Controlling the Significance Levels of Prediction Error Tests for Linear Regression Models*

Leslie G. Godfrey† and Chris D. Orme‡

revised, May 2000

Abstract

This paper provides evidence on problems associated with using standard tests for predictive failure when the errors of a linear regression model are not normally distributed. The ability of a simple bootstrap procedure to give a useful degree of control over the significance levels is examined.

1. Introduction

Tests for predictive failure are used in many applied studies involving the least squares estimation of a linear regression model. Outcomes of such tests are often reported for cases in which the number of predictions is rather smaller than the number of observations used to estimate the regression model. Indeed, the number of predictions is not infrequently smaller than the number of regressors in the estimated model. In such cases, the test of prediction errors described in Chow's (1960) seminal article is widely used and a modification of this test due to Hendry (1980) is sometimes employed. However, the validity of both of these procedures requires the normality of the error term in the regression model that, under the null hypothesis, applies to all the data. Asymptotic analysis shows that the tests of prediction errors proposed by Chow and Hendry are both vulnerable

*We are grateful to Karim Abadir, Mike Veall, two referees and Richard Smith for their constructive comments.

†Department of Economics and Related Studies, University of York, YO10 5DD, United Kingdom.

‡School of Economic Studies, University of Manchester, M13 9PL, United Kingdom.

to nonnormality, even when the size of the estimation sample is large; see, for example, Stewart and Gill (1998, pp. 82-83) for a discussion of Chow's procedure.

In this paper, we present Monte Carlo evidence that illustrates that: (i) nonnormality can have serious effects on the rejection rates of standard prediction error tests; and (ii) using a nonparametric bootstrap can give a useful degree of control over finite sample significance levels.

The plan of this paper is as follows. Section 2 contains details of the model, test procedures and asymptotic analysis. The Monte Carlo experiments are described in Section 3, and the results that they produce are discussed in Section 4. Finally, Section 5 contains some concluding remarks.

2. The Model and Test Procedures

Let y_t be a random dependent variable and x_t a $(k \times 1)$ vector of strictly exogenous variables with its first element equal to unity for all t. It is assumed that the following linear regression model applies to a set of $n_1 > k$ observations

$$y_t = x_t'\beta + u_t, \tag{2.1}$$

in which β is a $(k \times 1)$ unknown coefficient vector. The error terms u_t are independently and identically distributed (iid) with zero mean and variance σ^2. Without loss of generality, let this set of observations be for $t = 1, \ldots, n_1$, and hereafter it will be referred to as the estimation sample.

In addition to the estimation sample, there is a prediction sample consisting of n_2 observations on (y_t, x_t'), $t = n_1 + 1, \ldots, n$, where $n = n_1 + n_2$. The estimate of β in (2.1) derived from the estimation sample is used to predict the values of y_t, conditional upon x_t, $t = n_1 + 1, \ldots, n$. The predictive validity of the model is assessed by checking the joint significance of the differences between actual and predicted values of y_t, i.e., the prediction errors, for $t = n_1 + 1, \ldots, n$.

It is useful, at this point, to introduce some additional notation and assumptions. Let X_1 be the $(n_1 \times k)$ matrix with typical row x_t', $t = 1, \ldots, n_1$. It is assumed that $\text{rank}(X_1) = k < n_1$. A matrix-vector version of (2.1) for the estimation sample is

$$y_1 = X_1\beta + u_1 \tag{2.2}$$

in which y_1 and u_1 are both $(n_1 \times 1)$ having typical elements y_t and u_t, respectively, $t = 1, \ldots, n_1$. The coefficient vector estimator derived by Ordinary Least Squares (OLS) estimation of (2.2) is $\hat{\beta}_1 = (X_1'X_1)^{-1} X_1'y_1$, and the usual error variance estimate is $s_1^2 = \hat{u}_1'\hat{u}_1/(n_1 - k)$, where $\hat{u}_1 = y_1 - X_1\hat{\beta}_1$ is the residual vector.

The data for the dependent variable and regressors in the prediction sample are the elements of the $(n_2 \times 1)$ vector y_2 and $(n_2 \times k)$ matrix X_2, respectively. Let $y = (y_1', y_2')'$ and $X = (X_1', X_2')'$. If (2.1) applies to all observations, we can write

$$y = X\beta + u \tag{2.3}$$

with $u' = (u_1, u_2, ..., u_n)$.

Chow proposes testing the joint significance of the prediction errors which are the elements of $(y_2 - X_2\hat{\beta}_1)$ using

$$P = \frac{(\tilde{u}'\tilde{u} - \hat{u}_1'\hat{u}_1)/n_2}{(\hat{u}_1'\hat{u}_1)/(n_1 - k)}, \tag{2.4}$$

in which $\tilde{u} = y - X\tilde{\beta}$, the OLS residual vector obtained when all n observations are used for estimation, with $\tilde{\beta} = (X'X)^{-1}X'y$. Provided $y \sim N(X\beta, \sigma^2 I_n)$, P has the $F(n_2, n_1 - k)$ distribution with large values of this test statistic indicating predictive failure. Thus, when the significance of (2.4) is assessed using right-hand tail critical values of the $F(n_2, n_1 - k)$ distribution, the null model under test includes the assumption that the errors u_t are normally distributed.

There are alternatives to Chow's test. Hendry (1980) has proposed a large sample test of predictive failure that, like Chow's procedure, requires the errors to be normally distributed. Under normality, it is asymptotically valid (as $n_1 \to \infty$) to compare sample values of

$$H = \frac{(y_2 - X_2\hat{\beta}_1)'(y_2 - X_2\hat{\beta}_1)}{(\hat{u}_1'\hat{u}_1)/(n_1 - k)} \tag{2.5}$$

to right-hand-tail critical values of the $\chi^2(n_2)$ distribution; see Hendry (1980, p.222) and Kiviet (1986, Section 4).

The assumption of normality that justifies both the exact test P and the large sample test H is convenient, but it may not provide a very good approximation in practical situations. The effects of nonnormality therefore merit consideration.

2.1. Nonnormality and asymptotic analysis

The total sample of n observations is made up of n_1 observations in the estimation sample and n_2 observations in the prediction sample. It is necessary to decide what to assume about the separate behaviour of n_1 and n_2 as their sum tends to infinity. Previous studies, in which asymptotic results are obtained for predictive

failure tests, have assumed $n_1 \to \infty$ and $n_2 \to \infty$, e.g., with $n_2 = \tau n_1$, τ fixed; see Hoffman and Pagan (1989), and West and McCracken (1998). However, in published empirical work using estimated linear regression models, it is often the case that n_2 is small relative to n_1 and is sometimes smaller than k. In order to generate approximations relevant to such cases, we carry out an asymptotic analysis in which $n_1 \to \infty$ and n_2 is fixed.

The following assumptions are made.

(a) The error terms u_t are *iid* having common *cumulative distribution function* (*cdf*), $\Im$, with zero mean, variance σ^2, and finite fourth moment.

(b) If the regressors are not stochastic, they satisfy the weak requirement that $x_t'(X_1'X_1)^{-1}x_t$ is $o(1)$ as $n_1 \to \infty$ for all t. Under this assumption, Huber's (1973) condition for the asymptotic validity of standard OLS inference is met for both $\hat{\beta}_1$ and $\tilde{\beta}$. Moreover, this requirement, combined with the Cauchy-Swartz inequality, implies the asymptotic irrelevance of estimation effects that justifies Hendry's (1980) test H.

(c) If the regressors are stochastic, the required primitive conditions usually vary from one setting to another, e.g. cross-section or time series, but the following conditions are assumed to hold: $(X_1'X_1)^{-1}X_1'u_1$ and $(X'X)^{-1}X'u$ are $o_p(1)$ as $n_1 \to \infty$, as is $x_t'(X_1'X_1)^{-1}x_t$ for all t. Given these assumptions, $\hat{\beta}_1$ and $\tilde{\beta}$ are both consistent for β whilst s_1^2 is consistent for σ^2, when $n_1 \to \infty$ with n_2 fixed. White (1980) provides assumptions that ensure the consistency of OLS in a class of models relevant to cross-section analysis. For models estimated using time series data, the various sets of regularity conditions provided by Hamilton (1994, Chapter 8) may be of interest. If the linear regression equation being used is an autoregressive distributed lag model involving integrated variables, the regularity conditions and results of Pesaran and Shin (1999) can be employed.

Under the null hypothesis of model constancy and our assumptions, it is clear that Chow's prediction error test statistic P of (2.4) is asymptotically equivalent to

$$P^* = \frac{u_2'u_2}{n_2\sigma^2}, \tag{2.6}$$

where $u_2 = y_2 - X_2\beta$, when $n_1 \to \infty$ and n_2 is fixed. Equation (2.6) implies that the conventional Chow test statistic P is, under the assumptions of this paper, asymptotically equivalent to the average squared value of the last n_2 elements of the standardized vector $v = \sigma^{-1}u$. If the assumption of normality is added to (2.3) , the asymptotic distribution of P is $\chi^2(n_2)/n_2$; where $\chi^2(q)$ denotes a chi-squared distribution with q degrees of freedom. It is straightforward to show that

Hendry's (1980) statistic H of (2.5) is asymptotically equivalent to $n_2 P^*$ and so, under normality and model stability, it is asymptotically distributed as $\chi^2(n_2)$.

The assumption of normality is crucial here since P^* is not proportional to a $\chi^2(n_2)$ random variable when the errors are not normally distributed. Thus, in practice, the case of nonnormal regression errors renders the asymptotic distributions of H and P unknown. These test statistics are not asymptotically pivotal in the sense of Beran (1988) because, although invariant to β and σ^2, their large sample distributions, under the null hypothesis that (2.1) holds for all observations, are determined by the distribution of $u_2'u_2$ which depends upon the cdf $\Im$. This dependence has implications for simulation-based methods for determining critical values which are discussed in the next sub-section.

Equation (2.6) is valid under the null hypothesis. The corresponding equation derived under the fixed alternative

$$y_t = x_t'\beta + \xi_t + u_t,\ u_t\ iid(0,\sigma^2),\ t = n_1+1, ..., n, \tag{2.7}$$

in which $\xi_t \neq 0$ for some t, may serve to provide some indication of the power of the tests. Under such alternatives, (2.6) must be replaced by

$$P_A^* = \frac{(u_2+\xi)'(u_2+\xi)}{n_2\sigma^2},$$

in which ξ is the $(n_2 \times 1)$ vector with typical element ξ_{t+n_1}, $t = 1, \ldots, n_2$. As with P^*, the distribution of P_A^* is unknown when the error distribution is unspecified.

2.2. Simulation-based tests

If the (nonnormal) error distribution were known, it would be possible to use the parametric bootstrap. More precisely, given knowledge of the true error distribution, many artificial samples could be generated, conditional upon the observed regressor values. The values of the prediction error test statistic (H or P) calculated from these artificial samples could then be used to estimate finite sample critical values or *p-values*. The parametric bootstrap tests are akin to the Monte Carlo type tests of Dufour and Kiviet (1997, 1998). However, in contrast to the types of situation considered by Dufour and Kiviet in which the wrong choice of the *cdf* $\Im$ does not imply asymptotically invalid tests, Monte Carlo tests and parametric bootstrap tests suffer from a major drawback when applied to the prediction error test statistics H and P.

As noted above, the distributions of H and P, under the null hypothesis, depend upon the *cdf* $\Im$ for fixed n_2, even when $n_1 \to \infty$. Hence, an incorrect choice of $\Im$ for either the parametric bootstrap or a Monte Carlo-type test will lead to invalid inferences for fixed n_2. It seems unlikely that applied workers will have access to very accurate information about the general form of the error distribution, let alone have perfect knowledge. The parametric bootstrap is, therefore, unlikely to provide a reliable tool for assessing the significance of observed values of the prediction error test statistics H and P.

The nonparametric bootstrap, in which randomly selected residuals from OLS estimation of the null model are used to generate artificial samples, does not require specification of the error distribution. Furthermore, the nonparametric bootstrap provides a basis for valid inferences when $n_1 \to \infty$ and n_2 is fixed, even though H of (2.5) and P of (2.4) are not asymptotically pivotal; see Beran (1988, Section 3).

If the regressors of (2.3) are strictly exogenous, the nonparametric bootstrap can be implemented by generating artificial observations using

$$y_t^* = x_t'\tilde{\beta} + u_t^*, \quad t = 1, \ldots, n \tag{2.8}$$

in which the bootstrap errors $(u_1^*, u_2^*, \ldots, u_n^*)$ are a simple random sample drawn with replacement from the empirical distribution of the observed residuals; i.e., from

$$\widetilde{\Im} : \text{ probability } \frac{1}{n} \text{ on } \tilde{u}_t, \qquad t = 1, \ldots, n. \tag{2.9}$$

The use of $\tilde{\beta}$, the OLS estimator derived under the imposition of the null hypothesis, and the associated *empirical distribution function* (*edf*), $\widetilde{\Im}$, as estimators of β and $\Im$ satisfies Beran's (1988) conditions for the bootstrap test to have an error in rejection probability that tends to zero as $n_1 \to \infty$.

If the regressors include lagged values of the dependent variable, as well as strictly exogenous variables, (2.8) is inappropriate; see Li and Maddala (1996) for a recent review of bootstrapping time series models, including regression equations with an autoregressive component. Suppose that (2.1) is generalized to allow for the inclusion of the first p lags of y_t in the regressors, so that

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + x_t'\beta + u_t.$$

Standard regularity conditions require that the roots of $(1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p) = 0$ are outside the unit circle and it assumed that the OLS estimates $\widetilde{\phi}_j$ are consistent with these restrictions. Equation (2.8) is then replaced by

$$y_t^* = \widetilde{\phi}_1 y_{t-1}^* + \widetilde{\phi}_2 y_{t-2}^* + ... + \widetilde{\phi}_p y_{t-p}^* + x_t'\tilde{\beta} + u_t^*, \quad t = 1, \ldots, n,$$

where the bootstrap sample starting values are set equal to the actual estimation sample starting values; see Li and Maddala (1996, Section 2.3). The u_t^* are, as in the static case, generated using (2.9).

It is worth noting that the use of the terms y_{t-j} has implications for the standard Chow test applied to models with normal errors. If the regressors were all strictly exogenous and the errors u_t were independent $N(0, \sigma^2)$ variables, a critical value from the relevant F distribution would be more appropriate than the corresponding bootstrap estimate. (In this special case, the former involves no error while the latter has an error that is $o(1)$.) However, when the regressors include lagged values of the dependent variable, the F test is only asymptotically valid under normality and so, like the bootstrap test, has an error in rejection probability that is $o(1)$.

The bootstrap is not the only method available for controlling the significance levels of prediction error tests in the presence of nonnormal disturbances. Dufour, Ghysels and Hall (1994) consider predictive tests in the context of nonlinear dynamic simultaneous equation models. Their results can be specialized to the case of the linear regression model. They examine two approaches: the use of Markov inequalities to give bounds on desired rejection probabilities; and the use of semi-nonparametric (SNP) methods to model the error distribution so that an estimated error distribution can be employed to calculate rejection rates.

The Markov inequality approach, like the nonparametric bootstrap, is applicable in quite general situations and involves the generation of artificial samples. The resampling is done to estimate the expected value of the test statistic under the null hypothesis; this estimate being needed for the calculation of the Markov upper bound. However, the Markov procedure does not offer the same kind of control over the significance level as the nonparametric bootstrap. The difference between the estimated Markov bound and the true rejection probability is $O(1)$ (so it is not asymptotically negligible) and can be large relative to the latter. For example, if W has the $\chi^2(2)$ distribution, so that $E(W) = 2$, the Markov inequality gives

$$prob(W \geq 6) \leq \frac{2}{6},$$

i.e., a bound of about 0.33, when the true probability is slightly less than 0.05.

The SNP approach overcomes the problem of producing only conservative probability statements, even asymptotically. Dufour et al. (1994, Section 5) discuss estimating the distribution of the error term using SNP approximations. A practical problem is that a *probability density function* (*pdf*) must be specified as a lead term and, as acknowledged by Dufour et al., the choice of this term may be influential on the outcome of the test. Sensitivity analyses could be carried out to examine the effects of different choices for the lead term, but this task would involve a substantial computational burden. As will be seen below, the nonparametric bootstrap can work quite well and does not need any choices about lead term *pdf's* to be made.

3. Monte Carlo Design

Results are obtained from two sets of experiments. For each set of experiments, we use all six combinations of (n, n_2) from $n = 50, 80, 100$ and $n_2 = 3, 6$.

3.1. Set 1: static data generation processes

The regression model used in the first set of experiments is

$$y_t = \sum_{j=1}^{6} x_{tj}\beta_j + u_t, \tag{3.1}$$

in which: $x_{t1} = 1$; x_{t2} is drawn from a uniform distribution with parameters 1 and 31; x_{t3} is drawn from a log-normal distribution with $\ln(x_{t3}) \sim N(3, 1)$. Unlike x_{t2} and x_{t3}, the remaining regressors are serially correlated with

$$\begin{aligned} x_{t4} &= 0.9x_{t-1,4} + v_{t4}, \\ x_{t5} &= 0.6x_{t-1,5} + v_{t5}, \\ x_{t6} &= 0.3x_{t-1,6} + v_{t6}, \end{aligned}$$

with v_{ts} being independently normally distributed, such that $E[x_{ts}] = 0$ and $var[x_{ts}] = 1$, for $s = 4, 5, 6$.[1] This group of regressors includes as special cases the types of regressors used in several earlier Monte Carlo studies of tests, e.g. for heteroskedasticity. It allows for serially correlated regressors and nonnormality of the regressors.

[1] All pseudo-random numbers are obtained using routines from the NAG library.

All regression coefficients β_j are set equal to zero. This involves no loss of generality, since Breusch's (1980) invariance results imply that the same results would be obtained whatever the values of the coefficients β_j. The error terms, u_t, of (3.1) are *iid* $(0,1)$ in all experiments.[2] Since sensitivity to nonnormality is of considerable practical interest, the disturbances u_t are obtained by standardizing pseudo-random variables drawn from several distributions. These disturbance distributions are: normal, denoted N; Student's t with 5 degrees of freedom, t_5; uniform, over the unit interval, U; chi-square with 2 degrees of freedom, $\chi^2(2)$; and, log-normal, Λ.

3.2. Set 2: dynamic data generation processes

The classical assumption that regressors are either fixed in repeated sampling or strictly exogenous is inappropriate in the many applied articles based upon models that include lagged values of the dependent variable in the regressor set. Moreover, in recent years, emphasis has been placed upon the possibility that the dependent variable and some regressors are unit root processes. In order to examine the performance of the conventional and nonparametric bootstrap tests under these two departures from classical assumptions, a second set of experiments is carried out using a dynamic data generation process employed by Giersbergen and Kiviet (1996) .

The autoregressive-distributed lag (ADL) model differs from that of Giersbergen and Kiviet (1996) only in that we allow for nonnormality of errors. The ADL is, therefore, written as

$$y_t = \theta_1 y_{t-1} + \theta_2 z_t + \theta_3 z_{t-1} + \theta_4 + u_t \tag{3.2}$$

in which the u_t are *iid* $(0, \sigma^2)$, z_t is integrated of order one and $|\theta_1| < 1$; so that y_t and z_t are cointegrated CI(1,1). The marginal model for z_t is given by

$$\Delta z_t = \pi \Delta y_{t-1} + \gamma \Delta z_{t-1} + \eta_t, \tag{3.3}$$

in which the terms η_t are independently and normally distributed with zero mean and variance σ_η^2, and η_t and u_s are independent for all t and s. Nonzero values of π in (3.3) imply that z_t is predetermined, but not strictly exogenous.

The choice of parameters for the experiments follows Giersbergen and Kiviet very closely; see Giersbergen and Kiviet (1996, Appendix) for a discussion of

[2] The restriction that $var[u_t] = 1$ causes no loss of generality since the results are invariant with respect to the value of the common variance, under the null hypothesis.

computational details. The coefficients of (3.2) are selected to be similar to those found in empirical studies of the consumption function with

$$(\theta_1, \theta_2, \theta_3) = (0.8, 0.5, -0.3),$$

and, without loss of generality, $\theta_4 = 0.0$. The variance parameter σ^2 is specified so that the error correction model corresponding to (3.2) has a population R^2 equal to 0.8, which again is similar to values found in practical situations. We use $(0.0, 0.8)$ and $(-0.4, 0.4)$ for (π, γ) in (3.3); see Giersbergen and Kiviet (1996, p. 641, Table 1) for the implied roots of the dynamic process. The variance of the error term of (3.3) is, without loss of generality, set equal to 1. The errors u_t are drawn from the five distributions used for static data generation processes. Given a choice of coefficients and error distribution, starting values of y and z are set equal to zero and then "observed data" are generated for $t = -49, ..., 0, 1, ..., n$. The first 49 observations are discarded to reduce the impact of the fixed initial conditions. The values y_0 and z_0 are regarded as start-up values for the sequence $\{(y_t, z_t); t = 1, ..., n\}$.

When $\pi = 0$ in (3.3), the sequence $\{z_t\}$ can be treated as fixed over bootstraps and the bootstrap samples are obtained using

$$y_t^* = \widetilde{\theta}_1 y_{t-1}^* + \widetilde{\theta}_2 z_t + \widetilde{\theta}_3 z_{t-1} + \widetilde{\theta}_4 + u_t^*, \quad t = 1, \ldots, n \tag{3.4}$$

in which: $y_0^* = y_0$; $(\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\theta}_3, \widetilde{\theta}_4)$ is the vector of OLS estimates for (3.2) based upon using all n observations; and the bootstrap errors $(u_1^*, u_2^*, \ldots, u_n^*)$ are a simple random sample drawn with replacement from the empirical distribution of the corresponding residuals; i.e., from

$$\widetilde{\Im} : \text{ probability } \frac{1}{n} \text{ on } \tilde{u}_t, \qquad t = 1, \ldots, n. \tag{3.5}$$

The OLS estimators, $\tilde{\theta}_j$, derived under the null hypothesis are consistent, as required for Beran's (1988) results to be applicable. The results of Sims, Stock and Watson (1990) imply that the slope estimators are not only consistent but also have conventional asymptotic marginal distributions with $\sqrt{n}(\widetilde{\theta}_j - \theta_j)$ converging to a normal distribution with zero mean and finite variance, for $j = 1, 2, 3$.

It might be thought important to simulate $\{y_t, z_t\}$, using (3.2) and (3.3) with estimated coefficients as a system, to mimic data generation processes in which π of (3.3) is not zero. However, like Giersbergen and Kiviet (1996), we find that this approach yields results that are very similar to those for the simple scheme of

(3.4) and (3.5) in which the sequence $\{z_t\}$ is not varied over bootstrap samples. All results reported below are for the simple scheme.

3.3. Estimation of critical values and rejection probabilities

The performance of the predictive failure tests is considered using 5000 replications of sample data generated using either (3.1) or (3.2), and various methods of approximating the critical values for nominal significance levels of 5% and 10%. First, the $F(n_2, n_1 - k)$ (resp. $\chi^2(n_2)$) distribution is used to obtain critical values of the Chow (resp. Hendry) test, which is appropriate only if the disturbances are normally distributed. Second, the approximate critical values are derived from the non-parametric bootstrap (*npbs*). Finally, four parametric bootstrap methods are employed to obtain critical values. The distributions used are t_5, U, $\chi^2(2)$ and Λ, as used to generate the sample data. Note that this method will yield appropriate critical values only if the choice of distribution used in the bootstrap procedure matches that of the disturbances in the data generation process, in which case there should be very close agreement between desired and actual significance levels. In both the parametric and non-parametric bootstrap procedures, 500 bootstrap samples are generated in order to estimate the true critical values.

Before discussing the results that are obtained from the experiments, it is useful to examine what guidance is available from asymptotic theory to assist understanding and interpreting such results. The asymptotic rejection rates of the conventional H and P tests are the rejection probabilities of P^* of (2.6), based on the assumption that P^* is distributed as $\chi^2(n_2)/n_2$. Therefore knowledge of the true distribution of P^* can be used to predict the possible behaviour of H and P, under non-normality. In certain special cases this distribution can be obtained analytically. For example, consider the simple case where u_2 is uniformly distributed with zero mean and $n_2 = 1$. Here, u_2/σ has distribution function $F(c) = \frac{c+\sqrt{3}}{2\sqrt{3}}$, $-\sqrt{3} < c < \sqrt{3}$, from which it follows that $\Pr(P^* > d) = 1 - \sqrt{\frac{d}{3}}$, $0 < d < 3$. Incorrectly assuming that P^* is $\chi^2(1)$ would yield actual asymptotic significance levels of 5%, 0% and 0% at the *nominal* 10%, 5% and 1% levels, respectively; i.e., $\Pr(P^* > 2.706) = 0.05$, $\Pr(P^* > 3.841) = 0$, $\Pr(P^* > 5.024) = 0$. Since H and P are both asymptotically equivalent to P^* when $n_2 = 1$, asymptotic theory suggests that, in this special case, conventional tests will be undersized when the errors are uniformly distributed.

More general combinations of n_2 and the error distribution are not tractable as far as exact results are concerned, but very accurate estimates of the rejec-

tion probabilities of P^* can be obtained by simulation for any given combination. This strategy was adopted in an attempt to use asymptotic theory to assist in the interpretation of Monte Carlo results. Table 1 contains estimates of the rejection probabilities of P^* based upon one million replications for each combination of nominal significance level, n_2, and nonnormal error distribution used in the experiments.

Table 1: Estimated rejection probabilities of P^*

(a) Nominal significance level is 5%

distribution is:	t_5	U	$\chi^2(2)$	Λ
$n_2 = 3$	6.92	0.15	8.34	6.69
$n_2 = 6$	8.32	0.25	10.61	9.10

(b) Nominal significance level is 10%

distribution is:	t_5	U	$\chi^2(2)$	Λ
$n_2 = 3$	10.54	2.51	11.09	8.22
$n_2 = 6$	12.03	2.26	13.76	10.80

These results suggest that P and H will both be over-sized under t_5, $\chi^2(2)$ and Λ errors, with the distortion increasing as n_2 increases from 3 to 6. On the other hand, the prediction is that these tests will be severely undersized, asymptotically, when the regression errors are uniformly distributed, which is consistent with the analytical result obtained above.

4. Monte Carlo Results

In this section, we examine in detail the reliability in finite samples of the various critical values relative to a desired nominal 5% significance level in the context of both the static and dynamic model.[3] Rather than presenting tables containing many estimates, we use figures to convey the main features pictorially and also report tests of hypotheses that restrict true rejection rates to be in an interval

[3]Results for the 10% level are not discussed in detail since they reveal similar features. However, a full set of the Monte Carlo results is available upon request.

centred on the nominal size. This sort of hypothesis seems more relevant from a practical point of view than the null hypothesis that the actual size equals the nominal size. With a large number of replications, differences of little real importance between actual and nominal size could lead to rejection of the latter type of null hypothesis with high probability.

Figures 1-5 plot the estimated significance levels of four variants of the prediction error test against the six possible pairs of (n, n_2), from $n = n_1 + n_2 = 50, 80, 100$ and $n_2 = 3, 6$, for each of the five error distributions in the static model (3.1) as described in Section 3.1. The four tests are: the Chow statistic using either *npbs* or $F(n_2, n_1 - k)$ critical values (denoted *Chow-npbs* and *Chow*, respectively); and Hendry's variant using either *npbs* or $\chi^2(n_2)$ critical values (denoted *Hendry-npbs* and *Hendry*, respectively). Figures 6-11 illustrate the corresponding rejection rates for the ADL model (3.2) of Section 3.2 with $\pi = 0$ and $\gamma = 0.8$ in (3.3).

In order to check if the evidence is consistent with the claim that the actual and true sizes differ by at most some specified amount, e.g. 0.5%, we proceed as follows. Let the hypothesis to be tested be that the true rejection rate r^* satisfies $H_r : r_1 \leq r^* \leq r_2$, where: $0 < r_1 < r_2 < 1$, with $r_1 - r_2$ being $O(1)$, not $o(1)$; and $r_1 + r_2 = 2\alpha$, where α is the desired size. Let the required asymptotic (as the number of Monte Carlo replications, R, tends to infinity) significance level of the test of H_r be $\varkappa$ and c^* be such that $prob(N(0,1) < -c^*) = \varkappa$. We use $\varkappa = 5\%$ in all such tests so that the corresponding value of c^* is 1.645. Asymptotically valid inferences are made by rejecting H_r if the estimated size of the prediction error test falls outside the interval

$$r_1 - 1.645\sqrt{\frac{r_1(1-r_1)}{R}}, \quad r_2 + 1.645\sqrt{\frac{r_2(1-r_2)}{R}},$$

as $R \to \infty$. We have $R = 5000$ in all of our experiments.[4] Thus, for example, with $\alpha = 5\%, r_1 = 4.5\%$ and $r_2 = 5.5\%$, H_r restricts the true size to be within 0.5% of the desired size of 5% and the acceptance range is 4.02% to 6.03%.

The Figures used to summarize the estimates also summarize the outcomes of tests of H_r. In each Figure, if an estimate lies within the outer pair of dashed horizontal lines, it is consistent with the claim that the true significance level is within 1% of the nominal value; this will be referred to as satisfactory agreement

[4] With 5000 replications, it seems reasonable to treat the estimated significance levels as normal variables. Also the significance levels of the test with $R = 5000$ are likely to be very closely approximated by those approached as $R \to \infty$.

between estimated and desired significance levels. If an estimate lies within the inner pair of dashed lines, it is consistent with the claim that the true significance level is within 0.5% of the nominal value and this shall be referred to as good agreement between estimated and desired significance levels.

In the case of the static model (3.1), the main features of the evidence concerning the behaviour of the Chow and Hendry tests, with the various critical values, are as follows.

(a) With normally distributed regression errors the Chow test, which uses critical values from $F(n_2, n_1 - k)$, is exact and this is reflected in the close agreement between estimated and desired significance levels, for all n_2 and n. The Hendry variant which employs critical values from $\chi^2(n_2)$ is only asymptotically valid, since it ignores asymptotically negligible estimation effects, and can exhibit quite poor sampling behaviour; e.g., for $n = 50$, the estimated significance levels are 9.48% and 10.40% for $n_2 = 3$ and 6, respectively, whilst the corresponding estimates for the Chow test are 4.62% and 4.84%, respectively. Kiviet (1986), using a dynamic regression model for stationary data, also obtains Monte Carlo evidence that Hendry's test is oversized.

(b) When the regression errors are not normally distributed the standard practice of using F critical values with the Chow statistic, or χ^2 critical values with Hendry's statistic, often leads to substantial differences between estimated and desired significance levels; see Figures 2-5. In particular, the tests are consistently oversized under t_5, $\chi^2(2)$ and Λ errors with the overrejection being larger when $n_2 = 6$. The size distortions can be seen for all three sample sizes and are especially marked for Hendry's variant. For example, with $\chi^2(2)$ errors and $(n = 50, n_2 = 3)$, the estimated significance level for the Hendry variant is 11.46% whilst for the Chow variant it is 8.22%. These results are consistent with the analysis of Section 3.3 and Kiviet's (1986) evidence on the finite sample impact of neglected estimation effects. When the errors are uniformly distributed, the tests are consistently undersized, as the asymptotic analysis of Section 3.3 suggested they would be, except for the H test when $n = 50$ where asymptotic theory appears not to provide a reasonable guide to finite sample behaviour. However, it was noted in (a) that neglect of estimation effects renders Hendry's test oversized and, since the rejection rates do indeed fall below well below 5% for $n = 80$ and

$n = 100$, it therefore seems likely that this effect can dominate the effect of uniform errors in finite samples.

(c) The nonparametric bootstrap enjoys considerable success in controlling the significance levels when either the Chow or Hendry statistic is applied to models with nonnormal errors. Indeed, the estimated significance levels for both variants of the test, using this technique, are remarkably similar across all cases and so the following conclusions apply to both tests. Very good results are obtained when $n_2 = 3$ with most estimates being consistent with the claim that the true significance level is within 0.5% of the nominal value, and all being consistent with the claim that the former is within 1% of the latter. There is also strong evidence in favour of the nonparametric bootstrap when $n_2 = 6$ and the error distribution is symmetric (whether normal or not). Unfortunately, with $n_2 = 6$ and either the $\chi^2(2)$ or log-normal distributions supplying the regression errors, the nonparametric bootstrap leads to oversized tests, although the degree of over-rejection decreases as n increases. Moreover, in these cases, the use of the conventional critical values leads to a rather higher degree of overrejection than the use of nonparametric bootstrap critical values.

(d) The message from the estimates using parametric bootstraps is very clear and full details are not reported here. Perfect knowledge of the error distribution leads to excellent agreement between estimated and desired significance levels, as one might expect. All estimates, for both the Chow and Hendry test, with correct matching of assumed and actual error distribution are consistent with the claim that the true significance level is within 0.5% of the nominal value. However, perfectly accurate information about the error distribution is most unlikely. Incorrect matching involving the use of an inappropriate parametric bootstrap frequently provides evidence of substantial differences between estimated and nominal significance levels. For example, with $n = 50$ and $n_2 = 3$ and uniform errors, the estimated significance levels for the three versions of the Chow test which use t_5, $\chi^2(2)$ and Λ parametric bootstrap critical values are 0.58%, 0.2%, and 0.0% respectively. Thus, as argued above, since information about the error distribution is only vague, the parametric bootstrap is not a reliable method.

The qualitative nature of the remarks made in (a)-(d), above, also apply to the ADL model (3.2) of Section 3.2 with $\pi = 0$ and $\gamma = 0.8$ in (3.3), although

the Hendry variant of the test used in conjunction with the standard $\chi^2(n_2)$ critical values performs slightly worse than in the static model. The Chow test with $F(n_2, n_1 - k)$ critical values still produces excellent estimated significance levels when the regression errors are normally distributed (despite now being only asymptotically valid). When the errors are non-normal, the performance of both these procedures is, in general, extremely poor. The non-parametric bootstrap works well, except for $n_2 = 6$ and asymmetric errors where the distortions obtained are qualitatively similar to those found in the static model. Finally, and as revealed for the static model, the parametric bootstrap can only be relied upon when there is perfect knowledge of the error distribution. Therefore the non-parametric bootstrap emerges as extremely effective at controlling the significance level of both the Chow and Hendry forms of the prediction error test, in the ADL model, with the rejection rates for both of these variants being remarkably similar across all experiments.

The non-parametric bootstrap employed in the ADL model assumed a strictly exogenous z_t process with (3.4) being used for resampling. If z_t were only predetermined, then this approach might cause finite sample problems, since, strictly speaking, joint simulation of the conditional and marginal models would be required to mimic the true data generation process. On the other hand, incorrect specification of the marginal model for z_t when using the non-parametric bootstrap with joint simulation could have far more serious consequences. In their study, Giersbergen and Kiviet (1996) conclude that "*...finite sample distortions do not depend heavily on the presence or absence of a feedback mechanism ... The benefits of taking the full system into account are moderate and sometimes counter productive*". This conclusion, although suggestive, was in a rather different inferential context. Thus, as a partial check on the efficacy of the proposed non-parametric bootstrap procedure based on (3.4), some experiments were performed in which $\pi = -0.4$ and $\gamma = 0.4$; so that z_t was only predetermined. The results were almost indistinguishable from the case of strictly exogenous z_t and, like Giersbergen and Kiviet, we conclude in this case that "*Bootstrap inference turns out be quite accurate, even if the feedback mechanism is ignored in the resampling scheme.*"

In the results for the static model discussed thus far, k is 6 and n_2 is either 3 or 6, so that $n_2 \leq k$. This is the inequality that defines situations in which Chow recommended the use of his prediction error test. If $n_2 > k$, Chow's other test (involving full sample and separate subsample estimations of the model) can be applied; see, e.g., Greene (1997, pp. 349-351) for details. However, the applied literature contains several examples of the prediction error test being used when

$n_2 > k$, e.g., Otto (1994) reports results with $n_2 = 20, 40$ and $k = 6$. In order to investigate the usefulness of the nonparametric bootstrap when $n_2 > k$, $n_2 = 12$ is used with the static model experimental design of Section 3.2. The qualitative features of the results obtained are similar to those for $n_2 = 6$. Agreement between estimated and nominal significance levels is good with symmetrically distributed regression errors. For example, the Chow test employing the nonparametric bootstrap produces oversized tests with the asymmetric error distributions considered, but when it gives a bad outcome the conventional F critical value gives a much worse one. For example, with log-normal disturbances, a nominal size of 5%, $n_1 = 88$ and $n_2 = 12$, the estimate from the nonparametric bootstrap is 7.84% and that from the F critical value is 16.34%. For consistently good performance of the nonparametric bootstrap over all error distributions considered, a value of n greater than 100 is required. This implies that applied workers wishing to carry out predictive failure tests with values of n_2 between 10 and 20 may, therefore, require access to moderately large estimation samples.

Since the finite sample size results presented in this paper suggest that the nonparametric bootstrap can perform better when the error distribution is symmetric, applied workers might place more confidence in prediction error tests using nonparametric bootstrap critical values if there were also evidence supporting symmetry of the error distribution. An asymptotically valid test of symmetry of nonnormal regression disturbances is described by Godfrey and Orme (1991).

Finally in this section, given the similarity and good size properties of the Chow and Hendry non-parametric bootstrap tests, it is appropriate to investigate their relative power. Experiments were carried out for both the static and dynamic model using the data generation process (2.7), incorporating all 5 error distributions and where for each model a common value $\xi_t = \xi^*$ was chosen so as to give an interesting range of power estimates. More precisely, values of ξ^* were chosen so that these estimates were in the range $40\% - 70\%$ for different combinations of n and n_2 under normal regression errors. Over all the distributions considered, the range of power estimates obtained was $28\% - 90\%$. These experiments revealed Hendry's variant to be slightly more powerful than the Chow form in every case (although the differences are not large) and, as expected, power increases with n_2.

5. Conclusions

The finite sample significance levels of the prediction error test proposed by Chow (1960) and of a modification of Chow's test proposed by Hendry (1980) have been discussed. In the static linear regression model, Chow's test is exact when the hypothesis of primary interest, i.e., the predictive validity of the model, is combined with the assumption that the regression errors are normally distributed. Normality is also required for Chow's test to be asymptotically valid when the regressors include lagged dependent variables and for Hendry's variant to be asymptotically valid in either static or dynamic regression models. However, there seems to be little reason, in general, to suppose that normality is an accurate assumption. When the distribution of the regression errors is not normal, not even asymptotic theory can be used to justify the standard Chow or Hendry test procedures.

The Monte Carlo evidence presented in this paper reveals that, when the errors are normal, asymptotic theory does provide a reasonable guide to the finite sample behaviour of the Chow test procedure in dynamic regression models. This is not so for Hendry's test for which there is evidence of overrejection in both the static and dynamic regressor case with normal errors. This evidence is consistent with the findings of Kiviet (1996). Consequently, if the errors were known to be normal, it would seem reasonable to regard Chow's test as reliable (in terms of size) but not Hendry's procedure. However, when the restrictive assumption of normality is relaxed, the Monte Carlo results indicate that neither of the standard tests can be assumed to have finite sample significance levels that are close to the required values. As an alternative to relying upon critical values derived assuming normal errors, simple non-parametric bootstrap methods are described which appear to give useful control of the significance levels of prediction error tests without requiring precise information about the error distribution.[5]

The Monte Carlo results on the agreement between the nominal and estimated significance levels associated with the nonparametric bootstrap reported in section 4 above are very encouraging. Moreover, use of the nonparametric bootstrap scheme yields a remarkable similarity between the behaviour of the Chow and Hendry tests. Given their similar good size properties, a Monte Carlo investigation of their relative power was undertaken and this shows Hendry's variant to be the slightly more powerful, although the differences are not large.

[5] With n_1 fixed as $n \to \infty$, the application of the methods of Dufour and Kiviet (1996) for dynamic models to prediction error tests requires correct knowledge of the error distribution.

References

[1] Beran, R., 1988, Prepivoting test statistics: a bootstrap view of asymptotic refinements, Journal of the American Statistical Association 83, 687-697.

[2] Breusch, T.S., 1980, Useful invariance results for generalised regression models, Journal of Econometrics, 13, 327-40.

[3] Chow, G., 1960, Tests of equality between sets of coefficients in two linear regressions, Econometrica, 28, 591-605.

[4] Dufour, J.-M., E. Ghysels and A. Hall, 1994, Generalized predictive tests and structural change analysis in econometrics, International Economic Review, 35, 199-229.

[5] Dufour, J.-M. and J.F. Kiviet, 1996, Exact tests for structural change in first-order dynamic models, Journal of Econometrics, 70, 39-68.

[6] Dufour, J.-M. and J.F. Kiviet, 1997, Exact tests in single equation autoregressive distributed lag models, Journal of Econometrics, 80, 325-353.

[7] Dufour, J.-M. and J.F. Kiviet, 1998, Exact inference methods for first-order autoregressive distributed lag models, Econometrica, 66, 79-104.

[8] Giersbergen, N.P.A. van and J.F. Kiviet,1996, Bootstrapping a stable AD model: weak vs strong exogeneity, Oxford Bulletin of Economics and Statistics, 58, 631-656.

[9] Godfrey, L.G. and C.D. Orme, 1991, Testing for skewness of regression disturbances, Economics Letters, 37, 31-34.

[10] Greene, W.H., 1997, Econometric Analysis (3rd Edition), Prentice-Hall International, Inc.

[11] Hamilton, J.D., 1994, Time Series Analysis, Princeton University Press: Princeton.

[12] Hendry, D.F., 1980, Predictive failure and econometric modelling in macroeconometrics: the transaction demand for money, in Modelling the UK Economy, edited by P. Omerod, Heinemann: London.

[13] Hoffman, D. and A. Pagan, 1989, Post-sample prediction tests for generalised methods of moments estimators, Oxford Bulletin of Economics and Statistics, 51, 333-343.

[14] Huber, P.J., 1973, Robust regression: asymptotics, conjectures and Monte Carlo, Annals of Statistics, 1, 799-821.

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

[16] Li, H. and Maddala, G.S., 1996, Bootstrapping time series models, Econometric Reviews, 15, 115-158.

[17] Otto, G., 1994, Diagnostic testing: an application to the demand for M1, in Cointegration for the Applied Economist, edited by B. Bhaskara Rao, St. Martin's Press: New York.

[18] Pesaran, M.H. and Y. Shin, 1999, An autoregressive distributed lag modelling approach to cointegration analysis, in Econometrics and Economic Theory in the 20th Century: the Ragnar Frisch Centennial Symposium, edited by S. Strom, Cambridge University Press: Cambridge.

[19] Sims, C.A., J.H. Stock, and M.W. Watson, 1990, Inference in linear time series models with some unit roots, Econometrica, 58, 113-144.

[20] Stewart, J. and L. Gill, 1998, Econometrics (2nd Edition), Prentice-Hall Europe.

[21] West, K.D. and M.W. McCracken, 1998, Regression-based tests of predictive ability, International Economic Review, 39, 817-840.

[22] White, H., 1980, A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity, Econometrica, 48,421-448.

Figures 1-10

All the following diagrams depict the size estimates (rejection rates), in percentages, of the Chow and Hendry forms of the predictive failure test statistics as defined in the main text.

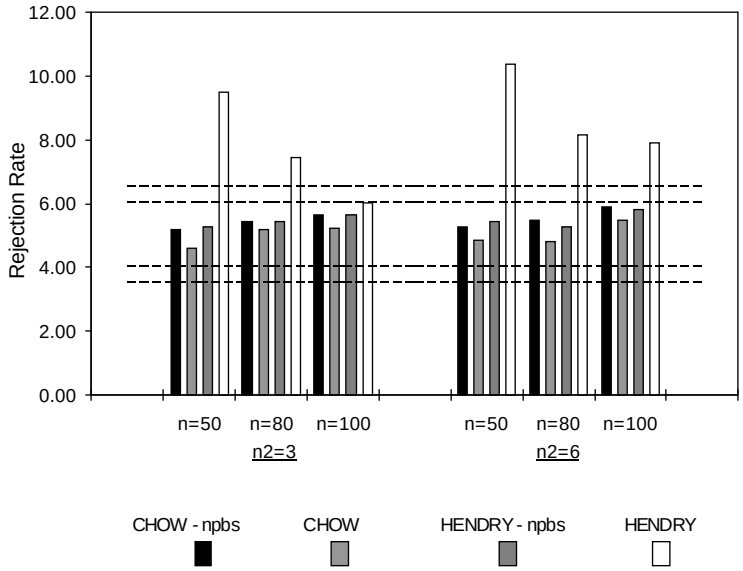

Figure 1: Static Model - Normal errors

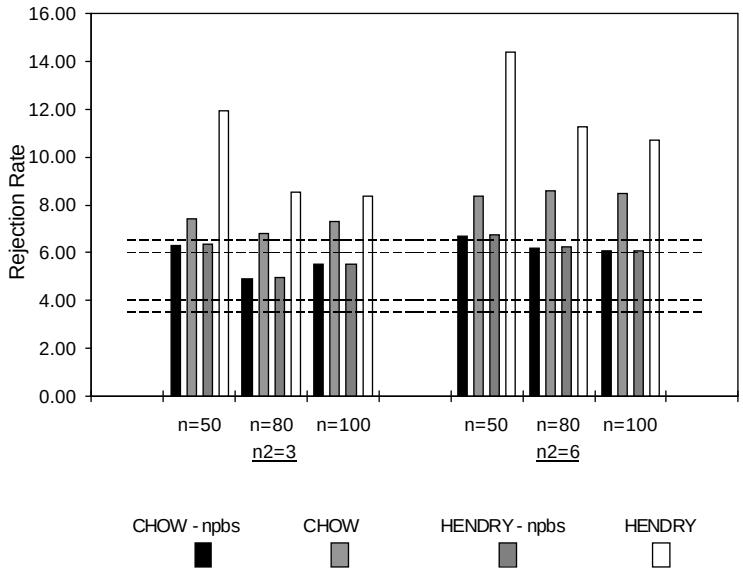

Figure 2: Static Model - Student t errors

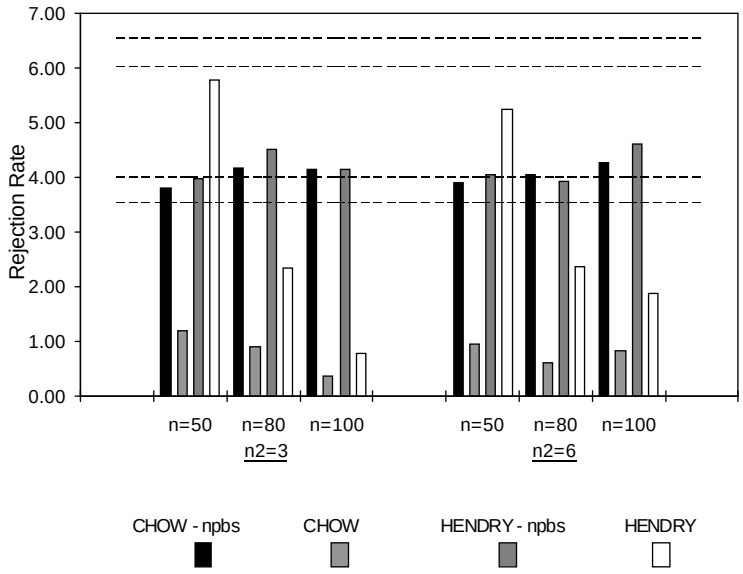

Figure 3: Static Model - Uniform errors

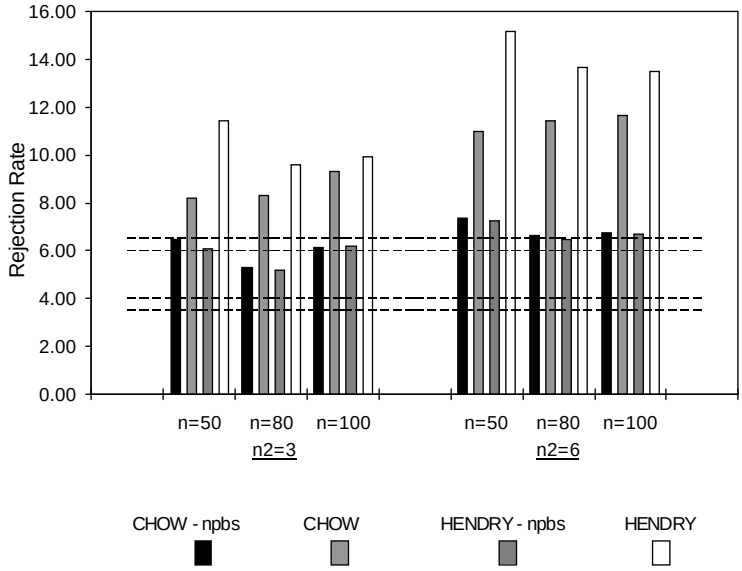

Figure 4: Static Model - Chisquare errors

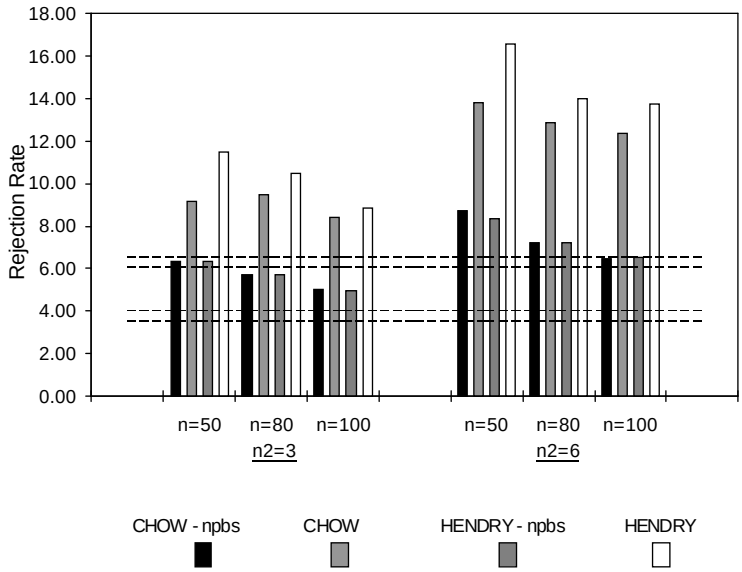

Figure 5: Static Model - Lognormal errors

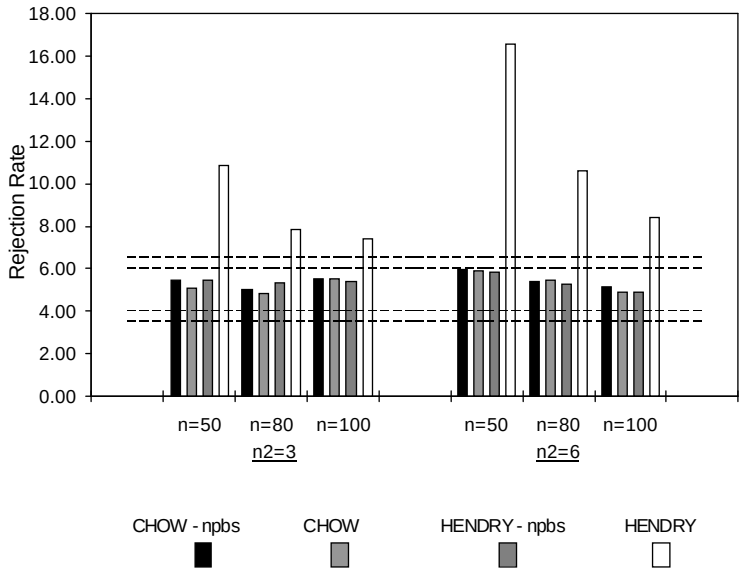

Figure 6: ADL Model - Normal errors

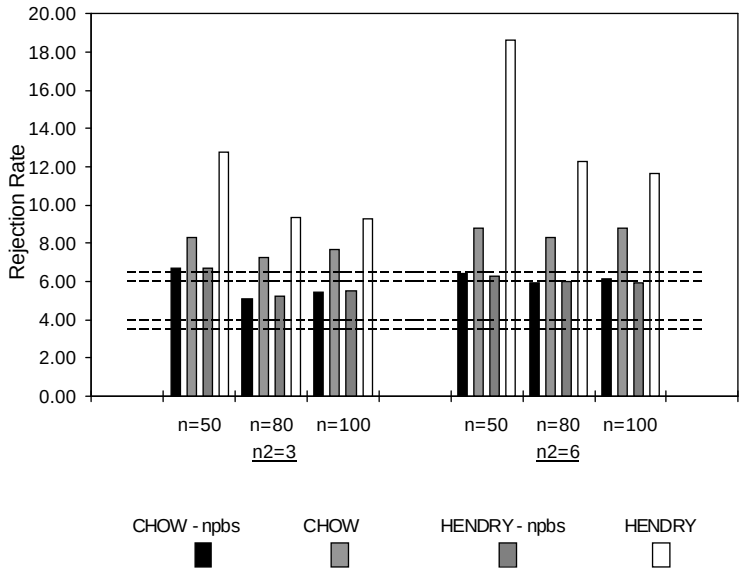

Figure 7: ADL Model - Student t errors

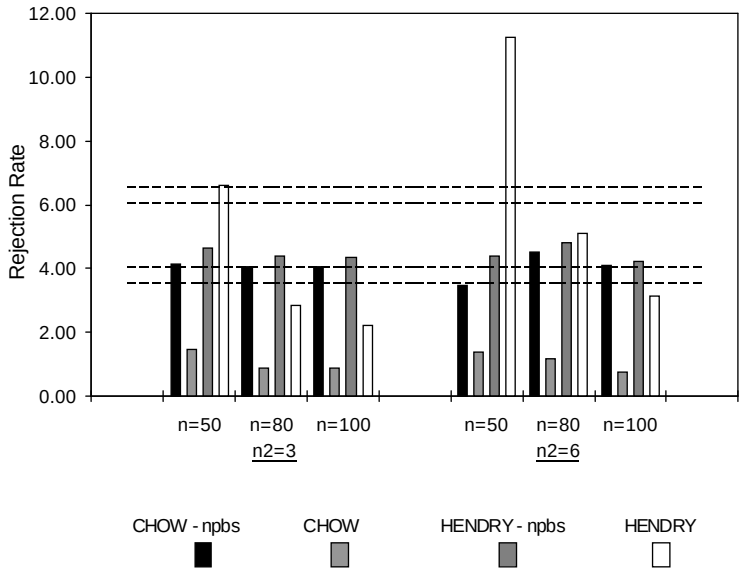

Figure 8: ADL Model - Uniform errors

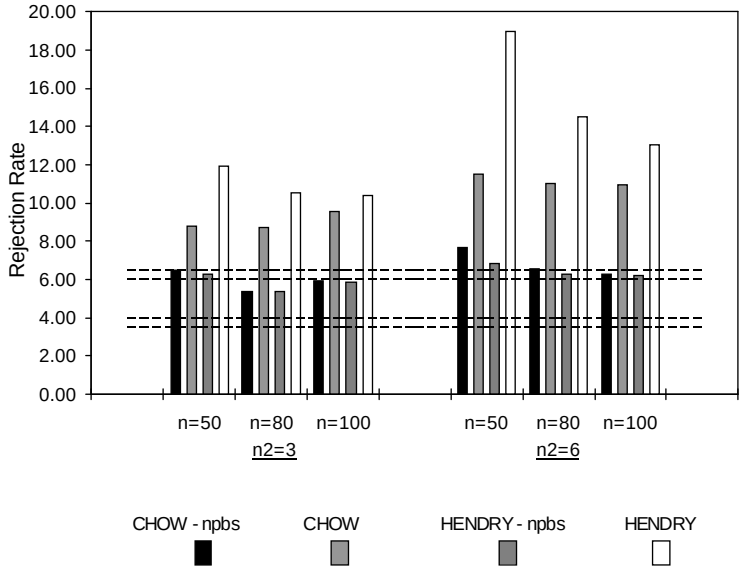

Figure 9: ADL Model - Chisquare errors

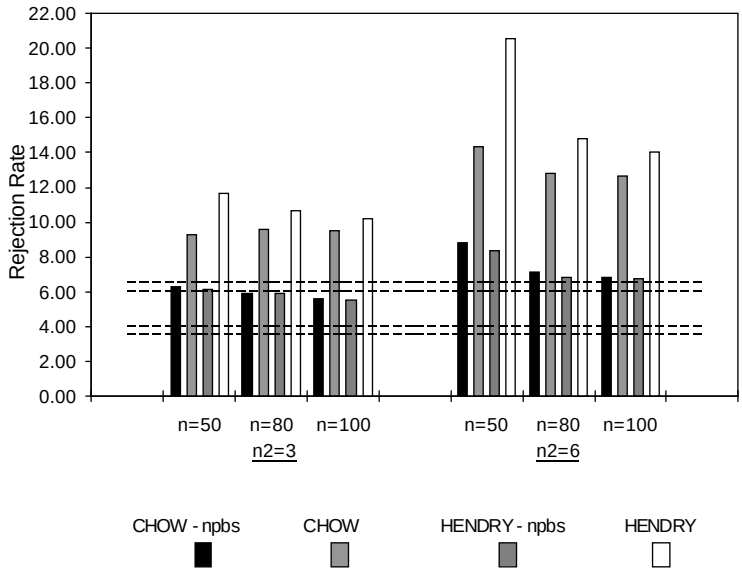

Figure 10: ADL Model - Lognormal errors