NON-NESTED MODEL SELECTION/VALIDATION:

MAKING CREDIBLE POSTSAMPLE INFERENCE FEASIBLE*

by

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

April, 1995

ABSTRACT

Effective, credible inference with respect to the postsample forecasting performance of time series models is widely held to be infeasible. Consequently, the model selection and Granger-causality literatures have focussed almost exclusively on in-sample tests, which can easily be biased by typical specification-search activity. Indeed, the postsample error series generated by competing models are typically cross-correlated, serially correlated, and not even clearly gaussian; thus, postsample inference procedures are necessarily only asymptotically valid. As a result, a postsample period large enough to yield credible inferences is perceived to be too costly in terms of sample observations foregone. This paper describes a new, re-sampling based, approach to postsample inference which, by explicitly quantifying the inferential uncertainty caused by the limited length of the postsample period, makes it feasible to obtain credible postsample inferences using postsample periods of reasonable length. For a given target level of inferential precision – e.g., significance at the 5% level – this new approach also provides explicit estimates of both how strong the postsample forecasting efficiency evidence in favor of one of two models must be (for a given length postsample period) and how long a postsample period is necessary, if the evidence is of given strength. These results indicate that postsample model validation periods substantially longer than the 5 to 20 periods typically reserved in past studies are necessary in order to credibly detect 20% - 30% MSE reductions. This approach also quantifies the inferential impact of different forecasting efficiency criterion choices – e.g., MSE vs. MAE vs. asymmetric criteria and the use of expected loss differentials (as in Diebold and Mariano(1994)) vs. ratios of expected losses. The value of this new approach to postsample inference is illustrated using postsample forecasting error data from Ashley, Granger, and Schmalensee(1980), in which evidence was presented for unidirectional Granger-causation from fluctuations in aggregate U.S. consumption expenditures to fluctuations in U.S. aggregate expenditures on advertising.

---

*Economics Department Working paper #E95-07. The author gratefully acknowledges the support of NSF grant #SES-8922394.

1. Introduction: The Desirability and (Perceived) Infeasibility of Credible Postsample Inference

In-sample model validation/selection procedures – which by definition use the same data for model identification (i.e., formal or informal specification search), for parameter estimation, and for model validation – are known to give frequently-misleading results. That, of course, is most of why econometric models so commonly yield dramatically poor out-of-sample forecasts. The reason for these misleading results is simple and well-known: while the best of the in-sample procedures do adequately account for the fact that one model may fit the sample data better than another because it has a larger number of free parameters, none of the in-sample procedures takes into account the intensity or breadth of specification search activity ("data mining") that went into producing the specification for each model[1]. This is a particularly pernicious problem in comparing the sample fit of two non-nested models, where the degree of specification search utilized in producing each model can be dramatically unequal.

In-sample specification search activity is neither inefficient or nor somehow immoral. We clearly need to mine the sample data, to one extent or another, in order to obtain an estimatable specification. However, our very efforts in doing so invalidate the statistical inference tools we then employ to quantify and test the adequacy of the resulting model. The problem is not that data-mining techniques are unable to uncover approximately correct model specifications – i.e. "good" models. After all, as Hendry & Mizon (1985) point out, the correctness (or otherwise) of a model specification inheres in the model itself, not in the process used to obtain it. The problem is that these techniques also produce many "bad" models **and**, in the process, inherently obliterate much of the power of our diagnostic techniques to detect their "badness."

Model selection/validation procedures based on the models' relative ability to generate accurate postsample forecasts are generally immune to this problem, since model specifications obtained by "torturing the sample data into submission" virtually always produce extremely poor postsample forecasts. And whether one of the models under consideration happens to be nested in the other is completely irrelevant for a postsample procedure, since all that such procedures use is the sequence of (usually one-step-ahead) postsample forecast errors made by each model, rather than the models themselves. Indeed, postsample model selection/validation

[1]Nor is the modeller's candor sufficient – he/she may honestly report fitting only j models to the sample data, but how many specifications were fit to this data by others in literature he/she is familiar with? Ashley(1994) further discusses the relative desirability (and unpopularity) of postsample versus in-sample inferential methods.

procedures are still applicable even when one or both of the models used to obtain the postsample forecast error data is unavailable or simply infeasible to analyze – e.g., a large macroeconometric model or a neural net model, as in Rhee(1994) or LeBaron and Weigend(1994).[2]

However, the postsample inferential procedures in the existing literature {e.g., Ashley, Granger and Schmalensee(1980), Ashley(1981), Meese and Rogoff(1988), and Diebold and Mariano(1994)} have problems of their own. In particular, they all suffer from the following two defects, the second of which is quite serious:

The first defect is inherent in any postsample procedure: the opportunity cost of each additional postsample observation is one sample observation foregone. This defect is overwhelming where the total amount of data available is really small – e.g., 10 to 30 observations – since an attempt to validate models at the cost of not being able to sensibly produce them in the first place obviously makes no sense. In such instances, however, it is well to explicitly recognize that one simply does not have enough data to both produce and credibly validate a model. On the other hand, with larger samples – 80 to 120 observations, say – it does seem reasonable to suppose that one could "afford" to allocate perhaps 10 to 20 observations to a postsample model-validation period. After all, the standard deviations of one's (in-sample) parameter estimates using the remaining N observations typically fall off like $N^{-.5}$, so in this case the last 10 to 20 observations in the total data set are probably not (on average) crucially important for model specification/estimation purposes.

But is 10 to 20 postsample observations is "enough" data to credibly utilize a postsample model selection/validation procedure? How can one "know"? This question brings up the second generic defect of postsample inferential methods. This defect is so serious as to have substantially limited the applicability (and the frequency-of-use, where they are applicable) of the postsample model selection/validation procedures available in the existing literature. Indeed, the amelioration/elimination of this second defect is the *raison d'être* of the postsample inference procedure developed in Ashley(1992,1994) and presented below.

This second generic defect of previous postsample inferential procedures arises due to the "messiness" of typical postsample forecast error series. Since the postsample forecast errors made by the models under

[2]West(1994)'s analysis of the postsample predictive ability of econometric models is exceptional in its heavy dependence on information about the models themselves. No actual postsample forecasts are used in his framework, so it is actually an *in-sample* procedure for inferring the postsample forecasting effectiveness of the models under consideration.

consideration are typically strongly cross-correlated, often significantly serially correlated, and rarely known to be gaussian, exact statistical tests are never available – thus, all known postsample inference procedures are only asymptotically valid.[3]

This is a particularly serious defect since, as noted above, the fact that each observation included in the postsample period reduces the amount of data available for model specification and estimation purposes typically leads to quite short postsample periods. Thus, existing postsample inference procedures merely replace the inferential uncertainty of the in-sample procedures (due to data mining) with inferential uncertainty due to the application of asymptotically-justified procedures to short postsample periods.

However, postsample inferential methods have one additional advantage over in-sample methods, an advantage which is intensively exploited in the postsample inference procedure described below: whereas the inferential uncertainty in the in-sample methods due to prior specification searches is, in practice, unknowable; the inferential uncertainty in a postsample inference due to the finite length of the postsample period can be estimated nonparametrically using a variation on the bootstrap.

The remainder of this paper describes a practical implementation of this idea. The nature and limitations of the ordinary bootstrap are illustrated in the first portion of Section 2 via an application to a very simple problem: inference on the population mean of an i.i.d. random variate. The remainder of Section 2 describes the simple (but crucial) innovation which allows the variation on the bootstrap used here to provide useful estimates of the small-sample uncertainty in its own inferences. In Section 3 this approach is extended in a straightforward way to provide inferences on ratios of (expected) functions of a pair of correlated (and serially correlated) postsample forecast error series; in Section 4 it is used to re-analyze 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. In addition to providing substantially more credible inference results on the postsample mean square forecasting error ratios obtained in that study, the inference approach developed here is also used to provide Granger-causation inferences under

[3]The "exact" tests given in Diebold and Mariano(1994) are valid only if the loss differential series is serially uncorrelated. Their suggested "partial remedy" for this limitation (via Bonferroni bounds) would hardly be reasonable in small samples. The simulation results they report on their explicitly asymptotic test (S_1) indicate that it is quite strongly mis-sized at N=16. S_1 looks much better for $N \geq 32$, but their results are specific to the loss function/error distribution combination used in these particular simulations. The Meese and Rogoff(1988) test requires gaussianity even asymptotically; simulations reported in Diebold and Mariano(1994) indicate that even N=512 is problematic for this test.

alternative loss functions on forecast errors (such as the absolute error loss function or an asymmetric piecewise-quadratic loss function) and to generate credible small-sample inferences using the scaled mean loss differential statistic introduced (as part of an asymptotic test) in Diebold and Mariano(1994). This new inference approach can also be used to estimate how strong the differential forecasting effectiveness evidence would need to be (with the postsample period length actually used) or how large a postsample period would be needed (with evidence of similar strength to that actually observed) in order to obtain a given target level of inferential precision, such as significance at the 5% level. Results of this nature – which clearly indicate that postsample model validation periods substantially longer than the 5 to 20 periods typically reserved in past studies are necessary in order to credibly detect 20% - 30% MSE reductions – are discussed in the final portion of the paper.

2. A Variation on the Bootstrap Yielding Credible Small-Sample Inferences

It seems evident that the crucial impediment to postsample statistical inference has been the fact that it is often infeasible to reserve enough data for a postsample inference period as to make the use of the usual asymptotic inferential methods credible. But bootstrap-based inference methods are only asymptotically valid also. How, then, can they possibly resolve this problem? The answer lies in an almost universally un-utilized feature of the bootstrap: the fact that it can be used to quantify the uncertainty its own large-sample approximation imposes on the inference results. In this Section the bootstrap is described in a very simple setting so as to (a) convey its essence, (b) contrast it with traditional inference methods, and (c) clarify the implementation of this rarely-utilized feature.

Consider the simple problem of making inferences about the population mean, μ, of a random variable x. Suppose that N independent realizations of x are available, where N is not very large. What can one say about the (fixed but unknown) value of μ? Since μ is unknown, our statements regarding its value are necessarily probabilistic – e.g., "a specified interval contains μ with given probability" or "a specified hypothesis about μ will be wrongly rejected with given probability," etc.

A typical non-bootstrap approach to this inference problem would begin by constructing an estimator of μ and continue by deriving the sampling distribution of this estimator. In this case the sample mean, $\bar{x}$, is an obvious choice for the estimator. The sample mean is gaussian if the x's are gaussian but, since $\bar{x}$ is known to be asymptotically gaussian so long as the x's are sufficiently close to being i.i.d., most analysts would in this

instance routinely assume that $\bar{x}$ is "close enough" to gaussian because the x's are "close enough" to gaussian. This assumption is necessarily counterfactual in practice (although it may often be a reasonable approximation) since gaussian variates are continuously distributed and have infinite range, whereas economic variates (as measured) are typically discretely distributed over a finite range. Asymptotically valid confidence intervals for μ and hypothesis tests concerning μ would then be derived on the assumption that the sampling distribution of $\bar{x}$ is gaussian.

In contrast, using the bootstrap approach one begins by making a different – also counterfactual – assumption about the population distribution of the x's. Instead of assuming that this distribution is essentially gaussian, using the bootstrap one assumes that the population distribution of the x's is essentially identical with the empirical distribution of the x's, which places equal probability mass on each observed value. This could conceivably be true, but it is hardly likely. After all – and especially for smaller samples – the empirical distribution is obviously "lumpier" than one would expect the population distribution to be. Moreover, it is inherently unreasonable to expect that the range of the sample data is as large as the range of the population from which this finite (and usually rather small) sample was drawn.

This counterfactual assumption (like the counterfactual gaussianity assumption of the non-bootstrap approach) can be shown to make no difference asymptotically. (See, for example, Singh(1980), Bickel, P. J. and D. A. Freedman (1981), and Beran, R. (1986) for proofs.) And, like gaussianity, it is an extremely useful assumption. Having made it, one can then create as much "new" data as one likes by merely picking at random out of the empirical distribution.[4] Thus, one could easily create, say, 2000 "new" N-samples on x and thereby obtain estimates of the c.d.f.'s of x and $\bar{x}$. These are most commonly used (1) to estimate moments (or functions of moments) of x – e.g., μ or var(x) – or (2) to obtain a $(1-\alpha)$% confidence interval for μ with α given.[5]

---

[4]More explicitly, if the original sample data is denoted x(1) … x(N), a "new" N-sample can be obtained (via "re-sampling from the empirical distribution") by generating N random integers $\{j_1 \ldots j_N\}$ which are independent draws from the (discrete) uniform distribution which puts equal weight on each integer in [1, N]. The resulting "new" N-sample is
$x(j_1) \ldots x(j_N)$.

[5]The theory of bootstrap-based confidence intervals has seen enormous development in the last decade – e.g., see DiCiccio and Romano(1988). Nevertheless, in the present context of model validation/choice based on comparing a measure (e.g., MSE or MAE) of the postsample forecasting errors made by two competing models, it turns out that expressing the inference problem in terms of an hypothesis test confers a crucial advantage; this point is taken up later in this Section.

Alternatively, the bootstrapped c.d.f. of $\bar{x}$ can be used to estimate the significance level (α) at which the (composite) null hypothesis H_o: $\mu \leq 5$ can be rejected in favor of the alternative hypothesis H_a: $\mu > 5$.

Since both approaches make counterfactual (albeit asymptotically valid) assumptions about the true population distribution from which the observed data were picked, why might one prefer the bootstrap? The bootstrap has three major advantages: (1) the bootstrap is often easier to apply than alternative methods, although it does require substantially larger amounts of computational resources,[6] (2) the bootstrap is often more accurate for small samples, occasionally {as in Freedman and Peters (1984)} dramatically so, and, finally, (3) the bootstrap approach lends itself to checking its own accuracy.

This last advantage of bootstrap-based inference is the most relevant feature for the present purpose. Suppose that – instead of generating 2000 new N-samples and computing the proportion of them for which $\bar{x} \leq 5$ – one initially generates just 100 new N-samples ("starting samples") and then uses each one to initiate the computation of a bootstrap inference. That is, one generates a group of 2000 new samples out of the empirical distribution of x implied by each of these 100 starting samples and then computes the proportion of the 2000 N-samples for which $\bar{x} \leq 5$. Since $\bar{x}$ is unbiased for μ, one minus this proportion is the significance level at which H_o: $\mu \leq 5$ can be rejected ("the inference probability on μ" or just "the inference on μ" below) using this group of 2000 N-samples generated from the ith starting sample. Thus, at a cost of drawing 200,000 N-samples instead of 2000, one obtains 100 inference probabilities on μ instead of just one. Presuming that 2000 N-samples is "enough" – which is easily checked by comparing the inferences from the first 1000 N-samples to those from the remaining 1000 N-samples – the dispersion of these 100 inferences provides an estimate of the uncertainty one should attach to the median[7] of the 100 inferences due to the finite size of N. One could take this dispersion as an estimate of the "fragility" of the inference, in the spirit of Leamer(1985).[8]

---

[6]This comment presupposes the availability of appropriately accessible implementing software. MS-DOS software implementing the approach described below is available from the author and from the *JBES* ftp server.

[7]Since the distribution of the inference probabilities necessarily becomes quite non-gaussian when the median inference probability becomes small, the median and interquartile range 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 as the median inference ($Q_{.50}$) plus either its interquartile range ($Q_{.75} - Q_{.25}$) or its empirical 50% confidence interval, $[Q_{.25}, Q_{.75}]$, where Q_α is the α% fractile of the 100 inference probabilities.

[8] Of course, this uncertainty estimate is itself only asymptotically valid, but this, too, can be checked. For example, one can (and should) immediately use this dispersion estimate to check that the median inference does not significantly differ from the inference one obtains from the ordinary bootstrap, in which 2000 new samples are drawn from the empirical distribution of the observed sample data. Monte-carlo simulations, providing a more detailed check at a tremendously larger cost, are reported below in footnote 24.

There are three reasons why this simple check on the finite-sample validity of bootstrap-based inferences is almost never implemented. {Freedman and Peters (1984) provides a rare, perhaps unique, exception.} First of all, the computational burden involved is obviously substantial. That isn't nearly as important a constraint as it was ten years ago, but even now it is unreasonable to assume that a desktop computer will be able to calculate ca. 100 bootstrap inferences on a time scale commensurate with interactive data analysis. Secondly, as noted above[5], the main thrust of bootstrap inference theory has focussed on the construction of confidence intervals of given coverage. One could, of course, calculate 100 such intervals, each one based on one of the 100 generated starting samples, but it is by no means evident how to utilize the resulting 100 confidence intervals to coherently quantify the small-sample uncertainty in the bootstrap inference. Thus, while several second-level ("double") bootstrapping proposals superficially similar to that outlined above have been advanced – e.g., Beran(1987) – the second level of bootstrapping in these proposals is used to improve the small-sample coverage accuracy of the resulting confidence interval, not to quantify the small-sample uncertainty in the inference results.

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

In the preceding example, 100 new N-samples ("starting samples") are picked from the empirical distribution of the original data and then 2000 "new" samples are picked from the empirical distribution of each of these 100 new N-samples. Thus, $\bar{x}$ is actually computed 200,000 times: 2000 times for each new "starting sample." 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 new starting samples; the population mean of this distribution could sensibly be denoted μ_1. (Note that μ_1 must precisely equal $\bar{x}_1$, 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 new starting samples; the population mean of this distribution could sensibly be denoted μ_2 – which must precisely equal $\bar{x}_2$ – and so forth. Clearly, these 100 population means ($\mu_1 \ldots \mu_{100}$) will vary substantially for small N, introducing substantial additional dispersion in the resulting 100 inferences, dispersion which is extraneous in that it is not due to the bootstrap approximation of using the empirical distribution of the original sample data to replace the true 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 were drawn, monte-carlo fashion, from the true population distribution.

In contrast, 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. How does this work? If the observed sample mean (using the actual sample data) is $\bar{x}_o$, then H_o: $\mu \leq 5$ implies that $\bar{x}_o$'s sampling error exceeds

$\bar{x}_o - 5$. Since the population mean of the distribution used to generate the 2000 N-samples (and 2000 sample means) from the ith starting sample is known – as noted above, it is just the sample mean of the ith starting sample – these 2000 sample means yield 2000 observations on the sampling error distribution of $\bar{x}$. Thus, the fraction of these 2000 sampling errors which exceed $\bar{x}_o - 5$ is the inference probability on μ from the ith of the 100 starting samples. Note, however, that 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 ($\bar{x}_i - \mu_i$) is much more stable across the starting samples than is the distribution of the $\bar{x}_i$.

The median of the 100 inference probabilities obtained from these 100 "bootstrapped" sampling error distributions thus provides a reasonable, asymptotically valid, estimate of the probability that $\mu \leq 5$. And, most importantly, the dispersion (interquartile range or empirical 50% confidence interval[7]) of these 100 inference probabilities explicitly quantifies the uncertainty in the median inference due to the finite-sample impact of the bootstrap approximation of replacing the population distribution of the x's by their observed empirical distribution. In this way – by explicitly estimating the small-sample uncertainty in the bootstrap inference – one can obtain credibly useful bootstrap inferences in a small-sample setting.

3. Adapting the Bootstrap to Postsample Inference

The procedure described in the previous Section provides credible small-sample inferences on the population mean of the distribution from which an i.i.d. N-sample was drawn. In essence, small-sample credibility is achieved by generating ca. 100 independent bootstrap inferences, whose dispersion quantifies the small-sample uncertainty in the median inference. In this Section this procedure is extended in a straightforward way to the problem of testing whether the expected size of the postsample forecasting errors from one model significantly exceeds that of another, based on a single observed sequence of N postsample forecasting errors from each model.[9]

These two postsample forecasting error series (denoted x_t and y_t below) may be both contemporaneously and/or serially correlated. More precisely, it is assumed that the (x_t, y_t) are sufficiently covariance stationary (and that their serial dependence is sufficiently linear in nature) that their generating mechanism can be sensibly represented as a bivariate VAR process:

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

This covariance stationarity assumption can and should 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 by examining scatterplots of x_t and y_{t+k} for k = 0, ±1, ±2 … (looking for evidence of nonlinearities in the relationships and checking for the occasional overly-influential observation[10]) – but formal testing is beside the point in this context: if N were large enough for such testing to be justified, one could use available asymptotic methods to obtain the 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*

[9]See Ashley(1992) for details on importance sampling, non-bootstrap re-sampling schemes and other elaborations; while implemented in the software as options, these are not described here because they are not all that useful in this context.

[10]Since gaussianity is not assumed, an "outlier" which is not overly influential can be tolerated; such an observation can sensibly be viewed as an ordinary realization from a nongaussian (x_t, y_t) distribution.

these models; it is merely a descriptive parameterization of the possible serial correlation and likely crosscorrelation structure of the forecast errors (x_t and y_t) made by these two models. Thus, covariance stationarity of the VAR model 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, one need not 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 details (or lack of same) in these two models is irrelevant.[11]

The coefficients in the distributed lag polynomials $\{\phi_{11}(B), \phi_{12}(B), \phi_{21}(B), \text{and } \phi_{22}(B)\}$, the intercepts ($\mu_x$, 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 order (maximum lag in) each of the four lag polynomials. In ordinary VAR modelling, 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.[12]

At first glance the required specification of upper bounds on the orders of the lag polynomials in the VAR representation of (x_t, y_t) seems a bit problematical. In practice, however, it is not all that difficult to obtain usably accurate upper bounds on these lag polynomial orders by running a few linear regressions and culling out the clearly insignificant terms. This is because the inference results from this procedure turn out to be quite insensitive to minor over-elaboration in the specification of the upper bounds on these lag polynomial orders for the VAR model; simulations illustrating this point are given at the end of this Section.

How then, might one proceed to calculate ca. 100 inference probabilities in this setting? Figure 1 provides a schematic description of the calculation of ρ_{37}, the probability that a specified relative accuracy

---

[11]If one were privy to the "true" data generating mechanism, one could, in principle, derive the asymptotic distribution of (x_t , y_t) without having to actually doing the postsample forecasting, as in West(1994). Of course, the validity of one's conclusions are then directly conditional on the accuracy of one's information on the "true" data generating mechanism.

[12]This assumption will fail if (x_t, y_t) is related to its own past in a substantially **nonlinear** way {Ashley(1994, fn. 13)} but postsample forecasting periods are ordinarily so short that any consideration of serial dependencies substantially more complex than the low-order VAR mechanism used here is out of the question in any case.

criterion, r, is less than or equal to some given value, τ, based on the 37th starting sample. (Most commonly, τ will equal one.) The population value for r *could* be the ratio of the two MSE's

$$r_{MSE} = \frac{MSE(x_t)}{MSE(y_t)}$$

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

$$r_{MAE} = \frac{MAE(x_t)}{MAE(y_t)} = \frac{E(|x_t|)}{E(|y_t|)}$$

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

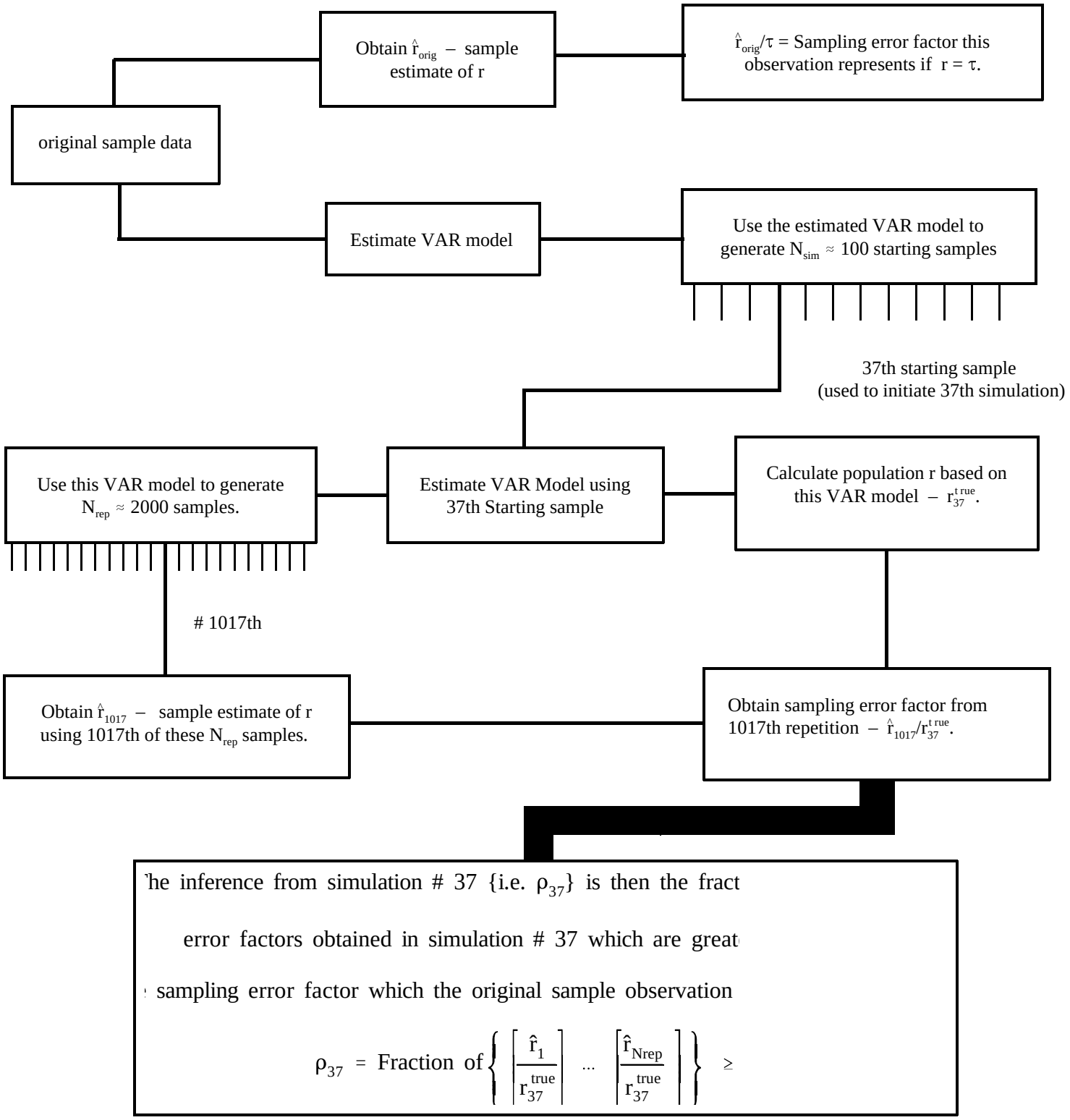
$$r_{asy} = \frac{E(s(x_t)x_t^2)}{E(s(y_t)y_t^2)}$$

where $s(z) = 1$ for $z \geq 0$ and $s(z) = 2$, say, for $z < 0$. Alternatively, one might prefer to measure the superiority of the y_t forecast error series over the x_t error series using the studentized expected loss differential criterion proposed by Diebold and Mariano(1994):

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

Figure 1.

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



where $f_d(0)$ is the spectral density of the numerator at frequency zero.[13] Loss(·) is a situationally appropriate loss function; the absolute value function is used here. I have defined r_{DM} as the exponential of Diebold and Mariano's criterion so that, for all four criteria used here, $r \in (1, \infty)$ and $r = 1$ for equivalent forecasts.

Returning to Figure 1, r is the (unknown) population value of whichever forecast accuracy criterion one has chosen. The object is to test the null hypothesis H_o: $r \leq \tau$ against the alternative hypothesis that $r > \tau$ based on a single observed N-sample: $(x_1, y_1) \ldots (x_N, y_N)$. As Figure 1 indicates, this sample data is only used twice. First, it is used to obtain $\hat{r}_{orig}$, a consistent sample estimate of r. {E.g., for r_{MSE}, $\hat{r}_{orig}$ is just the ratio of the sample $MSE(x_t)$ to the sample $MSE(y_t)$.} The resulting $\hat{r}_{orig}/\tau$ figure is the sampling error *factor* this sample estimate represents if H_o is barely true.[14] Second, the original sample data is used to obtain OLS estimates of the parameters in the VAR model, Equation 1.[15] At this point the original sample data has yielded (1) an observed sampling error factor (under H_o) and (2) an estimated data generating mechanism (the estimated VAR model and its fitting errors) which can be used (as described below) to both estimate the distribution from which the observed sampling error factor was drawn and to quantify the small-sample uncertainty in the inference based on this estimated sampling error factor distribution.

At this point the 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 $\{(\epsilon_1, \eta_1) \ldots (\epsilon_N, \eta_N)\}$ were drawn is identical to the empirical distribution of the fitting errors $\{(\hat{\epsilon}_1, \hat{\eta}_1) \ldots (\hat{\epsilon}_N, \hat{\eta}_N)\}$ which places probability mass 1/N on each of these observed 2-vectors.[16] If the maximum lag in the VAR is p, 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

---

[13]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.

[14]The population value of r is defined as an expected loss ratio so as to be independent of the units in which x_t and y_t are expressed. Consequently, it is preferable to compare the sampling error factor $\hat{r}_{orig}/\tau$ to the distribution of sampling error factors rather than to compare the sampling error itself ($\hat{r}_{orig} - \tau$) to the distribution of sampling errors, $\hat{r}_{orig} - r$. Most importantly, this choice makes the inference results independent of which error series is chosen to appear in the numerator of r – i.e. it ensures that Prob$\{r \leq \tau\}$ precisely equals Prob$\{r^{-1} \geq \tau^{-1}\}$.

[15]Since corr(ϵ_t, η_t) can be substantial, SUR would be preferable in some cases; typically, N will not be large enough to justify its use in this context, however. It *is* often useful to correct the OLS parameter estimates for small-sample bias; this issue is taken up in the next Section.

[16]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. What *is* new here is the use of the double bootstrap to quantify the uncertainty in the inference results induced by the boostrap assumption and the sampling errors in the estimated coefficients of the autoregressive model.

and the "next" innovation 2-vector, $(\hat{\epsilon}_j, \hat{\eta}_j)$, where j is a randomly chosen integer in the interval [1, N]. One can use the last p sample values of (x_t, y_t) as "start-up" values or the sample means of x_t and y_t or even zeroes – after a sequence of 50 to 100 observations have been generated in this way (and discarded) the influence of the "start-up" values becomes negligible. In this way, the algorithm generates $N_{sim} \approx 100$ "new" N-samples on (x_t, y_t).[17]

Each of these N_{sim} "starting samples" – Figure 1 explicitly uses the 37th – is then used to "bootstrap" an estimate of the distribution of sampling error factors $(\hat{r}_j / r_{37}^{true})$ for $j = 1 \ldots N_{rep}$, where $N_{rep} \approx 2000$. Figure 1 describes how this is done. A VAR model is estimated using the 37th starting sample and then used to generate (a) $j = 1 \ldots N_{rep}$ N-samples $\{(x_1, y_1) \ldots (x_N, y_N)\}$, each of which yields a sample estimate of r, $\hat{r}_j$ and (b) a single huge sample of length 10,000N $\{(x_1, y_1) \ldots (x_{10000N}, y_{10000N})\}$, which yields a large-sample estimate of r $\{r_{37}^{true}\}$ which is essentially equal to the population value of r for this 37th VAR process.[18] Presuming that N_{rep} is sufficiently large that the observed distribution of the N_{rep} values of $(\hat{r}_j / r_{37}^{true})$ adequately characterizes the sampling error factor distribution implied by the 37th starting sample, $\hat{\rho}_{37}$ (the 37th estimate of the probability that $r \leq \tau$) is just the fraction of the N_{rep} "observed" sampling error factors that exceeds $\hat{r}_{orig}/\tau$.

Note that the N_{sim} sampling error factor 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. And second, even if the fixed VAR coefficients (obtained from the original sample data) used in generating all N_{sim} of the starting samples were given, for finite N the empirical distribution of the residuals implied by the ith starting sample is not identical to the population distribution {the fitting errors from the VAR model estimated over the original data, $(\hat{\epsilon}_1, \hat{\eta}_1) \ldots (\hat{\epsilon}_N, \hat{\eta}_N)$, each with equal weight} from which these residuals were drawn. Thus, the dispersion in the inference probabilities obtained from these N_{sim} sampling error factor distributions – i.e., the dispersion in $\hat{\rho}_1 \ldots \hat{\rho}_{Nsim}$ – quantifies the inferential uncertainty caused by sampling errors in the estimation of the VAR model *and* by the use of the (asymptotically valid) bootstrap approximation of replacing the population distribution of the VAR model fitting errors by their empirical distribution.

---

[17]Actually, 2100 new N-samples are drawn – the last 2000 yield an estimate of the distribution of sampling error factors (and hence an inference on r) using the original data as the "starting sample." This inference is ordinarily quite close the median inference based on the N_{sim} generated starting samples. A significant discrepancy indicates that something is amiss; for example, the OLS estimates of the parameters in Equation 1 may need to be corrected for small-sample bias.

[18]The $r_1^{true} \ldots r_{Nsim}^{true}$ are analogous to the 100 population means $(\mu_1 \ldots \mu_{100})$ of Section 2.

The only specification information which is required in order to produce the empirical distribution of inferences $\{\hat{\rho}_1 \ldots \hat{\rho}_{Nsim}\}$ is a reasonably tight upper bound on the order (maximum lag in) each of the four distributed lag polynomials $\{\phi_{11}(B), \phi_{12}(B), \phi_{21}(B), \text{and } \phi_{22}(B)\}$ in Equation 1. The sensitivity of this distribution of inferences to errors in specifying these bounds was investigated by generating a number of (x_t, y_t) 50-samples from the VAR model:

$$x_t = .001 + \epsilon_t \qquad \epsilon_t \sim \text{NIID}(0, .001)$$

$$y_t = .001 + .6\, y_{t-1} - .2\, y_{t-2} + \eta_t \qquad \eta_t \sim \text{NIID}(0, .0005) \qquad \text{corr}(\epsilon_t, \eta_t) = .20$$

Two of these 50-samples were selected for further analysis. The first, denoted "Strong Model" below, yields an estimated t ratio of 2.63 for the estimated y_{t-2} coefficient; the second, denoted "Marginal Model" below, yields an estimated t ratio of 1.43 for this coefficient.

Table 1 summarizes the inference results obtained using these two samples to estimate the probability that $\text{var}(x_t)/\text{var}(y_t) \leq \tau$, using two different values of τ and a selection of different specifications for the maximum lags in the $\phi_{i,j}(B)$ polynomials. These specifications are denoted $N_x : N_y$ in the Table, where N_x is the order of $\phi_{11}(B)$ – i.e., the maximum lag with which x_t enters the x_t equation – and N_y is the order of $\phi_{22}(B)$. Thus, the true specification is denoted 0:2; the specification denoted 0:2:1 includes y_{t-1} in the x_t equation. Two values of τ are used, so as to examine the sensitivity of the inferences both where the null hypothesis is supported ($\tau = 1.1$) and where the null hypothesis is not supported ($\tau = 1.7$), leading to rejection at the 5% to 8% level.

Median inference results which lie outside the empirical 50% confidence interval[7] for the 100 inferences obtained using the correct specification are entered in Table 1 in bold. These results are typical: the choice of an over-elaborate specification is inconsequential, but an under-elaborate specification significantly affects the inference results if (and typically only if) the estimated coefficient on the incorrectly omitted term is statistically significant in the VAR model estimated using the observed sample data. Thus, wrongly omitting a term from Equation 1 significantly biases the inference results only when the mis-specification is sufficiently serious as to be evident in the estimated VAR model. Note also that the dispersion in the inferences is not appreciably affected by minor over-elaboration of the VAR specification, so there is no penalty for "playing it safe" and including a marginal term in the specification. In essence, explicitly quantifying the inferential uncertainty caused by likely sampling errors in estimating the coefficients in the VAR (and the additional uncertainty due to

the bootstrap approximation of substituting the empirical distribution of the fitting errors for the population distribution of the innovations) makes what would otherwise be a crucial specification choice inconsequential.

Table 1
Sensitivity of Inferences to Errors in Specifying VAR Model

	Strong Model: t = 2.63				Marginal Model: t = 1.43			
	τ = 1.1		τ = 1.7		τ = 1.1		τ = 1.7	
	Median Inference	IQR	Median Inference	IQR	Median Inference	IQR	Median Inference	IQR
Correct Specification: {50% confidence interval underneath}								
0 / 2	.427 {.408, .450}	.043	.060 {.048, .078}	.030	.436 {.412, .469}	.058	.064 {.045, .082}	.036.
Over-Elaborate Specification:								
1 / 2	.422	.036	.064	.022	.430	.055	.064	.028
0 / 3	.424	.033	.064	.031	.446	.067	.072	.048
0 / 2 / 1	.430	.034	.061	.031	.425	.053	.050	.036
Under-Elaborate Specification:								
0 / 0	**.344**	.023	**.017**	.017	**.328**	.023	**.010**	.008
0 / 1	**.481**	.080	**.081**	.058	.452	.069	.067	.050

4. 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?[19] 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 one most credibly test hypotheses about Granger causation between a pair of time series? Here AGS break new ground by explicitly proposing that one can most credibly test the direction of Granger-causation between two time series by comparing the *postsample forecasting effectiveness* of models for each series based on nested information sets.[20]

In particular, suppose that the postsample forecasts of, say, 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, one can conclude that aggregate advertising expenditures Granger-cause aggregate consumption spending. Thus, AGS "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 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 20 period postsample period, 1970I to 1975IV. AGS propose a procedure for testing whether this MSE reduction is statistically significant but, in common with all other postsample inference procedures,[3] their procedure is only asymptotically valid. 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

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

[20]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 in this Section.

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, where r is r_{MSE} or r_{MAE} or r_{ASY} or r_{DM}, depending on one's loss function with respect to forecast errors.

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 despite the fact that the largest observation on each series happens to be the last one. Histograms of the data clearly indicate that the only unusual feature to this maximal observation pair is its position at the end of the postsample period – its size is unremarkable even under the hypothesis that the population distributions are gaussians truncated at $\pm 2\sigma$. Both series are serially correlated, however; OLS regression yields:

$$\begin{array}{lcccccccc} x_t & = & 7.38 & - & .032\, x_{t-1} & - & .452\, x_{t-2} & + & \epsilon_t \\ & & (1.12) & & (.11) & & (1.52) & & \end{array} \qquad \begin{array}{l} R^2 = .134 \\ DW = 1.92 \end{array}$$

$$\begin{array}{lcccccccc} y_t & = & .76 & + & .332\, y_{t-1} & - & .428\, y_{t-2} & + & \eta_t \\ & & (.14) & & (1.23) & & (1.53) & & \end{array} \qquad \begin{array}{l} R^2 = .193 \\ DW = 1.61 \end{array}$$

where the figures in parentheses are estimated t ratios and both fitting error series appear to be serially uncorrelated. The coefficients on x_{t-2} and y_{t-2} are hardly compelling, but are included in the VAR since the results from the inference procedure are known (Table 1) to be quite insensitive to modest over-elaboration in the VAR specification. Table 2 summarizes the inference results which, with $N_{sim} = 100$ and $N_{rep} = 2000$, tied up a desktop computer for ca. 30 minutes:

Table 2
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 {ratios defined in Section 3}	.738	.934	.803	.877
Significance level for asymptotic test	.092	unknown	unknown	.278
Bootstrap inference results: {Results without bias correction in brackets}				
Ordinary bootstrap on original data	.215 {.214}	.375 {.356}	.302 {.290}	.323 {.319}
Median of N_{sim} inferences ($Q_{.50}$)	.226 {.236}	.380 {.383}	.310 {.314}	.322 {.329}
Intequartile range ($Q_{.75}$ - $Q_{.25}$)	.080 {.078}	.034 {.039}	.073 {.064}	.044 {.050}
Empirical 50% interval [$Q_{.25}$, $Q_{.75}$]	[.184, .263]	[.361, .396]	[.271, .344]	[.306, .350]
Sample ratio needed for 5% result (discussed in Section 5)	.58	.73	.55	.65

Interpreting these results:

1. The "sample" r_{MSE} = .738 figure means that including past consumption spending in the information set for forecasting advertising expenditures yields a 26% reduction in (observed) postsample MSE. Thus, r < 1 is evidence for consumption Granger-causing advertising. All four ratios are less than one; but are they *significantly* less than one?

2. The (asymptotic) inference procedure given by AGS indicates that the MSE reduction is significant at the 9% level; the (asymptotic) inference procedure given by Diebold and Mariano(1994) indicates that the observed expected loss differential is significantly negative (so that r_{DM}< 1) at the 27.8% level.[21]

Turning to the results obtained using the inference procedure described here:

3. Correcting the OLS estimates of the VAR parameters for small-sample bias does not change the inference results very much; as one might expect, however, it does move the median inference noticeably (albeit not significantly) closer to the (single) bootstrap inference obtained directly from the original data. The small-sample bias in the OLS parameter estimates evidently biases the inferences somewhat; consequently, the bias-corrected

[21]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.

inference results are preferable.[22]

4. Since the empirical 50% interval (the endpoints of the "middle half" of the N_{sim} inference probabilities) is [.184, .263], one can reasonably conclude that the observed 26% MSE drop is only significant at the 18-26% level. This range of significance levels quantifies the small-sample uncertainty in the inference significance level due to the fact that the form of the distribution from which these data were drawn and its serial correlation structure is not arbitrarily assumed.

5. Similarly, the observed Diebold-Mariano postsample mean loss differential is significantly negative at the 31-35% level where, again, the range of significance levels quantifies the small-sample uncertainty in the inference result.

6. Even though both the AGS and the Diebold-Mariano inference procedures yield inferences which are substantially different from the results obtained using the inference procedure proposed here, *they are not "wrong."* Indeed, one of the $N_{sim} = 100$ starting samples yields a bootstrap inference probability of less than .092 for $r_{MSE} \leq 1$ and four of the starting samples yield bootstrap inference probabilities of less than .278 for $r_{DM} \leq 1$. So the the AGS and the Diebold-Mariano inference results are very probably misleading, but they are not actually incorrect.

7. But if the results of AGS and the Diebold-Mariano inference procedures are not "wrong," why go to all this trouble? The reason is that these other asymptotic inference procedures give no hint as to how accurate their results are – users of these procedures in essence ask their intended audience to take it "on faith" that the sample size (while small) is large "enough."[23] In contrast, the procedure presented here provides explicit, quantitative estimates as to how trustworthy the inference significance levels it produces are for the particular data at hand. Therefore, while each of the inferences produced by the procedure described here is a bit "fuzzier" than the analogous AGS or Diebold-Mariano result – since it is expressed as a range of significance levels rather than as a single value – the inference results given above in points #4 and #5 are substantially more *credible* than the

[22]Similar results with other data reported in Ashley(1992) indicate that these results are typical and that the inference results are quite insensitive to the details of how the bias correction is done. Here the OLS-estimated VAR model is used to generate 50 new samples. The small-sample OLS bias in each coefficient is estimated from the average discrepancy between the resulting 50 OLS estimates of the parameter and the value of the parameter used in the generating model.

[23]Or that the mechanism which generated their data is similar "enough" to that used in a particular simulation study.

analogous AGS or Diebold-Mariano results.[24]

5. Conclusions And Comments on the Feasibility and Desirability of Postsample Inference

Credible statistical inference on postsample forecast error series is now quite feasible – even easy – using the bootstrap-based inference procedure described above.[6] The inferences from this procedure

- avoid the pre-test biases which data mining induces in all in-sample tests,
- explicitly and satisfactorily allow for the "messiness" (i.e., the serial and/or crosscorrelations and the outliers/nongaussianity) commonly found in postsample time series data,

and

- are credible even for the small samples typically available for postsample inference, because the procedure explicitly quantifies the uncertainty in the inference caused by the limited sample size.

However, now that credible postsample inference is feasible, it is no longer either necessary or proper to remain vague about either the actual strength of our postsample inferences or about the amount of postsample data which is needed for effective inference.

For example, Table 2 of the previous Section clearly indicates that the 26% MSE improvement obtained by AGS(1980) over a 20quarter postsample period is simply not significant at even the 10% level. How large an MSE improvement would have sufficed to yield a 50% interval containing .05? The results reported in the last row of Table 2, obtained by testing $r \leq \tau$ with increasingly large values of τ, show that an MSE improvement of over 40% or an MAE improvement of over 30% is needed, given the distribution and serial/cross correlation structure of these data.

Alternatively, one can increase the length of the generated samples until a desired inferential precision is achieved. Such calculations, reported in Ashley(1992), indicate that a postsample model validation period must typically be ca. 25 to 45 periods long in order to detect a 30% MSE drop at the 5% level of significance or must

[24]But – with such a small sample – are these empirical 50% intervals themselves credible? To check this, 200 20-samples on (x_t, y_t) were generated in monte carlo fashion, taking the parameter estimates {and estimated variance(ϵ_t, η_t) matrix} from equation 2 as the "true" values and drawing the (ϵ_t, η_t) as serially independent picks from a bivariate gaussian distribution truncated at $\pm 2\sigma$. Scaling x_t so that each of the 200 generated samples yields the same sample MSE ratio (.738) actually observed and then applying the inference procedure to each of the resulting 200 samples yields 200 sets of N_{sim} inference results, each of which is comparable to the "r_{MSE}" column of Table 2. The median of these 200 "median inference" results is .245; the median of these 200 "interquartile range" results is .081. These are quite close to the Table 2 values of .226 and .080. Taking .245 to be the "true" value of ρ, 58% of the 200 50% intervals contain ρ, which seems like excellent agreement in view of the small sample size and the remaining uncertainty as to the true value of ρ.

be more like 50 to 100 periods long to detect a 20% MSE drop. Evidently, a model which cannot provide at least a 20-30% MSE improvement over that of a competing model is simply not going to "test out" as significantly better than its competitor over postsample periods of reasonable size. And the 5 to 20 periods that have in the past been allocated to postsample model validation/inference (when it was done at all) are apparently quite inadequate in length to detect MSE improvements of the sizes one ordinarily sees. Yet retention of a postsample model validation period much in excess of 30 to 40 periods seems rather impractical in many econometric contexts.

What is to be done? One possibility 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 improvements which are significant at the 10% 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 explicitly revise upward our estimate of the relative importance of model validation and thereby revise downward our notion of just how impractical a 40 to 80 period postsample model validation/inference period actually is.

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, why do it? This issue is discussed at some length in Ashley(1992) and briefly addressed in Section 1 above. Here, the best response to this question is: "because the alternative approach of in-sample model validation/inference (over the same data used for specifying and estimating the model) makes it **too** easy to obtain supportive results."

Indeed, I would argue that this is the principal reason why the economics community has, over the years, accepted for use (in both theory-testing and policy evaluation/forecasting settings) so many badly-misspecified models. Had we been able (through the use of tools such as the inference procedure proposed here) and willing (still in doubt!) 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.[25]

6. 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.

Ashley, R. (1994), "Postsample Model Validation and Inference Made Feasible," V.P.I. & S.U. Economics Department Working Paper #E94-15.

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. (1994), "Comparing Predictive Accuracy," unpublished manuscript.

[25]Note that the use of postsample model validation methods in no way precludes the application of in-sample model validation procedures to the model specification/estimation sample or (at the very end of one's analysis) to the entire available data set. However, one must explicitly recognize that the specification search activity which routinely takes place using the specification/estimation sample typically renders the asymptotic distribution of the usual in-sample test statistics a misleadingly poor approximation to their actual sampling distributions. Consequently, these in-sample test statistics are actually appropriate for use only as descriptive statistics, not in testing hypotheses. Finally, I should add that it is by no means impossible to invalidate the postsample inference machinery via pre-test biases induced by repeated use of postsample inference to "mine" the postsample period for a model specification.

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.

Hendry, D.F. and Mizon, G.E. (1985), "'Procrustean Econometrics: Or Stretching and Squeezing Data," (in Granger, C.W. J., (1990), **Modelling Economic Series**, Clarendon Press, Oxford.)

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

LeBaron, B. and Weigend, A. S. (1994), "Evaluating Neural Network Predictors by Bootstrapping," Paper #9447, Social Systems Research Institute, University of Wisconsin-Madison.

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.

Rhee, M. J. (1994), "Forecasting Stock Market Indices with Neural Networks," unpublished manuscript.

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

West, K. D. (1994), "Asymptotic Inference About Predictive Ability," unpublished manuscript.