USING THE BOOTSTRAP TO EVALUATE FORECASTING EQUATIONS

BY

STEPHEN C. PETERS AND DAVID A. FREEDMAN

TECHNICAL REPORT NO. 20
REVISED OCTOBER 1984

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY

RESEARCH PARTIALLY SUPPORTED BY
THE ENERGY INFORMATION ADMINISTRATION
THROUGH THE NATIONAL BUREAU OF STANDARDS

To appear in
J. of Forecasting

Using the Bootstrap to Evaluate Forecasting Equations[†]

Stephen C. Peters
M.I.T. Center for Computational Research
in Economics and Management Science
Cambridge, Massachusetts 02139
617-253-8416

David A. Freedman
Statistics Department
University of California, Berkeley
California 94720
415-642-4272

Abstract

The bootstrap, like the jack-knife, is a technique for estimating standard errors. The idea is to use Monte-Carlo simulation, based on a non-parametric estimate of the underlying error distribution. The bootstrap will be applied to an econometric equation describing the demand for energy by industry, to determine multi-period forecasting error, and choose among competing specifications. The delta-method for estimating forecast errors turns out to be too optimistic by a factor of two.

Keywords and phrases: /forecasting errors/confidence interval-predictions: simultaneous system/least squares/regression-seemingly unrelated regressions/econometric models/estimation-generalized regression-model choice/evaluation-theory of/

Running head: Evaluating a forecasting equation.

Received March 1983
Accepted August 1984

[†] We would like to thank Persi Diaconis, M. L. Eaton, Bradley Efron, Douglas Hale, John Herbert, Edwin Kuh, George Lady, Thomas Permutt, Thomas Rothenberg, and H. Theil for useful discussions and other support. Authors usually find referees' reports quite annoying, but our referee made a lot of very good suggestions. Computer facilities were provided by the Center for Computational Research in Economics and Management Science at M.I.T. The project was financed by the Office of Statistical Standards, Energy Information Administration, Department of Energy, through the National Bureau of Standards.

After graduating from MIT with a B.Sc. in computer science, Stephen Peters got his Ph.D. at Stanford in 1983; his thesis was on the bootstrap in econometric models. He is now a research associate at the Center for Computational Research in Economics and Management Science, Sloan School, MIT. His main interest is statistical computing.

David Freedman took a B.Sc. in mathematics at McGill, and a Ph.D. in probability at Princeton, graduating in 1960. His thesis was on de Finetti's theorem. He is now a professor of statistics at the University of California, Berkeley. His interests include the foundations of statistics, the behavior of standard statistical procedures under nonstandard conditions, and the use of statistical models in litigation.

Table of Contents

§1 Introduction

The bootstrap is a relatively new statistical technique which permits the assessment of variability in an estimate using just the data at hand; see Efron (1979, 1982). The idea is to resample the original observations in a suitable way to construct "pseudo-data" on which the estimator of interest is exercised. More specifically, the theoretical distribution of a disturbance term is approximated by the empirical distribution of a set of residuals. Measures of variability, confidence intervals, and even estimates of bias may then be calculated.

In the regression case, the bootstrap is useful for investigations when mathematical analysis can give only asymptotic results. Within the scope of the bootstrap are non-normal errors, lag structures, and generalized least squares with estimated covariance matrices. A previous paper (Freedman and Peters, 1984a) compared the performance of conventional asymptotic estimates of standard error of coefficients to the performance of a bootstrap procedure in the setting of a single econometric equation: also see Daggett and Freedman (1984). This paper will indicate how to use the bootstrap to develop standard errors for multi-period forecasts, and to choose among competing equations. The main finding is that in a real example, the delta-method for estimating forecast errors was too optimistic by a factor of two.

The balance of this paper is organized as follows. Section 2 gives a brief review of the bootstrap idea, in the context of linear econometric models. Section 3 gives an even briefer review of generalized least squares. The bootstrap is used to attach standard errors to multi-period forecasts in Section 4, and to choose among competing equations in Section 5. Section 4 also compares the bootstrap to the method of Schmidt (1974, 1977). Conclusions will be drawn in Section 6.

The approach may be distinguished from the classical work of Brown (1954), or Goldberger, Nagar, and Odeh (1961): the bootstrap uses simulation rather than asymptotics based on Taylor series, and applies to multi-period forecasts. The work of Fair (1979, 1980) is closer in spirit to the bootstrap, but somewhat different in detail: Fair assumes that the disturbance terms follow a multivariate normal distribution, and that the parameter estimates follow their multivariate normal limiting distribution. The bootstrap is distribution-free, and develops the appropriate finite-sample behavior for the estimates.

§2 The bootstrap

The bootstrap is described by Efron (1979, 1982). Related papers, which provide a theoretical basis for using the bootstrap are Bickel and Freedman (1981, 1983), Freedman (1981). The bootstrap is a procedure for estimating standard errors by re-sampling the data in a suitable way. First, an informal overview of the idea. In brief, the model has been fitted to data, by some statistical procedure; and there are residuals, namely the difference between observed and fitted values. Some stochastic structure was imposed on the stochastic disturbance terms, explicitly or implicitly, in the fitting. The key idea is to resample the residuals, preserving this stochastic structure, so the model is tested against its own assumptions.

Assuming the model and the estimated parameters to be right, the resampling generates "pseudo-data," both for the past and for the future. Now the model can be re-fitted to the pseudo-data for the past, and used to "forecast" the pseudo-data for the future. In this artificial world, the errors of forecast are directly observable. The Monte-Carlo distribution of such errors can be used to approximate the distribution of the unobservable errors in the real forecasts. This approximation is the bootstrap.

A variation on this idea can be used to compare two models, testing each one against the assumptions of the other. This involves generating the pseudo-data with one model; the second model is re-fitted to the pseudo-past, and used to predict the pseudo-future. Then the roles of the two models are interchanged. In this way, the bootstrap can sometimes be used to select forecasting equations which are relatively insensitive to specification error. An example will be presented in Section 5.

A more explicit, but still informal, description is as follows. Consider a dynamic linear model, of the form

$$\underset{1\times q}{Y_t} = \underset{1\times q}{Y_{t-1}} \underset{q\times q}{B} + \underset{1\times p}{X_t} \underset{p\times q}{C} + \underset{1\times q}{\epsilon_t} \tag{1}$$

In this equation, B and C are coefficient matrices of unknown parameters, to be estimated from the data, subject to identifying restrictions; Y_t is the vector of "endogenous" variables at time t; X_t is the vector of "exogenous" variables at time t; and ϵ_t is the vector of disturbances at time t. The endogenous variables are determined within the model, the exogenous variables by some external process. Technically, endogenous variables may be correlated with ϵ, exogenous variables are not correlated with ϵ. The following standard condition is imposed on the error distribution: given the $X's$, the $\epsilon's$ are independent and identically distributed with mean 0.

Data are available for $t = 1, \ldots, n$ and Y_0 is available too. The coefficient matrices are estimated as $\hat{B}$ and $\hat{C}$ by some well-defined statistical procedure, like generalized least squares (Sections 3 and 4). The $X's$ are forecast for time $t = n+1, \ldots, n+m$, by some procedure external to the model. The $Y's$ are forecast from the model for this time interval by substituting $\hat{B}$ and $\hat{C}$ for B and C in (1), setting $\epsilon_{n+1}, \ldots, \epsilon_{n+m}$ equal to 0, and solving iteratively. Call these forecasts $\hat{Y}_{n+1}, \ldots, \hat{Y}_{n+m}$. (This forecasting procedure is standard, but not optimal.)

In algebraic terms, $\widehat{Y}_n = Y_n$, and for $n \le t < n+m$,

$$\begin{aligned}\widehat{Y}_{t+1} &= \widehat{Y}_t\widehat{B} + \widehat{X}_{t+1}\widehat{C}\\ &= \widehat{Y}_{t-1}\widehat{B}^2 + \widehat{X}_t\widehat{C}\widehat{B} + \widehat{X}_{t+1}\widehat{C}\\ &\;\vdots\\ &= Y_n\widehat{B}^{t+1-n} + \widehat{X}_{n+1}\widehat{C}\widehat{B}^{t-n} + \ldots + \widehat{X}_t\widehat{C}\widehat{B} + \widehat{X}_{t+1}\widehat{C}\end{aligned}$$

Here, $\widehat{X}_t$ denotes the forecast for X_t at time $t > n$. Likewise,

$$\begin{aligned}Y_{t+1} = {} & Y_nB^{t+1-n} + X_{n+1}CB^{t-n} + \ldots + X_tCB + X_{t+1}C\\ & + \epsilon_{n+1}B^{t-n} + \ldots + \epsilon_tB + \epsilon_{t+1}\end{aligned}$$

The forecast error $\widehat{Y}_{n+m} - Y_{n+m}$ is due to:

- The difference between $\widehat{B}$, $\widehat{C}$ and B, C.
- The term involving $\epsilon_{n+1}, \ldots, \epsilon_{n+m}$, which is dropped in making the forecast.
- The difference between $\widehat{X}_t$ and X_t.

For an analytic treatment of the second component in the linear case, see Findley (1984).

Coming now to the bootstrap, when $\widehat{B}$ and $\widehat{C}$ are computed, residuals are defined:

$$\hat{\epsilon}_t = Y_t - Y_{t-1}\widehat{B} - X_t\widehat{C} \tag{2}$$

These are estimates for the true disturbances ϵ_t in the model (1). Let μ be the empirical distribution of the residuals, assigning mass $1/n$ to each of $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$. To avoid trivial complications, assume the equations have intercepts, so the residuals have mean zero.

Consider next a model like (1), but where all the ingredients are known:

- Set the coefficients at $\widehat{B}$ and $\widehat{C}$ respectively.
- Make the disturbance terms independent, with common distribution μ.

The exogenous $X's$, past and future both, are kept as before, as is Y_0. Using this simulation model, pseudo-data can be generated for the past, namely periods $t = 1, \ldots, n$. These pseudo-data will be denoted by stars: $Y_0^*, \ldots, Y_n^*$. Likewise, the pseudo-future can be generated: $Y_{n+1}^*, \ldots, Y_{n+m}^*$. The construction is iterative: $Y_0^* = Y_0$, and for all $t = 1, \ldots, n+m$,

$$Y_t^* = Y_{t-1}^*\widehat{B} + X_t\widehat{C} + \epsilon_t^* \tag{3}$$

the ϵ^*'s being independent with the common distribution μ

Now pretend the pseudo-data come from a model like (1), with unknown coefficient matrices. Using the previous estimation procedures, estimate these coefficients from the pseudo-data; denote the estimates by $\widehat{B}^*$ and $\widehat{C}^*$. Likewise, use the previous forecasting procedure to generate a forecast for period $n+m$; denote this by $\widehat{Y}_{n+m}^*$. The distribution of the pseudo-errors $\widehat{Y}_{n+m}^* - Y_{n+m}^*$ can be computed, and used to approximate the distribution of the real errors $\widehat{Y}_{n+m} - Y_{n+m}$ This approximation is the bootstrap. It is emphasized that the calculation assumes the validity of the model (1), and accurate forecasts of the exogenous variables. The distribution of the pseudo-errors can be computed, *e.g.*, by Monte Carlo, simply repeating the procedure some number of times and seeing what happens.

The procedure can be modified to take into account stochastic errors in forecasting the $X's$ as well as measurement errors in past $X's$. In effect, on each run, the $X's$ can be perturbed to mimic any assumed error structure. That is, X_t for $t > n$ can be replaced by a forecast value $\widehat{X}_t$ involving a stochastic error; likewise, X_t for $t \leq n$ can be replaced by $\widetilde{X}_t$, which differs from X_t by measurement error. The distribution of these errors must be specified, and this is a very delicate thing to do. In this paper, the X_t will be held fixed. For an application of the idea, however, see Finke, Flood, and Theil (1984).

Turn now to model comparisons. Suppose there are two competitive models for a certain data series. Suppose each fits reasonably well. Suppose that each, if an adequate representation of reality, is likely to forecast reasonably well. How to choose between the two models? This can be investigated by a variation on the foregoing:

- Use model #1 to generate the pseudo-data and pseudo-future; but fit and forecast with model #2; compute the pseudo-error of forecast $\widehat{Y}^*_{n+m} - Y^*_{n+m}$.
- Now interchange the two models.

If, *e.g.*, model #1 forecasts well both on its assumptions and on those of model #2, while model #2 fails in the world of model #1, then model #1 is more robust and is to be preferred on this ground. A case of interest is where model #1 is a simplification of model #2, and there is a trade-off between precision and bias.

Another procedure, more in the spirit of Efron (1983), involves creating a "neutral" model, as a linear combination of the two competing models. This can be fitted to the original data, and used to generate pseudo-data for both past and future. In this neutral simulation world, the forecasting performance of each competing model can be studied, by fitting it to the pseudo-past and using it to forecast the pseudo-future. We do not pursue this idea here. There is still another approach described in Cox (1982), using likelihood ratio tests.

§3 Generalized least squares

Consider the model

$$Y = X\beta + \epsilon, \quad E(\epsilon) = 0, \quad \mathrm{cov}(\epsilon) = \Sigma \tag{4}$$

With Σ known, the generalized least squares (*gls*) estimate is

$$\hat{\beta}_{gls} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \tag{5}$$

As usual,

$$E(\hat{\beta}_{gls}) = \beta \tag{6}$$

$$\mathrm{cov}(\hat{\beta}_{gls}) = (X^T\Sigma^{-1}X)^{-1} \tag{7}$$

When Σ is unknown, statisticians routinely use (5) and (7) with Σ replaced by some estimate $\hat{\Sigma}$. Iterative procedures are often used, as follows. Let $\hat{\beta}^{(0)}$ be some initial estimate for β, typically from a preliminary ordinary least squares (*ols*) fit. There are residuals $\hat{\epsilon}^{(0)} = Y - X\hat{\beta}^{(0)}$. Suppose the procedure has been defined through stage k, with residuals

$$\hat{\epsilon}^{(k)} = Y - X\hat{\beta}_{gls}^{(k)}$$

Let $\hat{\Sigma}_k$ be an estimator for Σ, based on $\hat{\epsilon}^{(k)}$. Then

$$\hat{\beta}_{gls}^{(k+1)} = (X^T\hat{\Sigma}_k^{-1}X)^{-1}X^T\hat{\Sigma}_k^{-1}Y \tag{8}$$

This procedure can be continued for a fixed number of steps, or until $\hat{\beta}_{gls}^{(k)}$ settles down. To ease notation, $\hat{\beta}_{gls}^{(k)}$ will be refered to as the (gls,k)-estimator. This paper only considers the (gls,1) estimator.

§4 Bootstrapping RDFOR

The object of this section is to illustrate the bootstrap procedure for determining forecast errors in a real example, the Regional Demand Forecasting Model (RDFOR). This is a system of econometric equations designed to forecast demand through 1995, for various fuel types, by consumption sector and geographical region, as a function of prices and other exogenous variables. The focus here is on that part of the model concerned with the industrial sector demand for fuel. For more detailed discussions of RDFOR, see Freedman, Rothenberg, and Sutch (1983), and Kuh *et al* (1982).

The model distinguishes ten geographical regions, indexed here by r. The equation for total demand by the industrial sector in geographical region $r = 1, \ldots, 10$ and year $t = 1961, \ldots, 1978$ is taken as

$$q_{rt} = a_r + b\, c_{rt} + c\, h_{rt} + d\, p_{rt} + e\, q_{r,t-1} + f\, v_{rt} + \epsilon_{rt} \tag{9}$$

where in region r and year t: q_{rt} is the log of an index of fuel consumption, c_{rt} is the log of cooling degree days, h_{rt} is the log of heating degree days, p_{rt} is the log of a fuel price index, v_{rt} is the log of value added in manufacturing, ϵ_{rt} is a stochastic disturbance term, and a_r, b, c, d, e, and f are parameters to be estimated. This particular equation is the one reported by Kuh *et al* (1982). Notice that the coefficients b, c, d, e, and f are constant across regions; however, the intercepts a_r are region-specific. The equation is dynamic in the sense that the lagged endogenous variable $q_{r,t-1}$ appears on the right hand side.

The assumptions on the stochastic disturbance terms ϵ_{rt} are as follows:

$E(\epsilon_{rt}) = 0$ for all r and t. (10a)

The ϵ_{rt} are stochastically independent of the c_{rt}, h_{rt}, p_{rt}, and v_{rt}. (10b)

The vectors $\epsilon_t = (\epsilon_{1,t}, \ldots, \epsilon_{10,t})$ are independent and identically distributed in time. (10c)

This model is outside the framework of standard regression theory because of the dynamics: q_{rt} is correlated with $\epsilon_{r,t-1}$. It is outside the framework of standard multivariate theory because the coefficients are constrained to equality across regions. However, equation (9) does fit into the framework (1) with $q = 10$ and $p = 5 \times 10 + 1 = 51$; the coefficient matrices are subject to numerous constraints.

Historical data for estimating this regression relation were taken from the SEDS (State Energy Data System) data base. SEDS was previously called FEDS. This data base is reviewed in Freedman, Rothenberg, and Sutch (1983). This data base contains the annual data required for the period 1960 through 1978. The fitting period, however, runs from 1961 to 1978: a year of data is lost due to the lag term.

The main object of this section is to indicate how the bootstrap can be used to develop standard errors for multi-period forecasts. The idea is that in the simulation world of the bootstrap, we can observe both the actual value and the predicted value in the forecast year, say 1995; hence we can observe the difference, which is the forecast error. This idea will be illustrated on equation (9), with a historical period of 1960–78, and a forecast period of 1979–95. In the equation, the variables c, h, p, and v are exogenous. They are forecast over the period 1979–95 by some procedure external to the equation. (Value-added was forecast by the US Department of Energy; heating and cooling degree days were set at the historical average values; regional prices were forecast to increase at 5% per year. The forecast values will be denoted by hats. The values of the exogenous variables, for the forecast period and the historical period, will be held fixed in all the simulations described below. Only the forecast year 1995 will be considered.)

The equation (9) is used to forecast the endogenous variable q_{rt} as follows:

- Estimate the coefficients using the historical data from 1960 to 1978.
- Set the disturbances ϵ_{rt} to their expected value of zero over the forecast period, and solve the equation iteratively: $\hat{q}_{r,1978} = q_{r,1978}$ and for $t = 1979, \ldots, 1995$

$$\hat{q}_{rt} = \hat{a}_r + \hat{b}\,\hat{c}_{rt} + \hat{c}\,\hat{h}_{rt} + \hat{d}\,\hat{p}_{rt} + \hat{e}\,\hat{q}_{r,t-1} + \hat{f}\,\hat{v}_{rt} \qquad (11)$$

To get started on the bootstrap, estimate the parameters in the model (9) using the data from 1960 to 1978, obtaining the $(gls,1)$ estimates $\hat{a}_r$, $\hat{b}$, $\hat{c}$, $\hat{d}$, $\hat{e}$, and $\hat{f}$, and calculate the residuals $\hat{\epsilon}_{rt}$. Let $\hat{\epsilon}_t$ be the 10-vector $(\hat{\epsilon}_{1,t}, \ldots, \hat{\epsilon}_{10,t})$ of residuals, and μ the empirical distribution of $\{\hat{\epsilon}_t : t = 1961, \ldots, 1978\}$. Make thirty-five independent draws ϵ^*_t for $t = 1961, \ldots, 1995$ from μ. Let ϵ^*_{rt} denote the r^{th} component of ϵ^*_t. Construct a starred data set with the resampled residuals: $q^*_{r,1960} = q_{r,1960}$ and for $t = 1961, \ldots, 1995$

$$q^*_{rt} = \hat{a}_r + \hat{b}\,c_{rt} + \hat{c}\,h_{rt} + \hat{d}\,p_{rt} + \hat{e}\,q^*_{r,t-1} + \hat{f}\,v_{rt} + \epsilon^*_{rt}$$

For $t = 1961, \ldots, 1978$, the q^*_{rt} are simulated historical data. For $t = 1979, \ldots, 1995$, the q^*_{rt} are simulated future "actuals." Now make the forecasts by the standard procedure, but using the simulated historical data instead of the real data. In particular, the parameters $\hat{a}^*_r$, $\hat{b}^*$, and so forth are re-estimated from the starred historical data by the $(gls,1)$ regression of q^*_{rt} on c_{rt}, h_{rt}, p_{rt}, $q^*_{r,t-1}$, and v_{rt}. The forecasts are made iteratively as in (11): $\hat{q}^*_{r,1978} = q^*_{r,1978}$ and for $t = 1979, \ldots, 1995$

$$\hat{q}^*_{rt} = \hat{a}^*_r + \hat{b}^*\,c_{rt} + \hat{c}^*\,h_{rt} + \hat{d}^*\,p_{rt} + \hat{e}^*\,\hat{q}^*_{r,t-1} + \hat{f}^*\,v_{rt}$$

The result is a set of simulated actuals q^*_{rt} and forecasts $\hat{q}^*_{rt}$ for $t = 1979, \ldots, 1995$. Note that q^*_{rt} $\hat{q}^*_{rt}$ because $\hat{\beta}$ $\hat{\beta}^*$ and because q^*_{rt} incorporates ϵ^*_{rt} while $\hat{q}^*_{rt}$ does not. The difference between the future q^*_{rt} and the forecast $\hat{q}^*_{rt}$ is the forecast error. This procedure can be repeated, getting new starred disturbances on each repetition, to develop the distributions of

- the simulated actual demand $q^*_{r,1995}$
- the simulated forecast $\hat{q}^*_{r,1995}$
- the simulated forecast error $q^*_{r,1995} - \hat{q}^*_{r,1995}$.

Table 1 summarizes 100 replications of the bootstrap forecasting experiment just described. Coefficients were estimated by $(gls,1)$. Column 1 of the table displays the sample mean of the 100 simulated actuals; column 2, the sample mean of the 100 simulated forecasts; and column 3, the standard deviation of the 100 simulated forecast errors. Column 2 is very close to column 1, indicating negligible bias in the forecasting procedure. The standard deviations in column 3 are the bootstrap measures of random error in the forecasts, one for each region. They are fairly high, compared with the mean values. The forecasts are subject to large random error.

Table 1. Bootstrap forecast experiment for equation (9). Estimation is by one-step *gls*. There are 100 bootstrap replications.

	(1)	(2)	(3)	(4)	(5)	(6)
	Sample Mean	Sample Mean	Standard Deviation		RMS	RMS
Region	Actuals $q^*_{r.1995}$	Forecasts $\hat{q}^*_{r.1995}$	$q^*_{r.1995} - \hat{q}^*_{r.1995}$	Delta SE	Delta SE	Bootstrap SE
1	.40	.41	.12	.063	.053	.094
2	.12	.14	.14	.050	.055	.11
3	.31	.32	.12	.045	.055	.090
4	.77	.78	.11	.068	.055	.093
5	.44	.44	.11	.043	.053	.079
6	.96	.97	.14	.073	.059	.10
7	.69	.69	.12	.065	.054	.095
8	.94	.94	.11	.084	.058	.091
9	.55	.56	.13	.098	.057	.097
10	.63	.63	.11	.061	.051	.087

Schmidt (1974) uses the delta-method to attach a standard error to a multi-period forecast. Applied to the original RDFOR data, his method gives the results shown in column 4. As can be seen, column 4 is much smaller than column 3, so the delta-method and the bootstrap give different results. Which is better?

Column 5 of Table 1 reports the results of a simulation experiment showing that the delta-method is seriously biased downward. To do the bootstrap, we have set up a fully-defined simulation world, where the parameters and the distribution of the disturbances are all known. In this world, the size of the random error in the forecasts was determined empirically, and reported in column 3.

The experiment involved generating 100 starred data sets, which we index by $i = 1, \ldots, 100$. For the i^{th} such data set, we can use Schmidt's formula to estimate the standard error of forecast. For example, in region 1, let SE_i be that estimate. Then

$$\sqrt{\frac{1}{100}\sum_{i=1}^{100} \text{SE}_i^2} \approx .053$$

is reported in column 5 of Table 1. This is smaller than the SD in column 3, by a factor of about two. And likewise for the other regions.

The comparison between columns 3 and 5 is fair, being made within the same coherent simulation model. In RDFOR, the delta-method substantially underestimated the random error in the forecasts, and of course misses any bias which may be present.

Column 6 presents the results from a similar test of the bootstrap standard errors. The bootstrap does better than the delta-method, but is still biased downward. The details for column 6 may be a bit complicated, but the idea is straightforward. We check the bootstrap by trying it out in our simulation world, where we know the answer. Column 3 in Table 1 shows the "real" size of the random error in the forecasts. Column 5 shows the size of these errors as estimated by the delta-method, in a typical starred data set. Likewise, column 6 shows the size as estimated by the bootstrap.

The experiment involves a nested iteration: at the "outer loop" starred data-sets are built up one after another in the way described earlier, and presented to an "inner loop" bootstrap for an estimate of the standard error of forecast. The outer loop quantities q_{rt}^*, a_r^*, ... and so forth are defined as before. Let $\hat{\epsilon}_{rt}^* = q_{rt}^* - \hat{q}_{rt}^*$ be the residuals. Let $\hat{\epsilon}_t^*$ be the 10-vector $(\hat{\epsilon}_{1,t}^*, \ldots, \hat{\epsilon}_{10,t}^*)$ of residuals for year t. Let μ^* be the empirical distribution of $\{\hat{\epsilon}_t^* : \ t = 1961, \ldots, 1978\}$. So μ^* will change on each pass through the outer loop. On each pass through the inner loop generate ϵ_t^{**} for $t = 1961, \ldots, 1995$ as thirty-five independent draws from μ^*. Construct a doubly starred data set: $q_{r,1960}^{**} = q_{r,1960}$ and for $t = 1961, \ldots, 1995$

$$q_{rt}^{**} = \hat{a}_r^* + \hat{b}^* c_{rt} + \hat{c}^* h_{rt} + \hat{d}^* p_{rt} + \hat{e}^* q_{r,t-1}^{**} + \hat{f}^* v_{rt} + \epsilon_{rt}^{**}$$

Obtain the doubly-starred parameter estimates $\hat{a}_r^{**}, \hat{b}^{**}, \ldots, \hat{f}^{**}$ by the $(gls,1)$ regression of q_{rt}^{**} on c_{rt}, h_{rt}, p_{rt}, $q_{r,t-1}^{**}$ and v_{rt}. Compute the forecasts and the forecast errors in the conventional way. Repeating the "inner loop" gives the bootstrap estimate of the standard error of forecast computed from one starred data set. The "outer loop" may be repeated to develop the distribution of these bootstrap estimates.

Column 6 of Table 1 summarizes an experiment with 100 passes through the outer loop, and at each pass there were 100 passes through the inner loop. Column 6 gives the root mean square of the 100 bootstrap estimates for the standard error of forecast, each such estimate being itself the standard deviation of 100 doubly-starred forecast errors. Consider, for example, region 1. Let i index the outer loop, and j index the inner loop. On pass i through the outer loop and pass j through the inner loop, a doubly-starred forecast error $q_{1,1995}^{**} - \hat{q}_{1,1995}^{**}$ is computed; call this value e_{ij}. On pass i, the bootstrap standard error of forecast is the standard error of the 100 numbers $\{e_{ij} : \ j = 1, \ldots, 100\}$: call this SD_i. Then column 6 of Table 1 reports

$$\sqrt{\frac{1}{100} \sum_{i=1}^{100} \mathrm{SD}_i^2} \approx .094$$

This is the typical standard error of forecast for region 1 estimated by the bootstrap method, in the simulation world. The "real" size of the random error is displayed in column 3 and is .12. Column 6 is uniformly smaller than column 3, indicating downward bias in the bootstrap procedure. But the bootstrap is closer to the mark than the results from Schmidt's delta-method. Indeed the bootstrap is off by 20 to 30 percent; the delta-method, by factors ranging from 2 to 2.5.

This finding does not diminish the interest in Schmidt's formula, which applies for n large. The only conclusion is that some care is needed in using the delta-method on finite samples, and the bootstrap may be more robust. Schmidt (1977) already indicates some need for caution, with a simulation study using artificial data and normal errors. For similar results in a model without dynamics, see Freedman and Peters (1984b).

After seeing a draft of this paper, Professor H. Theil asked how much of the difference between columns 3 and 5 was due to the estimation of the 10 by 10 inter-regional covariance matrix for the $\epsilon's$ in (9), and how much was due to the linearization of $\widehat{Y}$, which is a polynomial of degree 17 in $\widehat{B}$ and $\widehat{C}$. Simulations indicate that the two sources make roughly equal contributions. Of course, if the covariance matrix for the $\epsilon's$ is known, the bootstrap performs very well indeed.

§5 Using the bootstrap to choose an equation

The bootstrap can sometimes be used to choose between two equations. For example, in equation (9), consider relaxing the constraint that the coefficients be constant across regions. The new model is

$$q_{rt} = a_r + b_r c_{rt} + c_r h_{rt} + d_r p_{rt} + e_r q_{r,t-1} + f_r v_{rt} + \epsilon_{rt} \tag{12}$$

Assumption (10) on ϵ_{rt} remains in force. Relaxing the constraints may reduce bias — but increase variance. The bootstrap can be used to assess the trade-off: as it turns out, for the year 1995 at any rate, the reduction in bias is almost exactly offset by the increase in variance. (The method outlined in this section can also be used to compare non-nested models.)

Table 2 summarizes a bootstrap experiment which evaluates the forecasting performance for 1995 of the model (12) on its own assumptions, as in Section 4. In Table 2, the standard deviations of the forecast errors are more variable and on the whole larger than those appearing in Table 1, which gave a similar analysis for equation (9). The regional patterns in Tables 1 and 2 are quite different: the constraints matter.

Table 2. Bootstrap forecast experiment for equation (12). Estimation is by one-step *gls*. There are 100 bootstrap replications.

Region	Sample Mean Actuals $q^*_{r.1995}$	Sample Mean Forecasts $\hat{q}^*_{r.1995}$	Standard Deviation $q^*_{r.1995} - \hat{q}^*_{r.1995}$
1	.71	.71	.13
2	-.076	-.073	.074
3	.53	.52	.048
4	.93	.91	.090
5	.94	.92	.17
6	.83	.85	.23
7	1.24	1.18	.27
8	1.31	1.31	.097
9	.35	.39	.39
10	.94	.91	.16

The equation (12) can also be tested on the assumptions of (9). This involves constructing the simulated historical data and the simulated future actual data according to (9), but re-estimating parameters and forecasting using (12). The results of this bootstrap experiment are displayed in Table 3. The standard deviations of the forecast errors displayed in the last column of Table 3 are quite large. This is because the model (12) requires that sixty parameters be estimated. The variability in each estimated coefficient contributes to the variability of the forecast.

Table 3. Bootstrap model robustness experiment. The generating equation is (9), the forecasting equation is (12). Estimation is by one-step *gls*. There are 100 bootstrap replications.

Region	Sample Mean Actuals $q^*_{r,1995}$	Sample Mean Forecasts $\hat{q}^*_{r,1995}$	Standard Deviation $q^*_{r,1995} - \hat{q}^*_{r,1995}$
1	.40	.45	.24
2	.12	.21	.39
3	.31	.36	.25
4	.77	.79	.19
5	.44	.49	.26
6	.96	.96	.22
7	.69	.69	.37
8	.94	.89	.56
9	.55	.60	.32
10	.63	.60	.28

The roles of (9) and (12) can now be interchanged: construct the simulated historical and the simulated future actual data according to (12), but re-estimate and forecast with (9). Table 4 presents the results of a bootstrap experiment conducted this way. Here, the forecasts significantly undershoot the 1995 actuals. However, the standard deviations are quite small.

Table 4. Bootstrap model robustness experiment. The generating equation is (12), the forecasting equation is (9). Estimation is by one-step *gls*. There are 100 bootstrap replications.

Region	Sample Mean Actuals $q^*_{r,1995}$	Sample Mean Forecasts $\hat{q}^*_{r,1995}$	Standard Deviation $q^*_{r,1995} - \hat{q}^*_{r,1995}$
1	.71	.47	.095
2	−.076	.22	.084
3	.53	.39	.089
4	.93	.82	.10
5	.94	.50	.091
6	.83	1.02	.12
7	1.24	.74	.11
8	1.31	.97	.10
9	.35	.62	.11
10	.94	.67	.095

A direct comparison of the performance of (9) and (12), when each is tested on the assumptions of the other, can be made by calculating the root mean square forecast error. This criterion takes into account both forecast bias and variability. Denote by FE_i the forecast error calculated in the i^{th} bootstrap repetition. The RMS forecast error is

$$\sqrt{\frac{1}{100}\sum_{i=1}^{100}\mathrm{FE}_i^2}$$

Table 5 displays these values. Neither specification appears to hold an advantage. Although model (12) tracks the future with less bias than does model (9), the large variability of the forecast errors from (12) introduced by the many additional parameters nullifies that gain.

Table 5. Comparison of specifications (9) and (12). RMS forecast error.

Region	Generating Equation (9) Forecasting Equation (12)	Generating Equation (12) Forecasting Equation (9)
1	.25	.27
2	.40	.31
3	.25	.16
4	.19	.15
5	.27	.45
6	.22	.22
7	.37	.51
8	.56	.35
9	.32	.30
10	.28	.29

§6 Conclusions

This paper demonstrates the use of the bootstrap to attach standard errors to multi-period forecasts and to select between alternate model specifications in the context of a dynamic energy demand model fitted by generalized least squares. By means of a simulation experiment we have shown that the bootstrap SE's are more reliable than the asymptotics based on the delta-method. This finding stands in agreement with other results we have obtained concerning the quality of asymptotic formulae for SE's.

References

Bickel, P. J. and Freedman, D. A., "Some asymptotic theory for the bootstrap", Annals of Statistics, 9 (1981), 1196-1217.

Bickel, P. J. and Freedman, D. A., "Bootstrapping regression models with many parameters", A Festschrift for Erich Lehmann, P. Bickel, K. Doksum, and J. L. Hodges, editors, Belmont, California: Wadsworth, 1983, 28-48.

Brown, T. M., "Standard errors of forecast of a complete econometric model", Econometrica, 22 (1954), 178-192.

Cox, D. R., "Further results on tests of separate families of hypotheses", JRSS B, 24 (1982), 400-424.

Daggett, R. S. and Freedman, D. A., "Econometrics and the law: a case study in the proof of antitrust damages", Technical report no. 23, Department of Statistics, University of California, Berkeley (1984). To appear in the Proceedings of the Neyman-Kiefer Conference, L. Le Cam, editor. Belmont, California: Wadsworth, 1985.

Efron, B., "Bootstrap methods: another look at the jackknife", Annals of Statistics, 7 (1979), 1-26.

Efron, B., "The jackknife, the bootstrap, and other resampling plans", CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. Society for Industrial and Applied Mathematics, Philadelphia (1982).

Efron, B., "Comparing non-nested linear models", Technical report no. 197, Department of Statistics, Stanford University (1983).

Fair, R., "An analysis of the accuracy of four macro-economic models", Journal of Political Economy, 87 (1979), 701-718.

Fair, R., "Estimating the expected predictive accuracy of econometric models", International Economic Review, 21 (1980), 355-378.

Findley, D. F., "On the use of the bootstrap to obtain estimates of mean square error for multi-step-ahead forecasts of short time series from autoregressive processes", Technical Report, U.S. Census Bureau (1984).

Finke, R., Flood, L. R. and Theil, H., "The budget share of food in 1990: bootstrapping for distribution free prediction intervals", Technical Report, Graduate School of Business, University of Florida, Gainesville (1984).

Freedman, D. A., "Bootstrapping regression models", Annals of Statistics, 9 (1981), 1218-1228.

Freedman, D. A., "On bootstrapping two-stage least-squares estimates in stationary linear models", Annals of Statistics, 12 (1984), 827-842.

Freedman, D. A. and Peters, S. C., "Bootstrapping a regression equation: some empirical results", JASA, 79 (1984a), 97-106.

Freedman, D. A. and Peters, S. C., "Bootstrapping an econometric model: some empirical results", Journal of Business and Economic Statistics , 2 (1984b), 150-158.

Freedman, D. A., Rothenberg, T. and Sutch, R., "On energy policy models", Journal of Business and Economic Statistics, 1 (1983), 24-36.

Goldberger, A., Nagar, A. L. and Odeh, H. S., "The covariance matrices of reduced-form coefficients and of forecasts for structural econometric models", Econometrica, 29 (1961), 556-573.

Kuh, E., Lahiri, S., Minkoff, A., Swartz, S. and Welsch, R., "Analysis of the validity of the coefficient estimates and forecasting properties of the RDFOR models--a summary report", Technical report, MIT Center for Computational Research in Economics and Management Science (1982).

Schmidt, P., "The asymptotic distribution of forecasts in the dynamic simulation of an econometric model", Econometrica, 42 (1974), 303-309.

Schmidt, P., "Some small sample evidence on the distribution of dynamic simulation forecasts", Econometrica, 45 (1977), 997-1005.