Diagnostic Testing After 45 Years: The Impact on Empirical Research in Economics

C. R. McKenzie, Osaka School of International Public Policy, Osaka University
Mikiyo Kii, Osaka School of International Public Policy, Osaka University

Abstract After providing a short history of when various diagnostic tests were developed for different estimation techniques, this paper surveys all the empirical literature that has been published in twelve economics journals in 1994. This survey tries to determine the impact of the theoretical literature on diagnostic testing on empirical research. As a general conclusion, the paper finds that notwithstanding the vast theoretical and Monte Carlo literature on diagnostic tests, their impact on empirical work has been quite small. Some reasons for this outcome are suggested.

"The three golden rules of econometrics are test, test, and test."
Hendry (1980, p. 403)

"There are too many diagnostics. One is too many as far as I am concerned."
Leamer (1988, p. 332)

"...it is silly to indulge in diagnostic testing to the point where it becomes counterproductive in terms of data reduction...it seems to me to be even more silly to ignore the useful evidence that good diagnostics can impart."
Phillips (1988, p. 349)

"A test that is never used has zero power. The power of a popular test is irrelevant."
McAleer (1994, p. 334)

1. INTRODUCTION

In 1950, Durbin and Watson (1950) developed their famous diagnostic test for testing whether the disturbance of a linear regression model is serially correlated or not. The test assumes the model has been estimated by ordinary least squares, there is no other deviation from the assumptions of the standard linear regression model and the disturbances are normally distributed. Although it is now possible to compute the exact critical values relatively easily and quickly, the test is typically used in association with the upper and lower bounds for the critical values. In the 45 years that have passed since that paper was published, hundreds of papers have been published proposing other diagnostic tests for serial correlation and other deviations from the standard linear regression model and other more general models, and evaluating their small-sample properties using Monte Carlo simulations (for a summary of some of this work see Pagan and Hall (1983a, b), McAleer (1987, 1995), Godfrey (1988), Pagan (1990) and MacKinnon (1992)).

As the four quotes indicate, there is an extremely wide range of views on the appropriate use of diagnostic tests in empirical work amongst econometricians ranging from never using them (Leamer), using them in moderation (Phillips), to requiring their use as a matter of course (Hendry). However, a quick glance through any issue of nearly any economics journal suggests that this theoretical work on diagnostic tests has not had an impact on empirical work in economics that would reflect the pages devoted to the theoretical developments, and that many applied economists have been following Leamer's prescription. Following McAleer's logic, this would mean that many (if not most) diagnostic tests have no power. The first purpose of this paper is to see if the casual observation on the lack of usage of diagnostic tests is supported by a more thorough and systematic evaluation of the available empirical evidence, and whether there are any differences in the extent to which use is made of diagnostic tests across journals. A second purpose of the paper is to try and explain why such little use is made of diagnostic tests in empirical work.

Before analysing the evidence on the use of diagnostic tests, a brief case is made for why diagnostic tests might be used (section 2). After providing a short history of when various tests were developed for different estimation techniques (section 3), this paper surveys all the empirical literature that has been published in twelve economics journals in 1994 (section 4). Section 5 provides a discussion and evaluation of some of the reasons why diagnostic tests have not met with general acceptance. Section 6 contains some concluding remarks.

2. WHY USE DIAGNOSTIC TESTS?

Consider the linear regression model

$$y = X\beta + u \qquad (1)$$

where y is a Tx1 vector of observations on the dependent variable, X is a Txk matrix of observations on the explanatory variables, β is a kx1 vector of unknown parameters, and u is a Tx1 vector of unknown disturbances. Given the observations on y and X, the ordinary least squares (OLS) estimator of $\beta, \hat{\beta} = (X'X)^{-1}X'y$, can easily be computed. To say anything about the properties of this or any other estimator, however, some additional statistical assumptions are required. In the standard linear regression model, the typical assumptions are:

A1: $E(u) = 0$.

A2: $E(uu') = \sigma^2 I_T$.

A3: X is non-stochastic.

A4: $u \sim N$.

Given these assumptions, it is well-known that $\hat{\beta}$ is the best linear unbiased estimator (BLUE); is a consistent estimator and, moreover, is efficient. Define the OLS estimator of σ^2 as

$$s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(T-k),$$

and denote the ith element of β and $\hat{\beta}$ by β_i and $\hat{\beta}_i$, respectively, and the standard error of $\hat{\beta}_i$ by SE($\hat{\beta}_i$), then given A1-A4

$$t_{\beta_i} = (\hat{\beta}_i - \beta_i)/SE(\hat{\beta}_i) \sim t_{T-k}. \qquad (2)$$

The null hypothesis $H_0: R\beta = r$ where R is a qxk matrix and r is a qx1 vector of constants, can be tested using the quantity

$$F_{R\beta} = (R\hat{\beta} - r)'(R(X'X)^{-1}R')^{-1}(R\hat{\beta} - r)/qs^2$$

since $F_{R\beta} \sim F_{(q,T-k)}$ under H_0 given A1-A4.

Assume the model in (1) holds for observations T+1,..,T+p, that is,

$$y_p = X_p\beta + u_p \qquad (3)$$

where y_p is a px1 vector of the additional observations on the dependent variable, X_p is a pxk matrix of the additional observations on the explanatory variables, and u_p is a px1 vector of unknown disturbances. Given $\hat{\beta}$, a predictor of y_p, $\hat{y}_p$, is easily computed as $\hat{y}_p = X_p\hat{\beta}$. To determine the properties of this predictor, we need some assumptions about u and u_p. Assuming A1-A4 hold for u and similar assumptions hold for u_p, and that u and u_p are not serially correlated, then $\hat{y}_p$ can be shown to the best linear unbiased predictor (BLUP) of y_p and it can be shown that the quantity

$$F_p = (y_p - \hat{y}_p)'[I_p + X_p(X'X)^{-1}X_p']^{-1}(y_p - \hat{y}_p)/ps^2$$

is distributed as $F_{(p,T-k)}$.

It is well-known that all these properties of the OLS estimator and OLS based-predictor are strongly dependent on some or all of the assumptions A1-A4 being satisfied. For example, if a relevant variable is excluded from (1), so that $E(u) = \mu \neq 0$, then all these properties are, in general, lost. It is this dependency of the properties of the OLS estimator on the assumptions A1-A4, that provides the general justification for the use of diagnostic tests and estimators in place of OLS. For example, if u is serially correlated (and or heteroscedastic), OLS is, in general, inefficient and the OLS formula for the variance of the parameter estimates is no longer an unbiased (or consistent) estimator of the true variance. Diagnostic tests are used to determine whether the estimated model is misspecified and if it exhibits deviations from the assumptions relied on for conducting hypothesis tests or for using a particular estimator. A brief examination of Leamer's extreme bounds analysis (Leamer (1978) and Leamer and Herman (1983)) or his elucidation diagnostics (Leamer (1992)) soon indicates that these diagnostic tools strongly depend on A1-A4 as well, and are really only concerned about fragility in certain directions.

3. WHAT DIAGNOSTIC TESTS ARE AVAILABLE?

Tables 1 and 2 provide the results of an incomplete survey of the theoretical literature on diagnostic tests. The information in Tables 1 and 2 is basically the same but presented in two different formats. Table 1 provides the information in chronological order according to when the paper was published. Table 2 provides the information according to a two way classification: the type of problem the test was designed to detect; and the appropriate estimation technique (or model). It should be noted that the survey was limited to single equation methods (except for Johansen's cointegration tests) and only a small number of the available diagnostic tests are actually listed in either Table.

While the first page of Table 1 is dominated by tests for models estimated by OLS, it is a little surprising that in the 1950s some econometricians were interested in developing diagnostic tests for models estimated by the limited information maximum likelihood method. Serial correlation was the principal focus of attention in the 1950s. As time passes, the focus moves away from models estimated by OLS to more sophisticated estimation techniques, and away from serial correlation to other problems. The number of papers cited in Table 1 alone is 76 providing one indicator of the number of journal pages devoted to these theoretical problems.

One point that can be deduced from Table 2 is that for misspecification errors typically alluded to in standard undergraduate econometrics courses on the standard linear regression model, serial correlation, heteroscedasticity, structural change, functional form and normality, diagnostic tests also are available when a more sophisticated estimation technique is used like instrumental variables, general method of moments, or some limited dependent variable estimator. For this class of model, a lack of usage of diagnostic tests cannot be because there is no test. As Pagan and Hall (1983b) demonstrate, diagnostic tests for models with a non-linear regression component can also be easily developed. Recently, some diagnostic tests have also been developed for models estimated using non-parametric techniques (see, for example, Rilstone (1992), Gozalo (1993) and Delgado and Stengos (1994)).

4. WHAT USE IS MADE OF DIAGNOSTIC TESTS?

In order to determine the impact that the theoretical diagnostic test literature has had on empirical work in econometrics, a survey of empirical papers in the journals listed in Table 3 was implemented. Most people would probably have little argument with the choice of most of the journals but the choice of two, *Applied Economics* and *Economic Studies Quarterly* requires a little explanation. *Applied Economics* was included because of its emphasis on empirical work and the journal's claim that it "hopes to foster quantitative studies, the results of which promise to be of use in the practical field and help bring economic theory nearer to the realities of life". In 1994, *Economic Studies Quarterly* was the top ranked domestic journal in Japan and was included because both authors are affiliated with a Japanese university (*Economic Studies Quarterly* has now replaced by the *Japanese Economic Review*). The reason for choosing 1994 was to provide the most up-to-date information on the usage of diagnostic tests in economics.

Some summary information about the survey are also provided in Table 3: editors of the journal; the number of articles published in the year; the number of empirical papers; and the number of papers having at least one diagnostic test. Editors were listed to determine their location as well as their field of expertise. The proportion of empirical papers reporting at least one diagnostic test ranges from 0% for *Economic Studies Quarterly* to 56% for the *Review of Economic Studies*. For the rest of the journals the proportion lies somewhere between 20% and 50%, with the American-based journals (*American Economic Review*, *Journal of Political Economy* and *Journal of Finance* but not *Review of Economics and Statistics*) being on the low side and the English-based journals (*Applied Economics*, *Economic Journal* and *Journal of Applied Econometrics*) being on the high end. In contrast, *Econometrica*, the journal of the Econometrics Society, has very few empirical papers to begin with and seems to set an example of not requiring diagnostic tests.

Even though a liberal definition of a paper with a diagnostic test is used, the results in Table 3 are potentially misleading because for some of the estimation techniques used diagnostic tests may not be available. This is particularly important for the econometric theory journals, *Econometrica* and the *Journal of Econometrics*. To control for this bias, only papers containing models estimated by OLS are surveyed in Table 4 and a listing of the sorts of diagnostic tests reported are presented. *Economic Studies Quarterly* and the *Review of Economic Studies* do not appear because the former has no paper reporting a diagnostic test and the latter has no paper with a model estimated by OLS. Table 5 suggests that at least one diagnostic test is far more likely to be reported for models estimated on time-series data than on cross-section data but, to some extent, this reflects the fact that the most commonly used diagnostic tests, the Augmented Dickey-Fuller test followed by the Durbin-Watson test, are only appropriate for time-series data. However, even for time-series data the record cannot be said to be good. The records of the *Journal of Finance* and the *Journal of Political Economy* (and, to a lesser extent, the *American Economic Review*) are especially poor.

5. WHY ARE DIAGNOSTIC TESTS USED SO LITTLE?

In this section, several issues relevant to the question of why diagnostic tests are used so little are discussed: the purpose of empirical work; computation; pretesting; significance levels; interpretation; small-sample properties of tests; robustness of results; rationality of paper writers; and the idea market. Economists are used to using cost benefit analysis so in the case of diagnostic tests it must be the case that the expected costs of using diagnostic tests are perceived to outweigh the expected benefits of their use for journal editors, journal referees and paper writers. Applying Phillips' (1988, p. 345) proposition that the most successful paradigms are the ones that survive and multiply, then the use of diagnostic tests as a means to improve economic models must be seen as a paradigm that has dismally failed.

5.1 Purpose of empirical work.

Why do we do empirical work? Although this might appear to be a silly question on its face, it is extremely important. For example, there is a view that stylised facts, a few figures or summary statistics are really all that we need in many cases (Keuzenkamp (1995, p. 238)) or that formal statistical tests have contributed almost nothing to changes in views about key macroeconomic or microeconomic questions (Summers (1991)). An alternative view is that there are very few sharp hypotheses to be tested in economics (for example, Leamer (1988) but see the response by Phillips (1988)). Keuzenkamp and Magnus (1995) contains a variety of views on the significance of testing in econometrics. These views play down the role of hypothesis testing and as indicated in section 2 that is one of the areas where the assumptions made about the disturbances and the regressors can be crucial. In contrast, all the papers surveyed in section 4 report the results of hypothesis tests of one sort or another.

5.2 Computation

In the 1950s, 1960s and the 1970s, the difficulty of computing diagnostic tests may have provided a valid excuse for not reporting them but this is no longer the case. If we look at MICROFIT, SHAZAM and TSP we see that these software now automatically compute various diagnostic tests (see Pesaran and Pesaran (1991), White *et al.* (1990) and Wago and Ban (1994)). For example, for an equation estimated by ordinary least squares or instrumental variables, MICROFIT automatically computes tests for serial correlation, heteroscedasticity, normality and functional form (see Pesaran and Pesaran (1991, pp. 65-69)). Korosi *et al.* (1993) provide a more detailed summary of the diagnostic tests available in some of the commonly used econometric packages.

5.3 Pretesting and Significance Levels

Two issues often raised in the application of diagnostic testing are the questions of pretesting and the overall significance level to be adopted. The importance of pretesting is that the model to be estimated (or the estimation technique to be used) depends on the outcome of one or more hypothesis tests (or diagnostic tests), so the true distribution of the estimator is much more complicated than often suggested. In the simplest case, the linear regression model and the ideal assumptions, it is possible to derive some useful results on this question but not generally (see Giles and Giles (1993) for a recent survey of the literature). When several diagnostic tests are used, the overall significance level is difficult if not impossible to determine (although there are exceptions, see Pagan and Hall (1983b)).

A consideration of the problem of testing for unit roots indicates that these problems are just as severe, if not more severe, in the unit root testing case. Consider the Augmented Dickey Fuller test procedure based on

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{j=1}^{s} \delta_j \Delta y_{t-j} + \varepsilon_t$$

where y_t is the variable being investigated, ε_t is a disturbance, and the hypothesis of interest is H_0: $\gamma=0$, H_1: $\gamma<0$. The choice of s, whether to set $\beta=0$, the possibility of structural change in α or β, whether it is y_t rather than say $\log y_t$ that is of interest, and whether ε_t is serially uncorrelated are all questions that must be answered prior to carrying out the test. Typically decisions are made on the basis of some hypothesis or diagnostic test, that is, there is a problem of pre-test testing (heteroskedasticity does not affect the distribution of a wide class of unit root tests: Phillips (1987)). The widespread usage of this type of unit root tests suggests it is not the problems of pretesting or significance level that are leading applied economists not to use diagnostic tests.

5.4 Small-Sample Properties

As soon as we relax the assumption A4 about the normality of the disturbances or allow for the regressors to be stochastic, the results that can typically be derived are asymptotic in nature. The small sample size and power properties of the diagnostic tests then become important. As Kiviet (1986, 241) argues effective misspecification tests should have correct significance levels irrespective of the true parameters and any redundant regressors in the model and reasonable power against a wide class of alternative specifications. Studies like his provide a good means for at least determining tests that are unlikely to perform well in practice, for example, despite its widespread use the Durbin-h statistic is found to have poor small sample properties. Again it is difficult to argue that the small sample properties of unit root and cointegration tests (or the Durbin-Watson test) are far better that those of other diagnostic tests (for some Monte Carlo results for (a) unit root tests see Phillips and Perron (1988), Schwert (1989) and De Jong *et al.* (1992); (b) cointegration see Engle and Granger (1987), Hakkio and Rush (1991) and Haug (1993); and (c) other diagnostic tests: Bera and Jarque (1982), Thursby (1982, 1989) and Kiviet (1985, 1986)).

5.5 Idea Market and Education

The Meccas of diagnostic tests have been the London School of Economics and the Australian National University, and this is strongly reflected in the individuals associated with the theoretical development of diagnostic tests described in Tables 1 and 2, and in the journals edited out of the United Kingdom (*Applied Economics, Economic Journal*, *Journal of Applied Econometrics* and the *Review of Economic Studies*). An analysis of the location of the editors, co-editors and associate editors indicates a heavy preponderance of individuals working at North American institutions particularly for the *Journal of Political Economy* (5 out of 5), *Journal of Finance* (31.5 out of 32), *American Economic Review* (36 out of 37), *Review of Economics and Statistics* (45.5 out of 49), and *Journal of Econometrics* (21 out of 29). The proportion of editors (and associate editors) educated in the United States where they are unlikely to have been exposed to a significant dose diagnostic testing is also extremely high.

5.6 Robustness and Test Interpretation

The development and wide spread use of Whites (1980) heteroskedasticity-consistent covariance matrix and Newey-West's (1987) heteroskedasticity and autocorrelation consistent covariance matrix would seem to undercut the necessity for diagnostic testing associated with serial correlation and heteroscedasticity. However, this belief is based on a misinterpretation of the meaning of a rejection with a diagnostic test. Despite the common perception that a significant Durbin-Watson test implies serial correlation (see Leamer (1989, p. S20) and Giles and Giles (1993, p. 146)) and so can be corrected by robust standard errors, there are many other possible interpretations (see McAleer (1994)) that would imply correcting the standard errors is not the appropriate course of action. In addition, there is Monte Carlo evidence to suggest that these corrections can make matters worse rather than better (Mishkin (1990)).

5.7 Rationality

In a comment on an earlier version of this paper, Professor Hatanaka suggested that applied economists are merely being rational in their failure to use diagnostic tests. That is, they know their results will not pass even a simple battery of diagnostic tests so they do not use them. The results in Kramer *et al.* (1985) indicating that the eleven empirical papers examined in their sample fail the diagnostic tests more often than can be explained by chance is indirect evidence in support of Professor Hatanaka's conjecture. Furthermore, as the literature on replication in economics indicates it is unlikely that someone will attempt a replication of the empirical results per se let alone subject it to some diagnostic tests (see, for example, Kane (1984), Dewald *et al.* (1984, 1986), Mittelstaedt and Zorn (1984), Merrick (1988), Cartwright (1991), Collins (1991), Hubbard and Vetter (1991) and Tomek (1993)).

6. CONCLUSION

This paper has attempted to document the usage of diagnostic tests by applied economists as reflected in the major economists journal. A comparison of the use of unit root tests with the lack of use of diagnostic tests suggests that the latter can be attributed to the dominance of US institutions in the ideas market and the likelihood that authors will find problems with their simple models if they apply diagnostic tests.

7. ACKNOWLEDGEMENTS

The authors wish to thank Mototsugu Fukushige, Michio Hatanaka, Michael McAleer, Hashem Pesaran, Hiro Toda and participants at the third meeting of the Kansai Quantitative Study Group held in Osaka, May 1995 for their helpful suggestions. An earlier version of this paper was presented at the Western Meeting of the Japan Association of Economics and Econometrics, Fukuoka University, June 1995. The first author would like to acknowledge the financial support of the University of Western Australia and a travel grant provided by the Japanese Ministry of Education. Both authors wish to acknowledge the financial support of the Asset Management Services Industry Fund, Osaka School of International Public Policy, Osaka University.

8. REFERENCES

Bera, A.K., and C.M. Jarque, Model specification tests: a simultaneous approach, *Journal of Econometrics*, 20, 59-82, 1982.

Cartwright, N., Replicability, reproducibility, and robustness: comments on Harry Collins, *History of Political Economy*, 23(1), 143-151, 1991.

Collins, H.M., The meaning of replication and the science of economics, *History of Political Economy*, 23(1), 123-142, 1991.

De Jong, D.N., J.C. Nankervis, N.E. Savin and C.H. Whiteman, The power problems of unit root tests in time series with autoregressive errors, *Journal of Econometrics*, 53, 323-343., 1992

Delgado, M.A., and T. Stengos, Semiparametric specification testing of non-nested econometric models, *Review of Economic Studies*, 61, 291-303, 1994.

Dewald, W.G., J.G. Thursby and R.G. Anderson, Replication in empirical economics: The Journal of Money, Credit and Banking project, *American Economic Review*, 76, 587-603, 1986.

Dewald, W.G., J.G. Thursby and R.G. Anderson, Replication in empirical economics: The Journal of Money, Credit and Banking project: reply, *American Economic Review*, 78, 1162-1163, 1988.

Durbin, J., and G.S. Watson, Testing for serial correlation in least squares regression I, *Biometrika*, 37, 409-428, 1950.

Engle, R.F., and C.W.J. Granger, Co-integration and error correction: representation, estimation, and testing, *Econometrica*, 55, 251-276, 1987.

Giles, J.A., and D.E.A. Giles, Pre-test estimation and testing in econometrics: recent developments, *Journal of Economic Surveys*, 7, 145-197, 1993.

Godfrey, L.G. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*, Cambridge University Press, Cambridge, 1988.

Gozalo, P.L. A consistent model specification test for nonparametric estimation of regression function models, *Econometric Theory*, 9, 451-477, 1993.

Hakkio, C.S., and M. Rush, Co-integration: how short is the long run?, *Journal of International Money and Finance*, 10, 571-581, 1991.

Haug, A.A., Residual based tests for cointegration: a Monte Carlo study of size distortions, *Economics Letters*, 41, 345-351, 1993.

Hendry, D.F. , Econometrics - alchemy or science?, *Economica*, 47, 387-406, 1980.

Hubbard, R., and D.E. Vetter, Replications in the finance literature: an empirical study, *Quarterly Journal of Business and Economics*, 30(4), 70-81, 1991.

Kane, E.J., Why journal editors should encourage the replication of applied econometric research, *Quarterly Journal of Business and Economics*, 23, 3-8, 1984.

Keuzenkamp, H.A., The econometrics of the holy grail - a review of econometrics: alchemy or science?: essays in econometric methodology, *Journal of Economic Surveys*, 9, 233-248, 1995.

Keuzenkamp, H.A. and J.R. Magnus (eds), The significance of testing in econometrics, special issue of the *Journal of Econometrics*, 67(1), 1995.

Kiviet, J.F., Model selection test procedures in a single linear equation of a dynamic simultaneous system and their defects in small samples, *Journal of Econometrics*, 28, 327-362, 1985.

Kiviet, J.F., On the rigour of some misspecification tests for modelling dynamic relationships, *Review of Economic Studies*, 53, 241-261, 1986.

Korosi, G., L. Matyas and I. Szekely, Comparative review of some econometric software packages, *Journal of Economic Surveys*, 7, 105-118, 1993.

Kramer, W., H. Sonnberger, J. Maurer and P. Havlik, Diagnostic checking in practice, *Review of Economics and Statistics*, 67, 118-123, 1985.

Leamer, E.E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley and Sons, New York, 1978.

Leamer, E.E., Things that bother me, *Economic Record*, 64, 331-335, 1988.

Leamer, E.E., Planning, criticism, and revision, *Journal of Applied Econometrics*, 4, S5-S27, 1989.

Leamer, E.E., Bayesian elicitation diagnostics, *Econometrica*, 60, 919-942, 1992.

Leamer, E.E., and L. Herman, Reporting the fragility of regression estimates, *Review of Economics and Statistics*, 65, 306-317, 1983.

MacKinnon, J.G., Model specification tests and artificial regressions, *Journal of Economic Literature*, 30, 102-146, 1992.

McAleer, M., Specification tests for separate models: a survey, in *Specification Analysis in the Linear Model*, edited by M.L. King and D.E.A. Giles, Routledge & Kegan Paul, New York, 1987.

McAleer, M., Sherlock Holmes and the search for truth: a diagnostic tale, *Journal of Economic Surveys*, 8, 317-370, 1994.

McAleer, M., The significance of testing empirical non-nested models, *Journal of Econometrics*, 67, 149-171, 1995.

Merrick, J.J., Replication in empirical economics: the Journal of Money, Credit and Banking project: comment, *American Economic Review*, 78, 1160-1161, 1988.

Mishkin, F., Does correcting for heteroscedasticity help?, *Economics Letters*, 34,351-356, 1990.

Mittelstaedt, R.A. and T.S. Zorn, Econometric replication: lessons from the experimental sciences, *Quarterly Journal of Business and Economics*, 23(1), 9-15, 1984.

Pagan, A.R., and A.D. Hall, Diagnostic tests as residual analysis, Working Paper in Economics and Econometrics No. 087, Australian National University, 1983a.

Pagan, A.R., and A.D. Hall, Diagnostic tests as residual analysis, *Econometric Reviews*, 2, 159-254, 1983b (with discussion).

Pesaran, M.H., and B. Pesaran, *MICROFIT 3.0: An Interactive Econometric Software Package*, Oxford University Press, Oxford, 1991.

Phillips, P.C.B., Time series regression with a unit root, *Econometrica*, 55, 277-301, 1987.

Phillips, P.C.B., Reflections on econometric methodology, *Economic Record*, 64, 344-359, 1988.

Phillips, P.C.B., and P. Perron, Testing for a unit root in time series regression, *Biometrika*, 75, 335-346, 1988.

Rilstone, P., A simple Bera-Jarque normality test for nonparametric residuals, *Econometric Reviews*, 11, 355-365, 1992.

Schwert, G.W., Tests for unit roots: a Monte Carlo investigation, *Journal of Business and Economic Statistics*, 7, 147-159, 1989.

Summers, L.H., The scientific illusion in empirical macroeconomics, *Scandinavian Journal of Economics*, 93, 129-148, 1991.

Thursby, J.G., Misspecification, heteroscedasticity, and the Chow and Goldfeld-Quandt tests, *Review of Economics and Statistics*, 64, 314-321, 1982.

Thursby, J.G., A comparison of several specification error tests for a general alternative, *International Economic Review*, 30, 217-230, 1989.

Tomek, W.G., Confirmation and replication in empirical econometrics: a step toward improved scholarship, *American Journal of Agricultural Economics*, 75, 6-14, 1993.

Wago, H., and K. Ban, *TSP ni yoru keizai deeta no bunseki [dai 2 han] (Applied Econometric Analysis Using TSP [2nd Edition])*, University of Tokyo Press, Tokyo, 1994.

White, H., A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, 48, 817-838, 1980.

White, K.J., S.D. Wong, D. Whistler and S.A. Haun, *SHAZAM Econometric Computer Program: Users Reference Manual Version 6.2*, McGraw-Hill, Toronto, 1990.

TABLE 1: THE DEVELOPMENT OF DIAGNOSTIC TESTS IN ECONOMETRICS - CHRONOLOGICAL ORDER

YEAR	TEST	PURPOSE	METHOD	REFERENCES
1950	Durbin-Watson	Serial Correlation [AR(1)]	OLS	Durbin-Watson (1950, 1951)
		Overidentification	LIML	Anderson-Rubin (1950)
1957		Serial Correlation	LIML	Durbin (1957)
1959		Serial Correlation	IV	Sargan (1959)
1960	Chow	Structural Change (known break point)	OLS	Chow (1960)
	Predictive Failure	Structural Change	OLS	Chow (1960) Salkever (1976)
		Overidentification	2SLS	Basmann (1960)
1964		General Misspecification	IV	Sargan (1964)
	Common factor	Serial Correlation as a Common Factor	OLS	Sargan (1964), Hendry-Mizon (1978)
	Box-Cox	Linear vs Log-linear	NL	Box-Cox (1964)
1965	Goldfeld-Quandt	Heteroscedasticity	OLS	Goldfeld-Quandt (1965)
1969	RESET	Functional Form	OLS	Ramsey (1969)
	Glejser	Heteroscedasticity	OLS	Glejser (1969)
1970	Durbin-h	Serial Correlation [AR(1)]	OLS	Durbin (1970)
	Box-Pierce	Serial Correlation[general]	OLS	Box-Pierce (1970)
1971		Serial Correlation	k-class	Bouman (1971)
1973	Durbin-Wu -Hausman	Exogeneity of Explanatory Variables	OLS	Durbin (1954), Wu (1973) Hausman (1978)
1974	N	Non-Nested Models	OLS	Pesaran (1974)
1975	CUSUM CUSUMSQ	Structural Change (unknown break point)	OLS	Brown *et al.* (1975)
1976		Serial Correlation	IV	Godfrey (1976)
1977		Structural Change	2SLS	Harvey-Phillips (1977, 1989)
1978	LM	Serial Correlation [AR(p)]	OLS	Breusch (1978)
			GLS	Godfrey (1978a, b)
	Ljung-Box	Serial Correlation[general]	OLS	Ljung-Box (1978)
1979	Breusch-Pagan	Heteroscedasticity (known form)	OLS	Breusch-Pagan (1979) Godfrey (1978c)
1979	Dickey-Fuller	Unit Root	OLS	Dickey-Fuller (1979)
1980	Jarque-Bera	Normality	OLS	Jarque-Bera (1980) Bera-Jarque (1981)
		Serial Correlation	2SLS	Harvey-Phillips (1980)
1981	J	Non-nested models	OLS	Davidson-MacKinnon (1981)
		Linear vs Log-linear	OLS	Godfrey-Wickens (1981)
		Heteroscedasticity (LR)	Tobit	Petersen-Waldman (1981)
		Heteroscedasticity	2SLS	Harvey-Phillips (1981)
		Exogeneity	2SLS	Spencer-Berk (1981, 82)
		Non-normality	Tobit	Nelson (1981)
1982	Encompassing	Non-nested models	OLS	Deaton (1982)
	ARCH	Autoregressive Conditional Heteroscedasticity	OLS	Engle (1982)
		Heteroscedasticity	2SLS	Kelejian (1982)
		Normality	Tobit	Jarque-Bera (1982)
		Heteroscedasticity	Tobit	Jarque-Bera (1982)
	Differencing	Misspecification	OLS	Plosser *et al.* (1981) Davidson *et al.* (1985)
	J	Overidentification	GMM	Hansen (1982)

YEAR	TEST	PURPOSE	METHOD	REFERENCES
1983	PE	Linear vs Log-Linear	OLS	MacKinnon *et al.* (1983)
		Structural change	2SLS	Erlat (1983)
		Exogeneity	Tobit	Smith-Blundell (1983, 1986)
		Non-Nested Models	IV	Godfrey (1983)
				Ericsson (1983)
		Normality	IV	Pagan-Hall (1983a)
	RESET	Functional Form	IV	Pagan-Hall (1983b)
1984		Normality	LDEP	Bera et al. (1984)
		Heteroscedasticity	DDEP	Davidson-MacKinnon (1984)
		Misspecification	Tobit	Fin-Schmidt (1984)
1985		Structural change	2SLS	Lo-Newey (1985)
1987	Engle-Granger	Cointegration (1 vector)	OLS	Engle-Granger (1987)
	Predictive Failure	Parameter Constancy	LDEP	Anderson (1987)
1988	Johansen	Cointegration (s vectors)	VAR(p)	Johansen (1988)
1989	BM	Linear vs Log-Linear	OLS	Bera-McAleer (1989)
	Predictive Failure	Parameter constancy	GMM	Hoffman-Pagan (1989)
				Ghysels-Hall (1990a)
		Functional Form, Normality, Heteroscedasticity, Serial Correlation	GMM	Pagan-Vella (1989)
		Distribution	LDEP	Smith (1989)
1990		Heteroscedasticity (LM)	Tobit	Greene (1990)
		Non-Nested Models	GMM	Ghysels-Hall (1990b)
				Smith (1992)
1992		Serial correlation	GMM	Cumby-Huizinga (1992)
1993		Heteroscedasticity	GMM	Pagan-Pak (1993)

Note:

In the METHOD column, OLS = ordinary least squares estimator; LIML = limited information maximum likelihood estimator; 2SLS = two stage least squares estimator; IV= instrumental variable estimator; GLS = generalized least squares estimator; NL= non-linear least squares; GMM = generalized method of moments; k-class = k-class estimator; LDEP = limited dependent variable estimator; DDEP = discrete dependent variable estimator; VAR(p) = pth order vector autoregression estimator; and Tobit = Tobit estimator.

TABLE 2: THE DEVELOPMENT OF DIAGNOSTIC TESTS IN ECONOMETRICS - PROBLEM TYPE

PROBLEM	ESTIMATION METHOD			
	OLS	**IV/2SLS**	**GMM**	**Tobit/Limited Dependent**
Serial Correlation	Durbin-Watson (50,51), Durbin (70) Box-Pierce (70), Breusch (78) Godfrey (78a,b), Ljung-Box (78)	Sargan (59), Godfrey (76) [Durbin (57), Bouman (71)] Harvey-Phillips (80)	Pagan-Vella (89) Cumby-Huizinga (92)	Petersen-Waldman (81)
Structural Change	Chow (60) [Salkever (76)] Brown *et al.* (75)	Harvey-Phillips (77), Erlat (83) Pagan-Hall (83b), Lo-Newey (85)	Hoffman-Pagan (89) Ghysels-Hall (90a)	Anderson (87)
Heteroscedasticity	Goldfeld-Quandt (65), Glejser (69) Godfrey (78c), Breusch-Pagan (79) Engle (82)	Harvey- Phillips (81) Kelejian (82), Pagan-Hall(83b)	Pagan-Vella (89) Pagan-Pak (93)	Jarque-Bera (82) Davidson-MacKinnon(84) Greene (90)
Functional Form	Ramsey (69) [Anscombe (61)]	Pagan-Hall(83b)	Pagan-Vella (89)	
Normality	Jarque-Bera (80), Bera-Jarque (81)	Pagan-Hall(83a)	Pagan-Vella (89)	Nelson (81), Bera *et al.* (82) Jarque-Bera (82), Smith (89)
Non-Nested Models	Pesaran (74) [Cox (61,62)] Davidson-MacKinnon (81), Deaton(82)	Godfrey (83), Ericsson (83)	Ghysels-Hall (90b) Smith (92)	
Overidentification		Basmann (60) [Anderson- Rubin(49,50)]	Hansen (82)	
General Misspecification	Plosser *et al.* (82), Davidson *et al.* (85)	Sargan (64), Sargan (76)		Finn-Schmidt (84)
Linear vs Log-Linear	Godfrey-Wickens (81) MacKinnon *et al.* (83), Bera-McAleer (89)			
Exogeneity	Wu (73), Hausman (78) [Durbin (54)]	Spencer-Berk (81,82)		Smith-Blundell (83, 86)
Unit Root	Dickey-Fuller(79)			
Cointegration	Engle-Granger (87)			
Common Factor	Sargan (64), Hendry-Mizon (78)			

TABLE 3: JOURNALS SURVEYED AND SURVEY RESULTS FOR 1994

JOURNAL	EDITORS[1]	NUMBER OF PAPERS	EMPIRICAL PAPERS[2]	PAPERS WITH DIAGNOSTIC TESTS[3]
American Economic Review (AER)	O. Ashenfelter R.H. Gordon R.P. McAfee K.D. West	92	45	13
Applied Economics (AE)	M.H. Peston	128	127	51
Econometrica (ECON)	G. Laroque D. Card D. Gale P. Robinson	35	5	1
Economic Journal (EJ)	J.D. Hey	81	41	18
Economic Studies Quarterly (ESQ)	T. Ihori K. Nishimura T. Yamamoto H. Yoshikawa	23	6	0
International Economic Review (IER)	W.J. Ethier H. Miyazaki	55	10	5
Journal of Applied Econometrics (JAE)	M.H. Pesaran J. Geweke A. Kapteyn N.M. Kiefer M. Watson	30	23	11
Journal of Econometrics (JoE)	T. Amemiya R. Blundell A.R. Gallant C. Hsiao A. Zellner	77	33	13
Journal of Finance (JF)	R.M. Stulz S.A. Buser D. Mayers	56	37	10
Journal of Political Economy (JPE)	G.S. Becker R.E. Lucas S. Rosen J.A. Scheinkman R. Topel	50	27	6
Review of Economic Studies (REStud)	C.R. Bean	40	9	5
Review of Economics Statistics (REStat)	R.E. Caves R.A. Moffitt J.H. Stock	80	68	31

Notes:

1. Editors are defined as those individuals listed in the journal's first issue of 1994 as either an editor or co-editor. For the *Review of Economic Studies*, the editor is taken as the Chairman of the Editorial Board.
2. Empirical paper is defined as any paper reporting at least one regression result.
3. Paper with a diagnostic test is defined as a paper where the results of a diagnostic test are explicitly presented or the results of it are discussed.

TABLE 4: DETAILS OF DIAGNOSTIC TESTS FOR MODELS ESTIMATED BY OLS

JOURNAL	DATA	NO	NONE	DW	Dh	ADF	Z	EG	JOH	CHOW	Q	HAUS	ARCH	LM	JB	RESET	OTHER
AER	TS	16	8	6	0	3	1	0	1	1	1	1	0	0	0	0	0
	XS	7	6	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	O	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AE	TS	22	2	8	3	12	2	1	3	3	4	0	4	4	3	2	8
	XS	11	7	1	0	0	0	0	0	0	0	1	0	0	0	0	3
	O	2	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1
ECON	TS	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	XS	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EJ	TS	9	0	2	1	6	1	1	3	2	0	1	1	2	0	1	2
	XS	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	7	3	0	0	1	1	0	0	0	0	3	0	0	0	0	0
IER	TS	2	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0
JAE	TS	3	0	3	1	2	2	0	0	2	0	0	0	2	1	1	3
	XS	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1
JoE	TS	2	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1
	XS	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JF	TS	11	8	0	0	0	2	0	0	0	0	0	0	0	0	0	1
	XS	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	6	5	1	0	0	0	0	0	0	0	0	0	0	0	0	1
JPE	TS	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	XS	7	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O	9	7	0	0	0	0	0	0	0	0	0	0	0	0	0	2
REStat	TS	12	6	0	0	4	2	0	1	0	1	0	0	2	1	1	0
	XS	9	7	0	0	0	0	0	0	0	0	0	1	0	0	0	1
	O	13	6	2	0	0	0	0	0	0	0	2	0	0	0	0	4

Notes: In the data column, TS=time-series data, XS = cross-section data and O = pool or panel data. NO: number of papers, NONE: number of papers with no diagnostics, DW: Durbin-Watson test, DH: Durbin h test, ADF: augmented Dickey-Fuller test, Z: Phillips-Perron test; EG: Engle-Granger cointegration test, JOH: Johansen test, CHOW: Chow test, Q: Box-Pierce or Ljung-Box test, HAUS: Hausman test, ARCH: ARCH test, LM: Lagrange Multiplier test for serial correlation, JB: Jarque-Bera normality test, RESET: Reset test, OTHER: other test.