Department of Statistics
University of California, Berkeley
Technical Report #1

How Many Variables Should Be Entered in a Regression Equation?

L. BREIMAN and D. FREEDMAN

Reprinted from the Journal of the American Statistical Association
March 1983, Volume 78, Number 381

How Many Variables Should Be Entered in a Regression Equation?

L. BREIMAN and D. FREEDMAN*

The optimal number of regressors is determined to minimize mean squared prediction error and is shown to be a small fraction of the number of data points. As the number of regressors grows large, the S_p criterion provides an asymptotically optimal rule for the number of variables to enter.

KEY WORDS: Regression; Stepwise regression; Best subsets regression; Prediction error.

1. INTRODUCTION

Consider the model

$$Y = \sum_{j=1}^{\infty} \beta_j X_j + \epsilon, \tag{1.1}$$

where the X's and ϵ are jointly Gaussian: ϵ is independent of the X's; the X's may be correlated among themselves, but only imperfectly; all variables have mean zero; and the sum converges in L_2. The $\{\beta_j\}$ are unknown, as is the covariance structure of the X's and the variance of ϵ.

A statistician is given n independent replicates of Y's and X's satisfying (1.1). More specifically, suppose

$$\{Y_i; X_{ij}, j = 1, 2, \ldots; \epsilon_i\}$$

are independent for $i = 1, \ldots, n$; for each i, these variables are distributed like $\{Y; X_j, j = 1, 2, \ldots; \epsilon\}$ of (1.1); in particular,

$$Y_i = \sum_{j=1}^{\infty} \beta_j X_{ij} + \epsilon_i.$$

The statistician chooses a positive integer p, enters the first p variables in the order preassigned above; that is, enters $X_1, X_2, \ldots, X_p$, regresses Y on these p variables, and gets ordinary least squares estimates $\hat\beta_1, \ldots, \hat\beta_p$. Abbreviate $Y = Y(n)$ for the column n-vector whose ith entry is Y_i; and $X = X(n, p)$ for the $n \times p$ matrix whose ijth entry is X_{ij}; and $\hat\beta = \hat\beta(n, p)$ for the column p-vector whose ith entry is $\hat\beta_i$. Then $\hat\beta = (X^TX)^{-1}X^TY$. To show the dependence on n and p, we write

$$\hat\beta(n, p) = [X(n, p)^TX(n, p)]^{-1}X(n, p)^TY(n).$$

Now an $(n + 1)$st copy of Y and $\{X_j\}$ is made, independent of the first n, to be denoted by $Y_{n+1}, X_{n+1,j}, j = 1, \ldots$. The statistician predicts Y_{n+1} by

$$\hat Y_{n+1} = \sum_{j=1}^{p} \hat\beta_j X_{n+1,j}.$$

(The dependence on p is suppressed in the notation.) Loss is measured by the squared prediction error $(Y_{n+1} - \hat Y_{n+1})^2$.

The following two notions are relevant. The conditional mean squared prediction error is defined as

$$M = M_{np} = E\{(Y_{n+1} - \hat Y_{n+1})^2 \mid Y_i \text{ and } X_{ij}, \text{ for all } j \text{ and } i = 1, \ldots, n\}. \tag{1.2}$$

The unconditional mean squared prediction error is

$$U = U_{np} = E\{M_{np}\}. \tag{1.3}$$

Thus, U is M averaged over the data. Asymptotically, as will be seen, $M \doteq U$ for nearly all configurations of the data: $\doteq$ means nearly equal and is used only informally.

The basic question of this article is how to choose p so as to minimize M or U. It is to be noted that the models are nested in p. Section 2 solves this problem from the point of view of an omniscient statistician who knows the parameters. To state the result, let

$$\sigma^2 = \text{var } \epsilon$$

$$\sigma_p^2 = \text{var}\left\{ \sum_{j=p+1}^{\infty} \beta_j X_j \;\middle|\; X_1, \ldots, X_p \right\}.$$

Since the X's are Gaussian, σ_p^2 is not random.

Theorem 1.1. Under the foregoing conditions, if $p \leqq n - 2$,

$$U_{np} = (\sigma^2 + \sigma_p^2)\left(1 + \frac{p}{n - 1 - p}\right). \tag{1.4}$$

There is a $p^* = p^*(n)$ minimizing this expression; for any such minimizer, $p^*(n)/n \to 0$ as $n \to \infty$. (If $p \geqq n - 1$, then $U_{np} = \infty$.)

If p is much smaller than n, then $p/(n - 1 - p) \doteq p/n$, so

$$U_{np} \doteq \sigma^2 + \sigma_p^2 + \sigma^2 p/n. \tag{1.5}$$

* L. Breiman and D. Freedman are Professors, Department of Statistics, University of California, Berkeley, CA 94720. The authors are indebted to Professor M.L. Eaton for the elegant proof of Theorem 1.3 and also thank Dr. Louis Gordon (EIA) for some useful discussions. Research was partially supported by National Science Foundation Grant MCS-80-02535.

Reprinted from: © Journal of the American Statistical Association
March 1983, Volume 78, Number 381
Theory and Methods Section

The first term in (1.5) measures the effect of ϵ_{n+1} on the prediction error and gives a fixed minimum for U_{np}. The second term measures the effect of the omitted variables X_j for $j > p$. This term decreases as p increases. The third term measures the effect of random error on the coefficient estimates $\hat{\beta}_j$. This term increases as p increases. It reflects the often ignored fact that putting additional variables into the equation introduces additional random error into the coefficient estimates. Since σ_p^2 decreases with p and $\sigma^2 p/n$ increases, there is an optimal p; this was denoted by p^* in the theorem.

Example. Take the X's independent with common variance v^2, and $\beta_j = j^{-\alpha}$ where $\alpha > \frac{1}{2}$. Then $p^*(n) = (nv^2/\sigma^2)^{1/(2\alpha)}$. For $\alpha = 1$ and $\sigma = v = 1$, the optimal U is nearly $1 + 2/\sqrt{n}$. If p is taken as $n/2$, a not uncommon choice in applied work, then U is nearly 2. With too many variables in the equation, the mean squared prediction error is unnecessarily large.

Since statisticians are seldom omniscient, the question arises how to estimate p^* without knowing the parameters of the equation. An answer is given in Sections 3 and 4, as is now outlined. The regression mean squared error, with the first p variables entered, is

$$R_{np} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

where as before $\hat{Y}_i = \sum_{j=1}^{p}\hat{\beta}_j X_{ij}$ depends on n and p. Notice that R_{np} estimates $\sigma^2 + \sigma_p^2$ in (1.4).

The S_p Criterion. Let $\hat{p}(n)$ be the smallest $p \leqq n/2$

$$\text{that minimizes } \hat{U}_{np} = R_{np}\left(1 + \frac{p}{n-1-p}\right). \quad (1.6)$$

Enter the first $\hat{p}(n)$ variables.

Notice that $\hat{p}(n)$ depends only on n and the data. Then the S_p criterion works almost as well as the optimal rule. To state this clearly, recall the conditional mean squared prediction error M_{np} from (1.2). Let $\tilde{p}(n)$ minimize M_{np}.

Theorem 1.2. Assume $\sigma_p^2 > 0$ for all p. As $n \to \infty$, in probability,

(a) $[M_{n\tilde{p}(n)} - \sigma^2]/[U_{np^*(n)} - \sigma^2] \to 1$
(b) $[M_{n\hat{p}(n)} - \sigma^2]/[U_{np^*(n)} - \sigma^2] \to 1$.

Informally, for most large Gaussian data sets, a statistician who uses the S_p criterion, estimating $\hat{p}$ from the data, gets just about the same conditional mean squared prediction error as an omniscient statistician who uses the optimal $\tilde{p}$; and the optimal conditional mean squared prediction error is about the same as the unconditional obtained using p^*.

Although the S_p criterion is asymptotically efficient, $\hat{U}_{n\hat{p}(n)} - \sigma^2$ need not be a good estimate of the optimal $M_{n\tilde{p}(n)} - \sigma^2$, $U_{np^*(n)} - \sigma^2$, or $M_{n\hat{p}(n)} - \sigma^2$. Indeed, in Theorem 1.4 the term

$$\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2)$$

creates a random error of order $1/\sqrt{n}$; fortunately, this error does not depend on p.

To prove Theorem 1.2, it is necessary to estimate M_{np} and R_{np}.

Theorem 1.3. Suppose $p \leqq n$. Then M_{np} is distributed as

$$(\sigma^2 + \sigma_p^2)(1 + \chi_p^2/\chi_{n-p+1}^2),$$

the two chi-squared variables being independent.

Theorem 1.4. Assume $\sigma_p > 0$ for all p. Then

$$R_{np} = \sigma^2 + \sigma_p^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2) + \theta_{np}(\sigma_p^2 + \sigma^2 p/n)$$

where the maximum of $|\theta_{np}|$ over $1 \leqq p \leqq \frac{1}{2}n$ tends to zero in probability as n tends to infinity.

The present results are restricted to the Gaussian case, although at least one of us believes that extension to non-Gaussian variables is possible. The restriction that $\sigma_p^2 > 0$ for all p is irksome, because it rules out the important special case in which $\beta_p = 0$ for $p \geqq p_0$. We have some hopes that this latter case can be treated by a backward sequence of F tests. Bootstrap and cross-validation techniques may also help. However, counterexamples show that Theorems 1.2 and 1.4 fail if $\sigma_p = 0$ for some p: then the S_p criterion does not pick off a nearly optimal p.

Our work is closely related to that of Thompson (1978) who derives Theorems 1.1 and 1.3 in slightly different form. Our proofs have been included since they are short and make the article self-contained. The criterion (1.6) is the S_p criterion, which is first given explicitly by Hocking (1976) and further explored by Thompson (1978). Both authors come upon S_p by the heuristic replacement of $\sigma^2 + \sigma_p^2$ in Theorem 1.1 by R_{np}. As we pointed out, it is almost accidental that this works. It does work, not because R_{np} is a good approximation to $\sigma^2 + \sigma_p^2$, but instead because the dominant error term does not depend on p.

There is some similarity in spirit between S_p and the C_p of Mallows (1973), but the two criteria are different. Krieger and Pickands (1981) suggest still another criterion equivalent to S_p but again without a direct proof of optimality. The criterion can be applied to models used in clinical trials; see Freedman and Moses (1981), where loss is measured by the variance of the estimator for the main effect.

The S_p criterion is related to work that has been done on selecting the order of an autoregressive model. The final prediction error of Akaike (1970) is asymptotically equivalent, as is the information criterion in Akaike (1974). The recent work of Shibata (1980) uses the prediction error criterion in a way similar to ours in this article and establishes asymptotic optimality for another expression asymptotically equivalent to S_p. Another related paper is Shibata (1981), where the design matrix is nonrandom, so the model changes as p increases. His criterion is equivalent to ours. Having a nonrandom de

sign matrix eliminates some distributional problems. Present techniques allow the consideration of a design matrix with some columns nonrandom and some columns random.

Our results connect to James-Stein (1961) shrinking, as follows. Consider estimating the infinite vector β by $\hat{\beta}$, with squared-error loss, but computing norms relative to the variance-covariance matrix of the X's. The estimators considered here take $\hat{\beta}_j = 0$ for $j > p$, and this shrinks the vector towards 0. The shrinking is rather abrupt, but our results indicate that for the optimal p, even when estimated from the data, the shrinking will reduce the mean squared error. With our loss function, the ordinary least squares estimates will not in general be admissible, even if there is an a priori upper bound on p; for example, it is given that $\beta_j = 0$ for $j > p_0$. For moderate n, the optimal p may be substantially less than p_0.

2. PROOFS

We begin by arguing that the X_j in (1.1) may be assumed independent and identically distributed, with mean 0 and variance 1. To this end, let X_{j+1}^* be the part of X_{j+1} orthogonal to $\{X_1, \ldots, X_j\}$, rescaled to have variance 1: by assumption, X_{j+1} is not a linear combination of $X_1, \ldots, X_j$. It is automatic that X_{j+1}^* has mean 0. Of course,

$$X_{k+1}^* = \sum_{j=1}^{k+1} c_{j,k+1} X_j. \tag{2.1}$$

Define $X_{i,k+1}^*$ the same way:

$$X_{i,k+1}^* = \sum_{j=1}^{k+1} c_{j,k+1} X_{ij}. \tag{2.2}$$

Of course, (2.1) is invertible: $X_1, \ldots, X_{k+1}$ can be expressed as fixed linear combinations of $X_1^*, \ldots, X_{k+1}^*$. Now there are β_j^* such that

$$Y_i = \sum_{j=1}^{\infty} \beta_j^* X_{ij}^* + \epsilon_i. \tag{2.3}$$

Of course, β_j^* is a fixed linear combination of $\beta_1, \ldots, \beta_j$. We can define coefficient estimates $\hat{\beta}^*(n, p)$, predicted values $\hat{Y}_{n+1}^*$, conditional and unconditional mean squared prediction errors M_{np}^* and U_{np}^*, as well as the regression mean squared error R_{np}^* in terms of the starred model (2.3). All these quantities, except for the coefficient estimates themselves, depend only on the column space of the design matrix. This can be stated formally as follows.

Lemma 2.1.

(a) The column space of $X(n, p)$ coincides with the column space of $X^*(n, p)$.

(b) The σ field generated by $\{Y_i$ and X_{ij} for all j and $i = 1, \ldots, n\}$ coincides with the σ field generated by $\{Y_i$ and X_{ij}^* for all j and $i = 1, \ldots, n\}$.

(c) $M_{np}^* = M_{np}$ and $U_{np}^* = U_{np}$ and $R_{np}^* = R_{np}$.

The routine proof is omitted. But now, we can drop the stars, and assume

The X_j in (1.1) are independent and identically distributed, with mean 0 and variance 1. (2.4)

Under this circumstance,

$$\sigma_p^2 = \sum_{j=p+1}^{\infty} \beta_j^2.$$

Let

$$\delta_{pi} = \left[\left(\sum_{j=p+1}^{\infty} \beta_j X_{ij}\right) + \epsilon_i\right] \Big/ (\sigma^2 + \sigma_p^2)^{1/2}$$

so the δ_{pi} have mean 0 and variance 1. Let δ_p be the n vector whose ith component is δ_{pi}. Recall that $X = X(n, p)$ is the $n \times p$ matrix whose ij entry is X_{ij}.

Proof of Theorem 1.3. Clearly

$$Y_{n+1} - \hat{Y}_{n+1} = \sum_{j=1}^{p} (\beta_j - \hat{\beta}_j) X_{n+1,j} + \sum_{j=p+1}^{\infty} \beta_j X_{n+1,j} + \epsilon_{n+1}.$$

Abbreviate $\hat{\beta} = \hat{\beta}(n, p)$. Then

$$M_{np} = \| \hat{\beta} - \beta \|^2 + \sigma^2 + \sigma_p^2.$$

As usual,

$$\hat{\beta} - \beta = \sqrt{\sigma^2 - \sigma_p^2}(X^T X)^{-1} X^T \delta_p.$$

Thus, under condition (2.4),

$$M_{np} = [\sigma^2 + \sigma_p^2][1 + \| (X^T X)^{-1} X^T \delta_p \|^2]. \tag{2.5}$$

Let S be the unique positive definite square root of $X^T X$, and $\psi = XS^{-1}$, an $n \times p$ matrix. Then ψ is orthonormal; $\psi^T \psi = I_{p \times p}$. And $X = \psi S$, so

$$\begin{aligned} \| (X^T X)^{-1} X^T \delta_p \|^2 &= \delta_p{}^T X (X^T X)^{-2} X^T \delta_p \\ &= \delta_p{}^T \psi S S^{-4} S \psi^T \delta_p \\ &= \eta^T S^{-2} \eta \\ &= \eta^T (X^T X)^{-1} \eta, \end{aligned}$$

where $\eta = \psi^T \delta_p$ is a p vector of independent $N(0, 1)$ variables, even conditionally on X, because ψ is orthonormal. Thus, η is independent of X; and $\eta^T (X^T X)^{-1} \eta$ can be recognized as Hotelling's T^2 statistic, which has the claimed distribution. See Hotelling (1931).

Proof of Theorem 1.1. The evaluation of $U_{np} = E\{M_{np}\}$ is immediate from Theorem 1.3, because $E\{\chi_p^2\} = p$ and $E\{1/\chi_{n-p+1}^2\} = 1/(n - p - 1)$. Also see Wijsman (1957). It is only left to show that $p^*(n)/n \to 0$. Note that $\sigma_p^2 \to 0$ as $p \to \infty$, so

$$\sigma^2 + \sigma_{p^*(n)}^2 + \sigma^2 p^*(n)/n \leqq U_{np^*(n)} \to \sigma^2.$$

We turn now to Theorem 1.4. The following lemma will be helpful; its proof is standard.

Lemma 2.2. Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed, with mean 0 and variance 1. Suppose the moment generating function $E\{\exp(h\xi_1)\}$ exists for all h in a proper neighborhood of 0. Then there is a positive constant c not dependent on k such that for all $x > 0$ and all k,

(a) $P\{\xi_1 + \cdots + \xi_k > x\} < \exp\{-x^2/(4k)\}$ if $x < ck$
(b) $P\{\xi_1 + \cdots + \xi_k > x\} < \exp\{-ck/4\}$ if $x > ck$

Proof. Let $\phi(h) = E\{\exp(h\xi_1)\}$, finite for $0 \leqq h \leqq h_0$, where h_0 is positive. By Chebychev's inequality,

$$P\{\xi_1 + \cdots + \xi_k > x\} < e^{-hx}\phi(h)^k.$$

Since the mean is 0 and the variance is 1, $\phi(h) = 1 + \frac{1}{2}h^2 + o(h^2)$. If $0 \leqq h \leqq h_0$ where h_0 is small enough,

$$\phi(h) \leqq \exp\{h^2\}$$

so the probability in question is bounded above by

$$\exp\{-hx + kh^2\}.$$

Choose $c = 2h_0$. To prove part (a), set $h = x/2k$. To prove part (b), set $h = h_0$.

Corollary 2.1. Let ζ_i be independent $N(0, 1)$ variables. For any positive A there is a finite k_A such that $k > k_A$ entails

$$P\left\{\left|\sum_{i=1}^{k}(\zeta_i^2 - 1)\right| > 3\sqrt{Ak\log k}\right\} < 1/k^A.$$

Let $H = H(n, p)$ be the usual projection onto the column space of $X = X(n, p)$:

$$H = X(X^TX)^{-1}X^T. \tag{2.6}$$

Let $Y = Y(n)$ be the n vector whose ith entry is Y_i, and let $S = S(n, p)$ be the sum of squares for error:

$$S = S(n, p) = \| (I - H)Y \|^2. \tag{2.7}$$

Thus, $R_{np} = S(n, p)/(n - p)$. Abbreviate ϵ for the n vector whose ith entry is ϵ_i, and δ for the n vector whose ith entry is

$$\delta_i = \sum_{j=p+1}^{\infty}\beta_j X_{ij}. \tag{2.8}$$

(This represents a change of notation.) Plainly,

$$S = \| (I - H)(\epsilon + \delta) \|^2 = S_1 + S_2 + 2S_3, \tag{2.9}$$

where

$$S_1 = \| (I - H)\epsilon \|^2 = \| \epsilon \|^2 - \| H\epsilon \|^2,$$

$$S_2 = \| (I - H)\delta \|^2 = \| \delta \|^2 - \| H\delta \|^2,$$

$$S_3 = \langle (I - H)\epsilon, (I - H)\delta \rangle = \langle \epsilon, \delta \rangle - \langle H\epsilon, H\delta \rangle, \tag{2.10}$$

with $\langle\ \rangle$ for inner product. The dependence of S_1, S_2, S_3 on n and p is suppressed. Clearly,

$$\| \epsilon \|^2/(n - p) = \left(1 + \frac{p}{n - p}\right)\left[\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2)\right]. \tag{2.11}$$

In outline, the balance of the argument is as follows:

$$\| H\epsilon \|^2/(n - p) \doteq p\sigma^2(n - p)$$

$$S_2/(n - p) \doteq \sigma_p^2$$

$$S_3/(n - p) \doteq 0,$$

where terms of smaller order than σ_p^2 or p/n can be dropped.

Lemma 2.3. Fix small positive numbers α and β. Then there is a large number $n_{\alpha\beta}$ such that for $n > n_{\alpha\beta}$,

$$\left|\frac{\| H\epsilon \|^2 - p\sigma^2}{n - p}\right| < \beta\left(\sigma_p^2 + \sigma^2\frac{p}{n}\right)$$

for all p with $1 \leqq p \leqq \frac{1}{2}n$

except on a set of probability α.

Proof. Recall that H projects onto the column space of the $n \times p$ matrix of $\{X_{ij}\}$'s; since ϵ is independent of X, a routine argument shows that $\| H\epsilon \|^2$ is distributed for each n and p as $\sigma^2\sum_{i=1}^{p}\zeta_i^2$, the ζ_i being independent $N(0, 1)$ variables. Now use Corollary 2.1 with $A = 2$:

$$P\{\,|\,\| H\epsilon \|^2 - p\sigma^2\,| > 3\sigma^2\sqrt{2p\log p}\} < 1/p^2$$

so

$$\left|\frac{\| H\epsilon \|^2 - p\sigma^2}{n - p}\right| < \frac{6\sqrt{2p\log p}}{n}\sigma^2$$

for all p with $p_0 \leqq p \leqq \frac{1}{2}n$

except on a set of probability $1/p_0$. Choose p_0 so large that $1/p_0 < \alpha/2$, and $6(2p\log p)^{1/2} < \beta p$ for $p_0 \leqq p \leqq \frac{1}{2}n$.

We must now deal with $p < p_0$. In this range, $\sigma_p^2/\sigma^2 > \gamma > 0$, so for all large n,

$$P\{\,|\,\| H\epsilon \|^2 - p\sigma^2\,| > \beta\sigma_p^2(n - p)\} < P\left\{\left|\sum_{i=1}^{p}(\zeta_i^2 - 1)\right| > \frac{1}{2}\beta\gamma n\right\} < 2\exp\{-c_0\beta\gamma n\},$$

where c_0 is an absolute constant, by Lemma 2.2(b). Then

$$P\{\,|\,\| H\epsilon \|^2 - p\sigma^2\,| > \beta\sigma_p^2(n - p) \quad \text{for some } p < p_0\} < 2p_0\exp\{-c_0\beta\gamma n\} \to 0.$$

Lemma 2.4. Fix small positive numbers α and β. Then there is a large number $n_{\alpha\beta}$ such that $n > n_{\alpha\beta}$,

$$|\,S_2 - (n - p)\sigma_p^2\,| < \beta(n - p)\sigma_p^2$$

for all p with $1 \leqq p \leqq \frac{1}{2}n$

except on a set of probability α.

Proof. As before, for each n and p, S_2 is distributed like $\sigma_p^2\sum_{i=p+1}^{n}\zeta_i^2$, so

$$P\{\,|\,S_2 - (n - p)\sigma_p^2\,| > 3\sigma_p^2\sqrt{2(n - p)\log(n - p)}\} < 1/(n - p)^2$$

nd

$$S_2 - (n - p)\sigma_p^2 |$$

$$< 6\sqrt{(\log n)/n}(n - p)\sigma_p^2 \quad \text{for all } p \text{ with } 1 \leqq p \leqq \tfrac{1}{2}n$$

xcept on a set of probability $2/n$.

Lemma 2.5. Fix small positive numbers α and β. Then here is a large number $n_{\alpha\beta}$ such that, for $n > n_{\alpha\beta}$,

$$S_3 | < \beta(n - p)\left(\sigma_p^2 + \sigma^2 \frac{p}{n}\right)$$

for all p with $1 \leqq p \leqq \tfrac{1}{2}n$

xcept on a set of probability α.

Proof. As before, for each n and p, S_3 is distributed ke $\sigma\sigma_p \sum_{i=p+1}^{n} \zeta_i\zeta_i'$, where the ζ_i, ζ_i' are independent $N(0, 1)$ variables. We must now estimate

$$\pi_{pn} = P\left\{ \left| \sum_{i=p+1}^{n} \zeta_i\zeta_i' \right| > \beta(n - p)\left[(\sigma_p/\sigma) + (\sigma/\sigma_p)\frac{p}{n}\right]\right\}. \quad (2.12)$$

et $0 < p_0 < \infty$, to be chosen later. Now $\sigma_p/\sigma > \gamma > 0$ or $1 \leqq p \leqq p_0$; in that range, for small β and γ, for all rge n,

$$\pi_{pn} < P\left\{ \left| \sum_{i=p+1}^{n} \zeta_i\zeta_i' \right| > \beta\gamma(n - p)\right] < 2\exp\left[-\frac{1}{4}\beta^2\gamma^2(n - p)\right]\right\} \quad (2.13)$$

ccording to Lemma 2.2(a). Next, take $p > p_0$. Abbreiate $y = \sigma_p/\sigma$. Then

$$y + \frac{p}{n}\frac{1}{y} > 2\sqrt{p/n}$$

nd

$$2(n - p)\sqrt{p/n} \geqq \sqrt{np} \geqq \sqrt{(n - p)p}$$

or $p \leqq \frac{1}{2}n$, so for β small and $p_0 < p \leqq \frac{1}{2}n$, by Lemma .2(a),

$$\pi_{pn} < P\left\{ \left| \sum_{i=p+1}^{n} \zeta_i\zeta_i' \right| > \beta\sqrt{(n - p)p}\right\} < 2\exp\{-\tfrac{1}{4}\beta^2 p\}.$$

Now

$$\sum_{p=1}^{n/2} \pi_{pn} = \sum_{p=1}^{p_0} \pi_{pn} + \sum_{p=p_0+1}^{n/2} \pi_{pn} < 2p_0\exp\{-\tfrac{1}{8}\beta^2\gamma^2 n\} + 2\sum_{p=p_0+1}^{\infty} \exp\{-\tfrac{1}{4}\beta^2 p\}.$$

The first sum goes to 0 as $n \to \infty$, and the second is small or p_0 large.

Proof of Theorem 1.4. This is immediate from (2.11) and Lemmas 2.3–2.5.

Proof of Theorem 1.2.

Claim (a). Since $\sigma_p^2 > 0$ for all p and $\sigma_p^2 \to 0$ as $p \to \infty$, it follows that $\tilde{p}(n) \to \infty$. Clearly, $M_{n\tilde{p}(n)} \to \sigma^2$ so $n - \tilde{p}(n) \to \infty$ too. For k fixed, $\max\{\chi_1^2, \chi_2^2, \ldots, \chi_k^2\}$ has some finite distribution, and $\min\{\chi_n^2, \chi_{n-1}^2, \ldots, \chi_{n-k+1}^2\} \to \infty$. Only convergence in probability is needed, for present purposes. Fix θ positive but small, p_0 and p_1 large but finite, $n > p_0 + p_1$. Then

$$(1 - \theta)\frac{p}{n - p + 1} < \chi_p^2/\chi_{n-p+1}^2 < (1 + \theta)\frac{p}{n - p + 1}$$

simultaneously for all p with $p_0 \leqq p \leqq n - p_1$, (2.14)

except on a set of small probability. This follows from Lemma 2.1. In particular, although this is not needed, $\tilde{p}(n)/n \to 0$. When (2.14) holds, some easy algebra gives

$$(1 - \theta)f(p) < g(p) < (1 + \theta)f(p)$$

for $p_0 \leqq p \leqq n - p_1$, (2.15)

where

$$f(p) = U_{np} - \sigma^2, \; g(p) = M_{np} - \sigma^2.$$

Now p^* minimizes f and $\tilde{p}$ minimizes g; both fall in the range p_0 to $n - p_1$, at least for n large, and with high probability. So

$$g(\tilde{p}) > (1 - \theta)f(\tilde{p}) \geqq (1 - \theta)f(p^*),$$

$$g(\tilde{p}) \leqq g(p^*) < (1 + \theta)f(p^*).$$

This completes the argument for Claim (a).

Claim (b). Again, $\hat{p}(n) \to \infty$. Fix $\theta > 0$ and p_0 large and $n \geqq 2p_0$. Let $S_n = (1/n)\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2)$. Then

$$| R_{np} - \sigma^2 - \sigma_p^2 - S_n | < \tfrac{1}{4}\theta\left[\sigma_p^2 + \sigma^2\frac{p}{n}\right]$$

for all p with $p_0 \leqq p \leqq \frac{1}{2}n$, (2.16)

except on a set of small probability. When (2.16) holds, and $| S_n | < \frac{1}{4}\theta\sigma^2$, a tedious calculation shows that

$$(1 - \theta)f(p) < \hat{f}(p) < (1 + \theta)f(p)$$

for $p_0 \leqq p \leqq \frac{1}{2}n$, (2.17)

where

$$f(p) = U_{np} - \sigma^2, \hat{f}(p) = \hat{U}_{np} - \sigma^2 - S_n.$$

Now $\hat{p}$ minimizes $\hat{f}$ and p^* minimizes f; both fall between p_0 and $\frac{1}{2}n$, at least for n large and with high probability. Thus, except on a set of small probability,

$$(1 + \theta)f(\hat{p}) > \hat{f}(\hat{p}) > (1 - \theta)f(\hat{p}) \geqq (1 - \theta)f(p^*), \quad (2.18)$$

$$(1 - \theta)f(\hat{p}) < \hat{f}(\hat{p}) \leqq \hat{f}(p^*) < (1 + \theta)f(p^*). \quad (2.19)$$

In particular, (2.19) implies

$$f(\hat{p}) < [(1 + \theta)/(1 - \theta)]f(p^*). \quad (2.20)$$

Again, p^* minimizes f, so by (2.15) and (2.20),

$$\begin{aligned}(1 - \theta)f(p^*) &\leqq (1 - \theta)f(\hat{p}) \\ &< g(\hat{p}) \\ &< (1 + \theta)f(\hat{p}) \\ &< \frac{(1 + \theta)^2}{1 - \theta} f(p^*).\end{aligned}$$

Thus

$$(1 - \theta)f(p^*) < g(\hat{p}) < \frac{(1 + \theta)^2}{1 - \theta} f(p^*).$$

[*Received March 1982. Revised August 1982.*]

REFERENCES

AKAIKE, H. (1970), "Statistical Predictor Identification," *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

——— (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

FREEDMAN, D., and MOSES, L. (1981), "Adjusting for Covari in Clinical Trials," Technical Report, Stanford University, Dept Statistics.

HOCKING, R.R. (1976), "The Analysis and Selection of Variable Linear Regression," *Biometrics*, 32, 1–49.

HOTELLING, H. (1931), "The Generalization of Student's Rati *Annals of Mathematical Statistics*, 2, 360–378.

JAMES, W., and STEIN, CHARLES (1961), "Estimation With Qu ratic Loss," *Proceedings of the Fourth Berkeley Symposium*, Vol U. of California Press.

KRIEGER, A.M., and PICKANDS, J. III (1981), "A Criterion for rameter Specification in Multivariate and Regression Analys Technical Report #5, University of Pennsylvania, Wharton Sch Analysis Center.

MALLOWS, C.L. (1973), "Some Comments on C_p," *Technometr* 15, 661–675.

SHIBATA, R. (1980), "Asymptotically Efficient Selection of the O of the Model for Estimating Parameters of a Linear Process," *American Statistician*, 8, 147–164.

——— (1981), "An Optimal Selection of Regression Variables," *Bi etrika*, 68, 45–54.

THOMPSON, M.L. (1978), "Selection of Variables in Multiple Reg sion," *International Statistical Review*, 46, 1–49 and 129–146.

WIJSMAN, R.A. (1957), "Random Orthogonal Transformations Their Use in Some Classical Distribution Problems in Multivar Analysis," *Annals of Mathematical Statistics*, 28, 415.