test were applied to the residuals obtained when estimating eighteen different algebraic forms of the Engel function for each of five major expenditure groups using Irish data [1]. The function forms are described in detail in Pratschke [13, Table 1]. In general, they involve linear,

logarithmic and inverse transformations of the variables *vi* (expenditure on *i*), *E* (total expenditure), and *N* (average household size)—when *vi*, or its transformation is dependent variable in all cases. Sixteen observations only were available from which to estimate the regressions. The observations were ordered in ascending order of *E*, the major determining variable.

The critical values of the Durbin-Watson *d* are given in Table 1 following:

TABLE 1. *Critical Values of Durbin-Watson d*

Number of Independent Variables	*Probability Level (One Sided)*	d_L	d_u
1	0·05	1·10	1·37
2		0·98	1·54
3		0·86	1·73
1	0·01	0·84	1·09
2		0·74	1·25
3		0·63	1·44

Source: Durbin and Watson *(op. cit.)*.

Thus, if $\hat{d} \leqslant d_L$, the test is significant; if $d_L < d \leqslant d_u$, the test is inconclusive; if $\hat{d} > d_u$, the null hypothesis of no serial correlation is accepted. The test may also be used to test for negative serial correlation, by calculating $4-d$ and using as d. In this instance, a one-sided test for positive serial correlation of errors was used.

The application of the Runs Test is complicated by the fact that $N = 16$; instead of using the normal approximation, the exact probability distribution of u, as tabulated by Swed and Eisenhart (*op. cit.*) was used. Their table is easily manipulated to give Table 2, which shows, for $1 < n_1 < 16$, given $N = 16$, the values of u_ϵ, where (i) u_ϵ is the largest integer u' for which $P(u \leqslant u') \geqslant \epsilon$ when $\epsilon < 0{\cdot}050$; and (ii) u_ϵ is the smallest integer u' for which $P(u \leqslant u') \geqslant \epsilon$ when $\epsilon \geqslant 0{\cdot}50$. The significance levels chosen are the conventional 0·05, 0·01, 0·95 and 0·99 per cent levels.

For the application of the Geary test, the cumulative point binomial distribution, for $N = 16$, is given in Table 3. It may be seen that the probability of having three of less sign changes is approximately 2 per cent. The discontinuity of the distribution makes it impossible to establish precise 1 per cent or 5 per cent probability levels.

TABLE 2: *Significance Levels of u*

$N=16$		u_ϵ			
n_1	n_2	0·01	0·05	0·95	0·99
1	15	—	—	—	—
2	14	—	2	5	5
3	13	2	3	7	7
4	12	3	4	9	9
5	11	3	4	11	11
6	10	3	5	11	13
7	9	4	5	12	13
8	8	4	5	12	13
9	7	4	5	12	13
10	6	3	5	11	13
11	5	3	4	11	11
12	4	3	4	9	9
13	3	2	3	7	7
14	2	—	2	5	5
15	1	—	—	—	—

Source: Swed and Eisenhart (*op. cit.*).

Notes:

When $\epsilon < 0{\cdot}50$, u_ϵ is the largest integer u' for which $P\left\{u \leqslant u'\right\} \leqslant \epsilon$

When $\epsilon > 0{\cdot}50$, u is the smallest integer u' for which $P\left\{u \leqslant u'\right\} \leqslant \epsilon$

TABLE 3: *Cumulative Point Binomial Distribution*

τ	P	τ	P
0	0·0000	8	·6964
1	0·0005	9	·8491
2	0·0037	10	·9408
3	0·0176	11	·9824
4	0·0592	12	·9963
5	0·1509	13	·9995
6	0·3036	14	1·0000
7	0·5000	15	1·0000

The convention adopted is to take $P(\tau \leqslant 2) \backsim 1$ per cent and $P(\tau \leqslant 4) \backsim 5$ per cent, since these are the values of τ whose probabilities most nearly approximate to the 1 and 5 per cent levels. Thus, especially at the 1 per cent level we are discriminating somewhat against τ—what we have styled the 1 per cent level is in reality the 0·004 per cent level. On the other hand, we are being a little "kind" to τ at the 5 per cent level by using $P = 0{\cdot}059$. In this way we are using Geary's τ test in a one-tailed version, to test for positive serial correlation.

The small number of observations also raises difficulties in the utilisation of the 2x2 table. With $N = 16$ (i.e. $N-1 = \Sigma_1 \Sigma_j a_{ij}$) the individual cell entries are generally small, and the correctness of using chi-squared in such cases has, of course, been the subject of some controversy. The Fisher Exact Probability Test was therefore used instead—see Fisher [6, § 21.02]. When the marginal totals of the 2x2 table are regarded as fixed, the probability of observing the set of frequencies in the table is given by

$$P = \frac{(a_{11}+a_{12})!(a_{21}+a_{22})!(a_{11}+a_{21})!(a_{12}+a_{22})!}{a_{11}!a_{12}!a_{21}!a_{22}!(\Sigma_i \Sigma_j a_{ij})!}$$

Because of the tedious nature of the computations required for this calculation, the table of significance levels presented by Finney [5] for the exact test has been used.

The results from the four tests on the residuals are set out in Table A (Appendix).

The concordance between the Durbin-Watson d-test, the Runs Test and the Geary sign-change test is quite good: the surprising result is the poor showing of the Fisher Exact test when applied to this type of data. Leaving aside the results of the Fisher Exact test, it is noteworthy that of the eight cases where d is significant at the 5 per cent level, τ is significant in five cases, (one of which at the 1 per cent level) and u is significant in five cases (of which two are at the 1 per cent level). In general, τ and u appear to give results that are broadly similar to those obtained using the d-test, as may be seen from Table 4, which is a summary of Table A.

Geary (*op. cit.*) using a Monte Carlo type of experiment to compare τ and d reaches the same conclusion, and takes it further by showing that, in his constructed example, there is no significant difference in the number of autocorrelated regressions identified by the two tests, at the 5 per cent level. The greater ease with which τ can be calculated, without access to computers, provides a simple test which appears to be as efficient as d.[1] It is also interesting to note the results for τ and u when d is inconclusive.

1. In a more recent study, Habibagahi and Pratschke [9], using Monte Carlo techniques extensively, found that, in general, the power of Geary's test is approximately as high as that of Durbin-Watson for 30 or more observations, and lower than D-W for less observations.

The close relationship between d and τ and u is not entirely surprising. In the τ and u tests, we are ignoring the arithmetic size of each $\hat{e}_j$. Thus, as Griliches *et alia* report, they are less likely to be unduly affected by single high values of $\hat{e}_j$.

TABLE 4: *Concordance of Results Using π, d, and u Statistics*

Significance Levels of				
		d	π	*u*
d	*u*			
0·01	·01	12	4 } 10	5 } 9
	·05		6 }	4 }
	n.s.		2	3
0·05	·01	8	1 } 5	2 } 5
	·05		4 }	3 }
	n.s.		3	3
Inconclusive	·01	17	3 } 9	4 } 7
	·05		6 }	3 }
	n.s.		8	10
Not Significant	·01	53	– } 5	– } 5
	·05		5 }	5 }
	n.s.		48	48
Total Frequency		90	90	90

Note: n.s. indicates not significant.

University of Waterloo, Ontario.

Table A: *Comparsion of Alternative Forms of the Engel Function Using Four Tests for Uniformity of Fit*

Function Number	Commodity Group: Food				Clothing				Fuel and Light				Housing				Sundries			
	Significance Appraisal																			
	d	u	τ	Fisher Exact	d	u	τ	Fisher Exact	d	u	τ	Fisher Exact	d	u	τ	Fisher Exact	d	u	τ	Fisher Exact
1					†	*	*													
2		*	*		†	*	*											*	*	
3					†								*	**	*					
4					**		*						†				**	**	**	*
5	†		*				*										**	**	**	*
6	*																			
7																		*	*	
8		*															†	*	*	
9	*	*	*		**				†	**	*		**				**	**	*	
10	†				†	**	**						**	*	*		**	*	*	
11	*	*	*		†		*						**	*	*		**	*	*	
12	*								*				†							
13	†	**	**	*	†				†	**	**	*					*	**	**	*
14	**	**	**	*					*	*	*						**	**	**	*
15																				
16																		*	*	
17					†								†							
18					†															

Notes: *Indicates significance at the 5 per cent level.
**Indicates significance at the 1 per cent level.
†Indicates an inconclusive d test.

REFERENCES CITED

[1] Central Statistics Office, *Household Budget Inquiry 1965/66*, Dublin: The Stationery Office, Prl. 266, 1969.

[2] Draper, N. R. and Smith, H., *Applied Regression Analysis*, New York: John Wiley & Sons, Inc., 1966.

[3] Durbin, J. and Watson, G. S., "Testing for Serial Correlation in Least Squares Regression, I", *Biometrika*, Vol. 37, p. 409, 1950.

[4] ——, "Testing for Serial Correlation in Least Squares Regression, II", *Biometrika*, Vol. 38 p. 159, 1951.

[5] Finney, D. J., "The Fisher-Yates Test of Significance in 2 × 2 Contingency Tables", *Biometrika*, Vol. 35, p. 145, 1948.

[6] Fisher, R. A., *Statistical Methods for Research Workers*, 10th ed. Edinburgh: Oliver and Boyd, 1946.

[7] Geary, R. C., "Relative Efficiency of Count of Sign Changes for Assessing Residual Auto-regression in Least Squares Regression", *Biometrika*, Vol. 57: 1, p. 123, 1970.

[8] Griliches, Z., Maddala, G. S., Lucas, R., and Wallace, N., "Notes on Estimated Aggregate Quarterly Consumption Functions", *Econometrica*, Vol. 30, No. 3, 1962.

[9] Habibagahi, H. and Pratschke, J. L., "A Comparison of the Power of the von Neumann Ratio, Durbin-Watson & Geary Tests", *Waterloo Economic Series*, No. 36, University of Waterloo, Ontario, Canada, 1971.
[10] Leser, C. E. V., *Econometric Techniques & Problems*, London: Charles Griffin & Co., 1966.
[11] Mood, A. M., *Introduction to the Theory of Statistics*, New York: McGraw-Hill Book Co., 1950.
[12] Prais, S. J. and Houthakker, H. S., *The Analysis of Family Budgets*, Cambridge U.P., 1955.
[13] Pratschke, J. L., *Income-Expenditure Relations in Ireland, 1965-1966*, Dublin: The Economic and Social Research Institute, Paper No. 50, 1969.
[14] Stevens, W. L., "Distribution of Groups in a Sequence of Alternatives", *Annals of Eugenics*, Vol. 9, Part I, January 1939.
[15] Swed, F. S. and Eisenhart, C., "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", *Annals of Mathematical Statistics*, Vol. 14, No. 1, March 1943.
[16] Wald, A. and Wolfowitz, J., "On a Test Whether Two Samples are From The Same Population", *Annals of Mathematical Statistics*, Vol. 11, No. 2, June 1940.