ON THE IMPACT OF VARIABLE SELECTION IN FITTING REGRESSION EQUATIONS

By

D.A. Freedman, W. Navidi and S.C. Peters
Statistics Department
University of California
Berkeley, Ca 94720

Technical Report No. 87
February 1987

Research supported by National Science Foundation DMS86-01634

Department of Statistics
University of California
Berkeley, California

ON THE IMPACT OF VARIABLE SELECTION IN FITTING REGRESSION EQUATIONS

D.A. Freedman, W. Navidi and S.C. Peters
Statistics Department
University of California
Berkeley, Ca 94720

Abstract.

Consider developing a regression model in a context where substantive theory is weak. Search procedures are often used to develop the equation: eg, fitting the equation, dropping insignificant variables, and refitting. As is well known, this can seriously distort the conventional goodness-of-fit statistics. Furthermore, the bootstrap and jackknife may not help in high-dimensional cases.

1. Introduction.

When regression equations are used in empirical work, the ratio of data points to parameters is often low. Further, the exact form of the equation is seldom known *a priori*, so investigators will often do some preliminary screening before settling on the final version of the equation. One stylized version of this strategy is as follows:

(i) Fit the equation with all variables included.

(ii) Screen out variables whose coefficients are insignificant at the 25% level. (This level is used to represent "exploratory" analysis.)

(iii) Refit the equation on the remaining variables.

Real investigators use more complicated – and subjective – screening procedures; the version just presented is mechanical, and therefore amenable to statistical analysis.

As is well known, screening procedures introduce substantial distortion into the conventional measures of goodness-of-fit, like R^2, t or F. See (Lovell, 1983) or (Freedman, 1983), and (Gong, 1986) on logistic regression. Perhaps the bootstrap or jackknife can be used to eliminate these distortions? This question will be investigated here by simulation.

Consider the basic linear model

$$Y = X\beta + \epsilon. \tag{1}$$

Here, X is an $n \times p$ matrix of iid N(0,1) variables; and ϵ is another $n \times 1$ vector of iid $N(0,\sigma^2)$ variables. These distributional facts are known to the investigator. The $n \times 1$ vector Y is computed from (1). The investigator observes X and Y, but not ϵ. The $p \times 1$ vector β of parameters is unknown, as is σ^2, and these are to be estimated from the data.

Two statistical tasks are considered:

(i) *Estimation.* The object is to estimate β_1; and $\beta_2,\ldots,\beta_p$ are introduced to control other sources of variation and improve the precision in estimating β_1. This is like a standard problem in clinical trials: β_1 is the treatment effect, and columns 2,3,... in X represent covariates.

(ii) *Prediction.* Let ξ be another $1 \times p$ row vector of iid N(0,1) variables, and δ an independent $N(0,\sigma^2)$ variable; δ is unobservable. Suppose

$$\eta = \xi\beta + \delta. \tag{2}$$

The β-vector here is the same as in (1), and is unknown to the statistician. The object is to predict η from ξ, using the β estimated from (1). The explanatory variables ξ are related to the dependent variable and should therefore help in predicting η. This is like a standard problem in econometrics.

Our setup is a statistical cartoon, but it has elements of realism. And in some respects, it provides a favorable environment for conventional methodology. After all, (1) is the textbook regression model: ordinarily, variables will not be normal nor regressions linear. In the simulations, we usually set $\sigma^2=1$, $n=100$ and $p=75$. The number of columns in X may seem large, but in practice an indefinitely large number of covariates present themselves to empirical workers. For example, in typical econometric macro-models, there will be several hundred equations to explain several hundred endogeneous variables, but only several dozen data points. The "constraints," including the decision as to which explanatory variables to put in each equation, are largely data-driven. Also see (Freedman, 1981a) or (Freedman-Rothenberg-Sutch, 1983).

We consider β's of the form $\beta_j=\gamma$ for $j=1,...,p_1$ and $\beta_j=0$ otherwise. The γ's of interest are those near the resolving power of the system, ie, of order $\sigma/\sqrt{n-p}$. Indeed, let V_j be the (j,j)-element of $(X^TX)^{-1}$. On our assumptions, V_j is distributed as $1/\chi^2_{n-p+1}$, and so is of order $1/n-p$.

Denote the columns of X by X_j, for $j=1,...,p$. The screening procedure selects a subset S of these columns to enter the equation, as follows:

Fit Y to X by OLS (ordinary least squares), so $\hat\beta=(X^TX)^{-1}X^TY$, while $\hat\epsilon=Y-X\hat\beta$ is the residual vector, and $\hat\sigma^2=||\hat\epsilon||^2/(n-p)$ is the usual unbiased estimate of σ^2. (3i)

Enter X_1 into the equation automatically. For $j=2,...,p$, enter X_j if $|\hat\beta_j|/\hat\sigma\sqrt{V_j}$ exceeds the 25%-point of the t-distribution with $n-p$ degrees of freedom: recall that V_j is the (j,j)-element of $(X^TX)^{-1}$. Write $j \in S$ if column j was entered. Then S is a random subset of $\{1,...,p\}$ and $1 \in S$. (3ii)

Let X_S be the matrix consisting of the columns of X which were entered in step (ii). Let p_S be the number of such columns. Refit Y on X_S by OLS, so $\hat{\hat\beta}=(X_S^TX_S)^{-1}X_S^TY$. Define $\hat{\hat\epsilon}=Y-X_S\hat{\hat\beta}$, and $\hat{\hat\sigma}^2=||\hat{\hat\epsilon}||^2/(n-p_S)$. For $j \notin S$, we set $\hat{\hat\beta}_j=0$. (3iii)

Now $\hat{\hat\beta}_1$ is an estimate of β_1. And $\xi\hat{\hat\beta}$ predicts the η of (2) from its ξ.

The main performance measures of interest are $MSE=E\{(\hat{\hat\beta}_1-\beta_1)^2\}$ and $MSPE=E\{(\eta-\xi\hat{\hat\beta})^2\}$, the mean square error of estimate and the mean square prediction error, respectively. These may be taken conditionally on X, or unconditionally (averaged over X).

We also consider a version of R^2. For any subset T of columns, let

$$\rho_T^2 = (\sum_{j\in T}\beta_j^2)/(\sigma^2+\sum_{j=1}^{p}\beta_j^2), \tag{4}$$

the true R^2 for a model based on columns in T. Let

$$\phi^2 = E\{\rho_S^2\}. \tag{5}$$

The expectation is over S, the random set of selected columns in (3).

Empirical workers often neglect the randomness in S, treating $\hat{\beta}$ and $\hat{\sigma}^2$ as OLS estimators. In other words, they take the model to be

$$Y = X_S\beta + \epsilon$$

where the ϵ_i's are iid $N(0,\sigma^2)$ variables – but S is the result of the search procedure. Then they use the conventional OLS formulas for MSE, MSPE, and R^2. That is, they estimate the MSE of $\hat{\beta}_1$ by

$$\text{naive MSE} \;=\; \hat{\sigma}^2 \cdot \text{the (1,1)-element of } (X_S^T X_S)^{-1}. \tag{6}$$

Likewise,

$$\text{naive MSPE} \;=\; \hat{\sigma}^2 \cdot \{1 + \text{trace}\,(X_S^T X_S)^{-1}\}. \tag{7}$$

And ρ_S^2 is estimated by $\overline{R}^2$, where

$$1 - \overline{R}^2 \;=\; \frac{n}{n-p_S}(1 - R^2) \;=\; \hat{\sigma}^2/(||Y||^2/n). \tag{8}$$

As will be seen, these estimators tend to be much too optimistic: in effect, they ignore the component of variance due to model selection.

Only the notation in (6-7-8) is unfamiliar. In the OLS context, $E\{\hat{\beta}_1 \mid X\}=\beta_1$. And $\text{var}\{\hat{\beta} \mid X\}=\sigma^2 \cdot (X^TX)^{-1}$ is estimated by putting $\hat{\sigma}^2$ in place of σ^2, giving (6). With respect to (7), if $\tilde{\beta}$ is any estimator for β based on X and Y,

$$E\{(\eta - \xi\tilde{\beta})^2|X\} \;=\; \sigma^2 + E\{||\tilde{\beta} - \beta||^2|X\}. \tag{9}$$

In the OLS case,

$$E\{||\hat{\beta} - \beta||^2|X\} \;=\; \sigma^2 \cdot \text{trace}\,(X^TX)^{-1}$$

and σ^2 is estimated by $\hat{\sigma}^2$. Formula (8) is close to standard, as in (Theil, 1971, p178): by (4), if $T = \{1, \cdots, p\}$,

$$1 - \rho_T^2 \;=\; \sigma^2/(\sigma^2 + \sum_{j=1}^{p} \beta_j^2)\,.$$

Numerator and denominator are estimated separately as $\hat{\sigma}^2$ and $||Y||^2/n$.

Coming now to the jackknife and cross validation, for each i let $Y^{(i)}$ and $X^{(i)}$ denote the result of deleting row i from the matrix. Let $\hat{\beta}^{(i)}$ denote the estimator of β obtained by the screening process (3) applied to the i^{th} reduced data set. Then

$$\text{jackknife MSE} \;=\; \frac{n-1}{n} \sum_{i=1}^{n} [\hat{\beta}_1^{(i)} - \hat{\beta}_1^{(\cdot)}]^2 \tag{10}$$

where

$$\hat{\beta}_1^{(\cdot)} \;=\; \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_1^{(i)}$$

(In principle, the jackknife is only considered to pick up the variance component of MSE.) For cross validation,

$$\text{cross validation MSPE} \;=\; \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{11}$$

where

$$\hat{Y}_i \;=\; (\text{row i of } X) \cdot \hat{\beta}^{(i)}$$

and $\cdot$ stands for inner product. (Despite the notation, $\hat{Y} \neq X\hat{\beta}$.) In particular, the screening process is applied separately to each of the reduced data sets.

Psychologists often use the "cross-validated R^2:" in the present context, this may be taken as

$$(Y \cdot \hat{Y})^2/||Y||^2\,||\hat{Y}||^2 \tag{12}$$

and viewed as an estimate of ρ_S^2 in (4). Here, $\hat{Y}$ is defined as for (11).

Consider next the bootstrap. The idea is to estimate performance characteristics in a simulation model estimated from the data, and two choices present themselves for the parameters: using $\hat{\beta}$ to generate the starred data, or $\hat{\hat{\beta}}$. We elected to use $\hat{\beta}$, and found the bootstrap did not perform well: $\hat{\hat{\beta}}$ would make things even worse; indeed, $\hat{\hat{\beta}}$ can become in effect a self-fulfilling prophecy. Other choices present themselves for the explanatory variables: the bootstrap can be run conditionally by keeping X fixed and resampling the disturbances; or unconditionally, resampling X as well from its distribution, which is known in the present case. The conditional bootstrap seems more interesting, and turns out to perform better, so we report that. In principle, we view the conditional version of the bootstrap as estimating the conditional MSE or MSPE given X. Of course, eg, $E\{E[(\hat{\hat{\beta}}_1 - \beta_1)^2 \mid X]\} = E\{(\hat{\hat{\beta}}_1 - \beta_1)^2\}$. So, if all went well, the conditional bootstrap would also give nearly unbiased estimates of the unconditional MSE or MSPE. A third option – resampling rows – is not available in this problem: there is a high probability of getting fewer than 75 distinct rows in the resampling, so the rebuilt cross-product matrix will usually not be invertible. In any case, the empirical distribution of 100 data points in R^{75} is not a good estimate of the theoretical distribution.

To spell out the bootstrap procedure in more detail, given X and ϵ let $\hat{\beta}$ be the OLS estimate of β in (1). Let

$$Y^* = X\hat{\beta} + \epsilon^* \tag{13}$$

where ϵ^* is an $n \times 1$ vector of iid $N(0,\hat{\sigma}^2)$ variables. In principle, we also consider

$$\eta^* = \xi^*\hat{\beta} + \delta^* \tag{14}$$

where ξ^* is another $1 \times p$ row vector of iid N(0,1) variables, and δ^* is $N(0,\hat{\sigma}^2)$. Pretend for a moment that $\hat{\beta}$ in the model (13) is an unknown parameter vector to be estimated by the selection procedure (3): run Y^* on X to get OLS estimates $\hat{\beta}^*$; let S^* be the set of significant columns, with $1 \in S^*$ by fiat; let $\hat{\hat{\beta}}^* = (X_{S^*}{}^T X_{S^*})^{-1} X_{S^*}{}^T Y^*$.

The bootstrap estimates of the performance measures are as follows:

$$\text{bootstrap MSE} = E_*\{(\hat{\hat{\beta}}_1^* - \hat{\beta}_1)^2\} \tag{15}$$

$$\text{bootstrap MSPE} = E_*\{(\eta^* - \xi^*\hat{\hat{\beta}}^*)^2\} = \hat{\sigma}^2 + E_*\{||\hat{\hat{\beta}}^* - \hat{\beta}||^2\} \tag{16}$$

$$\text{bootstrap } R^2 = E_*\{(\sum_{j \in S^*} \hat{\beta}_j^2)/(\sigma^2 + \Sigma_j\, \hat{\beta}_j^2)\}. \tag{17}$$

In these formulas, X and ϵ are held fixed. As will be seen, the bootstrap estimates of MSE and MSPE are too high. Paradoxically, so is the bootstrap R^2. References are given on the bootstrap, especially (Efron, 1979, 1982). For asymptotic theory, see (Beran, 1982), (Bickel and Freedman, 1981), (Freedman, 1981b); for applications, (Freedman and Peters, 1984abc).

2. Empirical results.

This section reports simulation results for the screening procedure $\hat{\hat{\beta}}$ defined by (3). The naive, bootstrap, and jackknife estimates of squared error will be compared, for estimation (MSE) and prediction (MSPE). The basic model is (1), with 100 rows and 75 columns, so

n=100 and p=75. And $\sigma^2=1$. Take $\beta_j=.2$ for $1\le j\le 25$ and $\beta_j=0$ for $26\le j\le 75$, so $\gamma=.2$ and $p_1=25$. For Table 1, we generated 100 basic data sets following (1): making the number of replicates equal to the number of rows was a matter of taste rather than necessity. The "true value" for $E\{(\hat{\hat{\beta}}_1 - \beta_1)^2\}$ is the empirical average

$$\frac{1}{100}\sum_{r=1}^{100}[\hat{\hat{\beta}}_1(r) - .2]^2$$

where $\hat{\hat{\beta}}_1(r)$ is the computed value of $\hat{\hat{\beta}}_1$ for the r^{th} data set. As shown in the table, this average is .031. The SD of the 100 values $\{\hat{\hat{\beta}}_1(r) : r=1,...,100\}$ is quite large, .039. Still, the SE for the average is .0039. So the instability in the Monte Carlo is small. For the naive MSE, we report the average and SD of the 100 values $\hat{\hat{\sigma}}^2(r)\cdot$ (1,1)-element of $[X(r)_{S(r)}{}^T X(r)_{S(r)}]^{-1}$, with $1\le r\le 100$. As before, $\hat{\hat{\sigma}}^2(r)$ is the value of $\hat{\hat{\sigma}}^2$ for the r^{th} data set, X(r) is the r^{th} matrix of explanatory variables, and S(r) is the set of columns selected by procedure (3) applied to the r^{th} data set. At .012, the naive MSE averages less than half what it should be. For the jackknife MSE, we report the average and SD of the 100 values

$$\text{jackknife MSE}(r) \;=\; \frac{n-1}{n}\sum_{i=1}^{n}[\hat{\hat{\beta}}_1^{(i)}(r) - \hat{\hat{\beta}}_1^{(-)}(r)]^2 \tag{18}$$

for r=1,...,100, which result from applying formula (10) to the r^{th} data set. On average, the jackknife is too high by a factor of about 8. Whether viewed as estimating the conditional or unconditional MSE, the jackknife is not estimating it well.

Finally, for the bootstrap MSE, we report the average and SD of the 100 numbers generated by applying formula (15) to the r^{th} data, for r=1,...,100. On average, the bootstrap is about 15% too high. And there is quite a lot of variability (from one data set to another) in the bootstrap estimate, as will be discussed later.

To approximate $E_*\{[\hat{\hat{\beta}}_1^*(r) - \hat{\beta}_1(r)]^2 \mid X(r)\}$ we generate 100 starred data sets according to (13), with X(r) and $\hat{\beta}(r)$ in place of X and $\hat{\beta}$. (The equality of the number of replications in the various processes is still a matter of choice.) Specifically, for each r we generate 100 vectors of errors, each having 100 iid $N(0,\hat{\sigma}(r)^2)$ components. Corresponding to the s^{th} vector $\epsilon(r,s)$ we make $Y(r,s)=X(r)\hat{\beta}(r)+\epsilon(r,s)$, run Y(r,s) on X(r) according to the screening procedure (3), and come up with $\hat{\hat{\beta}}_1(r,s)$: the vector Y(r,s) is 100×1 and the matrix X(r) is 100×75. Then the bootstrap estimate for the MSE of $\hat{\hat{\beta}}_1$ given X(r) is

$$E_*\{[\hat{\beta}_1^*(r) - \hat{\beta}_1(r)]^2 \mid X(r)\} \approx \frac{1}{100}\sum_{s=1}^{100}\{[\hat{\hat{\beta}}_1(r,s) - \hat{\beta}_1(r)]^2\} \tag{19}$$

The MSPE calculations are similar, and will not be recited in detail. On average, cross-validation does quite well, but the bootstrap is nearly 30% too high. Both show a lot of variability. In the R^2-column, the "true value" is an approximation to $\phi^2=E\{\rho_S^2\}$, obtained by averaging the values for the 100 data sets. As can be seen, the naive estimate is on average more than double the true value, and the bootstrap is worse. Cross validation is low, also by a factor of nearly 2.

Some benchmarks are shown at the bottom of the table. An old-fashioned statistician might elect to estimate β_1 by regressing Y on the first column of X, ie, neglecting the covariates: this procedure, $\hat{\beta}_{\text{no adjust}}$, does quite well, with a variance of .021. Another statistician might put in all the covariates: var $(\hat{\beta}_{OLS})$=.042. This is not so good.

The calculation for the variance of $\hat{\beta}_1$:

var $(\hat{\beta}_1 \mid X) \;=\; \sigma^2 V_1$,

Table 1. Simulation results for the jackknife and bootstrap applied to the screening estimator $\hat{\beta}$. The model specification: $\sigma^2=1$, n=100, p=75, $p_1=25$, $\gamma=.2$

	Estimates of MSE		Estimates of MSPE		Estimates of R^2	
	ave	SD	ave	SD	ave	SD
true	.031	.039	2.79	.63	.243	.064
naive	.012	.003	1.18	.25	.561	.128
jackknife	.233	.113	*	*	*	*
cross validation	*	*	2.70	.57	.152	.107
bootstrap	.036	.015	3.56	.84	.677	.088

$$\text{var}\,(\hat{\beta}_{\text{no adjust}}) = .021 \qquad \text{MSPE}\,(\bar{Y}) = 2.01$$

$$\text{var}\,(\hat{\beta}_1) = .042 \qquad \text{MSPE}\,(\xi\hat{\beta}) = 4.12$$

where V_1 is the (1,1)-element of $(X^TX)^{-1}$ and is distributed as $1/\chi^2_{n-p+1}$ with expectation 1/n-p-1. The computation for var $(\hat{\beta}_{\text{no adjust}})$ is similar, except that σ^2 must be revised upward to $1 + 24 \times (.2)^2 = 1.96$ to account for the omitted covariates, and p=1 not 75.

Why is OLS so bad? In principle, covariate adjustment should improve precision. But there is a tradeoff, since adding variables degrades the quality of the coefficient estimates: roughly speaking, adding an unnecessary variable is like throwing away a data point. See (Breiman and Freedman, 1983) or (Eaton and Freedman, 1982). By comparison, the screening procedure shrinks the estimated coefficients towards 0, and this improves the accuracy relative to OLS. In Table 1, however, the best strategy is still not to adjust at all.

Similar benchmarks are shown for the prediction problem: predicting $\eta=\xi\beta+\delta$ by $\bar{Y}$, ie, ignoring the covariates, has an MSPE of 2.01. In the circumstances, this is the best of the procedures we consider. Predicting η by $\xi\hat{\beta}$, the conventional OLS strategy using all the covariates, has an MSPE of 4.12. This is the worst.

How sensitive are these results to the selected value for β? To address this question we set the common value of β_i for $1 \le i \le 25$ to $\gamma=.1$, .5 and 1.0 as well as to .2. The results are not qualitatively different, except that for large values of γ the R^2's are all close to 1, and the balance tilts toward covariate adjustment. Of course, the independence assumption matters too.

The difficulties in Table 1 are mainly due to the fact that p/n is near 1. To illustrate the point, consider Table 2, where n=100 but p is reduced to 10; the first five β's are set at .2, the others at 0. When p is much smaller than n, the impact of the screening process (3) is small, since at most p/(n-p) of the degrees of freedom for error are being juggled. The naive, bootstrap and cross-validation procedures all give similar results for MSE and MSPE, although R^2 is still hard to estimate.

The jackknife estimate is still too big, by about 50%. We have no explanation to offer; on the other hand, we never understood why the jackknife was supposed to work, except as an approximation to the bootstrap (Efron, 1982, Chapter 6; and see Chapter 4 on bias in the jackknife). We also tried the jackknife on OLS, ie, to estimate var $(\hat{\beta}_1)$. Somewhat to our surprise, the jackknife was still about 20% too high; on the other hand, as theory predicts, the bootstrap came in right on the money.

Table 2. Simulation results for the jackknife and bootstrap applied to the screening estimator $\hat{\beta}$. The model specification. σ^2=1, n=100, p=10, p_1=5, γ=.2

	Estimates of MSE		Estimates of MSPE		Estimates of R^2	
	ave	SD	ave	SD	ave	SD
true	.0097	.0125	1.11	.062	.139	.024
naive	.0108	.0021	1.05	.153	.185	.066
jackknife	.0156	.0068	*	*	*	*
cross validation	*	*	1.14	.184	.103	.071
bootstrap	.0116	.0029	1.15	.170	.211	.068

$\text{var}(\hat{\beta}_{\text{no adjust}}) = .012$ $\quad$ MSPE $(\bar{Y}) = 1.21$

$\text{var}(\hat{\beta}_1) = .011$ $\quad$ MSPE $(\xi\hat{\beta}) = 1.11$

3. Reasons for bootstrap failure.

Although $\hat{\beta}$ is an unbiased estimator of β, there is bias in $||\hat{\beta}||^2$, conditionally or unconditionally:

$$E\{||\hat{\beta}||^2|X\} = ||\beta||^2 + \sigma^2 \cdot \text{trace}\,(X^TX)^{-1}$$

$$E\{||\hat{\beta}||^2\} = ||\beta||^2 + p\sigma^2/(n-p-1).$$

In other words, the bootstrap model (13) starts from a parameter vector with a much inflated length. In Table 1, for example, $||\beta||^2$=1 and $p\sigma^2/(n-p-1) \approx 3$. This explanation for bootstrap failure suggested deflating $\hat{\beta}$ by the appropriate factor (namely, $[1 + \hat{\sigma}^2 \cdot \text{trace}\,(X^TX)^{-1}/||\hat{\beta}||^2]^{1/2}$) to get its length about right, before resampling. For the model in Table 1, length adjustment does bring the bootstrap into better line: see Table 3.

We also tried a model with β_j=.2 for j=1,...,75 so p_1=75. See Table 4. In this case, adjustment makes things worse on MSE and MSPE: indeed, the raw bootstrap is already biased downward. For R^2, the adjustment helps. We do not recommend length adjustment without further analysis.

Table 3. Simulation results for the raw and length-adjusted bootstrap on the screening estimator $\hat{\beta}$: the model specification is as in Table 1.

	Estimates of MSE		Estimates of MSPE		Estimates of R^2	
	ave	SD	ave	SD	ave	SD
true	.031	.039	2.79	.63	.243	.064
raw bootstrap	.036	.015	3.56	.84	.677	.088
adjusted bootstrap	.025	.011	2.48	.50	.167	.157

Table 4. Simulation results for the raw and length-adjusted bootstrap on the screening estimator $\hat{\beta}$. Model specification: σ^2=1, n=100, p=75, p_1=75, γ=.2

	Estimates of MSE		Estimates of MSPE		Estimates of R^2	
	ave	SD	ave	SD	ave	SD
true	.057	.087	4.63	.81	.353	.065
raw bootstrap	.045	.022	4.21	.85	.800	.057
adjusted bootstrap	.037	.019	3.48	.56	.352	.154

With respect to the model in Table 1, denote the MSE given X by

$$\Phi(\beta,\sigma,X) \;=\; E\{(\hat{\beta}_1 - \beta_1)^2 \mid X\}. \tag{20}$$

The bootstrap approximates $\Phi(\beta,\sigma,X)$ by $\Phi(\hat{\beta},\hat{\sigma},X)$, and in effect our tables on MSE compare $E\{\Phi(\beta,\sigma,X)\}$ to $E\{\Phi(\hat{\beta},\hat{\sigma},X)\}$. In principle, Φ depends on all the coordinates of β, and in this respect screening differs from OLS, where $E\{(\hat{\beta}_1 - \beta_1)^2\}$ does not depend on β.

One explanation for bootstrap failure is strong nonlinear dependence of Φ on β. About the strongest we found was on $||\beta||^2$. To represent the data more conveniently, let

$$\Phi(\beta,\sigma) \;=\; E\{\Phi(\beta,\sigma,X)\} \;=\; E\{(\hat{\beta}_1 - \beta_1)^2\}\,. \tag{21}$$

This is the unconditional MSE. Figure 1 shows a plot of $\Phi(\beta,\sigma)$ against $||\beta||^2$ or σ^2. (The values of β and σ were drawn as a sample from the OLS distribution of $\hat{\beta}$ and $\hat{\sigma}$; computationally, we estimated $\Phi(\hat{\beta},\hat{\sigma})$ by running the unconditional bootstrap.) By regression,

$$\Phi(\hat{\beta},\hat{\sigma}) \;=\; .0035 \times ||\hat{\beta}||^2 + .026 \times \hat{\sigma}^2 + \text{residual}, \qquad R^2 = .70 \tag{22}$$

Since $||\hat{\beta}||^2$ tends to be too big, this does inflate the bootstrap estimate of MSE, as indicated at the beginning of the section – for the model in Table 1.

Switching now from estimation to prediction, a heuristic explanation for the bias in the bootstrap R^2 and MSPE runs as follows. Keeping σ^2 fixed, R^2 measures how big the β's are, and the MSPE measures how well they are estimated. The $\hat{\beta}$'s tend to be too big, inflating R^2. On the other hand, when a big $\hat{\beta}_i$ is estimated as 0 by the corresponding bootstrap $\hat{\beta}_i^*$, that is a big error.

To quantify the effect, for any subset H of columns let H^1=H while H^0 is the complement of H. Let J be the set of columns j with $1 \le j \le 25$, so $\beta_j = \gamma$ is positive for $j \in J^1 = J$ while $\beta_j = 0$ for $j \in J^0$. Recall the set S of selected columns from (3) and S^* from the discussion before (15). For a,b=0 or 1 let

$$E_{ab} \;=\; E\left\{\Sigma_{j\in J^a \cap S^b}(\hat{\beta}_j - \beta_j)^2\right\} \quad \text{and} \quad E^*_{ab} \;=\; E_*\left\{\Sigma_{j\in J^a \cap S^{*b}}(\hat{\beta}_j^* - \hat{\beta}_j)^2 \mid X\right\}. \tag{23}$$

Starting from equation (9),

$$\text{MSPE} \;=\; \sigma^2 + E_{11} + E_{01} + E_{10} + E_{00}.$$

Likewise from (16),

$$\text{bootstrap MSPE} \;=\; \hat{\sigma}^2 + E^*_{11} + E^*_{01} + E^*_{10} + E^*_{00}.$$

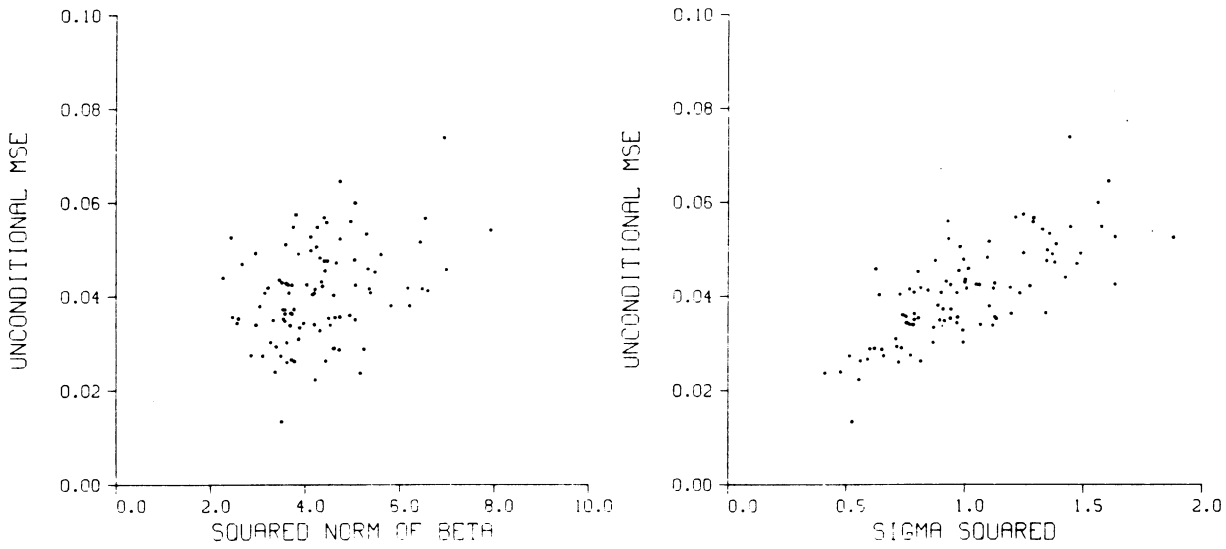

Figure 1. Plot of unconditional MSE against $||\beta||^2$ and σ^2, for a sample of β's and σ's.

The average values for these 5 components of MSPE are shown in Table 5. As will be seen, most of the bias in the bootstrap can be accounted for by the last row in the table corresponding to j's which have $\beta_j=0$ in the true model and are screened out of the bootstrap model: $26 \le j \le 75$ and $j \notin S^*$. Of course, such j's have $\hat{\beta}_j \neq 0$, and that is the problem in Table 1. Indeed, E_{00} is necessarily 0 while E_{00}^* is quite positive. By contrast, $E_{00} = E_{00}^* = 0$ in Table 4, where the bootstrap is biased downward.

Table 5. Simulation results for the components of MSPE and bootstrap MSPE: the model specification is as in Table 1.

	True	Bootstrap
variance	1.00	0.96
$1 \le j \le 25$ and j selected	0.40	0.53
$26 \le j \le 75$ and j selected	0.88	0.92
$1 \le j \le 25$ and j not selected	0.51	0.43
$26 \le j \le 75$ and j not selected	0.00	0.72
total	2.79	3.56

4. Other findings.

a) *The conditional MSE.* Table 1 shows the unconditional average and SD of $(\hat{\beta}_1 - \beta_1)^2$, as .031 ± .039. For each of the 100 data sets r=1,...,100 in the simulation, consider the conditional mean square error $E\{(\hat{\beta}_1 - \beta_1)^2|X(r)\}$. To estimate this conditional expectation, we generated for each r a set of 100 vectors of errors, each vector having 100 iid N(0,1) components. Corresponding

to the s^{th} vector $\epsilon(r,s)$, we made $Y(r,s)=X(r)\beta + \epsilon(r,s)$ and applied the screening process (3) to Y(r,s) and X(r), winding up with $\hat{\beta}(r,s)$. The conditional MSE of $\hat{\beta}_1$ given X(r) can now be estimated as

$$\text{MSE}(r) = \frac{1}{100}\sum_{s=1}^{100}[\hat{\beta}_1(r,s) - .2]^2. \tag{24}$$

These 100 conditional MSE's averaged out to .028, with an SD of .0084. The difference between .028 and .031 $\approx E\{(\hat{\beta}_1 - \beta_1)^2\}$ is sampling error, and the .028 is more reliable. Indeed, the difference between .0084 and .039 $\approx$ SD of $(\hat{\beta}_1 - \beta_1)^2$ shows how conditioning on X dramatically reduces the variability in $(\hat{\beta}_1 - \beta_1)^2$.

For each data set r, we previously computed in (19) the bootstrap estimate for the MSE of $\hat{\beta}_1$ given X(r), starting from $\hat{\beta}$ and $\hat{\sigma}$ rather than β and σ. A scatter plot of the bootstrap estimate against the conditional MSE across data sets is shown in the left hand panel of Figure 2; a similar plot for the jackknife is shown at the right. As will be clear, the bootstrap is consistently too high, by a little. The jackknife is an order of magnitude too big. Furthermore, the R^2 for the bootstrap is only 0.36; for the jackknife, 0.19. In addition to other troubles, these methods cannot discriminate very well between informative and uninformative data sets. (There is no real attenuation due to imprecision in the Monte Carlo.)

b) *Outliers.* As will be clear from Figure 2, the jackknife estimate has quite a long right hand tail. On the log scale in the right hand panel of Figure 3, the bias is still plain to see. The left hand panel gives a scatter plot for the log bootstrap; this looks quite normal, but R^2 is only 0.25. By regression,

$$\text{bootstrap estimate} = .43 \times (\text{true MSE given X})^{.71} \times \text{residual factor} \tag{25}$$

The small bias in the bootstrap can still be discerned; a majority of the points are above the 45-degree line.

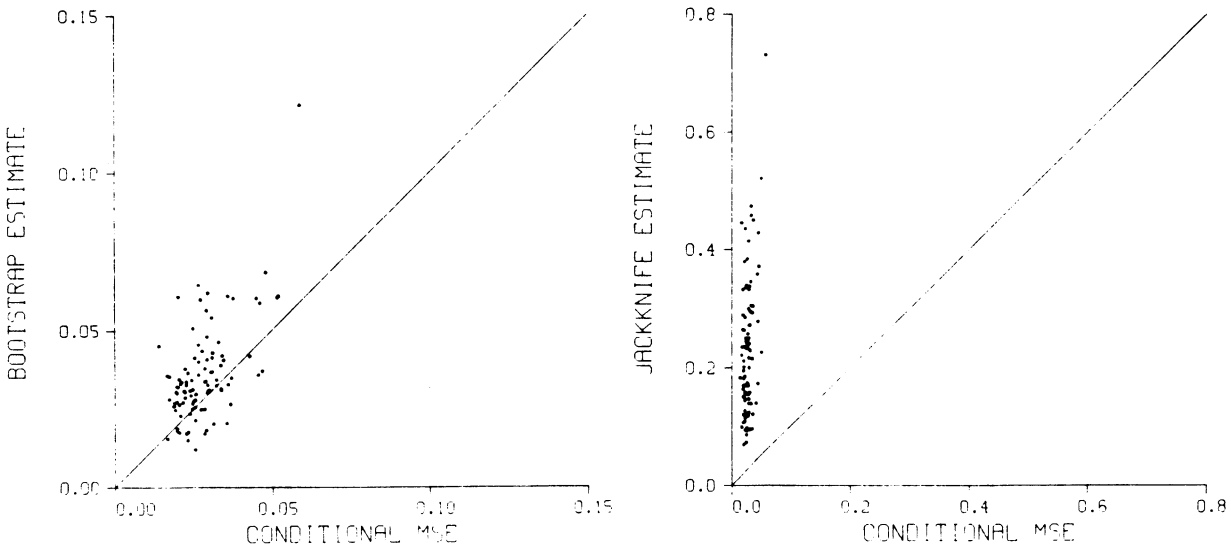

Figure 2. The left hand panel plots for each of 100 data sets the bootstrap estimate of mean square error against its true value (conditional on X). The right hand panel does the same for the jackknife. The scales differ. The 45-degree line is plotted for reference.

Table 6. Root mean square error for various estimates of MSE and MSPE: the model specification is as in Table 4.

	Estimates of MSE	Estimates of MSPE
naive	.018	1.77
jackknife	.232	*
cross validation	*	0.81
bootstrap	.014	1.36

c) *RMS error.* As another measure for the accuracy of the naive, jackknife and bootstrap MSE, we took the root mean square difference between each of these estimates and the true MSE conditional on X, over the 100 data sets in the simulation discussed in paragraph a). The results are shown in Table 6. The bootstrap is only a little better than the naive estimate: increased variability trades off against decreased bias. Table 6 also shows the results for MSPE. Here, the cross validation estimator is superior. The bootstrap estimates are not bad, on average (Table 1). But they are quite noisy: that is the message of this paragraph.

d) *Bias in the screening estimator.* When averaged over X, the screening estimator $\hat{\hat{\beta}}_1$ is unbiased by symmetry. Indeed, the first column of X is entered automatically; now project into its orthocomplement and use rotational invariance. However, $\hat{\hat{\beta}}_1$ is conditionally biased given X. For the simulation discussed in paragraph a), $E(\hat{\hat{\beta}}_1|X)$ averaged .20 with an SD of .038; the SD measures the conditional bias for a typical X as about 20% of the true value.

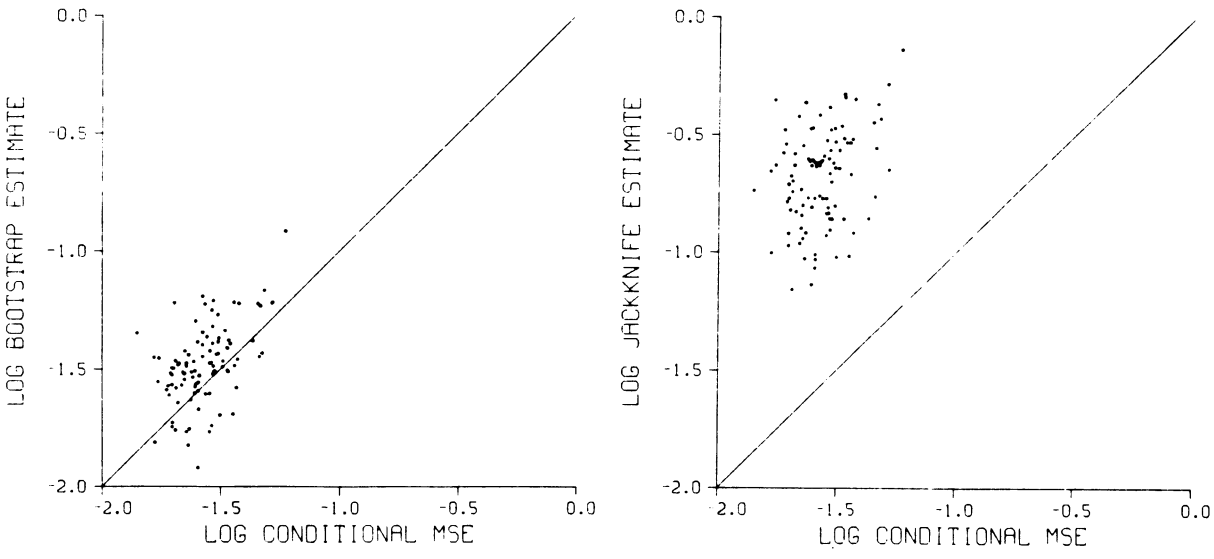

Figure 3. The left hand panel plots for each of 100 data sets the log of the bootstrap estimate of MSE against the log of the true value conditional on X. The right hand panel does the same for the jackknife. The logarithms are to base 10.

For columns 2 through 25, the screening estimator is biased toward 0, conditionally or unconditionally. For example, a hundred values of $\hat{\hat{\beta}}_2$ averaged .14 with an SD of .025 and an SE of .0025; again, the true value is .20.

e) *The effect of refitting.* Any two columns of X are nearly orthogonal, so the effect of refitting in (3) should be minimal: ie, $\hat{\hat{\beta}}_j \approx \hat{\beta}_j$ for $j \in S$. To test this idea, we compared $\Sigma_{j\in S}\, \hat{\hat{\beta}}_j^2$ with $\Sigma_{j\in S}\, \hat{\beta}_j^2$. In the simulation for Table 1, the first sum averages 2.2 and the second, 3.3. Refitting matters; there are a lot of pairs of columns and the non-orthogonality mounts up.

f) *F-tests for omitted variables.* It has been suggested that our results are due to model mis-specification, which could be detected by a routine F-test. We disagree. The explanation for our results is chance capitalization: data-snooping distorts conventional measures of goodness-of-fit. Indeed, the F-test cannot detect the miss-specification. To illustrate the point, consider the simulation for Table 1. This involved generating 100 data sets following the model (1); and for each, performing the screening operation (3), leading to a set S of selected columns. For each data set, we ran a naive F-test for adding *en bloc* the columns outside S. On the average, the F-statistic was .9, with 49.4 degrees of freedom in the numerator and 25 in the denominator. (Also see Table 7 below.) This would only confirm the value of the screening procedure. Of course, it is misleading to make F-tests this way, treating S as given rather than the result of data-snooping.

g) *How many variables get into the second pass?* In the simulation for Table 1, the coefficients of the first 25 columns were set to a common positive value; these will be called 1-columns. The remaining 50 columns had coefficients set to 0, and will be referred to as 0-columns. Let N_1 be the number of 1-columns which got into the second-pass regression. Likewise, let N_0 be the number of 0-columns which entered the second-pass regression. The bootstrap analogs will be denoted by stars: thus, N_0^* is the number of columns with $25 \le j \le 75$ which entered the second-pass bootstrap regression. (Of course, $\hat{\beta}_j \neq 0$ even for the 0-columns.)

Means for these N's are shown in Table 7, for a simulation involving 100 data sets. For example, we expect $.25 \times 50 = 12.5$ of the 0-columns to get in, and on the average 13.4 did: the difference is sampling error. (Since X is not exactly orthogonal, the $\hat{\beta}$'s are dependent, and the variability in N_0 is appreciably greater than binomial.)

On the average, 12.2 of the 1-columns got into the second-pass regression. This is only 49% of the 1-columns, which may seem disappointing, but in the present context even a test of size 25% does not have much power. The bootstrap estimates this quite well: $E(N_1^*)=13.8$. However, the bootstrap badly over-estimates the number of 0-columns: 21.4 versus 13.4. This is because the $\hat{\beta}$'s tend to be too large, so the $\hat{\beta}^*$'s are more likely to be significant.

Table 7. Simulation results for the number of variables entering the second pass: the model specification is as in Table 1.

	1-columns	0-columns	total
true	12.2	13.4	25.6
bootstrap	13.8	21.4	35.2

Dijkstra (as reported in these proceedings) had a sharper result for a smaller model. To replicate his work, we repeated our simulation for a model with five 1-columns and five 0-columns. The results are shown in Table 8: the bootstrap is over 50% too high on the 0-columns.

A small theoretical calculation might clarify matters. Consider the very simple regression model

$$Y_i = \beta x_i + \epsilon_i \tag{26}$$

where the ϵ_i are iid N(0,1) for i=1,...,n. Here, β is just a number. The x's are deterministic, and normalized so $\Sigma_1^n x_i^2 = n$. Fix a critical value c and let

$$\Phi(\beta) = \Pr\{|\hat{\beta}| > c/\sqrt{n}\}. \tag{27}$$

Of course, this Φ can be computed exactly from the normal distribution, since $\sigma^2=1$ is given:

$$\Phi(\beta) = \Pr\{|\beta\sqrt{n} + Z| > c\} \tag{28}$$

where Z is N(0,1). Indeed, $\hat{\beta}$ is distributed as $\beta + (Z/\sqrt{n})$.

Now we try to estimate Φ by the bootstrap:

$$\Phi(\hat{\beta}) = \Pr\{|\hat{\beta}\sqrt{n} + Z'| > c\} \tag{29}$$

where Z' is an independent N(0,1) variable, and Z is held fast. Finally,

$$E\{\Phi(\hat{\beta})\} = \Pr\{|\beta\sqrt{n} + Z + Z'| > c\} \tag{30}$$

where Z and Z' both vary. If β is of order $1/\sqrt{n}$ or smaller, the bootstrap will fail: $Z + Z'$ has fatter tails than Z, by a lot. If $\beta\sqrt{n} \to \infty$, then $\Phi(\beta)$ and $E\{\Phi(\hat{\beta})\}$ will both approach 1, but at different rates.

Table 8. Simulation results for the number of variables entering the second pass. The model specification: $\sigma^2=1$, n=100, p=10, $p_1=5$, $\gamma=.2$

	1-columns	0-columns	total
true	4.2	1.3	5.5
bootstrap	3.9	2.1	6.0

5. Computational details.

The program was written in FORTRAN, using LINPAK for the matrix algebra. The computations were done on a CRAY. Those for the model in Table 1, for example, took 10 minutes of CPU time. Among other things, there were a hundred 75×75 matrices to invert, and upwards of 50,000 regressions to run. (Cross-validation was done by updating X^TX: see Efron, 1982, p18). Some of calculations were replicated on a SUN workstation, in FORTRAN and in S. A few of them were replicated in True BASIC on a PC-XT. We therefore have some degree of confidence in the code. Too, exact distributions for many of the intermediate results can be computed and checked against observations. On the whole, this worked out quite well; there were a few small but highly significant anomalies. Of course, we are pushing the random number generator quite hard: Table 1 involves over a million calls.

6. Summary and conclusions.

In our simulations, when the number of variables is relatively large the bootstrap and particularly the jackknife have some trouble in dealing with uncertainty created by variable selection. It may not be possible on the basis of such techniques to develop a model and calculate its performance characteristics on the same data set. This would have gloomy implications for many kinds of modeling. Of course, an investigator can always develop the model on one data set and test it on another: replication is always a good idea.

In the classical setup, given some type of relationship among variables expressed in a well-specified statistical model, it is possible to estimate parameters or make predictions from a data set and put margins of error on the results. If you know what to look for, there is a way to find it. On the other hand, given any statistical procedure there will always be some kinds of relationships which will not be detected by that procedure. And someone who uses a variety of statistical procedures, taking many cuts at the data, is almost bound to find structure even when none exists. That is the trouble with data-snooping.

To illustrate the point that given some style of analysis there will be structure which escapes it, take linear regression analysis. Consider the time series x_t plotted against time $t=1, \cdots, 50$ at the left in Figure 4. This looks like pure noise, and fitting $x_t=a + bt + e_t$ isolates no trend. On the other hand, plotting x_t against x_{t-1} at the right shows this series to be perfectly deterministic: $x_t=f(x_{t-1})$, where

$$\begin{aligned} f(x) &= 2x \qquad \text{for } 0 \le x \le 1/2 \\ &= 2 - 2x \quad \text{for } 1/2 \le x \le 1. \end{aligned}$$

A major part of the problem in applications is the curse of dimensionality: there is a lot of room in high-dimensional space. That is why investigators need model specifications tightly deri-

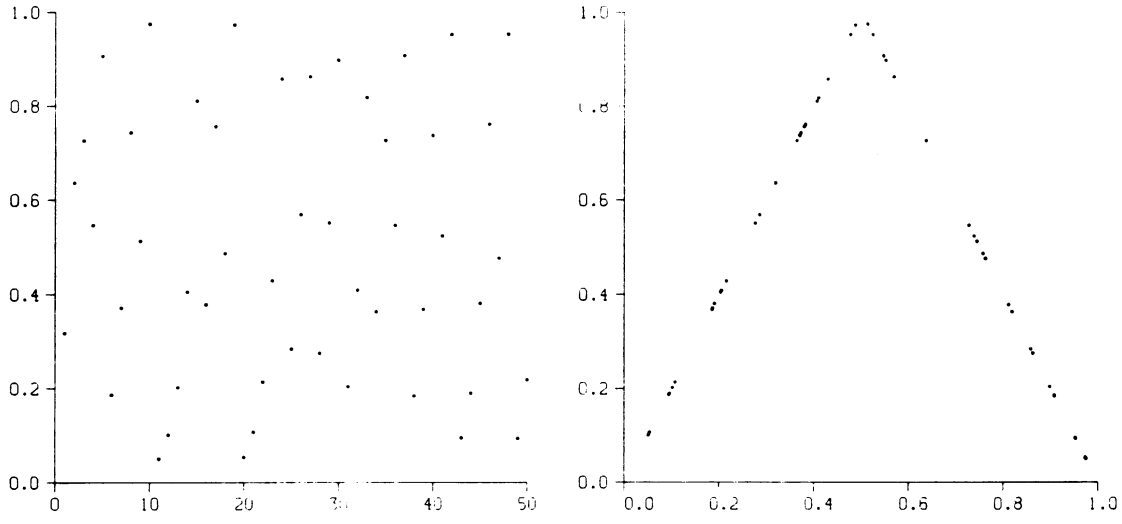

Figure 4. At the left, a time series with no linear regression structure. At the right, plotting x_t against x_{t-1} reveals the structure.

ved from good theory. We cannot expect statistical modeling to perform at all well in an environment consisting of large, complicated data sets and weak theory. Unfortunately, at present that describes many applications. References are given on modeling issues, eg, (Achen, 1982), (Baumrind, 1983), (Daggett and Freedman, 1985), (de Leeuw, 1985), (Freedman, 1985, 1986), (Freedman-Rothenberg-Sutch, 1983), (Hendry, 1980), (Leamer, 1983), (Ling, 1983), (McNees, 1986), (Zarnowitz, 1979).

Disclosures

Rudy Beran remarks that chance capitalization is a problem, even for bootstrap studies of chance capitalization. In principle, this is certainly right. However, in this paper we took our own advice about replication. We spent several months on free-style data snooping. Then we wrote a draft of the paper, with blank spaces for all the empirical numbers. Then we made a fresh set of computer runs and filled in those blanks. Finally, we ate all the words that had to be eaten.

Acknowledgements

We would like to thank Theo Dijkstra for his work in putting these proceedings together. He, Rudy Beran, Lincoln Moses and Jamie Robins made useful comments. Ani Adhikari provided lots of last-minute technical help. Our research was partially supported by NSF Grant DMS86-01634.

References

Achen, C. (1982). *Interpreting and using regression.* Beverly Hills, Calif.: Sage.

Baumrind, D. (1983). Specious causal attribution in the social sciences: the reformulated stepping-stone theory of heroin use as exemplar. *J. Pers. Soc. Psych.*, **45**, 1289-98.

Beran, R. (1984). Jackknife approximations to bootstrap estimates. *Ann. Statist.*, **12**, 101-118.

Bickel, P. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196-1217.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Am. Stat. Assoc.*, **78**, 131-136.

Daggett, R. and Freedman, D. (1985). Econometrics and the law: a case study in the proof of antitrust damages. In L. LeCam and R. Olshen (Eds.), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol I, 126-75. Belmont, Calif.: Wadsworth.

de Leeuw, J. (1985). Review of books by Long, Everitt, Saris and Stronkhorst. *Psychometrika*, **50**, 371-5.

Eaton, M. and Freedman, D. (1982). A remark on adjusting for covariates in multiple regression. Technical Report No. 11, Department of Statistics, University of California, Berkeley.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM.

Freedman, D. (1981a). Some pitfalls in large econometric models: a case study. *J. Bus.* **54**, 479-500.

Freedman, D. (1981b). Bootstrapping regression models. *Ann. Statist.*, **9**, 1218-1228.

Freedman, D. (1983). A note on screening regression equations. *Am. Stat.*, **37**, 152-5.

Freedman, D. (1985). Statistics and the scientific method. In W. Mason and S. Fienberg (Eds.), *Cohort Analysis in Social Research: Beyond the Identification Problem*, 345-390 (with discussion). New York: Springer.

Freedman, D. (1986). As others see us: a case study in path analysis. Technical report, Department of Statistics, University of California, Berkeley. To appear in *J. Ed. Stat.*.

Freedman, D. and Navidi, W. (1986). Regression models for adjusting the 1980 Census. *Stat. Sci.*, **1**, 1-39.

Freedman, D. and Peters, S. (1984a). Some notes on the bootstrap in regression problems. *J. Bus. Econ. Stat.*, **2**, 406-409.

Freedman, D. and Peters, S. (1984b). Bootstrapping a regression equation: some empirical results. *J. Am. Stat. Assoc.*, **79**, 97-106.

Freedman, D. and Peters, S. (1984c). Bootstrapping an econometric model: some empirical results. *J. Bus. Econ. Stat.*, **2**, 150-8.

Freedman, D. and Peters, S. (1985). Using the bootstrap to evaluate a forecasting equation. *J. Forecasting*, **4**, 251-262.

Freedman, D., Rothenberg, T. & Sutch, R. (1983). On energy policy models. *J. Bus. Econ. Stat.*, **1**, 24-36. (With discussion.)

Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *J. Am. Stat. Assoc.*, **81**, 108-113.

Hendry, D. (1980). Econometrics – alchemy or science? *Econometrica*, **7**, 387-406.

Leamer, E. (1983). Taking the con out of econometrics. *Am. Econ. Rev.*, **73**, 31-43.

Ling, R. (1983). Review of *Correlation and Causation* by Kenny. *J. Am. Stat. Assoc.*, **77**, 489-91.

Lovell, M. (1983). Data mining. *Rev. Econ. Statist.*, **LXV**, 1-11.

McNees, S.K. (1986). Forecasting accuracy of alternative techniques: a comparison of US macroeconomic forecasts. *J. Bus. Econ. Stat.*, **4**, 5-24. (With discussion.)

Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.

Zarnowitz, V. (1979). An analysis of annual and multiperiod quarterly forecasts of aggregate income, output, and the price level. *J. Bus.*, **52**, 1-34.