THE LITTLE BOOTSTRAP AND OTHER METHODS FOR DIMENSIONALITY SELECTION IN REGRESSION: X-FIXED PREDICTION ERROR

By

Leo Breiman

Research supported by National Science Foundation Grant DMS87-18362

Technical Report No. 169
August 1988
(revised August 1991)

University of California
Department of Statistics
Berkeley, California

The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error.

Leo Breiman*
Department of Statistics
University of California
Berkeley, California 94720

Abstract

When a regression problem contains many predictor variables, it is rarely wise to try and fit the data by means of a least squares regression on all of the predictor variables. Usually, a regression equation based on a few variables will be more accurate and certainly simpler. There are a variety of methods for picking "good" subsets of variables and programs that do such procedures are part of every widely used statistical package. The most common methods are based on stepwise addition or deletion of variables, and "best subsets". The latter refers to a search method that given the number of variables to be in the equation, say five, locates that regression equation based on five variables that has the lowest residual sum-of-squares among all five variable equations.

All of these procedures generate a sequence of regression equations, the first one based on one variable, the next on two variables, etc. Each member of this sequence is called a submodel and the number of variables in the equation is the dimensionality of the submodel. A complex problem is which submodel of the generated sequence to select. Statistical packages use a variety of ad hoc selection methods: F-to enter, F-to-delete, C_p, t-value cutoffs, etc. Our approach to this problem is through use of the criterion that a good selection procedure selects dimensionality so as to give low prediction error (PE), where the PE of a regression equation is its expected squared error over the points in the X-design.

Since the true PE is unknown, use of this criteria has to be based on PE estimates. We introduce a method called the little bootstrap which gives almost unbiased estimates for submodel PEs and use these to do submodel selection. Comparison is made to C_p and other methods by analytic examples and simulations. Little bootstrap does well -- C_p and, by implication, all selection methods not based on data reuse, give

---

* Supported by NSF Grant No. DMS-8718362

highly biased results and poor subset selection.

Key Words: variable selection, subset selection, best subsets, Mallows Cp.

The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error

Leo Breiman
Department of Statistics
University of California
Berkeley, California 94720

1. INTRODUCTION

In a regression problem with many predictor variables, data analysts often attempt to reduce the dimensionality of the model by running a procedure such as ‘‘best subsets’’, stepwise forward addition of variables or stepwise backwards deletion. These dimensionality reduction methods are among the most frequently used programs in packages such as SAS, SPSS, and BMDP.

Any one of these procedures produces a sequence of possible regression equations, each of which uses a subset of the predictor variables. Any such regression equation will be called a ‘‘submodel’’, and the dimensionality of a submodel will be the number of predictor variables it uses. The goal is to choose one out of this sequence of submodels as the preferred model.

From a theoretical point of view, submodel dimensionality selection is a trade-off between bias and variance. By decreasing the number of predictor variables in the model, its predictive capabilities will be enhanced because of the decrease in variance involved in parameter estimation. On the other hand, bias will be increased because the ‘‘true model’’ is usually not in the range of the lower dimensional models.

To get optimal prediction functions, we would like to balance the gain in variance against the loss in bias. There is additionally a desire to minimize the complexity of the model by reducing dimensionality. In going, say, from a 40 variable model to a 5 variable model, the apparent structure of the data is considerably simplified. Only the relationship between a few variables needs to be examined (although, in fact, this apparent simplicity can be quite deceptive).

Two major difficulties with these submodel procedures are:

(a) selecting the dimensionality of the submodel to be used.

(b) evaluating the model selected.

By this is meant choosing the dimensionality to get a near optimum balance between bias and variance, and then giving a realistic assessment of the predictive capability of the selected submodel.

In selection of dimensionality, a number of ad hoc methods are commonly used. In stepwise methods, use of F-to-enter, F-to-delete, and adjusted R^2 are prevalent. In "best subsets" the use of the Mallows C_P criterion has become common. Once the subset is selected, then another ad hoc figure of merit is attached to it, often the residual-sum-of-squares, R^2 or adjusted R^2, C_P, etc.

This usage has long been a quiet scandal in the statistical community. It is clear that selecting a sequence of submodels in terms of an optimum or suboptimum fit to the data can produce severe biases in all of the statistical measures used for the classical linear model. In recent years, with recognition of the shortcomings of the commonly used ad hoc methods, use of resampling methods such as bootstrap and cross-validation has been advocated. Their performance in the present context, however, has not been systematically explored.

My interest in this problem is when the data is thin compared to the number of variables--a common situation in many applied problems. For instance, in the simulation presented in Section 5 we go down to 60 cases with 40 variables. This is a land strange to asymptopia.

There is a substantial literature on this and related problems. Excellent reviews are given by Miller (1984, 1990) together with a complete list of references. These works, particularly the 1990 book, point out the biases inherent in the problem and look at the weaknesses of some of the standard procedures for dealing with it.

1.2 Criteria for Dimensionality Selection and Evaluation

We assume data of the form $(y_n, \mathbf{x}_n)$, $n = 1, \ldots, N$ where $\mathbf{x}_n$ is an M-variate vector. Suppose $\mu(\mathbf{x})$ is a prediction function for y in terms of $\mathbf{x}$. We need, at least, a conceptual definition of how good a model $\mu(\mathbf{x})$ is. The definition used in this paper is the x-fixed prediction error PE and the corresponding model error ME.

The x-fixed error measures are computed using the same values of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as in the data. Suppose that the true model is

$$y_n = \mu^*(\mathbf{x}_n) + \varepsilon_n$$

with $\{\varepsilon_n\}$ i.i.d with mean zero, variance σ^2. Once the model has been fitted to the existing data, consider the gedanken experiment of generating new data of the form

$$y_n^{new} = \mu^*(\mathbf{x}_n) + \varepsilon_n^{new}$$

with the $\{\varepsilon_n^{new}\}$ independent of the $\{\varepsilon_n\}$ but having the same distribution.

We use the notation

$$\mathbf{a} = (a_1, \ldots, a_N), \quad \mathbf{b} = (b_1, \ldots, b_N)$$

$$\|\mathbf{a}\|^2 = \sum_n a_n^2, \quad (\mathbf{a}, \mathbf{b}) = \Sigma_n a_n b_n.$$

Taking expectations only over the $\{\varepsilon_n^{new}\}$, define the prediction error by

$$PE = E\|y^{new} - \mu(\mathbf{x})\|^2 = N\sigma^2 + \|\mu - \mu^*\|^2.$$

This leads to the definition of x-fixed model error as

$$ME = \|\mu - \mu^*\|^2.$$

The prediction error is thus a sum of two components--a $N\sigma^2$ error due to the inherent noise level in the regression and the error in fitting the true model. Because there is a little that can be done with the $N\sigma^2$ term, we prefer to work directly with the model error.

The x-random definition of prediction error assumes $\{y_n, \mathbf{x}_n\}$ i.i.d. selected from some underlying distribution $(Y, \mathbf{X})$ and assesses the prediction error in a predictor μ as its expected squared error in predicting y^{new} from $\mu(\mathbf{x}^{new})$ where $(y^{new}, \mathbf{x}^{new})$ is selected from $(Y, \mathbf{X})$ independently of $\{y_n, \mathbf{x}_n\}$. The x-random definition and its consequences are explored in Breiman and Spector [1989].

Whether the x-fixed or x-random definition of PE is used leads to conceptual or methodological differences. For instance, cross-validation tries to estimate the x-random PE which is generally larger than the x-fixed PE (see Efron [1986]). In regression there are two versions of the bootstrap. The one commonly used (the unconditional bootstrap) gives x-random PE estimates. Another version (the conditional bootstrap) was developed for x-fixed estimates and is discussed in Section 2.2 (see Bickel and Freedman [1982]).

The x-fixed ME for the full model has expectation $M\sigma^2$, i.e. a penalty of σ^2 in variance is paid per coefficient estimated. The x-random definition leads to higher ME values, particularly for thin sample sizes and skewed long tailed **x**-distributions. Thus, it is important to distinguish between the two definitions and use appropriate methodology. We note that the x-fixed, x-random terminology was used in an earlier review article by Thompson [1978] where the difference was stressed. See also Copas [1983].

Other definitions of prediction error are possible and often desirable. The x-fixed definition used above assesses predictability only at the points $\{\mathbf{x}_n\}$ in the given X-design. Both referees point out that frequently the desidiratum is accurate prediction at x-points not in the X-design. Examination of the difference between x-fixed and x-random shows that the real distinction is whether the new data points $\{\mathbf{x}^{new}\}$ at which predictions are desired are known and fixed, not whether they are points in the present

X-design. This contrasts with the situation in which the future $\{x^{new}\}$ are random. Thus, a better terminology might be: *future* X *fixed* vs *future* X *random*. The conclusions of this present paper can be generalized to the *future* X *fixed* situation (see Section 4).

1.4 Notation and a more precise problem statement

Denote by ζ any subset of the indices $\{1, \ldots, M\}$; by H^{ζ} the projection matrix of any N-vector into the column space of $\{x_m;\ m \in \zeta\}$; by $\hat{\mu}(\zeta)$ the OLS predictor based on the variables $\{x_m;\ m \in \zeta\}$; and

$$\mathrm{RSS}(\zeta) = \|y - \hat{\mu}(\zeta)\|^2 \quad \mathrm{ME}(\zeta) = \|\hat{\mu}(\zeta) - \mu^*\|^2.$$

We assume that some well-defined procedure (best subsets, stepwise, etc.) has been applied to the data and resulted in a sequence of M + 1 submodels with variables having indices in

$$\zeta_0, \zeta_1, \ldots, \zeta_M \quad (\zeta_0 = \phi)$$

where $|\zeta_J| = J$, ($|\ | =$ cardinality). Associated with each OLS predictor $\hat{\mu}(\zeta_J)$ is the $\mathrm{ME}(\zeta_J)$ value. The sequence $\zeta_0, \ldots, \zeta_M$, the predictors $\hat{\mu}(\zeta_J)$, and the values $\mathrm{ME}(\zeta_J)$ are random, depending stochastically on the $\{\varepsilon_n\}$.

Define the best submodel in the sequence as the one which has the minimum value of $\mathrm{ME}(\zeta_J)$. Because the $\{\mathrm{ME}(\zeta_J)\}$ depend on the unknown μ^*, it is not obvious how to construct a submodel selection procedure that will produce low ME values. Our approach is to construct good estimates $\widehat{\mathrm{ME}}(\zeta_J)$ of $\mathrm{ME}(\zeta_J)$, and select the submodel having minimum $\widehat{\mathrm{ME}}(\zeta_J)$.

The exploration in this paper will be based throughout on the assumption of a classical linear model

$$y_n = \sum_m \beta_m^* x_{mn} + \varepsilon_n, \qquad n = 1, \ldots, N$$

with $\{\varepsilon_n\}$ i.i.d $N(0, \sigma^2)$. That is, the true prediction function is

$$\mu^*(x) = \sum_m \beta_m^* x_m$$

Sub-M will be used to denote full model values; i.e. $\hat{\mu}_M$ is the full model OLS predictor; RSS_M and ME_M the full model residual-sum-of-squares and model error.

1.5 Organization and results

Among methods currently in use or advocated as estimates of ME, the ones having some theoretical justification or rationale are Mallows C_p and the conditional bootstrap.

In Section 2 we look at some properties of the above estimates. While C_p is easy to compute, there is no reason why it should perform well in a dimensionality selection context. (Mallows [1973] points this out, but nevertheless the naive use of C_p persists.) We give examples, both analytic and simulated, to illustrate the potentially severe bias of this approach. It tends to select submodels of too high dimensionality and give ME estimates that are far too low.

We also give a simple example that shows that the conditional bootstrap can have considerable bias and give nonsensical results. In Section 3, we introduce the paradigm of the replicate data set. This procedure provides insight into the structure of the problem and is useful as a benchmark.

In Section 4 we introduce a procedure for estimating the $\{ME(\zeta_J)\}$ that we call the little bootstrap. It has some similarities to the conditional bootstrap, but also some interesting differences. We show that it gives almost unbiased estimates of the $\{ME(\zeta_J)\}$ when the submodels are generated by the commonly used methods of subset selection. This procedure also works in the more general future X fixed case.

Section 5 introduces the concept of rss-extreme. Given the sequence of submodels with indices in $\zeta_0, \ldots, \zeta_M$, a criterion is defined which designates some (usually a small fraction) of these to be rss-extreme.

Section 6 reports on an extensive simulation testing of little bootstrap using backwards variable deletion with forty variables and either 60, 160, or 600 cases with a variety of coefficients. It is compared with the use of C_p and a replicate data set. The results indicate that little bootstrap at the original sample size is almost competitive with the replicate data set method using double the sample size in evaluation, but is not quite as good in dimensionality selection. It also shows that, in selection, there is a gain in accuracy by restricting selection to the rss-extreme submodels.

Section 7 revisits the bias vs variance trade off in minimizing model error and gives some simulation results. Section 8 discusses what information is available after dimensionality selection, and section 9 gives brief conclusions.

2. C_p AND CONDITIONAL BOOTSTRAP DO NOT ALWAYS WORK

2.1 Mallows C_P

Let $\hat{\mu}(\zeta)$ be the OLS estimator on the subset ζ with $|\zeta| = J$. The C_P criterion is based on the following simple relation:

$$\begin{aligned} RSS(\zeta) &= \|y - \hat{\mu}(\zeta)\|^2 = \|\mu^* + \varepsilon - \hat{\mu}(\zeta)\|^2 \\ &= ME(\zeta) + \|\varepsilon\|^2 + 2(\varepsilon, \mu^* - \hat{\mu}(\zeta)). \end{aligned} \tag{2.1}$$

Now $\hat{\mu}(\zeta) = H^{\zeta}(\mu^* + \varepsilon)$, so the last term in (2.1) can be written as

$$2(\varepsilon, (I - H^{\zeta})\mu^*) \; - \; 2(\varepsilon, H^{\zeta}\varepsilon). \tag{2.2}$$

If $|\zeta| = J$ and the *choice of* ζ *does not depend on the data*, then the expectation of (2.2) is $-2J\sigma^2$. Thus, the C_P estimate of $\text{ME}(\zeta)$ is

$$\hat{\text{ME}}(\zeta) \;=\; \text{RSS}(\zeta) \;+\; (2J - N)\hat{\sigma}^2$$

where $\hat{\sigma}^2$ is estimated in the usual way from the full model.

If ζ depends on the data, this argument fails. To see what can happen, we repeat the example given by Mallows [1973]. Using an orthogonal design, $(\mathbf{x}_m, \mathbf{x}_{m'}) = \delta_{mm'}$, the OLS estimates of β_m^* are $\hat{\beta}_m = \beta_m^* + Z_m$, with $\{Z_m\}$ i.i.d $N(0, \sigma^2)$.

The best subset ζ_J of size J consists of those variables having the J largest values of $|\hat{\beta}_m|$. The subset selected by minimum C_P consists of all variables $\mathbf{x}_m$ such that $\hat{\beta}_m^2 \geq 2\hat{\sigma}^2$. For this subset ζ the C_P ME estimate is

$$\sum_m (\hat{\beta}_m^2 - \hat{\sigma}^2) \;-\; \sum_{m \in \zeta} (\hat{\beta}_m^2 - 2\hat{\sigma}^2). \tag{2.3}$$

Suppose all $\beta_m^* = 0$. Then the expectation of the first term in (2.3) is zero, while the second term is always negative. Assuming $\hat{\sigma}^2 = \sigma^2$, the expected value of the C_P ME estimate is $-.26M\sigma^2$ while the expected ME value is $.58M\sigma^2$. Furthermore, $E(|\zeta|) = .16M$.

2.2 The Conditional Bootstrap

As above, let $\hat{\mu}(\zeta)$ be based on ζ and let ζ_J be the best subset of dimension J, i.e.

$$\text{RSS}(\zeta_J) \;=\; \min\{\text{RSS}(\zeta);\; |\zeta| = J\}.$$

Then consider trying to estimate

$$E(\text{ME}(\zeta_J)) \;\equiv\; \phi_J(\beta_1^*, \ldots, \beta_M^*, \sigma).$$

With β^*, σ^2 unknown, one is tempted to compute the maximum likelihood estimate $\phi_J(\hat{\beta}_1, \ldots, \hat{\beta}_M, \hat{\sigma}^2)$. This latter is essentially what the conditional bootstrap does. Proceed as follows:

i) Fit the full model, getting

$$\hat{\beta}_1, \ldots, \hat{\beta}_M, \;\hat{\sigma}^2, \;\hat{\mu}_M(\mathbf{x}).$$

ii) Generate $\{\tilde{\varepsilon}_n\}$ i.i.d. $N(0, \hat{\sigma}^2)$ to get data

$$\tilde{y} \;=\; \hat{\mu}_M(\mathbf{x}) \;+\; \tilde{\varepsilon}.$$

iii) Using the $(\tilde{y}, \mathbf{x})$ data, find the best subset $\tilde{\zeta}_J$ of dimension J, and OLS predictor $\tilde{\mu}(\tilde{\zeta}_J)$.

iv) Estimate $ME(\zeta_J)$ by

$$\|\hat{\mu}_M - \tilde{\mu}(\tilde{\zeta}_J)\|^2.$$

v) Repeat many times and average.

Conditional bootstrap resamples residuals. Instead we have i.i.d. sampled from $N(0, \hat{\sigma}^2)$. With this minor modification, conditional bootstrap is seen as a Monte Carlo method for evaluating $\phi_J(\hat{\beta}, \hat{\sigma}^2)$. For M fixed and N large, conditional bootstrap should have all of the maximum likelihood asymptotic properties. But this is not applicable to the situation where N/M is of modest size.

To examine finite sample behavior, look at the orthogonal model used in section 2. To avoid complications, assume σ^2 known. Recall that $\hat{\beta}_m = \beta_m^* + Z_m$, $\{Z_m\}$ i.i.d. $N(0, \sigma^2)$. The bootstrap data is

$$\tilde{y} = \hat{\mu}_M + \tilde{\varepsilon},$$

and the estimated coefficients of the bootstrap model are

$$\tilde{\beta}_m = \hat{\beta}_m + \tilde{Z}_m, \qquad \tilde{Z}_m \text{ i.i.d. } N(0, \sigma^2)$$

and $\{\tilde{Z}_m\}$ independent of $\{Z_m\}$.

Let ζ_J be the indices of the J largest $|\hat{\beta}_m|$. Then, for the original data, ζ_J is the best subset of size J, and

$$ME(\zeta_J) = \sum_{m \in \zeta_J} Z_m^2 + \sum_{m \notin \zeta_J} \beta_m^{*2}.$$

Take all $\beta_m^* = 0$. Denoting by $R(Z_m)$ the rank of $|Z_m|$ in $|Z_1|, \ldots, |Z_M|$, and letting $I(A)$ be the indicator function of A,

$$ME(\zeta_J) = \sum_m Z_m^2 I(R(Z_m) \le J).$$

Now, letting $R(Z_m + \tilde{Z}_m)$ be the rank of $|Z_m + \tilde{Z}_m|$ among the values of $|Z_1 + \tilde{Z}_1|, \ldots, |Z_M + \tilde{Z}_M|$, the bootstrap estimate of $ME(J)$ is

$$\begin{aligned} \widehat{ME}(\zeta_J) &= \sum_m Z_m^2 I(R(Z_m + \tilde{Z}_m) > J) + \sum_m \tilde{Z}_m^2 I(R(Z_m + \tilde{Z}_m) \le J) \\ &= \sum_m Z_m^2 + \sum_m (\tilde{Z}_m^2 - Z_m^2) I(R(Z_m + \tilde{Z}_m) \le J). \end{aligned}$$

Thus

$$E(ME(\zeta_J)) = \sum_m E(Z_m^2 I(R(Z_m) \le J)),$$

$$E(\widehat{ME}(\zeta_J)) = M\sigma^2.$$

It is simple to verify that $\hat{ME}(\zeta_J)$ is always larger than $ME(\zeta_J)$ for $J < M/2$. Also, it is larger in a way that prevents effective subset selection. $ME(\zeta_J)$ decreases as J decreases and identifies the best subset as the empty one, but $\hat{ME}(\zeta_J)$ has constant expectation for all J.

This is admittedly a quite specialized example. But the Freedman et al. [1988] simulation results, in a less specialized case, also found that the conditional bootstrap has a large upward bias. This does not mean that bootstrapping doesn't work, but only that this method of applying it doesn't work. What does work comes in Section 4.

3. THE REPLICATE DATA SET PARADIGM

Conceptually, one method for doing dimensionality selection is to replicate the data. Use the first data set to do the model fitting and get the sequence of submodels. Then use the second set to get the PE and ME estimates for the submodels.

This procedure is hardly ever used in practice. But it is a useful paradigm for two reasons. First, the resulting analytic structure is fairly simple and can be understood more easily than that resulting from resampling methods. Secondly, it gives a measure against which to judge resampling methods. Resampling methods attempt to make the original data set do double service, first to fit with and then, under resampling, as ME estimators. How well they succeed can be measured against the yardstick of a replicate data set.

Denote the replicate data set by $y' = \mu^* + \varepsilon'$, $\{\varepsilon'\}$ independent of $\{\varepsilon\}$. Then for any submodel ζ, the replicate data set PE estimate is

$$\begin{aligned}\hat{PE}(\zeta) &= \|y' - \hat{\mu}(\zeta)\|^2 \\ &= \|\varepsilon'\|^2 + \|\mu^* - \hat{\mu}(\zeta)\|^2 + 2(\varepsilon', \mu^* - \hat{\mu}(\zeta)),\end{aligned}$$

so that

$$ME(\zeta) = \hat{PE}(\zeta) - \|\varepsilon'\|^2 - 2(\varepsilon', \mu^* - \hat{\mu}(\zeta)). \tag{3.1}$$

The second term has expectation $N\sigma^2$, and the last has zero expectation. But better estimates than $N\sigma^2$ of $\|\varepsilon'\|^2$ are available. Denote the full model PE estimate by $\hat{PE}_M$. Fit a full model to the $\{y'\}$, and denote the residual sum-of-squares by RSS_M' Then the estimate of $\|\varepsilon'\|^2$ given by

$$\frac{RSS_M' + \hat{PE}_M}{2}$$

has expected squared error of $2M\sigma^4$ (expectation over both $\{\varepsilon\}$, $\{\varepsilon'\}$) as compared to $2N\sigma^4$ using $N\sigma^2$ as the estimate. Thus, we use the ME estimate;

$$\hat{ME}(\zeta) = \hat{PE}(\zeta) - \frac{1}{2}(RSS_M' + \hat{PE}_M).$$

Going back to (3.1), note that

$$\mathrm{ME}_M \; - \; \mathrm{ME}(\zeta_J) \; = \; \widehat{\mathrm{PE}}_M - \widehat{\mathrm{PE}}(\zeta_J) + 2(\varepsilon', \hat{\mu}_M - \hat{\mu}(\zeta_J))$$

or

$$[\mathrm{ME}_M - \mathrm{ME}(\zeta_J)] \; - \; [\widehat{\mathrm{ME}}_M - \widehat{\mathrm{ME}}(\zeta_J)] \; = \; 2(\varepsilon', \hat{\mu}_M - \hat{\mu}(\zeta_J)). \qquad (3.2)$$

The term on the right has mean zero. Its variance, conditioned on $\{\varepsilon\}$, is $4\sigma^2 \, \|\hat{\mu}_M - \hat{\mu}(\zeta_J)\|^2$. To the extent that this term stays small, $\widehat{\mathrm{ME}}(\zeta_J)$ will track the changes in $\mathrm{ME}(\zeta_J)$ and give accurate estimates of the minimum ME submodel.

4. THE LITTLE BOOTSTRAP

In most practical situations, we have only one data set and no replication. What can be done? To temporarily simplify notation, let $\hat{\mu} = \hat{\mu}(\zeta_J)$, and start with

$$\begin{aligned} \mathrm{RSS}(\zeta_J) &= \|\mu^* + \varepsilon - \hat{\mu}\|^2 \\ &= \|\varepsilon\|^2 + \|\mu^* - \hat{\mu}\|^2 + 2(\varepsilon, \mu^* - \hat{\mu}). \end{aligned}$$

Therefore

$$\mathrm{ME}(\zeta_J) \; = \; \mathrm{RSS}(\zeta_J) - \mathrm{RSS}_M + \|\mu^* - \hat{\mu}_M\|^2 - 2(\varepsilon, \hat{\mu}_M - \hat{\mu}). \qquad (4.1)$$

The term $\|\mu^* - \hat{\mu}_M\|^2 = (\varepsilon, H\varepsilon)$ can be estimated by $M\hat{\sigma}^2$. The critical issue is estimating the last term. The C_p approximation is:

$$\begin{aligned} 2(\varepsilon, \hat{\mu}_M - \hat{\mu}) &= 2(\varepsilon, (H - H^{\zeta})(\mu^* + \varepsilon)) \\ &= 2(\varepsilon, (H - H^{\zeta})\mu^*) + 2(\varepsilon, (H - H^{\zeta})\varepsilon) \\ &\cong 2\hat{\sigma}^2 (M - J). \end{aligned}$$

As pointed out before, this cannot be accurate if ζ is data selected. The little bootstrap procedure, as given below, uses the data to compute a variable $B(\zeta)$ such that

$$E(B(\zeta_J)) \; \cong \; E(\varepsilon, \hat{\mu}_M \; - \; \hat{\mu})$$

when the sequence $\{\zeta_J\}$ is data selected using any of the common selection methods. Then the little bootstrap $\mathrm{ME}(\zeta_J)$ estimate is taken as

$$\widehat{\mathrm{ME}}(\zeta_J) \; = \; \mathrm{RSS}(\zeta_J) \; - \; \mathrm{RSS}_M \; + \; M\hat{\sigma}^2 \; - \; 2B(\zeta_J).$$

Note that there is no guarantee that $B(\zeta_J) \cong (\varepsilon, \hat{\mu}_M - \hat{\mu})$, but only that their expectations over $\{\varepsilon\}$ are nearly equal. The fact that resampling methods, in general, can at best recover only expectations of error rate corrections has been emphasized by Efron [1986] and Gong [1986].

To begin with we define the relevant class of submodel selection procedures. Denote the data by $\{y_n, \mathbf{x}_n\}$;

Definition 4.1. *The subset selection method is scale invariant if for each* J, $0 \le J \le M$, *the function* $f_J(\{y_n, x_n\})$ *such that*

$$\zeta_J = f_J(\{y_n, x_n\})$$

$$f_J(\{cy_n, cx_n\}) = f_J(\{y_n, x_n\})$$

satisfies for any constant $c \neq 0$.

All commonly used data dependent methods of submodel selection are scale invariant. For instance, in best subsets ζ_J is the minimizer of $RSS(\zeta)$, $|\zeta| = J$. With $y_n' = cy_n$, $x'_n = cx_n$, $RSS'(\zeta) = c^2 RSS(\zeta)$ and the same ζ_J minimizes $RSS'(\zeta)$, $|\zeta| = J$. It is easily verified that stepwise forward addition of variables and stepwise deletion are also scale invariant.

Use a scale invariant procedure to select the $\{\zeta_J\}$, and denote

$$\theta_J(\beta_1^*, \ldots, \beta_M^*, \sigma^2) = E(\varepsilon, \hat{\mu}_M - \hat{\mu}(\zeta_J)).$$

Assume σ^2 is known and generate data

$$\tilde{y} = y + \varepsilon_1$$

with $\{\varepsilon_1\}$ i.i.d $N(0, t^2\sigma^2)$, $t > 0$, and $\{\varepsilon_1\}$ independent of $\{\varepsilon\}$. Get the subsets $\tilde{\zeta}_J$ of dimension J, $J = 0, \ldots, M$, by applying the same selection procedure to the data $\{\tilde{y}_n, x_n\}$. Denote OLS predictors based on $\{\tilde{y}_n, x_n\}$ by $\tilde{\mu}$. Then

Theorem 4.2.

$$\frac{1}{t^2} E(\varepsilon_1, \tilde{\mu}_0 - \tilde{\mu}(\tilde{\zeta}_J)) = \theta_J\left[\beta_1^* / \sqrt{1 + t^2}, \ldots, \beta_M^* / \sqrt{1 + t^2}, \sigma^2\right].$$

Proof. See the technical appendix.

As a consequence of this theorem, for t small

$$\frac{1}{t^2} E(\varepsilon_1, \tilde{\mu}_M - \tilde{\mu}(\tilde{\zeta}_J)) \cong E(\varepsilon, \hat{\mu}_M - \hat{\mu}(\zeta_J)). \tag{4.3}$$

This result is used to get the little bootstrap ME estimate as follows:

i) Fit the full model getting RSS_M and $\hat{\sigma}^2$. Do variable selection, getting the sequence of subsets of indices $\zeta_0, \zeta_1, \ldots, \zeta_M$, and the values $RSS(\zeta_J)$.

ii) Generate $\{\varepsilon_{1n}\}$, $n = 1, \ldots, N$ as iid $N(0, t^2\hat{\sigma}^2)$ and form the new y-data

$$\tilde{y} = y + \varepsilon_1.$$

iii) Using the data $(\tilde{y}_n, x_n)$ find the subset sequence $\{\tilde{\zeta}_J\}$ using the same procedure as in i), and compute the predictors $\tilde{\mu}_M$ and $\tilde{\mu}(\tilde{\zeta}_J)$ based on the full model and $\tilde{\zeta}_J$

respectively.

iv) Calculate

$$\frac{1}{t^2}(\varepsilon_1, \tilde{\mu}_M - \tilde{\mu}(\tilde{\zeta}_J))$$

v) Repeat ii), iii), iv) a number of times and average the quantities computed in iv). Denote this average by $B_t(J)$.

vi) The little bootstrap estimate is

$$\hat{ME}(\zeta_J) = RSS(\zeta_J) - RSS_M + M\hat{\sigma}^2 - 2B_t(J).$$

The little bootstrap can also give almost unbiased estimates in the more general *future* X *fixed* context. Assume that the new data to be tested on a given linear regression equation $\mu(x)$ is $\{y_{n'}^{new}, x_{n'}^{new}\}$, $n' = 1, \ldots, N'$ where the X^tX matrix for the $\{x_{n'}^{new}\}$ is assumed known, say $V = X^tX$. Then define

$$\begin{aligned} PE &= E\,\|y^{new} - \mu(x^{new})\| \\ &= N'\sigma^2 + E\,\|\mu^*(x^{new}) - \mu(x^{new})\|^2 \\ &= N'\sigma^2 + (\beta - \beta^*)^t V (\beta - \beta^*) \end{aligned}$$

and the second term is defined to be the model error.

Let $\hat{\beta}_M$, $\hat{\beta}_J$ denote the OLS coefficients in $\hat{\mu}_M$ and $\hat{\mu}(\zeta_J)$ respectively. Denote also by A the matrix such that

$$\hat{\beta}_M = A\,y.$$

Then

$$ME_M = (\hat{\beta}_M - \beta^*)^t V (\hat{\beta}_M - \beta^*)$$

and

$$E(ME_M) = \sigma^2 \mathrm{Tr}(A^t V A). \tag{4.4}$$

Now

$$ME(\zeta_J) - ME_M = \hat{\beta}_J^t V \hat{\beta}_J - \hat{\beta}_M V \hat{\beta}_M - 2\beta^{*t} V (\hat{\beta}_J - \hat{\beta}_M).$$

Writing the 3rd term as $(\hat{\beta}_M - A\varepsilon)^t V (\hat{\beta}_J - \hat{\beta}_M)$ gives

$$ME(\zeta_J) = ME_M + (\hat{\beta}_J - \hat{\beta}_M)^t V (\hat{\beta}_J - \hat{\beta}_M) - 2\varepsilon^t A^t V (\hat{\beta}_M - \hat{\beta}_J). \tag{4.5}$$

The first term is estimated using (4.4). The second is calculable from the data. The third term is estimated using little bootstrap in a manner similar to the x-fixed case described above. Note that by taking $V = I$, we get estimates of $\|\hat{\beta}_J - \beta^*\|^2$.

Just because little bootstrap gives almost unbiased estimates of the submodel MEs does not necessarily imply that selecting the submodel that minimizes $\hat{\text{ME}}(\zeta_J)$ gives a good selection procedure. We rely on the simulations in Section 6 to give a picture of how well the little bootstrap estimates do in the selection and evaluation process.

5. RSS-EXTREME SUBMODELS

Assume a sequence of submodels $\zeta_0, \ldots, \zeta_M$, and denote

$$\text{RSS}(J) = \text{RSS}(\zeta_J).$$

Definition 5.1. *Call* ζ_J *a rss-extreme submodel if there is an* $\alpha \geq 0$ *such that*

$$\text{RSS}(J) + \alpha J = \min_{J' \geq 0} [\text{RSS}(J') + \alpha J'].$$

It is clear that the smallest and largest submodels, ζ_0, ζ_M, are always rss-extreme. The others are characterized as follows:

Proposition 5.2. ζ_J, $J \in (0, M)$, *is rss-extreme iff for every* $J' < J < J''$ *with* $J = tJ' + (1-t)J''$,

$$\text{RSS}(J) \leq t\,\text{RSS}(J') + (1-t)\text{RSS}(J''). \tag{5.3}$$

The proof is a simple convexity argument.

Proposition 5.2 characterizes the rss-extreme submodels as those for which their RSS is an extreme point of the lower convex envelope of the graph $\{k, \text{RSS}(k)\}$, $k = 0, \ldots, M$. The isotone regression algorithm can be adapted to give an efficient method for finding the rss-extreme submodels.

Note that the subset selected by C_P minimizes $\text{RSS}(J) + 2\hat{\sigma}^2 J$. Other candidate selection rules (see Thompson [1978]) are to minimize $\text{RSS}(J) + c\hat{\sigma}^2 J$, where c is larger than 2 and can range as high as 6 or 7. Guided by this, we restrict attention to rss-extreme submodels with α in the range $2\hat{\sigma}^2$ to $10\hat{\sigma}^2$. The number of such submodels is usually a small fraction of the total number M of submodels. Selecting from these gives a substantial savings in computations, and allows the analyst to focus on only a few competing submodels.

The question now is, if we select from only among the rss-extreme submodels, how much do we lose? The simulation results of the next section show that not only do we not lose, but in fact the restriction often helps matters.

6. SIMULATION EXPERIMENT

6.1 Description

The simulation was complex, so it will be described in stages.

a) For each run, the X-design was fixed, as were the coefficients of the full model. In each repetition, normal noise was generated and added to give the y-values. Backwards deletion was then carried out to give the sequence of submodels. There were always forty variables and either 60, 160, or 600 cases. In each run, there were 500 repetitions.

b) In each repetition the ME was computed for each submodel selected by the backwards deletion. ME estimates for each submodel were derived using a replicate data set and C_p.

c) In each repetition little bootstrap was applied. We tested to see how many repetitions of little bootstrap were necessary, by trying 20, 40, 80 iterations. We found that 40 was an improvement on 20, but that 80 gave only a marginal improvement over 40, so we stuck with 40 over the course of the simulation.

We were also unsure of the appropriate values for t, the multiplier of $\hat{\sigma}^2$. In all initial runs, we tried $t = .2, .6, 1.0$. In some initial runs we tried other values of t such as .5, .7, .8. We comment further on the results below.

d) Two general behavioral characteristics were observed. The first was the behavior of the ME estimates over the entire sequence of submodels. Since the MEs were known, the accuracy of the estimates could be computed and systematic errors noted. We call this the global behavior.

In the second type of behavior we looked at how well these estimates did in selecting dimensionality and evaluating the submodel selected. Knowing the MEs, we knew the optimal dimensionality.

Using the replicate data estimate, in each repetition we selected the subset having the minimum estimated ME. For this subset we computed its dimensionality and the value of its replicate data ME estimate. The selected dimensionality was compared with the optimal dimensionality. The replicate data ME estimate for this subset was also compared with the actual ME of the subset. This was repeated for the subset selected by C_p and by little bootstrap. We refer to these results as the submodel selection and evaluation behavior. We also did these computations for the rss-extreme subset having minimum little bootstrap ME estimate.

e) Details: The X-distribution was generated from a multivariate mean-zero normal with $E(X_iX_j) = \rho^{|i-j|}$, with $\rho = .7$. The generated X-design was then held fixed for all runs. In all cases $N(0,1)$ noise was added. The non-zero coefficients were in three clusters of adjacent variables with the clusters centered at the 10th, 20th, and 30th variables.

For the variables clustered around the 10th variable, the initial coefficients values were given by

$$\beta^*_{10+j} = (h-j)^2, \quad |j| < h.$$

The coefficient clusters at 20 and 30 had the same shape. All other coefficients were zero. The coefficients were then multiplied by a common constant to make the theoretical R^2 equal to .75 (theoretical $R^2 = (\beta^{*t} X^t X \beta^*)/(\beta^{*t} X^t X \beta^* + \sigma^2)$).

We used the h-values 1, 2, 3, 4. This gave, respectively, 3, 9, 15, 21 non-zero coefficients. For $h = 1$, there were three strong, virtually independent variables. At the other extreme, $h = 4$, each cluster contained 7 weak variables. These four different sets of coefficients are designated by H1, H2, H3, H4 in the tables and figures. The t-values for the coefficients are graphed in Figure 1 for the 3 sample sizes.

We also ran the case with all coefficients zero. This is designated by a Z in the tables, and figures. Many preliminary runs were done with other coefficients and X-designs before settling on the scheme for the final runs. Note that each run involved 500 repetitions, each with 41 sequences of 40 variable deletions. This required a non-trivial amount of CRAY-XMP time. My appreciation is due to Ludolf Meester who transferred my code to the CRAY, and did the graphs.

6.2 What Value Should t Have?

The smaller t, the less bias. But we suspected, (and our simulations confirmed) that the smaller t is, the larger the variance of the ME estimates. We did some preliminary runs to check the effects of different values of the parameter t. For each submodel of dimension J we averaged the values of the $\hat{ME}(\zeta_J)$ little bootstrap estimates over the 500 runs and compared these with average of the $ME(\zeta_J)$. The difference we refer to as the bias. Also, for each $\hat{ME}(\zeta_J)$ little bootstrap estimate we computed the RMS difference over the 500 runs between the estimate and the $ME(\zeta_J)$ value.

We used the t values .2, .6, 1.0. The bias generally increases slightly as we go from .2 to .6 and does not increase drastically even for $t = 1.0$. However the RMS error decreases markedly from $t = .2$ to $t = .6$ and is usually the lowest at $t = 1.0$.

In a set of preliminary runs at sample sizes 60 and 160, we used 4 different sets of coefficients (including Z) similar to, but not the same as, the coefficients described above. For each run of 500, we averaged the absolute value of the bias over J, as well as the RMS errors. Then we averaged over the 4 coefficient sets. The results are given in Table 6.1 below

Table 6.1

Although the best performance in terms of RMS error is given by $t = 1.0$, its theoretical justification is weak. Furthermore, in running a case at sample size 60 with X-design and coefficients different than those described above, we found that the bias and RMS error using $t = 1.0$ increased sharply at important values of J. For these reasons, we do not feel that we can recommend use of $t = 1.0$. Even when $t = 1.0$ yields lower RMS than $t = .6$, the improvement is small. For general use we prefer the .6-.8 range. The remainder of the simulation results are based on $t = .6$.

6.3 Global Comparison: Little Bootstrap, CP and Replicate Data

In our final runs we compared the little bootstrap procedure to C_p and use of a replicate data set. The average of the absolute values of the bias over J for little bootstrap and C_p are given in table 6.2 below (the replicate data bias is zero within limits of variability).

Table 6.2

The "average" over J of the RMS differences between the estimates and the $\mathrm{ME}(\zeta_J)$ values are given in table 6.3 below. The first row is the standard deviation of the $\mathrm{ME}(\zeta_J)$ over the 500 runs "averaged" over J (RD = replicate data).

The reason for quotes around the word average is this: for small J, $\mathrm{ME}(\zeta_J)$ becomes large except in the Z case. Then the RMS differences also become large (see figure 3). The average over all J would unduly reflect the RMSE for a few of the lowest J values. For this reason we averaged only over those J for which the 500 run average $\mathrm{ME}(\zeta_J)$ was less than the corresponding full model ME_M.

Table 6.3

In figure 2 we graph the averages over the 500 runs of the three different estimates of $\mathrm{ME}(\zeta_J)$ together with the $\mathrm{ME}(\zeta_J)$ values. Side by side we graph the RMS errors of the $\mathrm{ME}(\zeta_J)$ estimates together with the standard deviation of the $\mathrm{ME}(\zeta_J)$. Note that the C_p estimates are heavily biased downward. Surprizingly, this persists even for sample size 600.

For sample size 60, the test set estimates have substantially lower RMS values than little bootstrap. But at the two higher sample sizes, their overall RMS values are very comparable. Our approximate calculations show that little bootstrap and the test set would have comparable accuracies if the exact value of σ^2 were used in setting up the variance of the $\{\varepsilon_1\}$. We conjecture that the loss of accuracy at sample size 60 is due to the fact that only 20 degrees of freedom are available for the σ^2 estimate.

6.4 Dimensionality Selection and Evaluation Behavior

There are two aspects to this problem. First, is the procedure nearly picking out the optimum dimensionality? Second, is the estimated ME for the selected subset close to the actual ME for the subset? In this phase, we compared the replicate data, little bootstrap, and C_p procedures. The dimensionality selected was that at the minimum of the $\mathrm{ME}(\zeta_J)$ estimates. We also ran a modified little bootstrap, where the subset selected is that rss-extreme subset having minimum little bootstrap ME estimate. This procedure is designated as LB/E.

The most telling summary is the comparison of the average ME for the subset selected using the ME minimum and the average ME for the subset selected by the estimating procedure. This is given in table 6.4 below

Table 6.4

The next comparison is between the average dimension as selected using the actual ME's and by each of the estimates together with the RMS differences between them. In table 6.5 below, the figures in parentheses are the RMS differences except that the figures in parentheses following the average dimension selected by the actual ME's is the standard deviation over the 500 runs of the dimension selected.

Table 6.5

In terms of the ability of the estimate to evaluate the subset selected, we give two tables. The first (Table 6.6) compares the average estimated ME value for the subset selected by the estimate to the average ME value for the same subset. In table 6.6, the first number is the average *estimated* ME, the second is the average ME for the same subset. Note that both the RD and LB estimates have a downward bias, although over all J they are virtually unbiased. The reason is that the subset being evaluated was selected as the subset minimizing the RD estimates and LB estimates respectively.

Table 6.6

Table 6.7 gives the RMS differences between the true ME for the subset selected by the estimate and its estimated ME.

Table 6.7

6.5 Discussion of Results

These results, first at all, clearly indicate the salient difficulty in submodel selection. This is the presence of a number of weak variables whose estimated coefficients can be close to zero. These variables can be deleted sooner than variables with zero true coefficients, but estimated coefficients away from zero. When these former are deleted, their non-zero coefficient values make substantial contributions to the ME.

This difficulty can be made worse by substantial correlations between the weak variables and other variables, weak or strong. In this case, deletion of a weak variable can produce very little RSS increase since its predictive ability can be transferred to a correlated variable: Thus, the case of many weak correlated variables (case H4) continues to give high ME for the selected subsets even at sample size 600.

As to the behavior of the estimates; C_p is clearly very biased. This bias persists even at sample size 600. It selects models that are too large. If there are many weak variables, this is not too damaging because it will then retain some of the weak variables with nonzero coefficients. For this reason, C_p does slightly better than little bootstrap in some situations involving weak variables (see Table 6.4). But in terms of subset evaluation, the C_p estimates are out of the ball park.

Little bootstrap, in terms of selection, has difficulty with weak variables. It does not select dimensionality as well as the replicate data procedure, although on the average it selects about the right dimensionality. In terms of evaluation, it does quite well compared to use of a replicate data. It is somewhat less accurate at sample size 60. But accuracies are comparable at the two larger sample sizes.

The RMS errors in both the replicate data and little bootstrap methods are substantial. They average about 11-12 while the ME's we are trying to estimate have a maximum value of around 40. This seems to be inherent in the problem and I doubt if there is any method that could substantially increase this accuracy.

To illustrate, consider trying to estimate the full model error $(\varepsilon, H\varepsilon)$. The estimate we used above was $M\hat{\sigma}^2$. This, at best, is estimating a $\sigma^2\chi^2_M$ variable by its mean value $M\sigma^2$. The resulting variance is $2M\sigma^4$. The standard deviation is $\sigma^2\sqrt{2M}$; in the simulation this equals $\sqrt{80} \cong 9$.

Are better estimates of $(\varepsilon, H\varepsilon)$ available? The only things we have approximating the $\{\varepsilon\}$ are the residuals $\{r\}$. But the residuals are independent of $H\varepsilon$, so the best estimate of $(\varepsilon, H\varepsilon)$ we can get using the residuals cannot improve on using $M\sigma^2$ as an estimate. The essence of this problem is that we are forced to estimate unobservable random variables by their mean values. The result is substantial RMS error.

However, this error changes slowly across the sequence of submodels $\zeta_0, \ldots, \zeta_M$. As a result, the replicate data method is able to accurately select the minimum ME submodel. Little bootstrap does not do as well when weak variables are present, but it certainly improves on any other method currently being used.

The results also show that restricting selection to rss-extreme submodels uniformly improves the little bootstrap accuracy and significantly decreases the variability of the dimensionality selected. On the average, over all sample sizes and coefficients, about 5 out of the 41 submodels are rss-extreme. For H3 and H4 the average is around 6, while Z and H1 average around 4.

7. BIAS V.S. VARIANCE REVISITED

We earlier referred to submodel selection as a trade off between bias and variance. We can now make this more precise and give some results to quantify the trade off.

The submodel predictor $\hat{\mu}(\zeta)$ is not a predictor of μ^*, but rather of the reduced model $\mu^*(\zeta) = H^{\zeta}\mu^*$. In particular, the OLS coefficients of $\{x_m ; m \in \zeta\}$ in $\hat{\mu}(\zeta)$ are estimates of the corresponding coefficients in the reduced model $\mu^*(\zeta)$.

Now ME(ζ) can be split into two terms:

$$\|\mu^* - \hat{\mu}(\zeta)\|^2 = \|\mu^* - \mu^*(\zeta)\|^2 + \|\mu^*(\zeta) - \hat{\mu}(\zeta)\|^2.$$

The first term measures the minimum discrepancy between μ^* and *any* model based on $\{x_m ; m \in \zeta\}$. We call it the bias term. The second term measures the error in $\hat{\mu}(\zeta)$ as an estimate of $\mu^*(\zeta)$, and is called the variance term. This latter terminology is not, strictly speaking, correct in our present context.

7.1 Structure of the Variance Term

If ζ is not data selected, then $\hat{\mu}(\zeta)$ is an unbiased estimate of $\mu^*(\zeta)$, and $\|\mu^*(\zeta) - \hat{\mu}(\zeta)\|^2$ is correctly called variance. If $|\zeta| = J$,

$$E\|\mu^*(\zeta) - \hat{\mu}(\zeta)\|^2 = J\sigma^2.$$

But suppose ζ_J is a data selected subset of dimension J with $\zeta_J = \{m_1, \ldots, m_J\}$. Let $\beta^*_{m_j}$ be the coefficient of x_{m_j} in $\mu^*(\zeta_J)$ and $\hat{\beta}_{m_j}$ the OLS estimate in $\hat{\mu}(\zeta_J)$. The distribution of $\hat{\beta}_{m_j} - \beta^*_{m_j}$ may be quite complex. For instance, look at the orthogonal model with coefficients $\{\beta^*_m\}$. Suppose ζ_J is selected, then the distribution of $\hat{\beta}_{m_j}$, $j = 1, \ldots, J$ will depend on the relative magnitude of all of the $\{\beta^*_m\}$. For example, if $|\beta^*_1/\sigma| \geq 10$ and $|\beta^*_m/\sigma| \leq 1$, $m > 1$, then the 1st variable will almost always be in every ζ_J, $J \geq 1$ and $\hat{\beta}_1 - \beta^*_1$ will have an approximately normal distribution with mean zero and variance σ^2.

But now suppose there are numerous variables with $|\beta^*_m/\sigma|$ in the range 1-2. There is a competition for inclusion in ζ_J between the variables. The ones that win tend to have the largest values of $\hat{\beta}_m - \beta_m^*$ in the direction of the sign of β_m^*. For the weaker variables included in the model, the distribution of $\hat{\beta}_m - \beta_m^*$ given that they were selected will be skewed with non-zero means and inflated variances. In addition, if there are a large number of variables with $\beta_m^* = 0$, then some of these will have large $|\hat{\beta}_m|$ values and may be included in ζ_J, also resulting in inflated variances.

Thus, the variance component term $\|\mu^*(\zeta) - \hat{\mu}(\zeta)\|^2$ can reflect both the bias in coefficient estimates and inflated variance due to the selection process. The extent to which the expectation of this term exceeds $J\sigma^2$ is a measure of these selection biases.

7.2 Simulation Results

As a substudy in our simulation, in each iteration of a run, $\mathrm{ME}(J) = \mathrm{ME}(\zeta_J)$ was decomposed into the bias and variance components BIAS (J) and VAR(J). These were then averaged over the 500 iterations in the run. Graphs of these averages are given in Figure 3 for H1, H2, H3 and H4. Superimposed on the graphs is the straight line $J\sigma^2 (= J)$ for comparison with Av(VAR(J)).

To give a more quantifiable idea of how much effect the selection inflates the value of the variance term,, we also computed the ''excess''. In each iteration of a run, the dimensionality selected, $J_{\min}$, was defined by

$$\mathrm{ME}(J_{\min}) = \min_J \mathrm{ME}(J).$$

In this iteration, the excess was computed as

$$\frac{\mathrm{VAR}(J_{\min}) - J_{\min}}{J_{\min}}.$$

This quantity was then averaged over the 500 iterations. Table 7.1 below gives the values of this quantity (E) together with $V = \mathrm{Av}(\mathrm{VAR}(J_{\min})$ and $B = \mathrm{Av}(\mathrm{BIAS}(J_{\min}))$.

Table 7.1

The values of the excess are surprizingly low, compared to the higher excesses that show up in figure 3. For instance, if we look at the average excess at $J = 20$, we get

	ss 60	ss 160	ss 600
H1	.78	.83	.82
H2	.68	.69	.68
H3	.59	.62	.62
H4	.56	.53	.47

Looking at this table, and especially at H1, it is clear that the major source of the excess is in those variables selected in ζ_{20} that have zero or nearly zero true coefficients.

This also indicates that if the selected submodel has dimensionality close to the minimum ME submodel, then the variance inflation is not substantial. Of course, we can almost completely eliminate excess by always choosing submodels with small dimensionality, but only at the cost of increased ME.

8. WHAT CAN BE DONE AFTER SELECTION?

8.1 Do Confidence Intervals Make Sense?

For the classical linear model there are elegant conditional distributional results that give confidence intervals for the coefficients and significance levels for tests of hypotheses.

A nonsensical procedure that is often used in standard statistical packages is to do submodel selection, select (somehow) the best submodel and then to apply classical distributional theory to the coefficients by assuming that the other variables never existed.

That significance testing results in nonsense can be clearly seen from the orthogonal model with all $\beta_m^* = 0$. Say, for instance, that $M = 75$ and a model of size 4 is selected. Then, in 95% of the runs of this model, all four coefficients would be found significant at the 90% level (assuming $\hat{\sigma}^2 = \sigma^2$). *They are significant because they have been selected.*

What is the meaning here of confidence intervals? For instance, how can confidence intervals be defined for the coefficients of the variables deleted from the equation? Or consider the distribution of the estimated coefficients: Over many simulated runs of the model, each time generating new random noise, and selecting, say, a subset of size 4, the coefficient estimates of a given variable have a point mass at zero reflecting the probability that the variable has been deleted. In addition, there is a continuous mass distribution over those times when the variable showed up in the final 4 variable equation. The relation of this distribution to the original coefficients is obscure.

As pointed out in Section 6, the coefficients in $\hat{\mu}(\zeta)$ are not estimates of the coefficients of $\{x_m ; m \in \zeta\}$ in the full model μ^*, but rather estimates of the coefficients of the reduced model $\mu^*(\zeta) = H^{\zeta}\mu^*$. Suppose ζ_J is the selected subset of size J, $\zeta_J = \{m_1, \ldots, m_J\}$. Let $\beta^*_{m_j}$ be the coefficient of x_{m_j} in $\mu^*(\zeta_J)$. Then what we may want, in analogy to classical theory, is the distribution of $\hat{\beta}_{m_j} - \beta^*_{m_j}$ given that ζ_J is selected. As noted above, this distribution may be complicated, with skew and non-zero mean.

In general, even running a simulation to estimate these distributions, using known $\{\beta^*_m\}$, σ^2 seems formidable. One would have to repeatedly generate $\{\varepsilon\}$, set $y = \mu^* + \varepsilon$, look only at those outcomes in which ζ_J was selected, and using those outcomes construct some nonparametric estimate for the distribution of $\hat{\beta}_m - \beta^*_m$. The problem of estimating these distributions for $\{\beta^*_m\}$ unknown seem much more difficult. My opinion is that such an effort would be "love's labour lost". In particular how would such results be used?

8.2 Useful Information for Data Analysts

In my experience, the two most useful pieces of information about the structure of a problem involving submodel selection are first--some rough approximate idea of the relative importance of the variables still left in the equation. This can be gotten from deleting a variable still in the equation, computing the rise in the residual sum of squares, putting the variable back in and repeating this procedure with the next variable still in the equation. The sizes of these RSS increases give one measure of relative importance.

Second--an idea of the alternative subsets of the same dimensionality that have nearly the same residual-sum-of-squares. This information can give valuable insights into the structure of the problem. If the "best subsets" algorithm is used, this information can be easily supplied. But for more than 30 variables, this algorithm is too slow and stepwise methods must be used.

The advantage of resampling methods such as little bootstrap and cross-validation is that they form alternative sequences of submodels. In general, each application of little bootstrap will result in a different sequence of submodels than formed using the original data. As surprising as it may seem, in cross-validation, even the deletion of a single case often leads to a different sequence of submodels.

The fact that both little bootstrap and cross-validation can give alternative submodel sequences is the key to the fact that they can produce relatively unbiased PE and ME estimates for data selected submodels. Methods, such as C_p, by not providing for alternate submodel paths, cannot provide low-bias estimates.

This property can be used to advantage even when only stepwise deletion (or addition) is being used. For instance, suppose the analyst wants to look at alternative submodels containing five variables. In the, say, 40 iterations of little bootstrap, look at all subsets of size five selected in the 40 deletion procedures. Now run a regression (using the original data) on each distinct group of five variables selected in the little bootstrap deletions. The residual-sum-of-squares produced should be close to that of the subset produced by the original deletion process.

9. Conclusions

The issue of submodel selection and evaluation is a critical one in statistics. It occurs in analysis of variance, in analysis of discrete data, in generalized linear models, in time series, as well as in regression. In distinction to most theoretical results, which assume a predetermined sequence of submodels, in actual practice the sequence of submodels chosen is data dependent. Regardless of asymptotic optimality results, criterion or estimates such as C_P, AIC, BIC etc. are highly biased in finite data situations because they do not account for the data driven selection.

The simulation results emphasize again what many statisticians have long suspected--that the various ad hoc methods used to evaluate submodels when data driven selection occurs can be extremely optimistic.

Although the distributional assumptions are stringent, little bootstrap emerges as the only procedure to date that can give relatively unbiased estimates of the x-fixed ME or PE when data driven submodel selection is used and the number of cases relative to the number of variables is moderate.

An important subsidiary conclusion is that restricting selection to the class of rss-extreme submodels slightly improves model selection accuracy, while drastically reducing the number of candidates.

Little bootstrap has wider applicability than submodel selection in OLS regression. It works in contexts where the coefficient estimates are linear in the $\{y_n\}$. Thus, the theory and practice of little bootstrap generalizes to such situations as estimating optimum ridge parameters. However, that is another research story.

References

Bickel, P. and Freedman, D. (1982), "Bootstrapping regression models with many variables," A. Festschrift for E.L. Lehmann pp. 28-48.

Breiman, L., and Spector, P. (1989), "Submodel Selection and Evaluation in regression — the X-Random Case," Technical Report No. 197, Statistics Department, University of California at Berkeley. To be published, International Statistical Review.

Copas, J.B. (1983), "Regression, prediction, and shrinkage," J.R. Statist. Soc. B, 45 No. 3, 311-354.

Efron, B. (1986), "How biased is the apparent error rate of a prediction rule?" JASA, 81, 461-470.

Freedman, D., Navidi, W, and Peters, S. (1988), "On the impact of selecting variables in fitting regressions," (To appear).

Gong, S. (1986), "Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression," JASA, 81, 108-118.

Mallows, C.L. (1973), "Some remarks on C_p," Technometrics, 15, No. 4, 661-675.

Miller, A.J. (1984), "Selection of subsets of regression variables (with discussion)," J.R. Statist. Soc. UL A, 147, part 2, 398-425.

——— (1990), Subset Selection in Regression, Chapman and Hall, London.

Thompson, M.L. (1978), "Selection of variables in multiple regression: Part I. A review and evaluation," International Statistical Review, 46, 1-19.

Technical Appendix

Proof of Theorem 4.2.

Proof.

Consider the scaled response data $y' = y/\sigma$, $\mathbf{x}' = \mathbf{x}/\sigma$, and denote OLS predictors based on the (y',x') data by $\hat{\mu}'$. The estimates $\hat{\mu}(\zeta)$ and $\hat{\mu}'(\zeta)$ differ only by the scale factor σ. Assuming scale invariant subset selection, the same ζ_J are selected by both data sets. Denoting $\varepsilon' = \varepsilon/\sigma$, then

$$y' = \mathbf{x}(\beta^*/\sigma) + \varepsilon'$$

and

$$(\varepsilon, \hat{\mu}_M - \hat{\mu}(\zeta_J)) = \sigma^2(\varepsilon', \hat{\mu}'_M - \hat{\mu}'(\zeta_J)).$$

Therefore

$$\theta_J(\beta_1^*, \ldots, \beta_M^*, \sigma^2) = \sigma^2\theta_J\left[\beta_1^*/\sigma, \ldots, \beta_M^*/\sigma, 1\right]. \tag{A2}$$

Now, looking at the data

$$\tilde{y} = y + \varepsilon_1 = \mu^* + \varepsilon + \varepsilon_1,$$

note that $\tilde{\mu}_0 - \tilde{\mu}(\tilde{\zeta}_J)$ is a vector quantity that depends stochastically only on the random vector $\varepsilon + \varepsilon_1$. But for any n,

$$E(\varepsilon_{1n} | \{\varepsilon + \varepsilon_1\}) = (t^2/1 + t^2)(\varepsilon_n + \varepsilon_{1n}).$$

This implies that

$$E(\varepsilon + \varepsilon_1, \tilde{\mu}_0 - \tilde{\mu}(\tilde{\zeta}_J)) = (1 + \frac{1}{t^2})E(\varepsilon_1, \tilde{\mu}_0 - \tilde{\mu}(\zeta_J)). \tag{A3}$$

Putting (A2) and (A3) together proves the theorem.

Tables

Table 6.1. Bias and RMS Error for Different t-Values

	Sample Size 60		Sample Size 160	
t	Bias	RMS	Bias	RMS
.2	.7	16.4	.6	15.0
.6	.6	14.1	.8	11.9
1.0	.8	13.8	1.1	11.4

Table 6.2. Average Bias of the LB and CP Procedures

	Z	H1	H2	H3	H4
			Sample Size 60		
LB	.5	1.0	.9	.7	.5
CP	21.7	19.4	20.6	21.7	22.4
			Sample Size 160		
LB	.4	.4	.4	.7	.7
CP	23.1	20.8	20.9	21.1	22.7
			Sample Size 600		
LB	.2	.5	1.0	.8	1.1
CP	23.5	22.0	19.3	19.2	22.6

Table 6.3. ''Average'' RMS Error

	Z	H1	H2	H3	H4
Sample Size 60					
SD	7.4	8.2	9.0	9.0	8.9
RD	8.9	8.9	9.3	9.5	9.5
LB	12.9	13.2	13.6	13.8	13.8
CP	25.6	24.3	25.3	26.2	26.9
Sample Size 160					
SD	7.4	7.7	8.4	9.3	9.5
RD	9.1	9.2	9.6	9.8	9.8
LB	9.4	9.6	10.3	10.8	10.7
CP	25.4	23.9	24.6	24.9	25.7
Sample Size 600					
SD	7.5	7.7	9.3	9.0	9.6
RD	9.3	9.2	9.7	9.5	9.7
LB	8.8	8.9	10.1	9.6	10.0
CP	25.6	24.0	22.8	21.9	22.5

Table 6.4. Average MEs Produced by the Selection Procedure

Sample Size 60

	Z	H1	H2	H3	H4
ME(true)	.0	4.4	14.1	21.7	25.3
RD	1.4	5.9	16.5	24.4	28.5
LB	4.9	9.5	21.1	30.0	33.5
LB/E	4.6	8.5	19.7	28.6	32.4
CP	20.7	22.7	28.4	31.4	32.8

Sample Size 160

	Z	H1	H2	H3	H4
ME(true)	.0	3.1	17.2	24.7	29.1
RD	1.5	4.6	19.9	28.0	32.0
LB	1.9	5.0	22.5	34.3	37.7
LB/E	1.9	5.0	22.1	33.3	36.7
CP	22.6	23.8	30.3	32.6	35.3

Sample Size 600

	Z	H1	H2	H3	H4
ME(true)	.0	3.1	19.6	20.4	29.9
RD	1.3	4.3	22.0	22.6	32.7
LB	1.8	5.1	28.7	26.9	37.8
LB/E	1.8	4.9	27.9	26.1	36.7
CP	22.6	24.1	29.4	31.4	35.1

Table 6.5. Average Dimension Selected and RMS Difference to Dimension Selected by ME.

Sample Size 60

	Z	H1	H2	H3	H4
ME(true)	.0(.0)	3.2(1.3)	4.1(2.6)	6.1(3.6)	7.9(3.8)
RD	.6(2.1)	3.9(3.4)	5.5(4.1)	7.7(5.8)	9.4(6.5)
LB	1.7(6.1)	4.9(6.2)	6.8(8.4)	9.2(9.9)	11.0(10.7)
LB/E	1.2(4.0)	3.9(2.8)	5.1(3.8)	6.7(4.9)	7.9(5.4)
CP	6.8(8.3)	9.3(7.6)	10.6(8.1)	11.4(7.6)	11.7(6.8)

Sample Size 160

ME(true)	.0(.0)	3.0(.0)	4.5(1.9)	8.8(2.7)	11.6(3.8)
RD	.4(1.4)	3.4(1.4)	5.8(4.5)	9.8(5.1)	13.2(7.0)
LB	.3(1.0)	3.3(1.1)	5.5(5.9)	11.5(9.6)	15.4(11.3)
LB/E	.3(1.0)	3.3(1.1)	4.8(3.4)	9.1(5.1)	11.4(5.6)
CP	6.8(7.5)	9.1(6.8)	11.3(7.7)	13.0(5.8)	13.9(5.2)

Sample Size 600

ME(true)	.0(.0)	3.0(.0)	9.4(1.9)	10.0(1.7)	15.5(3.2)
RD	.3(1.1)	3.4(1.4)	10.2(3.9)	11.5(4.3)	17.5(6.6)
LB	.2(.7)	3.3(1.8)	10.9(6.4)	12.0(5.9)	19.0(9.4)
LB/E	.2(.7)	3.2(.7)	9.8(3.7)	11.1(3.5)	15.5(4.8)
CP	6.5(7.1)	9.0(6.6)	13.5(5.1)	15.2(6.0)	17.3(4.4)

Table 6.6. Estimated ME's for the Submodel Selected vs. Actual ME's.

	Z	H1	H2	H3	H4
			Sample Size 60		
RD	-.6(1.4)	3.7(5.9)	12.0(16.5)	18.2(24.4)	21.8(28.5)
LB	-.5(4.9)	5.5(9.5)	14.4(21.1)	21.2(30.0)	24.5(33.5)
LB/E	-.5(4.6)	5.8(8.5)	14.9(19.7)	21.9(28.6)	25.3(32.4)
CP	-9.2(20.7)	-5.4(22.7)	-4.0(28.4)	-3.1(31.4)	-2.8(32.8)
			Sample Size 160		
RD	-1.0(1.5)	1.9(4.6)	12.9(19.9)	20.8(28.0)	25.9(32.0)
LB	.1(1.9)	2.9(5.0)	15.3(22.5)	26.2(34.3)	30.7(37.7)
LB/E	.1(1.9)	2.9(5.0)	15.4(22.1)	26.6(33.3)	31.8(36.7)
CP	-8.9(22.6)	-5.4(23.8)	-2.0(30.3)	.2(32.6)	1.4(35.3)
			Sample Size 600		
RD	-.8(1.3)	2.3(4.3)	17.5(22.0)	18.5(22.6)	27.3(32.7)
LB	-.2(1.8)	2.8(5.1)	23.6(28.7)	22.9(26.9)	32.0(37.8)
LB/E	-.2(1.8)	2.8(4.9)	23.8(27.9)	23.1(26.1)	33.1(36.7)
CP	-9.3(22.6)	-5.8(24.1)	1.1(29.4)	3.4(31.4)	6.6(35.1)

Table 6.7. RMS Differencies Between Estimated and Actual ME's for the Submodels Selected.

	Z	H1	H2	H3	H4
		Sample Size 60			
RD	8.6	8.9	10.3	11.4	12.1
LB	15.1	13.7	14.6	15.6	15.3
LB/E	14.7	12.7	13.3	14.2	13.9
CP	31.6	29.9	33.9	35.8	36.9
		Sample Size 160			
RD	8.8	8.7	11.5	12.6	12.4
LB	10.2	10.4	12.7	15.1	14.0
LB/E	10.2	10.4	12.6	14.7	13.9
CP	32.7	30.5	33.4	33.9	35.3
		Sample Size 600			
RD	8.6	9.0	11.5	10.1	11.4
LB	10.7	10.3	14.1	11.2	12.2
LB/E	10.7	10.3	14.0	11.1	12.2
CP	32.9	30.9	29.9	29.3	30.0

Table 7.1. Bias, Variance and Excess at the Submodels Selected by ME.

	ss 60			ss 160			ss 600		
	B	V	E	B	V	E	B	V	E
H1	1.1	3.3	.00	0.0	3.1	.00	0.0	3.1	.00
H2	9.4	4.7	.06	12.2	5.0	.11	7.4	12.2	.17
H3	14.5	7.2	.15	12.6	12.1	.29	8.7	11.6	.09
H4	15.7	9.6	.18	13.7	15.4	.25	11.4	18.5	.16

Figure 1, T-VALUES FOR COEFFICIENTS

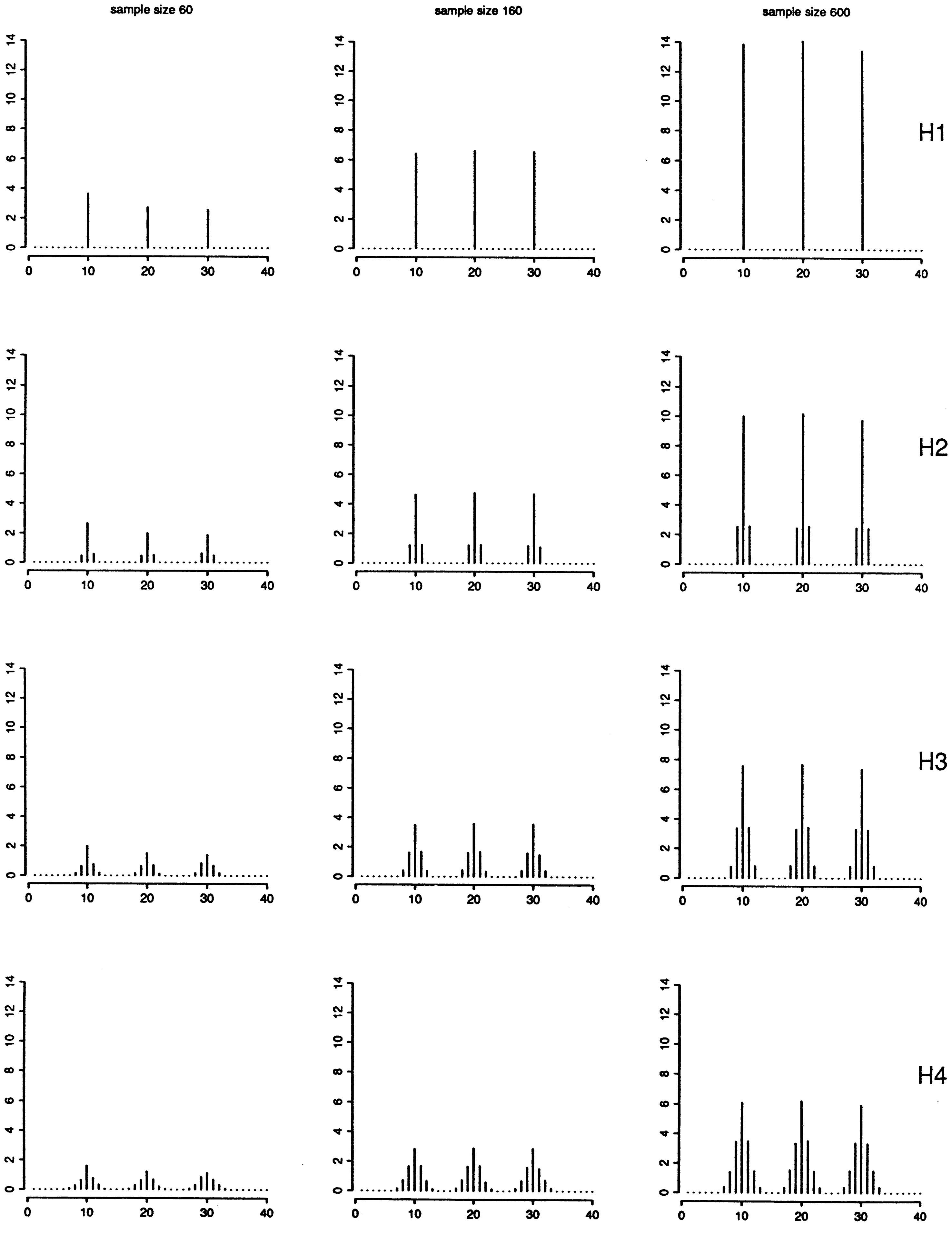


Figure 2

SAMPLE SIZE 60

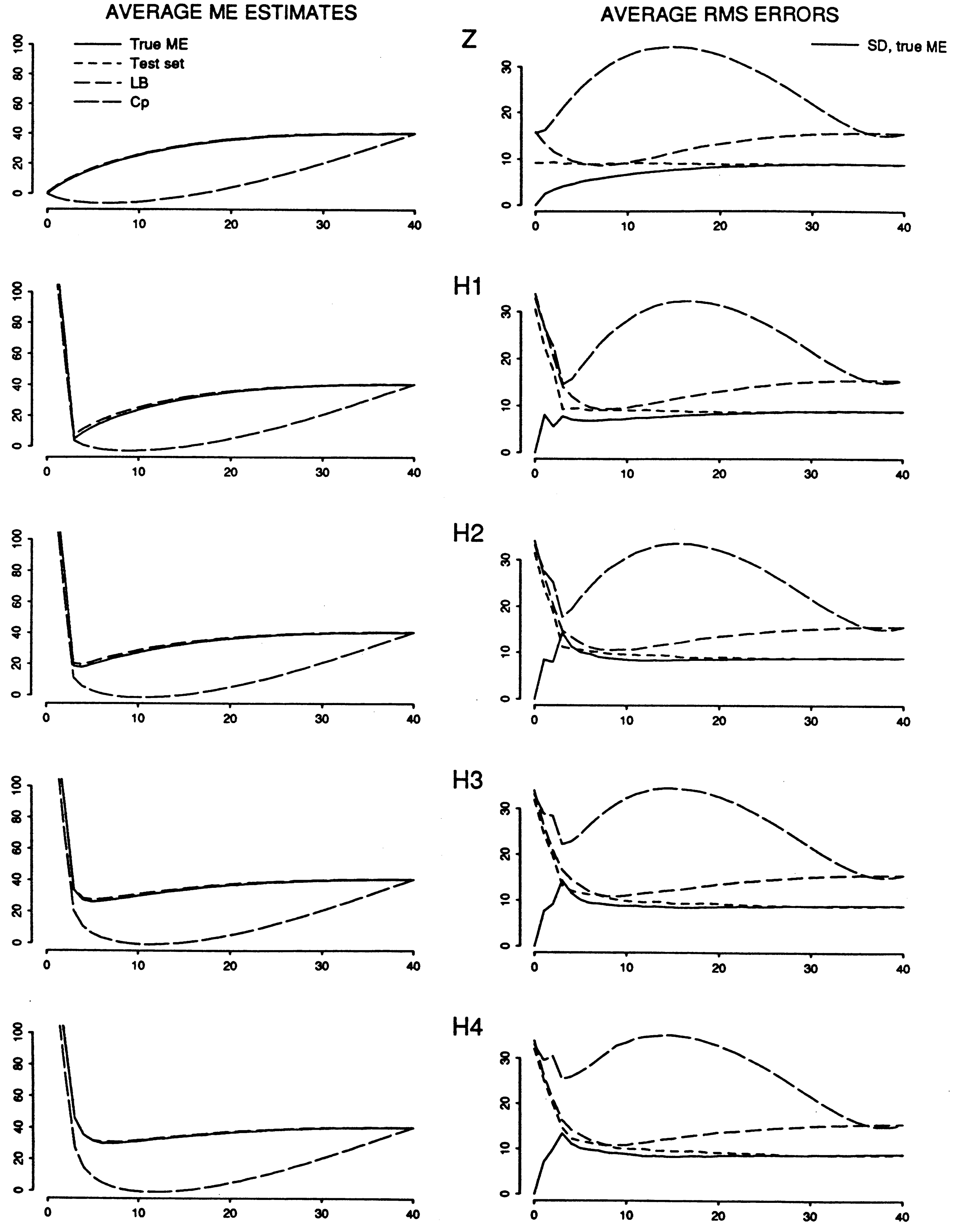


Figure 2 (continued)

SAMPLE SIZE 160

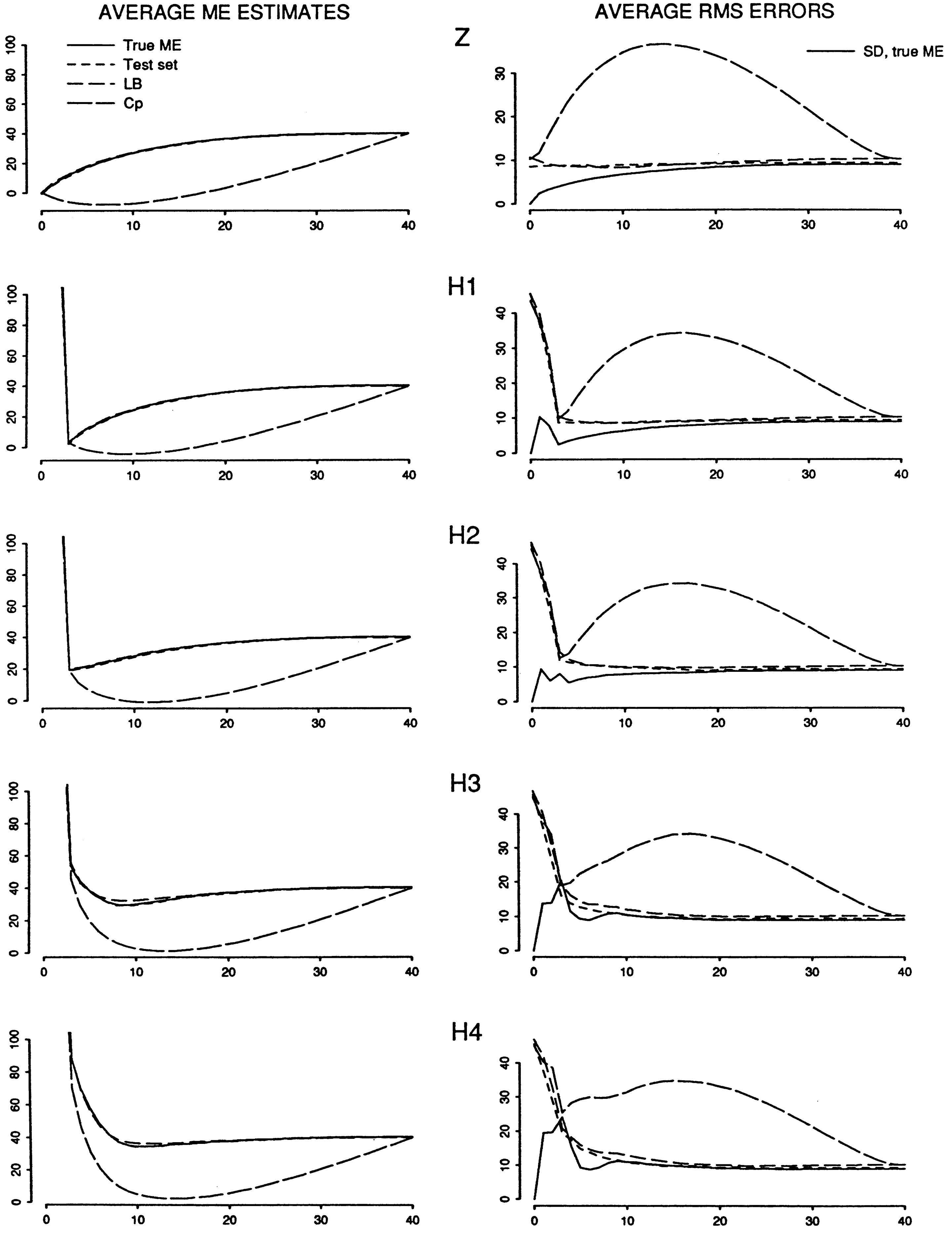


Figure 2 (continued)

SAMPLE SIZE 600

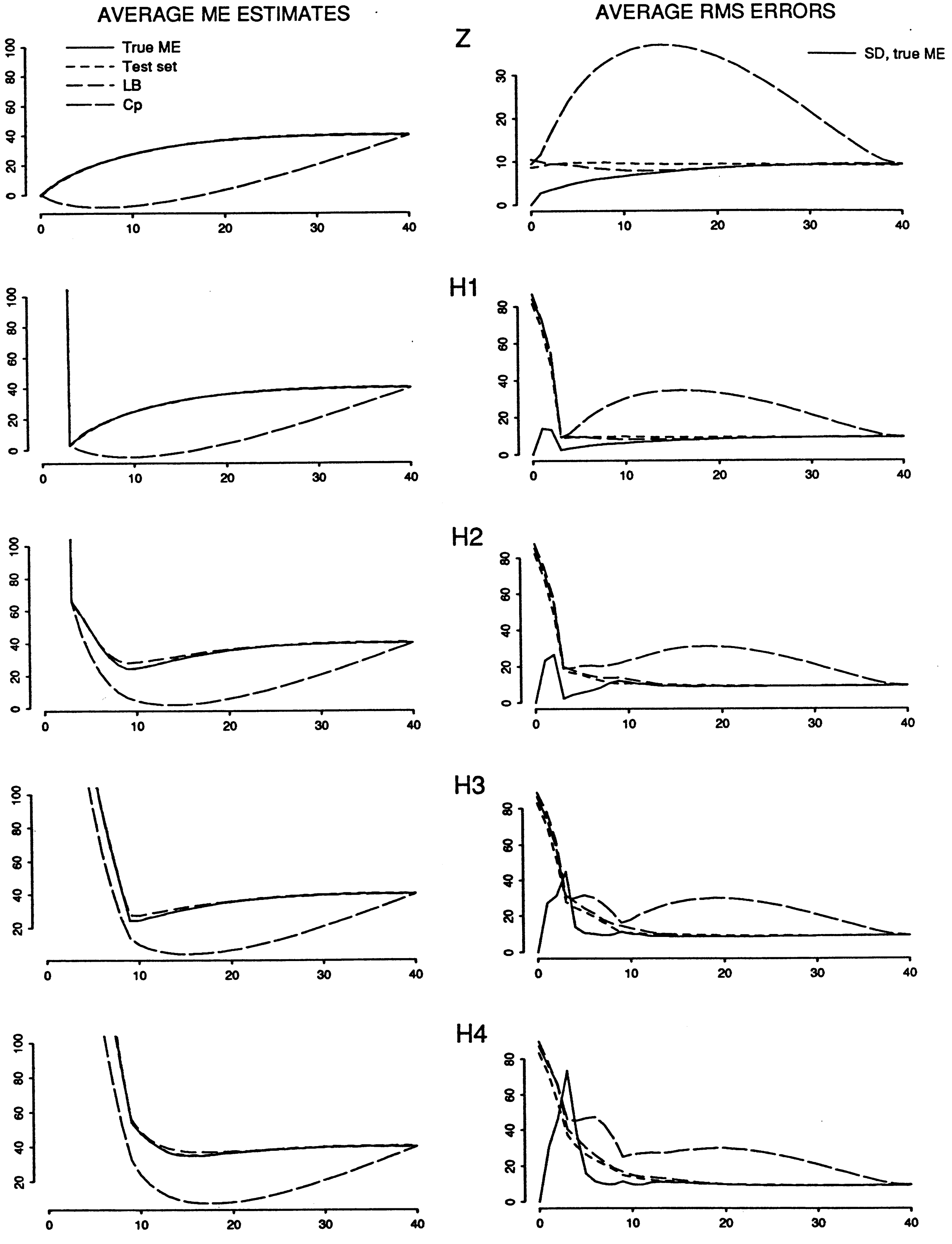


Figure 3

BIAS AND VARIANCE COMPONENTS OF ME

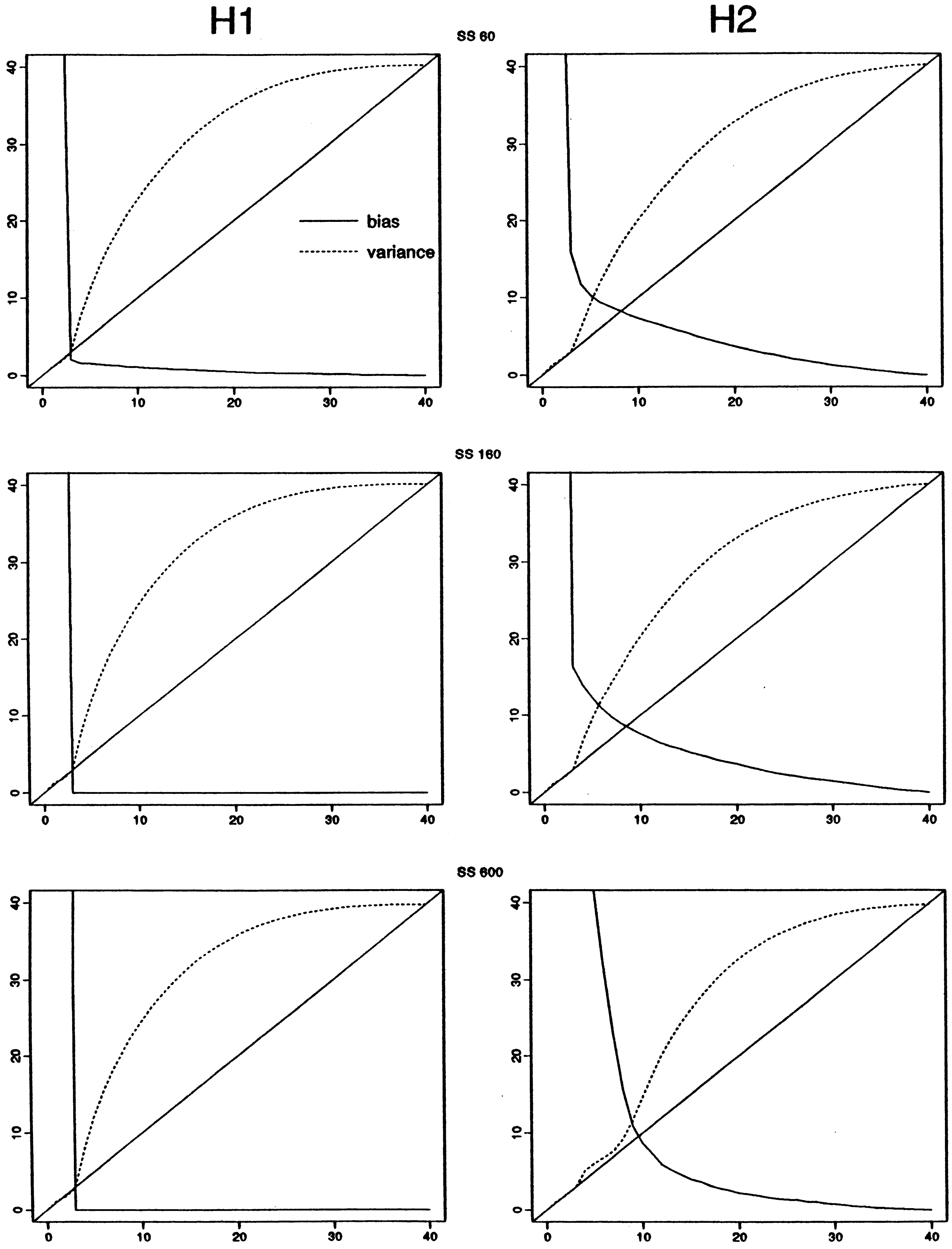
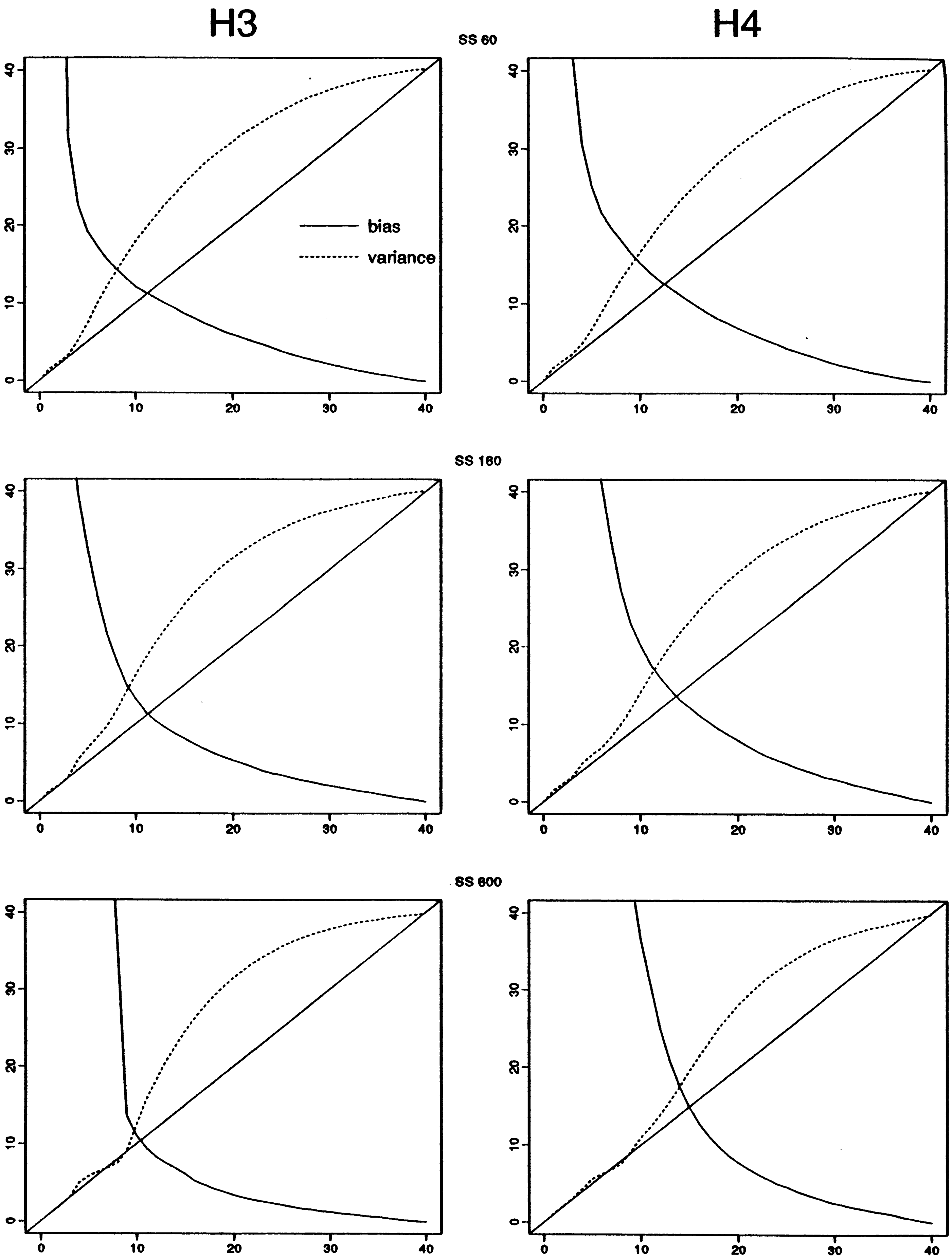


Figure 3 (continued)


H3
H4
SS 60
SS 160
SS 600
bias
variance