A NEW TECHNIQUE FOR POSTSAMPLE MODEL SELECTION AND VALIDATION[*]

by

Richard Ashley
Economics Department, Virginia Tech
Blacksburg, Virginia 24061
e-mail: ashleyr@vt.edu

February, 1997

The model selection and Granger-causality literatures have generally focused on insample rather than postsample hypothesis testing. In large part this is due to the fact that feasible postsample model validation periods are usually quite short, whereas large-sample methods are ordinarily required in order to deal with the serial correlation and crosscorrelation typically found in postsample forecast error series. This paper describes a re-sampling based postsample inference procedure which enhances the usefulness of the inference significance levels it produces by explicitly estimating the uncertainty which its own large-sample approximation induces in these levels.

For a given target level of inferential precision (such as significance at the 5% level) this procedure also provides estimates of both how strong the evidence in favor of one of two models must be for a given length postsample period and of how long a postsample period is necessary if the evidence is of given strength. These results indicate that a postsample model validation period substantially longer than the five to twenty periods typically reserved in past studies is necessary in order for a twenty to thirty percent MSE reduction to be significant at the 5% level. The value of the procedure is illustrated using postsample forecasting error data from Ashley, Granger, and Schmalensee (1980), in which evidence is presented for unidirectional Granger-causation from fluctuations in aggregate consumption expenditures to fluctuations in aggregate expenditures on advertising using U.S. data.

[*] Economics Department Working paper #E97-01. The author gratefully acknowledges the support of NSF grant #SES-8922394.

1. Introduction

This paper describes a new inference technique for assessing whether one sequence of postsample forecasting errors is smaller than another. Such inferences are particularly useful where substantial specification search activity has reduced the usefulness of insample procedures for model selection and/or validation.

Postsample forecast error series are typically strongly crosscorrelated and often significantly serially correlated as well. Under these conditions existing postsample inference techniques [such as Ashley, Granger and Schmalensee (1980), Ashley (1981), Meese and Rogoff (1988), and Diebold and Mariano (1995)] are only valid for large samples. This is awkward since postsample model validation periods are usually short.

The postsample inference technique proposed here is a variation on the bootstrap designed to mitigate this awkwardness by producing an explicit estimate of the uncertainty which its own large-sample approximation induces in the inference significance levels. In addition to accounting for any crosscorrelation or serial correlation in the two forecast error series being compared, this new technique also conveniently allows consideration of a variety of forecast error loss functions and provides an estimate of how much stronger the evidence would need to be or how much longer a postsample model validation period is necessary in order to obtain a given level of inferential precision.

The nature and advantages of bootstrap-based inference are briefly reviewed in the first portion of Section 2 in the context of a very simple problem: inference on the population mean of an i.i.d. random variate. The remainder of Section 2 describes the variation on the bootstrap introduced here. In Section 3 this new approach is extended to produce inferences on ratios of expected functions of pairs of correlated and/or serially dependent time series. The usefulness of the procedure in small sample settings is demonstrated in Section 4 using monte carlo simulations.

In Section 5 the new technique is applied to the postsample forecast error series generated in the Ashley, Granger, and Schmalensee (1980) study of the Granger-causal relation between aggregate U.S. consumption expenditures and aggregate U.S. advertising expenditures. This example illustrates the new procedure's ability to provide useful Granger-causation inferences in a small sample setting under alternative loss functions on forecast errors (such as the absolute error loss function or an asymmetric piecewise-quadratic loss function) and its ability to generate inferences using the scaled mean loss differential statistic introduced in Diebold and Mariano (1995).

The results from this example indicate that a postsample model validation period substantially longer than the five to twenty periods typically reserved in past studies is necessary in order to conclude that a twenty to thirty percent MSE reduction is significant at the 5% level. This issue is discussed in the final portion of the paper.

2. Enhancing the Usefulness of Bootstrap Inference in Small-Sample Settings

In this Section bootstrap-based inference is briefly described in a simple setting in order to clarify its nature and to show how a second level of bootstrapping can be used to quantify the small-sample uncertainty induced in the inferences by the bootstrap approximation itself.

Consider the problem of using a random sample of N observations on a variable x to test whether its population mean (μ) is less than some given value, μ_o. Non-bootstrap inference typically begins by choosing among candidate estimators for the unknown parameter based on their sampling properties. In this case the sample mean, $\overline{x}$, is the obvious choice. Since $\overline{x}$ is known to be asymptotically Gaussian, most analysts would routinely assume that the distribution of x is sufficiently close to being Gaussian that this asymptotic distribution provides a reasonable approximation to the finite-sample distribution of $\overline{x}$. Asymptotically valid confidence intervals for μ and hypothesis tests concerning μ are then derived on the assumption that the sampling distribution of $\overline{x}$ is Gaussian.

In the bootstrap approach, the population distribution of x is approximated by its empirical distribution, which places equal probability mass on each of the N observed values for x. Singh (1980), Bickel, P. J. and D. A. Freedman (1981), Beran, R. (1986) and others have shown that this approximation is inconsequential in large samples. Under it, as much additional data as is needed can be generated by sampling at random out of the empirical distribution of the observed sample.[1] Thus, the probability that $\mu < \mu_o$ can be estimated by generating, say, 2000 new

[1] More explicitly, if the original sample is denoted x(1)…x(N), a new N-sample can be obtained by drawing N integers $\{j_1 \dots j_N\}$ at random from the discrete uniform distribution which places equal weight on each integer in [1, N]. The resulting new N-sample is $x(j_1) \dots x(j_N)$.

N-samples on x and computing the proportion of the resulting 2000 realizations of $\overline{x}$ for which $\overline{x} < \mu_o$.

Both approaches are only asymptotically justified, but the bootstrap has three advantages. First, the bootstrap is often easier to apply than alternative methods, although it does require substantially larger amounts of computational resources. Second, inferences and parameter estimates obtained using the bootstrap are often more accurate in small samples– in some cases [e.g., Freedman and Peters (1984)] dramatically so. Finally, the bootstrap approach can in principle be used to quantify the sensitivity of its own inference results to the errors induced by the bootstrap approximation itself.

This last advantage of bootstrap-based inference is the most relevant feature for the present purpose. Instead of generating 2000 new N-samples and computing the proportion of them for which $\overline{x} < \mu_o$, consider generating a smaller number of N-samples (100, say) and proceeding with each one as if it were the original sample. That is, for each one of these 100 “starting samples,” 2000 new N-samples are generated and the proportion of the resulting 2000 realizations of $\overline{x}$ for which $\overline{x} < \mu_o$ is computed. In this way, each of the 100 starting samples yields an inference on μ in the form of a probability estimate. The dispersion of these 100 inference probabilities provides an estimate of the uncertainty in the bootstrap inference due to the finite size of N; it can be taken as estimate of the “fragility” of the inference, in the spirit of Leamer (1985).

Since the distribution of these inference probabilities necessarily becomes very non-Gaussian when the average inference probability becomes small, the median and interquartile range of the 100 inference probabilities provide more useful measures of the location and dispersion of this distribution than do the mean and standard deviation. Consequently, inference results are reported below in terms of the median inference (Q_{50}) and its empirical 50% confidence interval, [$Q_{.25}$, $Q_{.75}$]. Here Q_α is the α% fractile of the 100 inference probabilities, so the length of this interval is just the sample interquartile range.

This uncertainty estimate based on the interquartile range of the 100 probability estimates is itself only asymptotically justified, but it nevertheless conveys considerable information about the reliability of the median bootstrap inference. Alternative checks on the reliability of the median inference (and its dispersion) can be obtained at varying cost. For example, it is easy to verify whether or not 100 starting samples and 2000 bootstrap repetitions suffice to make the median inference insensitive to changes in the starting seed for the random number generator. Another inexpensive check is to compare the median inference probability to the inference one obtains from the ordinary bootstrap, in which 2000 new samples are drawn from the empirical distribution of the observed sample data. Finally, at tremendously larger cost, monte carlo simulations can be used to estimate the coverage of the empirical 50% interval on the inferences; the results of such calculations are reported in Section 4 below.

However, there are three reasons why this simple approach to verifying the small-sample reliability of bootstrap-based inferences is almost never implemented. [Freedman and Peters (1984) provides a rare exception.] First of all, the additional computational burden involved is obviously substantial – in the example above, estimating an empirical 50% confidence interval for the bootstrap inference requires 100 times as many bootstrap replications. Secondly, theoretical work on bootstrap inference [e.g., DiCiccio and Romano (1988)] has focused on improving the accuracy of bootstrap inference rather than on quantifying the degree to which it remains uncertain. Thus, while several second-level ("double") bootstrapping proposals have been advanced – e.g., Beran (1987) – the second level of bootstrapping in these proposals is used to improve the small-sample accuracy of the first level inference rather than to quantify its small-sample uncertainty.

Lastly, the straightforward dispersion calculation described above turns out to be subtly flawed in such a way as to substantially overstate the actual small-sample dispersion in the bootstrap inferences. Consequently, the initial results obtained in trying out an idea of this sort are apt to be so poor as to discourage further interest in the approach. The flaw in the straightforward dispersion calculation is fundamental, but simple and easily avoided once recognized.

Ordinarily, the distinction between the sampling distribution of an estimator (such as $\overline{x}$) and the distribution of its sampling errors ($\overline{x}$ - μ) is inconsequential since these two distributions differ only by a translation in the fixed amount μ. Thus, under the null hypothesis that $\mu = \mu_o$, the probability that $\overline{x} < 6$, say, is identical to the probability that the sampling error $\overline{x} - \mu_o < 6 - \mu_o$. But this distinction is *not* inconsequential in the present context.

In the preceding example, 100 starting samples are picked from the empirical distribution of the original data and then 2000 new N-samples are picked from the empirical distribution of each starting sample. Thus, a total of 200,000 N-samples are drawn; and each one is used to compute a value for $\overline{x}$. Notice, however, that these 200,000 N-samples are not all drawn from the same distribution. The first 2000 N-samples are all drawn from the empirical distribution of the first of the 100 starting samples. Letting μ_1 denote the population mean of the distribution from which these N-samples are drawn, note that μ_1 must precisely equal $\overline{x}_1$ (the sample mean of the first starting sample) since the empirical distribution gives equal weight to each of the N observations in the first starting sample. The second group of 2000 N-samples is drawn from the empirical distribution of the second of the 100 starting samples, with population μ_2 (which is equal to $\bar{x}_2$) and so forth. Clearly, these 100 population means ($\mu_1 \ldots \mu_{100}$) will vary substantially for small N, inducing substantial additional dispersion in the resulting 100 inferences. This additional dispersion is extraneous since it is not due to the bootstrap approximation of using the empirical distribution of the original sample data to replace the population distribution from which the original N-sample was drawn. Indeed, this source of inferential dispersion would be equally strong even if each of the 100 starting samples was drawn, monte-carlo fashion, from the actual population distribution of x.

This problem does not arise when the bootstrap approximation is used to generate 100 estimates of the *sampling error distribution* of $\bar{x}$, since all of these distributions have mean zero. Letting $\bar{x}_o$ denote the sample mean of the actual sample data, then H_o: $\mu < \mu_o$ implies that $\bar{x}_o - \mu$, the sampling error implied by $\bar{x}_o$, exceeds $\bar{x}_o - \mu_o$. Consequently, the ith starting sample can be used to estimate the probability that $\mu < \mu_o$ by computing the fraction of the 2000 N-samples for which the sampling error $\bar{x}_{ij} - \mu_i = \bar{x}_{ij} - \bar{x}_I$ exceeds $\bar{x}_o - \mu_o$, where $\bar{x}_{ij}$ is the sample mean of the jth N-sample drawn from the ith starting sample and $\bar{x}_I$ is the sample mean of the ith starting sample. Thus, the fraction of the 2000 sampling errors which exceed $\bar{x}_o - \mu_o$ is the inference probability on μ from the ith of the 100 starting samples. These 100 inference probabilities are much more stable across the starting samples than those obtained from the 100 sampling distributions of $\bar{x}$ itself, simply because the distribution of the sampling errors ($\bar{x}_{ij} - \mu_i$) is much more stable across the starting samples than is the distribution of $\bar{x}_{ij}$. In fact – presuming that the number of N-samples drawn from each starting sample is sufficiently large – the distribution of these sampling errors varies across the starting samples *only* because the empirical distribution implied by each starting sample is derived from an N-sample bootstrapped from the original sample data rather than from the original sample data itself. Thus, the dispersion across the starting samples of the inferences based on the sampling error distributions quantifies the uncertainty in the inferences caused by the bootstrap assumption itself.[2]

In summary, the median of the inference probabilities obtained from these bootstrapped sampling error distributions provides an estimate of the probability that $\mu < \mu_o$. And the dispersion of these inference probabilities across the starting samples quantifies the small-sample uncertainty in the median inference due to the bootstrap approximation of replacing the population distribution of x by its observed empirical distribution. In this way – by explicitly estimating the small-sample uncertainty in the bootstrap inference – it becomes feasible to obtain potentially convincing inferences from the bootstrap in a small-sample setting.

[2] The empirical distribution of the original sample data is not identical to the population distribution from which the original sample was drawn, however. Consequently, this dispersion estimate is itself valid only for large samples. This issue is examined using monte carlo simulations in Section 4 below.

3. Inference on Postsample Forecast Errors

The bootstrap-based inference approach described above is applied here to the problem of testing whether the expected size of the postsample forecasting errors from one model significantly exceeds that of another, based on an observed sequence of N postsample forecasting errors from each model.

These two postsample forecasting error series are denoted x_t and y_t below. Since postsample forecast errors are typically autocorrelated, it is not appropriate to sample directly from their empirical distribution. Instead, it is assumed here that (x_t, y_t) is covariance stationary and that its generating mechanism can be adequately represented as a bivariate VAR process:

$$\Phi(B)\begin{vmatrix} x_t \\ y_t \end{vmatrix} = \begin{vmatrix} \phi_{11}(B) & \phi_{12}(B) \\ \phi_{21}(B) & \phi_{22}(B) \end{vmatrix}\begin{vmatrix} x_t \\ y_t \end{vmatrix} = \begin{vmatrix} \mu_x \\ \mu_v \end{vmatrix} + \begin{vmatrix} \epsilon_t \\ \eta_t \end{vmatrix} \qquad \textbf{(1)}$$

so that new N-samples $\{(x_1, y_1) \ldots (x_N, y_N)\}$ can be obtained by sampling from the empirical distribution of the innovations, $(\epsilon_1, \eta_1) \ldots (\epsilon_N, \eta_N)$.

The assumptions underlying Equation 1 can and should be checked. Stationarity in mean and variance can be checked by examining time plots of x_t and y_t, looking for outliers and for evidence of substantial shifts or trends in mean or variance. And the linearity assumption inherent in Equation 1 can be checked by examining scatterplots of x_t and y_{t+k} for $k = 0, \pm 1, \pm 2$, etc. Since Gaussianity is not assumed, an outlier which is not overly influential can be tolerated; such an observation can be viewed as an ordinary realization from a non-Gaussian (x_t, y_t) distribution. However, formal testing is beside the point in this context: if N were large enough for such testing to be justified, available large-sample methods would suffice for obtaining the relevant postsample inference in the first place.

Note that x_t and y_t are the postsample *forecasting error* series produced by two different models whose relative forecasting effectiveness is being evaluated. *The VAR model given above as Equation 1 is neither of these models;* it is merely a descriptive parameterization of the serial correlation structure of the forecast errors (x_t and y_t) made by these two models. Thus, covariance stationarity of the (x_t, y_t) implies that the forecast horizon must be the same for all of the x_t – e.g., they are all h_x-step-ahead forecasts; similarly, all of the y_t must be h_y-step-ahead forecasts. But h_x need not equal h_y. Indeed, it is not necessary to observe or know *anything* further about the two forecasting models that generated the forecasting error series, x_t and y_t. These two models might be nested or they might not; they might be equally-complex constructs arising from differing schools of thought, or one of them might be quite naive compared to the other. Since all that is used from each of the two models is a sequence of postsample forecasting errors, the internal structure of these two models is irrelevant.

The coefficients in the distributed lag polynomials {$\phi_{11}(B)$, $\phi_{12}(B)$, $\phi_{21}(B)$, and $\phi_{22}(B)$}, the intercepts (μ_x, μ_y), and the distribution of (ϵ_t, η_t) in Equation 1 need not be supplied – the only specification information required is a reasonably tight *upper bound* on the maximum lag in each of the four lag polynomials.[3] Usably accurate upper bounds on these lag polynomial orders can be obtained by running a few linear regressions and eliminating the clearly insignificant terms. This suffices because the inference results are insensitive to minor over-elaboration in the specification of these upper bounds; monte carlo simulation results illustrating this point are given in Section 4.

[3] In ordinary VAR modeling, these orders are chosen to be sufficiently large that the innovation series (ϵ_t, η_t) is serially uncorrelated. Since the bootstrap makes independent picks from the observed (in general, non-Gaussian) innovation sequence, it must be assumed here that these orders are sufficiently large that (ϵ_t, η_t) is serially independent. However, this distinction is relevant if and only if (x_t, y_t) is related to its own past in a substantially nonlinear way; postsample forecasting periods are ordinarily so short that any consideration of serial dependencies more complex than the low-order VAR mechanism used here is out of the question in any case.

Figure 1 provides a schematic description of the calculation of the probability that a specified relative accuracy criterion, r, is less than or equal to some given value, τ; most commonly, τ will equal one. The population value for r could be the ratio of the two MSE's:

$$r_{MSE} = \frac{MSE_x}{MSE_y} = \frac{E\left(x_t^2\right)}{E\left(y_t^2\right)}$$

but other choices for r are possible and often preferable. For example, if the distribution of (ϵ_t , η_t) is fat-tailed, then

$$r_{MAE} = \frac{MAE_x}{MAE_y} = \frac{E\left(\left|x_t\right|\right)}{E\left(\left|y_t\right|\right)}$$

might be preferable; or, if negative errors are known to cause substantially higher losses, it would be preferable to use an asymmetric criterion, such as

$$r_{asy} = \frac{E\left(s\left(x_t\right)x_t^2\right)}{E\left(s\left(y_t\right)y_t^2\right)}$$

where $s(z) = 1$ for $z \geq 0$ and $s(z) = 2$, say, for $z < 0$. Alternatively, the superiority of the y_t forecast error series over the x_t error series can be quantified using the studentized expected loss differential criterion proposed by Diebold and Mariano (1995):

$$r_{DM} \quad = \quad \exp\left\{ \frac{E\left[loss\left(x_t\right) - loss\left(y_t\right) \right]}{\sqrt{2\pi f_d(0)}} \right\}$$

where $f_d(0)$ is the spectral density of the numerator at frequency zero. Loss(·) is a situationally appropriate loss function; the absolute value function is used in the calculations reported below in Section 5. Here r_{DM} is defined as the exponential of Diebold and Mariano's criterion so that r equals one for equivalent forecasts on all four criteria.[4]

Returning to Figure 1, r is the population value of whichever forecast accuracy criterion has been selected. The object is to test the null hypothesis H_o: $r \le \tau$ against the alternative hypothesis H_A: $r > \tau$, based on the observed N-sample: $\{(x_1, y_1) \dots (x_N, y_N)\}$. As Figure 1 indicates, this original sample data is only used twice.

First, it is used to obtain $\hat{r}_{orig}$, a consistent sample estimate of r; using r_{MSE}, for example, $\hat{r}_{orig}$ is the ratio of the sample mean of $(x_t)^2$ to the sample mean of $(y_t)^2$. The resulting $\hat{r}_{orig}/\tau$ figure is the sampling error factor this sample estimate represents if r equals τ, so that H_o is barely true. Comparing the observed sampling error factor, $\hat{r}_{orig}/\tau$, to the distribution of sampling error factors, $\hat{r}_{observed}/r_{true}$, rather than comparing the observed sampling error itself, $\hat{r}_{orig} - \tau$, to the distribution of sampling errors, $\hat{r}_{observed} - r_{true}$, makes the inference results independent of which error series is chosen to appear in the numerator of r – i.e. it ensures that $\text{Prob}\{r \le \tau\}$ precisely equals $\text{Prob}\{r^{-1} \ge \tau^{-1}\}$.

Second, the original sample data is used to obtain OLS estimates of the parameters in the VAR model, Equation 1. Since corr(ϵ_t, η_t) can be substantial, SUR would be preferable in

[4] Space limitations preclude an extensive discussion of their criterion here. They provide a consistent estimator of $\sqrt{N}\log(r_{dm})$ which is very easy to compute and is asymptotically distributed N(0,1) when the population loss differential is zero and the loss differential series (loss$\{x_t\}$ - loss$\{y_t\}$) is serially uncorrelated beyond a given lag.

principle, but N is not ordinarily large enough to justify its use. It *is* useful to correct the OLS parameter estimates for small-sample bias; monte carlo simulation results illustrating this point (and a description of the bias correction algorithm itself) are given in Section 4. At this point the original sample data has yielded (1) an observed sampling error factor (assuming that r just equals τ) and (2) an estimated data generating mechanism – the fitted VAR model and its residuals, {($\hat{\epsilon}_1$, $\hat{\eta}_1$) … ($\hat{\epsilon}_N$, $\hat{\eta}_N$).

Next, this estimated VAR model is used to generate new observations on (x_t, y_t) using the bootstrap assumption that the population distribution from which the innovation 2-vectors (ϵ_1, η_1) … (ϵ_N, η_N)} were drawn is identical to the empirical distribution of the fitting errors {($\hat{\epsilon}_1$, $\hat{\eta}_1$) … ($\hat{\epsilon}_N$, $\hat{\eta}_N$)}, which places probability mass 1/N on each of these observed 2-vectors.[5] If the maximum lag in the VAR is p, then the next observation on (x_t, y_t) follows directly from the previous p values of (x_t, y_t), the VAR model coefficient estimates and the next innovation 2-vector, ($\hat{\epsilon}_j$, $\hat{\eta}_j$), where j is a randomly chosen integer in the interval [1, N]. The p values of (x_t, y_t) needed to initiate the simulations can be sample values or even zeroes – after a sequence of 50 to 100 observations have been generated in this way (and discarded) their influence becomes negligible. In this way, the algorithm generates a number of new N-samples on (x_t, y_t). Since these N-samples are generated from the model estimated using the original sample data, they are analogous to the 100 "starting samples" discussed in Section 2; the number of these starting samples generated is denoted "N_{sim}" below.

Each of these N_{sim} starting samples is then used to obtain an estimate of the distribution of sampling error factors, $\hat{r}_{observed}/r_{true}$, via a large number of new samples generated using the

[5] Bootstrapping from the residuals of an estimated autoregressive time series model is not new – Efron and Tibshirani (1985, p.27) do this for a one-dimensional AR(1) model as one of their first applications of the bootstrap. Picking from the empirical distribution of 2-vectors asymptotically preserves the contemporaneous crosscorrelation structure of the original innovations.

bootstrap approximation. The number of new samples generated is denoted N_{rep} in Figure 1 and generally set equal to 2000 in the calculations reported below.[6]

Figure 1 describes how $\{(\hat{r}_j / r^{true}_{37}),\ j = 1 \ldots N_{rep}\}$, the distribution of sampling error factors obtained applying the bootstrap to the 37th starting sample, is obtained. First, a VAR model is estimated by applying OLS to the 37th starting sample. As with the estimated VAR model for the original sample data, the resulting parameter estimates are corrected for small sample bias using the algorithm described in Section 5. Then this estimated model is used as a data generating mechanism to generate:

(a) N_{rep} N-samples $\{(x_1, y_1) \ldots (x_N, y_N)\}$ which yield N_{rep} sample estimates of r, $\hat{r}_1 \ldots \hat{r}_{Nsim}$, and

(b) a single, large sample of length 100N $\{(x_1, y_1) \ldots (x_{100N}, y_{100N})\}$, which yields a large-sample estimate of r, r^{true}_{37}. This large sample ratio (r^{true}_{37}) is essentially equal to the population value of r for this 37th VAR process.[7] Presuming that N_{rep} is sufficiently large that the observed distribution of the N_{rep} values of $(\hat{r}_j / r^{true}_{37})$ adequately characterizes the sampling error factor distribution implied by this 37th data generating mechanism, the 37th estimate of the probability that $r \leq \tau$ (i.e., $\hat{\rho}_{37}$) is the fraction of the N_{rep} sampling error factors that exceed $\hat{r}_{orig}/\tau$.

The other N_{sim} - 1 sampling error factor distributions are obtained in a similar fashion.. These distributions are unequal to one another for two reasons: First, each starting N-sample is too small to precisely recover the single set of VAR coefficients (obtained using the original sample data) used to generate all N_{sim} starting samples. This imprecision is presumably similar to

[6] Thus, to fix the notation, there are N_{rep} = 2000 bootstrap "repetitions" for each of N_{sim} = 100 "simulations" and each simulation is initiated by using one of the N_{sim} starting samples generated from the original sample data. In practice, N_{rep} and N_{sim} are to be increased to the point where the results are no longer appreciably sensitive to their values.

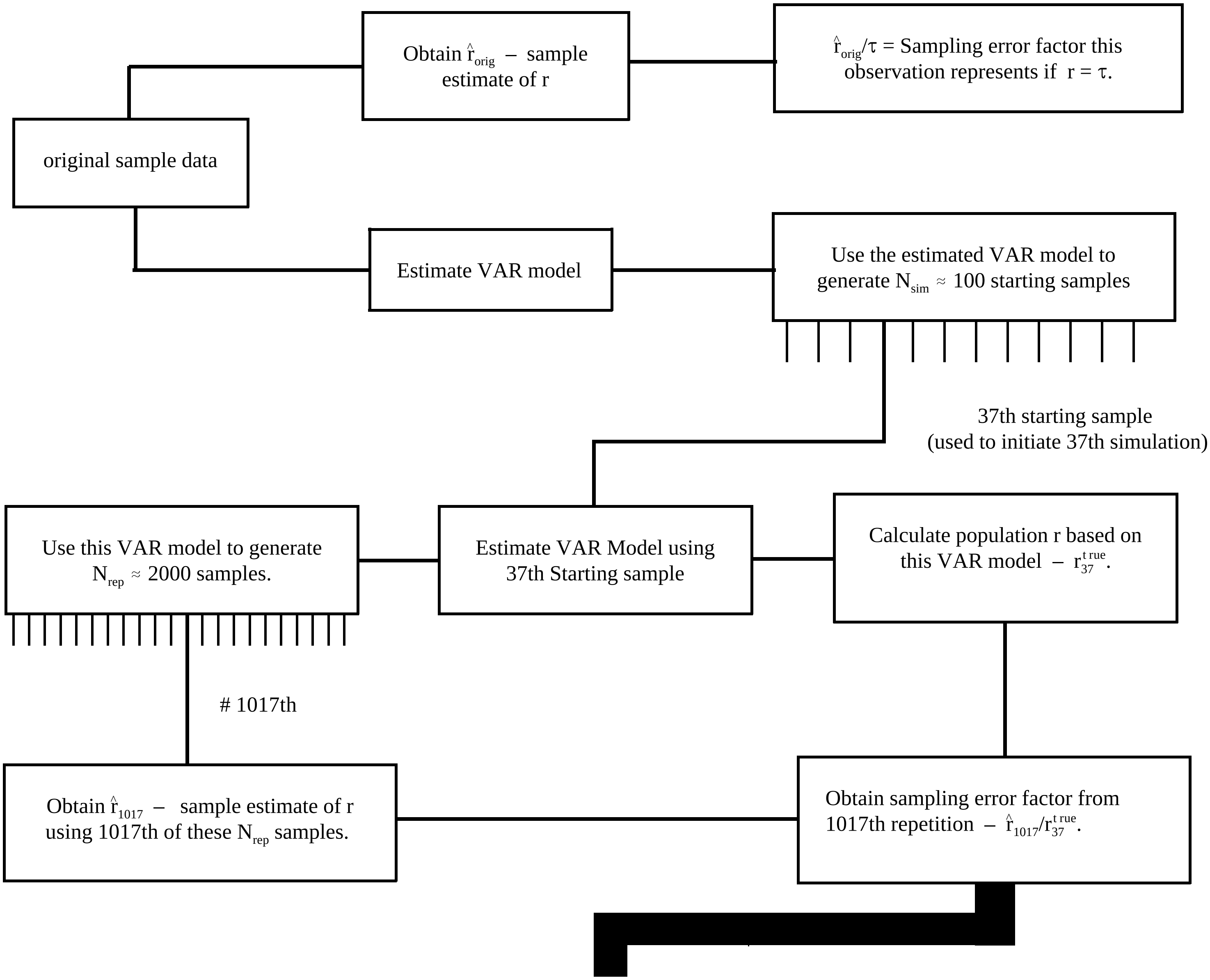
[7] The $r^{true}_1 \ldots r^{true}_{Nsim}$ are analogous to the 100 population means ($\mu_1 \ldots \mu_{100}$) of Section 2. For $r = r_{MSE}$ the population MSE ratio for the data generating mechanism derived from the 37th starting sample (r^{true}_{37}) can be obtained analytically. This is not feasible for the other criterion choices, so r^{true}_{37} is obtained by merely simulating a very large sample from the data generating mechanism.

that with which the true (population) VAR coefficients can be recovered from the single observed sample. And second, even if the true VAR coefficients could be used in generating all of the N_{sim} starting samples, the empirical distribution of the residuals implied by each of the starting samples is different for each starting sample because each one is only a bootstrap approximation to the population distribution from which they were all picked.[8] Thus, the dispersion in the inference probabilities obtained from these N_{sim} sampling error factor distributions – i.e., the dispersion in $\hat{\rho}_1$ … $\hat{\rho}_{Nsim}$ – quantifies the inferential uncertainty caused by sampling errors in the estimation of the VAR model *and* by the bootstrap approximation itself.

[8] This population distribution is just the empirical distribution of the residuals from the VAR model obtained using the original sample data.

Figure 1.

Calculation of ρ_{37} , the 37th Estimate of Probability that $r \leq \tau$


Obtain $\hat{r}_{orig}$ – sample estimate of r
$\hat{r}_{orig}/\tau$ = Sampling error factor this observation represents if $r = \tau$.
original sample data
Estimate VAR model
Use the estimated VAR model to generate $N_{sim} \approx 100$ starting samples
37th starting sample (used to initiate 37th simulation)
Use this VAR model to generate $N_{rep} \approx 2000$ samples.
Estimate VAR Model using 37th Starting sample
Calculate population r based on this VAR model – r_{37}^{true}.
# 1017th
Obtain $\hat{r}_{1017}$ – sample estimate of r using 1017th of these N_{rep} samples.
Obtain sampling error factor from 1017th repetition – $\hat{r}_{1017}/r_{37}^{true}$.

The inference from simulation # 37 {i.e. ρ_{37}} is then the fraction of the N_{rep} sampling error factors obtained in simulation # 37 which are greater than or equal to the sampling error factor which the original sample observation represents if $r = \tau$. Or:

$$\rho_{37} = \text{Fraction of} \left\{ \left[\frac{\hat{r}_1}{r_{37}^{true}}\right] \cdots \left[\frac{\hat{r}_{Nrep}}{r_{37}^{true}}\right] \right\} \geq \frac{\hat{r}_{orig}}{\tau}$$

4. Monte Carlo Simulation Results

The empirical 50% confidence interval, $[Q_{25}, Q_{.75}]$, is defined above in Section 2. By definition it contains the middle half of the N_{sim} inference probabilities ($\hat{\rho}_1 \ldots \hat{\rho}_{Nsim}$) described in Section 3. In this Section, monte carlo simulations are used to estimate the actual coverage of this 50% inference interval. The sensitivity of this coverage to sample size, mis-specification of the VAR model, and bias correction in the VAR coefficient estimates is examined.

Each monte carlo simulation is conducted as follows: First, N innovation vectors, $\{(\epsilon_t, \eta_t), t = 1 \ldots N\}$ are generated from a truncated bivariate gaussian distribution. Two truncation points are considered here: $\pm 3\sigma$, corresponding to an essentially gaussian distribution, and $\pm .5\sigma$, corresponding to a nearly uniform distribution. In view of the large contemporaneous crosscorrelations typically found among postsample forecasting errors obtained from different models, the innovations in the VAR model for these errors are generated with corr (ϵ_t, η_t) equal to .60. N-samples $\{(x_1, y_1) \ldots (x_N, y_N)\}$are generated from the model

$$\begin{aligned} x_t &= .5\, x_{t-1} + \epsilon_t \\ y_t &= .5\, y_{t-1} + \eta_t \end{aligned} \tag{2}$$

with t = 1 ... N.

Since the VAR model is symmetric in its treatment of x_t and y_t, the population values of r_{MSE} and r_{MAE} are both one. A large number of such N-samples are used to estimate the p = .05 and p = .01 critical points, $\tau_p^{mse}(N)$ and $\tau_p^{mae}(N)$, defined by:

$$\text{Prob}\left[\frac{\sum_{t=1}^{N} x_t^2}{\sum_{t=1}^{N} y_t^2} \quad \geq \quad \tau_p^{mse}(N) \right] \quad = \quad p$$

and

$$\text{Prob}\left[\frac{\sum_{t=1}^{N} |x_t|}{\sum_{t=1}^{N} |y_t|} \quad \geq \quad \tau_p^{mae}(N) \right] \quad = \quad p$$

for N = 5, 10, 20, and 40 and using both truncation points. Thus, for example, in 40,000 samples of length 20 simulated with truncation at $\pm 3\sigma$, the sample MSE ratio exceeds 1.48 only about 400 times; thus $\tau_p^{mse}(20)$ is 1.48 for p = .01 with this truncation point.

For each sample size and truncation point, 200 N-samples yielding MSE ratios in the interval [.99, 1.01] are generated; N-samples yielding MSE ratios outside this interval are rejected. Setting N_{sim} = 100 and N_{rep} = 2000, each of these 200 samples is then used as the sample data for the inference algorithm to obtain N_{sim} estimates of the probability that $r_{mse} > \tau_p^{mse}(N)$, first for p = .05 and then for p = .01. The largest and smallest of the middle 50 of these inference probabilities are the endpoints of the empirical 50% inference interval for each of the 200 monte carlo simulations. Since the true probability that $r_{mse} > \tau_p^{mse}(N)$ is p, the estimated coverage of this 50% interval is the fraction of these 200 50% inference intervals which contain p.

This process is then repeated, this time retaining only those N-samples whose sample MAE ratio is in [.99, 1.01] and computing the coverage of the 50% inference interval for the test that $r_{mae} > \tau_p^{mae}(N)$. The coverage results are not sensitive to minor changes in the sample MSE or MAE intervals for which starting samples are retained, but the coverage of the empirical 50% inference interval for the MSE ratio test *is* sensitive to whether the starting samples were conditioned on the observed MSE ratio or on the observed MAE ratio, and similarly for the MAE ratio test. This conditioning is necessary because the algorithm is bootstrapping the distribution of the sampling error factors ($\hat{r}_j$ / r_{37}^{true}, in the notation of Figure 1) rather than the distribution of sample ratio itself, $\hat{r}_j$. Indeed, when the algorithm is modified to bootstrap the distribution of the sample ratio itself rather than the distribution of its sampling error ratio, the coverage of the resulting (much wider) empirical 50% inference intervals is correct if and only if this conditioning is dropped.

Table 1 summarizes the coverage estimate results for the 1% and 5% tests on MSE and MAE ratios. These estimates are approximately normally distributed around .50 with a standard deviation of $\{.5(1 - .5)/200\}^{.5}$ with 200 monte carlo trials, so values in the interval [.43, .57] are insignificantly different from .50; values outside this interval are shaded in the tables. Evidently, N = 20 is sufficient for both distributions and truncation points, but N = 10 is not.[9]

Aside from the original sample data, the user need only specify the orders of the lag structures in the VAR model of Equation 1. The consequences for the coverage of the 50% inference interval of mistakenly choosing too large or too small an order are examined in the monte carlo simulations reported in Table 2. The "over-elaborate" rows of Table 2 correspond to incorrectly including y_{t-2} in the equation for y_t; the "under-elaborate" rows correspond to incorrectly omitting the y_{t-1} term from this equation. The true data generating mechanism,

[9] Similar results are obtained for p = .10 and p = .20. It seems likely that the coverage at N = 10 would be better for data which is more weakly autocorrelated.

Table 1

Coverage of 50% Interval

Gaussian Data Truncated at ± kσ – 200 monte carlo trials

		N = 5		N = 10		N = 20		N = 40	
k	test	p = .05	p = .01	p = .05	p = .01	p = .05	p = .01	p = .05	p = .01
3σ	MSE	.085	.145	.410	.405	.455	.460	.495	.500
3σ	MAE	.100	.110	.320	.315	.445	.445	.500	.495
.5σ	MSE	.100	.125	.480	.445	.525	.495	.550	.555
.5σ	MAE	.110	.085	.430	.375	.475	.475	.540	.500

Equation 2, is of course unchanged. As might be expected, this modest amount of over-elaboration is fairly inconsequential, whereas under-elaboration results in significant coverage distortions.[10]

Finally, a bivariate VAR model is estimated using the sample data each time this inference procedure is applied. This model is used to generate N_{sim} starting samples, so a total of $1 + N_{sim}$ bivariate VAR models are estimated. The simulation results reported in the last row of each section of Table 2 demonstrate that, if left untreated, small-sample bias in the OLS parameter estimates for these models yields substantial distortions in the resulting inference intervals. As implemented here, the bias correction procedure is a two step process. First, the bias in each coefficient estimator is estimated by using the model estimated using OLS to generate 100 new samples, estimating a model for each, and computing the average discrepancy in the slope

[10] The pair of “over-elaborate” results with N=20 on the MAE ratio test for which the coverage appears to deviate significantly from .50 are entirely consistent with chance given the number of cells in the “correct” and “over-elaborate” rows of Table 2.

estimates. The resulting bias estimates are then added onto the original OLS estimates, the intercepts are adjusted to force the sample mean of the fitting errors to zero, and 100 more samples are generated, yielding an improved estimate of the biases.

Overall, the monte carlo simulation results indicate that the inference procedure works quite well for $N \geq 20$. It might be possible to use the procedure with even smaller sample sizes by tuning the bias correction procedure, but this has not been established.

Table 2

Coverage of 50% Interval

Gaussian Data Truncated at $\pm 3\sigma$ – 200 monte carlo trials

	MSE ratio test				**MAE ratio test**			
	N = 20		N = 40		N = 20		N = 40	
	p = .05	p = .01	p = .05	p = .01	p = .05	p = .01	p = .05	p = .01
correct specification	.455	.460	.495	.500	.445	.445	.500	.495
over-elaborate VAR	.465	.440	.495	.485	.415	.385	.535	.545
under-elaborate VAR	.240	.220	.315	.300	.245	.220	.275	.245
without bias correction	.340	.315	.355	.370	.245	.230	.390	.380

5. An Illustrative Example: Testing for Granger-Causation Between Advertising and Aggregate Consumption Spending

Ashley, Granger and Schmalensee (1980) addresses two related questions. The first is a substantive empirical issue: do fluctuations in aggregate advertising expenditures Granger-cause fluctuations in aggregate consumption spending or does the causal relationship run in the other direction?[11] AGS describe the creation of a new aggregate advertising expenditures time series which can be brought to bear on this question. The second question is methodological: how can hypotheses about Granger causation between a pair of time series be tested most effectively? Here AGS break new ground by proposing a test of the Granger-causation between two time series based on an explicit comparison of the postsample forecasting effectiveness of models for each series based on nested information sets.[12]

In particular, suppose that the postsample forecasts of aggregate consumption spending generated by a forecasting model making optimal use of an information set including information on past aggregate advertising expenditures are demonstrably superior to those of an optimal model based on an otherwise-identical information set excluding past aggregate advertising expenditures. Then, so long as these information sets are sufficiently wide as to include any third variable which affects both consumption and advertising, AGS would conclude that aggregate advertising expenditures Granger-cause aggregate consumption spending. Thus, they reduce the analysis of Granger-causation to an assessment of whether one model for consumption spending provides better postsample forecasts than the other.

In fact, AGS find no evidence that aggregate advertising expenditures Granger-cause

[11] Or in both directions, yielding a feedback relationship.

[12] Note that "nested information sets" can and often do lead to non-nested forecasting models. Indeed, that is the case for the two AGS models (AC.2 and A.1) whose postsample forecasting errors are analyzed below.

aggregate consumption spending. They do, however, find that including past aggregate consumption spending in the information set for constructing a model to forecast aggregate advertising expenditures is quite helpful, reducing the postsample mean square forecasting error by 26% over the twenty period postsample period, 1970I to 1975IV. AGS propose a procedure for testing whether this MSE reduction is statistically significant but, as with virtually all postsample inference methods, their procedure is only valid in large samples. Consequently, with such a short postsample forecasting period, uncertainty as to the small-sample adequacy of their test substantially diminishes the additional credibility gained from assessing the relative forecasting effectiveness of the models over a postsample period.

Applying the inference procedure described in Section 3, let y_t denote the one-step-ahead postsample forecast errors from the model for advertising expenditures based on the wider information set (including past values of aggregate consumption spending); this is the ARMAX model denoted AC.2 (AGS, p. 1161); and let x_t denote the postsample forecast errors made by model A.1 (AGS, p. 1159), which excludes past consumption spending from its information set. Then $\rho = \text{Prob}\{r \leq 1\}$ is the significance level at which the null hypothesis that consumption spending Granger-causes advertising can be rejected. Depending on one's loss function with respect to forecast errors, r might be any one of the relative accuracy criteria given in Section 3: r_{MSE} or r_{MAE} or r_{ASY} or r_{DM}.

Time plots of x_t and y_t both look reasonably covariance stationary. In particular, neither series appears to be trended in either mean or variance. Both series appear to be serially correlated, however. OLS regression yields:

$$\begin{array}{llllllllll} x_t & = & 7.38 & - & .032\, x_{t-1} & - & .452\, x_{t-2} & + \epsilon_t & R^2 = .134 & \\ & & (1.12) & & (\ .11) & & (1.52) & & DW = 1.92 & \\ & & & & & & & & & \textbf{(2)} \\ y_t & = & .76 & + & .332\, y_{t-1} & - & .428\, y_{t-2} & + \eta_t & R^2 = .193 & \\ & & (\ .14) & & (1.23) & & (1.53) & & DW = 1.61 & \end{array}$$

where the figures in parentheses are estimated t ratios and both fitting error series appear to be serially uncorrelated. Formal hypothesis testing using these estimated t ratios is surely not justified with such small samples, but they still have value as descriptive statistics. On this basis the coefficients on x_{t-2} and y_{t-2} are hardly compelling, but since these coefficient estimates are negative for a variety of sub-samples and since the coverage of the empirical 50% inference interval is known to be sensitive to under-elaboration but insensitive to modest over-elaboration in the VAR model specification, the orders of the lag polynomials $\phi_{11}(B)$ and $\phi_{22}(B)$ in Equation 1 are set to two.

Table 3 summarizes the results. The "Sample r ratio" figure of .738 for r_{MSE} re-states the observation, noted above, that including past consumption spending in the information set for forecasting advertising expenditures yields a 26% reduction in the observed postsample MSE. Thus, r < 1 is evidence for consumption Granger-causing advertising. In fact, the sample ratios based on all four criteria are less than one – the question is whether or not they are *significantly* less than one.

The inference procedure given by AGS indicates that this postsample MSE reduction is significant at the 9% level; AGS interpret this as modest evidence that consumption spending Granger-causes fluctuations in advertising expenditures. However, the inference procedure given by Diebold and Mariano (1995) indicates that the observed expected loss differential is significantly negative (so that $r_{DM} < 1$) at only the 27.8% level.[13] These results disagree, but since there are only twenty observations and both procedures are justified only in large samples, it seems inappropriate to give much credence to either result.

The boostrap-based inference procedure described above was applied to these data using N_{sim} = 100 and N_{rep} = 2000; these calculations tied up a desktop computer for about five minutes. The median inference level exceeds .20 for all four criteria, suggesting that the 9% result obtained by AGS is an artifact caused by the small sample size. But this median inference is again the result of an asymptotically justified procedure applied to a sample of only twenty observations. Consequently, it is not by itself any more credible than the results obtained using the AGS or Diebold/Mariano procedures.

However, the bootstrap-based inference procedure settles the matter by explicitly quantifying the uncertainty in the median inference due to the small sample size. In particular, for the test on r_{MSE}, half of the N_{sim} = 100 generated starting samples yield bootstrap significance levels in the interval [.176, .268]. Thus, it is reasonable to conclude that the postsample MSE drop observed by AGS is only significant at the 18% to 27% level and to therefore reject the AGS assertion that they have obtained evidence for fluctuations in consumption spending Granger-causing fluctuations in advertising expenditures at the 10% level of significance – their evidence is now demonstrably weaker than this.

[13] The Diebold-Mariano S_1 statistic (a unit normal under the null hypothesis of zero expected loss differential) is .585 for this data set. Their recommended truncation lag {S(T)} is zero here since (x_t, y_t) are one-step-ahead forecast errors; consequently, $2\pi f_d(0)$ is just the sample variance of the average loss differential in this case.

Table 3

Inference Results Using Postsample Forecast Errors from

AGS(1980) Models for Aggregate Advertising Expenditures

	r_{MSE}	r_{MAE}	r_{ASY}	r_{DM}
Sample r ratio	.738	.934	.803	.877
Asymptotic test significance level	.092	unknown	unknown	.278
Bootstrap inference results:				
Median of N_{sim} inferences ($Q_{.50}$)	.237	.388	.319	.337
Empirical 50% interval [$Q_{.25}$, $Q_{.75}$]	[.176, .268	[.350, .419]	[.266, .348]	[.341, .368]
Sample ratio needed for 5% result	.58	.73	.55	.65

6. Conclusions

The postsample inference procedure proposed here

- avoids the pre-test biases which data mining induces in insample tests,
- allows for the contemporaneous crosscorrelation and serial dependence commonly found in postsample time series data, and
- yields inferences which are reasonably credible, even for the small samples typically available for postsample inference, by explicitly quantifying the uncertainty in the inference introduced by the bootstrap approximation itself.

However, now that it is possible to take postsample inference more seriously, it is no longer either necessary or proper to remain vague about the amount of postsample data which is needed for effective inference.

For example, the results given in Table 3 of Section 5 clearly indicate that the 26% MSE improvement obtained by AGS (1980) over a twenty quarter postsample period is simply not significant at even the 10% level. By repeating these calculations to test the null hypothesis that $r_{mse} \leq \tau$ for values of τ increasingly larger than one it is possible to explicitly estimate how large an MSE improvement would have sufficed in this case to yield a 50% inference interval containing .05. Such results are reported in the last row of Table 3; they show that an MSE improvement of over 40% or an MAE improvement of over 30% is needed, given the distribution and the correlation structure of these data.

Alternatively, the length of the generated samples can be increased until a desired level of inferential precision is achieved. Such calculations, reported in Ashley (1992), indicate that a postsample model validation period must typically be twenty five to forty five periods long in order to detect a 30% MSE drop at the 5% level of significance, or fifty to one hundred periods long to detect a 20% MSE drop.

Thus, a model which cannot provide at least a 30% MSE improvement over that of a competing model is not likely to appear significantly better than its competitor over postsample periods of reasonable size. And evidently the five to twenty periods that have in the past been allocated to postsample model validation/inference (when it was done at all) are quite inadequate to detect the modest postsample MSE reductions one ordinarily sees. Yet retention of a postsample model validation period much in excess of thirty to forty periods seems rather impractical in many econometric contexts.

One resolution of this dilemma is to explicitly recognize that, since experience indicates that postsample forecasting is quite a stringent test of the extent to which a model has captured a stable statistical regularity, perhaps we should be satisfied with postsample MSE or MAE improvements which are significant at the 10% or even the 20% level. This is analogous to our shared perception that a reasonable R^2 for a model estimated on cross-sectional data is substantially lower than that for a model estimated on time series data. Another possibility is to revise upward our estimate of the relative importance of model validation and to therefore allocate a substantially larger portion of the available data to a postsample model validation/inference period, perhaps pooling the data at the end once model choice/model validation is complete.

Still, if postsample model validation/inference requires more data than we have heretofore been willing to allocate to it in order to yield reasonably definitive results, then why do it at all? Perhaps the best response to this question is: "because the alternative approach of insample model validation/inference, over the same data used for specifying and estimating the model, makes it**too** easy to obtain supportive results."

It may well be that this is the principal reason why the economics community has produced and used so many badly-misspecified models. Had we been willing and able – through the use of tools such as the inference procedure proposed here – to routinely confront our models with an effective postsample model validation hurdle, I believe that we would have produced a significantly smaller number of econometric models and a significantly larger amount of actual progress in the resolution of both theoretical and applied economic controversies.

References

Ashley, R., Granger, C.W.J., and Schmalensee, R.L. (1980), "Advertising and Aggregate Consumption: An Analysis of Causality," **Econometrica 48**, 1149-68.

Ashley, R. (1981), "Inflation and the Distribution of Price Changes Across Markets: A Causal Analysis," **Economic Inquiry 19**, 650-60.

Ashley, R. (1992), "A Statistical Inference Engine For Small Dependent Samples With Applications to Postsample Model Validation, Model Selection, and Granger-Causality Analysis," V.P.I. & S.U. Economics Department Working Paper #E92-24.

Beran, R. (1986), "Simulated Power Functions," **Annals of Statistics 14**, 151-73.

Beran, R. (1987), "Prepivoting to reduce level error of confidence sets," **Biometrika, 74,** 456-68.

Bickel, P. J. and D. A. Freedman (1981), "Some Asymptotic Theory for the Bootstrap," **Annals of Statistics 9**, 1196-1217.

DiCiccio, T. S. and Romano, J. P. (1988), "A Review of Bootstrap Confidence Intervals," **J. Roy. Stat. Soc. B, 50**, 338-54.

Diebold, F. X. and Mariano, R.S. (1995), "Comparing Predictive Accuracy," **Journal of Business and Economic Statistics 13(3)**, 253-63.

Efron, B. and Tibshirani, R. (1985), "The Boostrap Method for Assessing Statistical Accuracy," Technical Report #101, Division of Biostatistics, Stanford University.

Freedman, D. A. and Peters, S. C. (1984), "Bootstrapping a Regression Equation: Some Empirical Results," **Journal of the American Statistical Association 79 (Theory and Methods Section)**, 97-106.

Leamer, E. E. (1985), "Sensitivity Analysis Would Help," **American Economic Review 75(3)**, 308-13.

Meese, R. A. and Rogoff, K. (1988) "Was it Real: The Exchange Rate - Interest Differential Relation Over the Modern Floating-Rate Period," **Journal of Finance, 43,** 933-48.

Singh (1980), " On Asymptotic Accuracy of Efron's Bootstrap," **Annals of Statistics 9**, 1187-1195.