BOOTSTRAPPING REGRESSION MODELS WITH MANY PARAMETERS

BY

P. J. BICKEL[1] AND D. A. FREEDMAN[2]

TECHNICAL REPORT NO. 7
JUNE 1982

RESEARCH PARTIALLY SUPPORTED
BY
[1]NAVAL RESEARCH GRANT N00014-80-C-0163
[2]NATIONAL SCIENCE FOUNDATION GRANT MCS-80-02535

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

BOOTSTRAPPING REGRESSION MODELS WITH MANY PARAMETERS

P. J. Bickel
D. A. Freedman
University of California, Berkeley

ABSTRACT

The regression model is considered, where the number of data points n and parameters p are both large. It is shown that the bootstrap approximation to the distribution of contrasts is valid, provided p/n is small. It is also shown that if p/n does not tend to zero, the bootstrap approximation is invalid. Similar results are obtained for the full p-dimensional distribution of the least squares estimates. Here, the relevant growth condition is that $p^2/n \longrightarrow 0$, and regularity conditions are needed on the tails of the error distribution.

KEY WORDS AND PHRASES: *Regression, Least squares, Bootstrap, Mallows metrics.*

1. INTRODUCTION

This paper is a sequel to Freedman (1981). It will develop some asymptotic theory for applications of the bootstrap to regression, where the number of parameters p and the number of data points n are both large. In more detail, let

$$Y = X\beta + \varepsilon \,, \tag{1.1}$$

where

Bickel's research was partially supported by the Office of Naval Research Grant N00014-80-C-0163. Freedman's research was partially supported by National Science Foundation Grant MCS-80-02535. AMS/MOS Classification Numbers: Primary 62E20, Secondary 62G05, 62G15.

β is a $p \times 1$ vector of unknown parameters, to be estimated from the data;

Y is an $n \times 1$ data vector;

X is an $n \times p$ data matrix, nonrandom, of full rank $p \leq n$;

ε is an $n \times 1$ vector of unobservables.

The main assumptions are on ε:

> The components $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ of ε are supposed to be independent random variables, with common distribution F, having mean 0 and variance $\sigma^2 > 0$; both F and σ^2 are unknown. (1.2)

Under the circumstances, the conventional least squares estimate $\hat{\beta}$ for β is $\hat{\beta} = (X^TX)^{-1}X^TY$. Let Q^2 be the cross-product matrix: $Q^2 = X^TX$. Then $\hat{\beta}$ has mean β and variance-covariance matrix σ^2/Q^2. If additional conditions are imposed, the distribution of the pivotal quantity $Q(\hat{\beta}-\beta)/\sigma$ is asymptotically normal, with mean 0 and variance-covariance matrix $I_{p\times p}$, the $p \times p$ identity matrix.

Notation

X^TX is positive definite, so it has a unique positive definite square root; this is Q. "Positive definite" is taken in the strict sense. And σ^2/Q^2 is to be interpreted as $\sigma^2Q^{-2} = Q^{-2}\sigma^2$.

It is convenient to separate the normalization by Q and by σ; only the first will be considered in detail. Let $\Psi_{np}(F)$ be the exact distribution of $Q(\hat{\beta}-\beta)$, with n data points, p parameters, and law F governing the disturbance terms. Similarly, if c is a $p \times 1$ coefficient vector, let $\Psi_{npc}(F)$ be the exact distribution of the contrast $c^T(\hat{\beta}-\beta)$, normalized so $c^T(\hat{\beta}-\beta)$ has variance σ^2; that is,

$$c^T(X^TX)^{-1}c = 1. \tag{1.3}$$

The "Mallows metrics" are defined in Bickel and Freedman (1981); also see Freedman (1981). In brief, let $\|\cdot\|$ stand for the Euclidean norm on R^p, and let $\alpha \geq 1$. Then $d_\alpha(\mu,\nu) = \inf E\{\|U-V\|^\alpha\}^{1/\alpha}$, where U has law μ and V has law ν;

the inf is over the joint distribution. Convergence in d_α is equivalent to weak convergence plus convergence of moments of order α or less.

There are useful inequalities connecting the Mallows metrics with other metrics for weak convergence. Let $d_{LP}(\mu,\nu)$ be the Levy-Prokhorov distance between μ and ν, namely, the inf of ε positive such that

$$\mu(K) \leqq \nu(K_\varepsilon) + \varepsilon \quad \text{and} \quad \nu(K) \leqq \mu(K_\varepsilon) + \varepsilon$$

for all compact K, where K_ε is the set of $x \in R^p$ whose Euclidean distance to K is ε or less. As may be seen from Chebychev's inequality,

$$d_{LP}(\mu,\nu) \leqq d_\alpha(\mu,\nu)^{\alpha/\alpha+1} . \tag{1.4}$$

In more detail,

$$\begin{aligned} \mu(K) &= P\{U \in K\} \\ &\leqq P\{V \in K_\varepsilon\} + P\{||U-V|| \geqq \varepsilon\} \\ &\leqq P\{V \in K_\varepsilon\} + \varepsilon^{-\alpha} E\{||U-V||^\alpha\} \end{aligned}$$

Choose U and V to minimize the expected value, and set $\varepsilon = d_\alpha(\mu,\nu)^{\alpha/\alpha+1}$. Likewise for the bound on $\nu(K)$. This completes the argument for (1.4); the tightest bound is obtained when $\alpha = 1$.

Similar comments apply to the "Bounded-Lipschitz" metric

$$d_{BL}(\mu,\nu) \leqq d_\alpha(\mu,\nu) \tag{1.5}$$

where

$$d_{BL}(\mu,\nu) = \sup_\phi \left| \int \phi d\mu - \int \phi d\nu \right|$$

the sup being over all ϕ such that

$$|\phi(x)| \leqq 1 \quad \text{and} \quad |\phi(x)-\phi(y)| \leqq ||x-y||$$

for all x and y.

The bootstrap estimates the distribution $\Psi_{np}(F)$ by $\Psi_{np}(G)$, where G is an estimate of F described below. Bounds will be given on $d_2[\Psi_{np}(F),\Psi_{np}(G)]$ and $d_2[\Psi_{npc}(F),\Psi_{npc}(G)]$ in terms of $d_2(F,G)$, for any F and G.

Theorem 1.1. Let F *and* G *be two possible laws for the disturbance terms* ε_i *in the model (1.1-2); it is assumed that both have mean* 0 *and finite variance. Then*

(a) $d_2[\Psi_{np}(F),\Psi_{np}(G)]^2 \leqq p \cdot d_2(F,G)^2$

(b) $d_2[\Psi_{npc}(F),\Psi_{npc}(G)]^2 \leqq d_2(F,G)^2$

Proof. (a) Let $\varepsilon_i(F)$ be independent, with law F; let $\varepsilon(F)$ be the $n \times 1$ column vector with components $\varepsilon_i(F)$; likewise for $\varepsilon_i(G)$ and $\varepsilon(G)$. Let A be an arbitrary $p \times n$ matrix. Let $\Psi_A(F)$ be the law of $A\varepsilon(F)$, and likewise for $\Psi_A(G)$. As in Bickel and Freedman (1981, Lemma 8.9),

$$d_2[\Psi_A(F),\Psi_A(G)]^2 \leqq (\text{trace } AA^T) \cdot d_2(F,G)^2 . \tag{1.6}$$

Clearly,

$$Q(\hat{\beta}-\beta) = Q(X^TX)^{-1}X^T\varepsilon = Q^{-1}X^T\varepsilon ,$$

because $Q^2 = X^TX$. Use (1.6), with $A = Q^{-1}X^T$; verify that

$$AA^T = Q^{-1}X^TXQ^{-1} = Q^{-1}Q^2Q^{-1} = I_{p\times p}$$

has trace p.

(b) The proof is similar, with $A = c^T(X^TX)^{-1}X^T$.

Compare the bounds in Theorem 1.1 (a) and (b): the first has an extra factor p on the right, due to the fact that it compares p-dimensional distributions.

In the setup of Freedman (1981), G is the empirical distribution $\hat{F}_n$ of the centered residuals. To be more specific, in a regression problem the fitted values are

$$\hat{Y} = X\hat{\beta} = HY , \quad \text{where } H = X(X^TX)^{-1}X^T . \tag{1.7}$$

The residuals are

$$\hat{\varepsilon} = Y - \hat{Y} = \Gamma\varepsilon , \quad \text{where } \Gamma = I_{n\times n} - H . \tag{1.8}$$

Let $\hat{\mu}_n = \frac{1}{n}\Sigma_{i=1}^n \hat{\varepsilon}_i$; this may be nonzero, for the constants need not be in the column space of X. Let $\hat{F}_n$ be the empirical distribution of the centered residuals, assigning mass 1/n to each $\hat{\varepsilon}_i - \hat{\mu}_n$.

Let F_n be the empirical distribution of $\varepsilon_1,\ldots,\varepsilon_n$.

To review the bootstrap operation: Given $Y_1,\ldots,Y_n$, let $\varepsilon_1^*,\ldots,\varepsilon_n^*$ be conditionally independent, with common distribution $\hat{F}_n$. And let

$$Y^* = X\hat{\beta} + \varepsilon^* .$$

Informally, ε^* is obtained by resampling the residuals. And Y^* is generated from the data, using the regression model with

$\hat{\beta}$ as the vector of parameters

$\hat{F}_n$ as the distribution of the disturbance terms ε

Now imagine giving the starred data (X,Y^*) to another statistician, and asking him or her to estimate $\hat{\beta}$. The least squares estimate is

$$\hat{\beta}^* = (X^TX)^{-1}X^TY^* .$$

Consider a contrast with coefficient vector c satisfying (1.3). The bootstrap principle is that the distribution of $c^T(\hat{\beta}^*-\hat{\beta})$, which can be computed directly from the data (e.g., by Monte Carlo), approximates the distribution of $c^T(\hat{\beta}-\beta)$. Likewise for the full p-dimensional distributions.

Of course, there is also the problem of estimating σ^2. Let $\hat{\varepsilon}^*$ be the residuals in the starred data set:

$$\hat{\varepsilon}^* = Y^* - X\hat{\beta}^* = \Gamma\varepsilon^* .$$

Let

$$s^2 = \frac{1}{n-p}\Sigma_{i=1}^n \hat{\varepsilon}_i^2$$

$$s^{*2} = \frac{1}{n-p}\Sigma_{i=1}^n \hat{\varepsilon}_i^{*2}$$

Division by $n-p$ rather than n is immaterial at the moment. The dependence of $\hat{\beta}$, $\hat{\beta}^*$, and $\hat{s}^*$ on n and p is suppressed in the notation. The next result justifies the use of the bootstrap to estimate the distribution of individual contrasts and t-statistics based on them.

Theorem 1.2. *Assume conditions (1.1-2) on the model. Let* c *satisfy (1.3). Condition on* $Y_1,\ldots,Y_n$. *Let* $n \to \infty$; *if* $p \to \infty$, *assume* $p/n \to 0$. *In probability:*

(a) The d_2*-distance between the conditional distribution of* $c^T(\hat\beta^*-\hat\beta)$ *and the distribution of* $c^T(\hat\beta-\beta)$ *goes to zero, uniformly in* c.

(b) The conditional distribution of $\hat s^*$ *converges weakly to point mass at* σ.

(c) The d_{LP}*-distance between the conditional distribution of* $c^T(\hat\beta^*-\hat\beta)/s^*$ *and the distribution of* $c^T(\hat\beta-\beta)/s$ *goes to zero, uniformly in* c.

Proof. Claim (a). Apply Theorem 1(b) to the conditional distribution of $c^T(\hat\beta^*-\hat\beta)$ given $Y_1,\ldots,Y_n$: put $\hat F_n$ in place of G. The bound is

$$d_2(F,\hat F_n)^2 \leq 2[d_2(F,F_n)^2 + d_2(F_n,\hat F_n)^2].$$

As shown in Freedman (1981, Lemma 2.2),

$$E\{d_2(F_n,\hat F_n)^2\} \leq \sigma^2(p+1)/n \to 0 .$$

On the other hand, $d_2(F,F_n) \to 0$ by Bickel and Freedman (1981, Lemma 8.4). Claim (a) follows.

Claim (b). This is argued as in Freedman (1981).

Claim (c) is immediate from (a) and (b), using (1.4). Of course, d_{LP} can be replaced by d_{BL}.

Evidently $p/n \to 0$ and (1.2) also suffice for bootstrapping the joint distribution of $c_i^T(\hat\beta-\beta)$, where $1 \leq i \leq r$ and r is fixed; the c_i are to satisfy (1.3). If p/n does not tend to 0, the bootstrap will fail for general contrasts: see sections 2 and 3. What happens to the distribution of $\hat s^{*2}$? A fairly technical argument gives

Theorem 1.3. *Assume conditions (1.1-2) on the model. Let* $n-p \to \infty$. *Then the conditional distribution of* $\hat s^*$ *converges weakly to point mass at* σ *in probability.*

Here is an interesting variation on the bootstrap idea. Suppose we want to use the resampled residuals from one model with $n_1 \times p_1$ design matrix X_1 to analyze a second model with $n_2 \times p_2$ design matrix X_2; the underlying distribution of the disturbance terms are assumed the same, and (1.1-2) in force. For instance, X_1 might be a submatrix of X_2. If $p_1/n_1 \longrightarrow 0$, and $c^T[X_2^TX_2]^{-1}c = 1$, the bootstrap can be used to estimate the law of $c^T(\hat{\beta}-\beta)$ in the X_2 model, without any conditions on n_2 and p_2. This is a consequence of Theorem 1.1(b) since $d_2(\hat{F}_1,F) \longrightarrow 0$ in probability, where $\hat{F}_1$ is the empirical distribution of the centered residuals from the X_1 model. Likewise for the law of the pivot $c^T(\hat{\beta}-\beta)/s$.

Let $\hat{Y} = X\hat{\beta}$ be the vector of fitted values. Of course, $E\{\hat{Y}\hat{Y}^T\} = \sigma^2H$. Huber (1973) has shown that asymptotic normality of all standardized contrasts is equivalent to the L_2 consistency of all the fitted values, which in turn is equivalent to $\max_i H_{ii} \longrightarrow 0$. Now trace $H = p$. So $\max_i H_{ii} \geq p/n$; and Huber's condition is stronger than that needed for bootstrapping contrasts, viz., $p/n \longrightarrow 0$. In particular, the bootstrap may work even when asymptotic normality fails. See Example 2.1.

There are plausible modifications of $\hat{F}_n$ as an estimator of F. One is to rescale the residuals by $\sqrt{n/(n-p)}$. Then s^2 is the conditional variance of the ε_i^* given $Y_1,\ldots,Y_n$. If $p/n \longrightarrow 0$, the theorems are unchanged. On the other hand, as shown in Sections 2 and 3, if p/n does not tend to 0, the rescaling does not help. Another modification is to use the standardized centered residuals $(\hat{\varepsilon}_i-\hat{\varepsilon}_\cdot)(\Gamma_{ii}-2\Gamma_{i\cdot}+\Gamma_{\cdot\cdot})^{-1/2}$, where $\Gamma = ||\Gamma_{ij}||$ is given in (1.8), and the dot is the averaging operator. It is easy to see that this works if (1.1-2) hold, and Huber's condition that $\max_i H_{ii} \longrightarrow 0$ is satisfied. With a little more effort, one can show that Huber's condition can be replaced by the weaker condition that $p/n \longrightarrow 0$. If F is normal but p/n does not converge to 0, standardizing the residuals this way may give a consistent estimate of F when our $\hat{F}_n$ fails. If F is not normal, standardizing the residuals does not help: see Theorem 3.2.

If p is fixed, simultaneous inference for all contrasts can be accomplished by the bootstrap, as in Freedman (1981). Another approach is to use Theorem 1.1(a):

Theorem 1.4. Assume conditions (1.1-2) on the model, suppose p *is fixed, and let* $n \longrightarrow \infty$. *Then the* d_2*-distance between the distribution of* $Q(\hat{\beta}-\beta)$ *and the conditional distribution of* $Q(\hat{\beta}^*-\hat{\beta})$ *given* $Y_1,\ldots,Y_n$ *tends to zero in probability. Likewise for the distribution of the pivotal quantities, with respect to the distance* d_{LP}.

If $p \longrightarrow \infty$ but $p^2/n \longrightarrow 0$, and $E\{d_2(F_n,F)^2\} = o(1/p)$, the distance between the distribution of $Q(\hat{\beta}-\beta)$ and the corresponding bootstrap distribution tends to 0. This has an interesting consequence for the Scheffé method of simultaneous inference. Consider bootstrapping $S = [(\hat{\beta}-\beta)^T Q^2(\hat{\beta}-\beta)]^{1/2}$ or S/s. Let $S^* = [(\hat{\beta}^*-\hat{\beta})^T Q^2(\hat{\beta}^*-\hat{\beta})]^{1/2}$.

Theorem 1.5. Assume conditions (1.1-2) on the model. If $n \longrightarrow \infty$, $p^2/n \longrightarrow 0$, *and* $E\{d_2(F_n,F)^2\} = o(1/p)$, *the* d_2*-distance between the distribution of* $S - p^{1/2}s$ *and the conditional distribution of* $S^* - p^{1/2}s^*$ *given* $Y_1,\ldots,Y_n$ *tends to zero in probability; likewise for the distributions of* $S/s - p^{1/2}$ *and* $S^*/s^* - p^{1/2}$, *with respect to the distance* d_{LP}.

Proof. Let $\Phi_{np}(F)$ be the joint law of $Q(\hat{\beta}-\beta)$ and $p^{1/2}s$. Arguing as in Theorem 1.1,

$$d_2[\Phi_{np}(F),\Phi_{np}(G)]^2 \leq 2p\, d_2(F,G)^2$$

because

$$s^2 = \frac{1}{n-p}\|\Gamma\epsilon\|^2 .$$

Let $\tilde{\Phi}_{np}(F)$ be the law of $\|Q(\hat{\beta}-\beta)\| - p^{1/2}s$. Then

$$d_2[\tilde{\Phi}_{np}(F),\tilde{\Phi}_{np}(G)]^2 \leq 2d_2[\Phi_{np}(F),\Phi_{np}(G)]^2 \leq 4pd_2(F,G)^2$$

This inequality can be applied to the bootstrap situation as in Theorem 1.2:

$$d_2[\tilde{\Phi}_{np}(F),\ \tilde{\Phi}_{np}(\hat{F}_n)]^2 \leq 4pd_2(F,\hat{F}_n)^2 \leq 8p[d_2(F,F_n)^2+d_2(F_n,\hat{F}_n)^2] .$$

Now $pd_2(F_n,\hat{F}_n)^2$ is of order p^2/n, which tends to zero by assumption, as does $pd_2(F,F_n)^2$.

We expect $E\{d_2(F_n,F)^2\} = O(1/n)$, in which case if $p = o(n)$, then $E\{d_2(F_n,F)^2\} = o(1/p)$. Indeed, by Bickel and Freedman (1981, Lemma 8.2),

$$E\{d_2(F_nF)^2\} = \int_0^1 E\{[F_n^{-1}(t)-F^{-1}(t)]^2\}dt .$$

Suppose F concentrates on an interval (a,b), and $f = F' > 0$, and f is continuous on $[a,b]$. Then

$$n\int_0^1 E\{[F_n^{-1}(t) - F^{-1}(t)]^2\}dt = \int_0^1 t(1-t)f^{-2}(F^{-1}(t))dt + o(1)$$

and our expectation is fulfilled. However, in general, there is a delicate dependence on the tails of F. Under the conditions of Lemma A2.3 of Albers et al (1976), if $a = n^{-1}$,

$$n\int_0^1 E\{[F_n^{-1}(t) - F^{-1}(t)]^2\}dt = \Omega\left(\int_a^{1-a} t(1-t)f^{-2}(F^{-1}(t))dt\right)$$

where Ω denotes exact order. For the normal distribution,

$$E\{d_2(F_n,F)^2\} = \Omega(\frac{\log\log n}{n}) = o(\frac{1}{p})$$

for $p=o(\sqrt{n})$. However, by taking $F^{-1}(t)\sim t^{-\frac{1}{2}+\varepsilon}$ as $t\to 0$, we get $E\{d_2(F_n,F)^2\} = \Omega(n^{-2\varepsilon})$.

2. EXAMPLES

We illustrate the range of behavior of the bootstrap in the one-way analysis of variance with $p-1$ treatment groups, all of the same size T, and a control group of size C. The model is

$$Y_{ij} = \beta_i + \varepsilon_{ij} \tag{2.1}$$

The treatment groups are indexed by $i = 1,\ldots,p-1$; the subjects

by $j = 1,\ldots,T$. The control group index is $i = 0$; the subjects in control are indexed by $j = 1,\ldots,C$. The disturbances ε_{ij} satisfy (1.2). So,

$$n = C + (p-1)T$$
$$\hat{\beta}_i = Y_{i\cdot} \quad \text{for} \quad 0 \leqq i \leqq p$$

The hat matrix H has diagonal entries $1/T$ and $1/C$; Huber's condition is satisfied iff

$$\text{both } C \text{ and } T \text{ tend to infinity.} \tag{2.2}$$

Everything works in this case. The variables $C^{1/2}(\hat{\beta}_0-\beta_0)$ and $T^{1/2}(\hat{\beta}_i-\beta_i)$ for $i \geqq 1$ are independent and asymptotically $N(0,\sigma^2)$ under F. The residuals $\varepsilon_{ij}-\varepsilon_{i\cdot}$ are automatically centered, and have empirical distribution $\hat{F}_n$ with $d_2(\hat{F}_n,F) \to 0$; so the conditional distributions of $C^{1/2}(\hat{\beta}_0^*-\hat{\beta}_0)$ and $T^{1/2}(\hat{\beta}_1^*-\hat{\beta}_i)$ given $Y_1,\ldots,Y_n$ have the same behavior.

On the other hand, $p/n \longrightarrow 0$ iff

$$\text{either } T \longrightarrow \infty \text{ or } C/p \longrightarrow \infty. \tag{2.3}$$

The difference between (2.2) and (2.3) makes trouble, as follows.

Example 2.1. The bootstrap may succeed even if some parameter estimates are neither consistent nor asymptotically normal. Suppose $p = 2$ and $C \longrightarrow \infty$ but T is fixed. Evidently $\hat{\beta}_1$ is neither consistent nor asymptotically normal. But $d_2(\hat{F}_n,F) \longrightarrow 0$ in probability. Given $Y_1,\ldots,Y_n$, the joint distribution of $C^{1/2}(\hat{\beta}_0^*-\hat{\beta}_0)$ and $T^{1/2}(\hat{\beta}_1^*-\hat{\beta}_1)$ is close to that of $C^{1/2}(\hat{\beta}_0-\beta_0)$ and $T^{1/2}(\hat{\beta}_1-\beta_1)$.

The next two examples show that, in general, the bootstrap will fail if p is of order n; i.e., there are too many parameters.

Example 2.2. Suppose $C = T = 2$, so $n = 2p$. Then $\hat{F}_n$ is the empirical distribution of $\pm\frac{1}{2}(\varepsilon_{i1}-\varepsilon_{i2})$, for $i = 1,\ldots,\frac{1}{2}n$. As $n \longrightarrow \infty$, this converges to the theoretical distribution $\tilde{F}$ of $\frac{1}{2}(\varepsilon_{11}-\varepsilon_{12})$. The conditional distribution of $\hat{\beta}_i^* - \hat{\beta}_i$ given

$Y_1,\ldots,Y_n$ goes to the theoretical distribution of $\frac{1}{4}(\varepsilon_{11}-\varepsilon_{12}+\varepsilon_{21}-\varepsilon_{22})$ whereas $\hat{\beta}_i-\beta_i$ has the distribution of $\frac{1}{2}(\varepsilon_{11}+\varepsilon_{12})$. The two distributions do not even have the same variance. If the residuals are rescaled to have the right variance before resampling, i.e., to $\pm(\varepsilon_{11}-\varepsilon_{i2})\sqrt{2}$, the resulting bootstrap distribution does have the right variance, but is still wrong unless F is normal. In this example, the projection matrix is constant at $1/2$ along the diagonal, and the behavior is prototypical for Theorem 3.2 below.

Example 2.3. Suppose $T = 2$ and $C \to \infty$ but $C/p \to \rho < \infty$. (For instance, given that $T = 2$, in the optimum allocation for estimating treatment effects $C = 2\sqrt{p-1}$, so $\rho = 0$.) As before, let $\tilde{F}$ be the distribution of $\frac{1}{2}(\varepsilon_{i1}-\varepsilon_{i2})$. Then $\hat{F}_n$ converges to a mixture $\tilde{\tilde{F}} = \frac{\rho}{2+\rho}F + \frac{2}{2+\rho}\tilde{F}$. The bootstrap distribution of $2(\hat{\beta}_i^*-\hat{\beta}_i)$ for $i \geq 1$ will have as a limit $\tilde{\tilde{F}}*\tilde{\tilde{F}}$, where $*$ denotes convolution. This will not agree with $F*F$ for most F, including F normal. If $\rho > 0$, rescaling by $\sqrt{n/(n-p)}$ will not help with the treatment means, although it will with the control mean. In this example, the projection matrix is variable along the diagonal, with $n\rho/(\rho+2)$ elements tending to 0 and the rest to $1/2$. The behavior is prototypical for Theorem 3.1 below. Rescaling the residuals before resampling does give the right answer for all contrasts if F is normal, but still goes wrong for other distributions.

In Examples 2.2 and 2.3, the estimates of β_i are not consistent; replacing β_i by $n\beta_i$ in (2.1), however, gives examples where the estimates are consistent.

3. THE FAILURE OF THE BOOTSTRAP

Consider a sequence of regression models satisfying the condition (1.2). In all these models, the disturbance terms ε are assumed to have the same distribution F. Suppose p/n, the ratio of parameters to data points, converges to a limit α, with $0 < \alpha < 1$. Let H be the projection matrix, and $\Gamma = I_{n\times n} - H$, as defined in

(1.7-8). The empirical measure λ_n of $\{\Gamma_{ii}: i = 1,\ldots,n\}$ sits on $[0,1]$ and has mean $(n-p)/n$. By compactness, suppose without loss of generality that λ_n converges weakly to a limit λ. Since $(n-p)/n \longrightarrow 1-\alpha$, clearly $\int x\lambda(dx) = 1 - \alpha$.

Theorem 3.1. Suppose λ *is nondegenerate:* $\lambda\{1-\alpha\} < 1$. *Suppose* F *is* $N(0,1)$. *Then the bootstrap fails. Indeed, the empirical distribution* $\hat{F}_n$ *of the residuals converges weakly in probability to a* λ*-scale mixture of normals, rather than normal. Furthermore, for each* n *there is an* $n \times 1$*-coefficient vector* c *satisfying (1.3), such that the conditional distribution of* $c^T(\hat{\beta}^* - \beta)$ *given* $Y_1,\ldots,Y_n$ *does not converge to normal. Scaling the residuals by* $\sqrt{(n-p)/p}$ *will not help.*

This is the general phenomenon exhibited in a special case in Example 2.3. The convergence is a bit perplexing. To be more specific, let $\tilde{\lambda}$ be the λ-scale mixture of normals. Thus, $\tilde{\lambda} = \int \Phi_c \lambda(dc)$ where Φ_c is normal with mean 0 and variance c. The assertion of the theorem is that for any weak neighborhood N of $\tilde{\lambda}$,

$$P[\hat{F}_n \in N] \longrightarrow 1 .$$

In particular, $\hat{F}_n$ becomes almost constant at $\tilde{\lambda}$. As it happens, $\tilde{\lambda}$ is a mixture. But this does not imply any asymptotic randomness in $\hat{F}_n$: there is none.

Proof. Let $\hat{\phi}_n(t)$ be the empirical characteristic function of the residuals:

$$\hat{\phi}_n(t) = \frac{1}{n} \Sigma_{i=1}^n \exp(\sqrt{-1}\, t\hat{\varepsilon}_i) .$$

Let ϕ be the characteristic function of the errors ε_i, so $\phi(t) = \exp(-\frac{1}{2}t^2)$. As before,

$$\hat{\varepsilon}_i = \Sigma_j \Gamma_{ij}\varepsilon_j$$

and

$$\Sigma_j \Gamma_{ij}\Gamma_{jk} = (\Gamma^2)_{ik} = \Gamma_{ik} .$$

Put $k = i$ and use the symmetry of Γ:

$$E\{\hat{\phi}_n(t)\} = \frac{1}{n} \Sigma_{i=1}^{n} \exp(-\tfrac{1}{2}\Gamma_{ii}t^2) \longrightarrow \int_0^1 \exp(-\tfrac{1}{2}ct^2)\lambda(dc) .$$

Next, it is claimed that for most pairs (j,k), the elements Γ_{jk} are nearly zero. Indeed

$$\frac{1}{n^2} \Sigma_{jk} \Gamma_{jk}^2 = \frac{1}{n^2} \Sigma_j \Gamma_{jj} = \frac{n-p}{n^2} \longrightarrow 0 . \tag{3.1}$$

Now

$$|\hat{\phi}_n(t)|^2 = \frac{1}{n^2} \Sigma_{jk} \exp[\sqrt{-1}\, t(\hat{\varepsilon}_j - \hat{\varepsilon}_k)]$$

and

$$\hat{\varepsilon}_j - \hat{\varepsilon}_k = \Sigma_\ell(\Gamma_{j\ell} - \Gamma_{k\ell})\varepsilon_\ell .$$

So

$$E\{|\hat{\phi}_n(t)|^2\} = \frac{1}{n^2} \Sigma_{jk} \exp[-\tfrac{1}{2}t^2 \Sigma_\ell(\Gamma_{j\ell} - \Gamma_{k\ell})^2] .$$

But

$$\Sigma_\ell(\Gamma_{j\ell} - \Gamma_{k\ell})^2 = \Sigma_\ell \Gamma_{j\ell}^2 + \Sigma_\ell \Gamma_{k\ell}^2 - 2\Sigma_\ell \Gamma_{j\ell}\Gamma_{k\ell}$$

$$= \Gamma_{jj} + \Gamma_{kk} - 2\Gamma_{jk}$$

and now

$$E\{|\hat{\phi}_n(t)|^2\} = \frac{1}{n^2} \Sigma_{jk} \exp[-\tfrac{1}{2}t^2(\Gamma_{jj} + \Gamma_{kk} - 2\Gamma_{jk})] .$$

However, for most pairs jk, $\exp[t^2\Gamma_{jk}]$ is essentially one by (3.1), so

$$E\{|\hat{\phi}_n(t)|^2\} \longrightarrow [\int_0^1 \exp(-\tfrac{1}{2}ct^2)\lambda(dc)]^2 .$$

Thus, $\hat{\phi}_n(t)$ has an asymptotic variance of zero. It follows that $\hat{F}_n$, the empirical distribution of the residuals, converges weakly

in probability to the λ-mixture of normals. See Freedman and Lane (1981) for details.

Turn to the contrasts. Since trace $H = p$, there is an i with $H_{ii} \geq p/n$. For each n, there will be a coefficient vector c (corresponding to a fitted value) satisfying (1.3), but

$$c^T(\hat{\beta}^* - \hat{\beta}) = \Sigma_{i=1}^n d_i \varepsilon_i^*$$

and $d_i \geq p/n$ for some i. The conditional characteristic function of the contrast is

$$\Pi_{i=1}^n \hat{\phi}_n(d_i t) .$$

Any limit here has a proper scale mixture of normals as a factor, and hence cannot be normal, by Cramér's theorem.

Here is the general phenomenon we exhibited in Example 2.2

Theorem 3.2. *Suppose* λ *is degenerate,* $0 < \alpha < 1$. *Suppose further,*

(a) F *is nondegenerate and is not representable as the convolution of the distribution of* $(1-\alpha)\varepsilon$ *and another distribution.*

(b) The Laplace transform

$$\psi(t) = \int e^{tx} dF(x)$$

is finite for t *in an interval about* 0.

Then,

(1) F *is not a limit point of the sequence* $\{\hat{F}_n\}$ *in probability. That is, there exists a weak neighborhood* N *of* F *such that* $P[\hat{F}_n \notin N] \longrightarrow 1$.

Suppose further

(c) F *is symmetric about* 0.

Then,

(2) For each n *we can find an* $n \times 1$ *coefficient vector* c *satisfying (1.3) such that the Lévy distance between the conditional distribution of* $c^T(\hat{\beta}-\beta)$ *stays bounded away from* 0 *in probability.*

Notes

1. As the statement of the theorem and Example 2.2 suggest, characterizing what $\hat{F}_n$ behaves like in general is difficult.
2. If F has $(1-\alpha)\varepsilon$ as a component, it must have a representation of the form $\Sigma_{k=1}^{\infty}(1-\alpha)^k\varepsilon_k$, where the ε_k are independent with common distribution F. In particular, the symmetric two-point distributions satisfy a-b-c.
3. Let ϕ be the characteristic function of F. If F has $(1-\alpha)\varepsilon$ as a component, then $\phi(t) = \phi[(1-\alpha)t]\psi(t)$ for some characteristic function ψ. In particular, the symmetric uniform distributions satisfy a-b-c.
4. The "balanced" designs with $\Gamma_{ii} = (n-p)/n$ are covered by the theorem.
5. The conclusions of this theorem apply even if the residuals are scaled or standardized.

The proof of this result requires two lemmas. The first extends the main argument in Theorem 3.1. Its rather technical proof is given in the appendix.

Let $\hat{\psi}_n$ be the empirical Laplace transform of the residuals, and ψ the theoretical Laplace transform of F. Thus

$$\hat{\psi}_n(t) = \frac{1}{n}\Sigma_{i=1}^{n} \exp(t\hat{\varepsilon}_i)$$

Let

$$\psi_n(t) = E\hat{\psi}_n(t) .$$

Lemma 3.1. Suppose ψ satisfies (b) of Theorem 3.2. Then $\log \psi$ *has a Taylor expansion convergent for* $|t| < a$, $a > 0$, *which we shall write*

$$\log \psi(t) = \Sigma_{r=1}^{\infty} \kappa_r \frac{t^r}{r!} . \tag{3.2}$$

And (1) $|\psi_n(t)| \leq \exp\{\Sigma_{r=2}^{\infty} \frac{|\kappa_r|}{r!}|t|^r\}$ *for* $|t| < a$.

(2) $\sup\{|\hat{\psi}_n(t)-\psi_n(t)| : |t| < \frac{a}{2}\} \to 0$ *in probability.*

Part (1) of Lemma 3.1 implies that not only are the distributions $E\hat{F}_n$ tight but also any limit point $\tilde{F}$ of a subsequence $E\hat{F}_{n_k}$ possesses a Laplace transform $\tilde{\psi}$ finite on $(-a,a)$, which is the limit of ψ_{n_k} on $(-a,a)$.

Lemma 3.2. Such an $\tilde{F}$ *may be represented as the convolution of the distribution of* $(1-\alpha)\varepsilon$, *and another distribution. Equivalently there exists a moment generating function* ρ *such that for* $|t| < a$,

$$\tilde{\psi}(t) = \psi((1-\alpha)t)\rho(t). \tag{3.3}$$

Proof. By (A.3) of the appendix,

$$\psi_n(t) = \frac{1}{n}\Sigma_{i=1}^{n} \psi(\Gamma_{ii}t) \Pi_{j\neq i} \psi(\Gamma_{ij}t) .$$

Since λ is degenerate, for $|t| < a$,

$$\frac{1}{n}\Sigma_{i=1}^{n} |\psi(\Gamma_{ii}t)-\psi((1-\alpha)t)| \to 0.$$

So $\{\psi_{n_k}\}$ behave like

$$\psi((1-\alpha)t)\frac{1}{n_k}\Sigma_{i=1}^{n_k} \Pi_{j\neq i} \psi(\Gamma_{ij}t) . \tag{3.4}$$

If the ψ_{n_k} converge to $\tilde{\psi}$, the second factor in (3.4) does also, to ρ (say). Since ρ is the limit of moment generating functions, the lemma follows.

Proof of Theorem. We argue by contradiction.

Claim (1). If $\tilde{F}$ is a weak limit point of $\tilde{F}_{n_k}$ in probability, it is by part (2) of Lemma 3.1 also a limit point of EF_{n_k}. Suppose that $\tilde{F} = F$. Then we would have a contradiction to assumption (a) by Lemma 3.2.

Claim (2). Since λ is degenerate we can find $\{i_n\}$ such $\Gamma_{i_n i_n} \to 1-\alpha$. Without loss of generality, suppose $i_n = 1$ for all n.

Take c^T to be the first row of X. We will argue that any weak limit of $c^T(\hat{\beta}^*-\hat{\beta})$ given $Y_1,\ldots,Y_n$ is different from the corresponding weak limit of $c^T(\hat{\beta}-\beta)$ and thus prove (2). Note first

$$c^T(\hat{\beta}^*-\beta) = \Sigma_{i=1}^n H_{li}\varepsilon_i^* \tag{3.5}$$

$$c^T(\hat{\beta}-\beta) = \Sigma_{i=1}^n H_{li}\varepsilon_i^* \tag{3.6}$$

where

$$H = X(X^TX)^{-1}X .$$

Arguing as in Lemma 3.1 and using (3.6), we get

$$E(\exp(c^T(\hat{\beta}-\beta))) \leq \exp \Sigma_{t=2}^{\infty}\{\frac{|\kappa_r|}{r!}|t|^r\} .$$

Hence the distributions of $c^T(\hat{\beta}-\beta)$ are tight and any weak limit must have cumulants of all orders which are the limits of the cumulants of $c^T(\hat{\beta}-\beta)$. Similarly, by (3.5), and the Marcinkiewicz inequality, for an absolute constant a_k,

$$\begin{aligned} E\{(c^T(\hat{\beta}-\beta))^{2k}|Y_1,\ldots,Y_n\} &\leq a_k \Sigma_{i=1}^n H_{li}^{2k}E\{[\varepsilon_i^*]^{2k}|Y_1,\ldots,Y_n\} \\ &\leq a_k\int x^{2k}d\hat{F}_n . \end{aligned} \tag{3.7}$$

By Lemma 3.1(1), the right-hand side of (3.7) is bounded in probability so any weak limit of the conditional distribution of $c^T(\beta^*-\beta)$ must also have cumulants which are the limits in probability of the conditional cumulants of $c^T(\hat{\beta}^*-\hat{\beta})$. But, then, if we let $\hat{\kappa}_r$ be the rth cumulant of $\hat{F}$ and $\hat{\kappa}_{rn}$ that of F_n, the cumulants of $c^T(\hat{\beta}-\beta)$ and the conditional cumulants of $c^T(\hat{\beta}^*-\hat{\beta})$ are respectively,

$$\left(\Sigma_{i=1}^n H_{li}^r\right)\kappa_r \quad \text{and} \quad \left(\Sigma_{i=1}^n H_{li}^r\right)\hat{\kappa}_{rn} . \tag{3.8}$$

Now consider a subsequence n_ℓ for which $\hat{F}_{n_\ell}$ converges to $\tilde{F}$ and $c^T(\hat{\beta}-\beta)$ converges weakly to G while $c^T(\hat{\beta}^*-\hat{\beta})$ converges weakly (conditionally) to G^*.

We claim that if $G = G^*$ we must have $F = \tilde{F}$ and hence arrive at a contradiction to part (1) of the theorem. If $\tilde{\kappa}_r$ are the cumulants of $\tilde{F}$ and $\hat{F}_{n_\ell}$ converges to $\tilde{F}$ we must have

$$\hat{\kappa}_{rn_\ell} \longrightarrow \tilde{\kappa}_r \tag{3.9}$$

in probability for each r. In view of (3.8) and (3.9), convergence of $c^T(\hat{\beta}-\beta)$ implies the existence of $\gamma_1,\gamma_2,\ldots$ such that

$$\Sigma_{i=1}^n H_{1i}^r \longrightarrow \gamma_r \tag{3.10}$$

for each r. Moreover, the cumulants of G and G^* are respectively $\kappa_r\gamma_r$ and $\tilde{\kappa}_r\gamma_r$. If r is even, $\gamma_r \geq \alpha^r > 0$, since the sum in (3.10) is no smaller than H_{11}^r. So if $G = G^*$,

$$\kappa_r = \tilde{\kappa}_r \quad \text{for} \quad r \quad \text{even} . \tag{3.11}$$

But if F is symmetric about 0, so are G and G^* and

$$\kappa_r = \tilde{\kappa}_r = 0 \quad \text{for} \quad r \quad \text{odd}. \tag{3.12}$$

Since F has a moment generating function (3.11) and (3.12) imply $F = \tilde{F}$ and we have our contradiction.

APPENDIX

Lemma 3.1

Before proving Lemma 3.1 we need the preliminary

Lemma A.1. *If* $\min(a,b) \geq 1$, $\max(a,b) \geq 2$

$$\Sigma_{i,j,k} |\Gamma_{ik}|^a |\Gamma_{jk}|^b \leq n^{3/2} . \tag{A.1}$$

Moreover,

$$\Sigma_{i,j} |\Gamma_{ij}| \leq n^{3/2} . \tag{A.2}$$

Proof. For (A.2) apply Cauchy Schwartz to get the bound

$$n^{1/2} \Sigma_i \left(\Sigma_j \Gamma_{ij}^2\right)^{1/2} = n^{1/2} \Sigma_i \Gamma_{ii}^{1/2} \leq n^{3/2} .$$

For (A.1) bound by

$$\Sigma_{i,j,k} \Gamma_{ik}^2 |\Gamma_{jk}| = \Sigma_{j,k} \Gamma_{kk} |\Gamma_{jk}| \leq \Sigma_{jk} |\Gamma_{jk}| \leq n^{3/2} \quad \text{by (A.2).}$$

Proof of Lemma 3.1. (1) Since

$$\hat{\varepsilon}_i = \Sigma_j \Gamma_{ij}\varepsilon_i$$

$$\psi_n(t) = \frac{1}{n}\Sigma_i \Pi_{j=1}^n \psi(\Gamma_{ij}t) . \tag{A.3}$$

By (3.1), if $|t| < a$,

$$\psi_n(t) = \frac{1}{n}\Sigma_{i=1}^n \exp\{\Sigma_{j=1}^n \Sigma_{r=2}^\infty \frac{\kappa_r}{r!}\Gamma_{ij}^r t^r\} . \tag{A.4}$$

(Note that by assumption $\kappa_1 = 0$.)

Take absolute values and change the order of summation in the exponent to get the bound

$$\frac{1}{n}\Sigma_{i=1}^n \exp\{\Sigma_{r=2}^\infty \frac{|\kappa_r|}{r!}|t|^r \Sigma_{j=1}^n |\Gamma_{ij}|^r\} .$$

Claim (1) follows since $\Sigma_{j=1}^n |\Gamma_{ij}|^r \le \Gamma_{ii} \le 1$ for $r \ge 2$.

(2) Compute

$$E(\hat{\psi}_n(t)-\psi_n(t))^2 = n^{-2}\{\Sigma_{i,j}[E\exp(t(\hat{\varepsilon}_i+\hat{\varepsilon}_j)) - E\exp(t\hat{\varepsilon}_i)\, E\exp(t\hat{\varepsilon}_j)]\}. \tag{A.5}$$

Arguing as for (A.3), the right-hand side of (A.5) reduces to

$$\frac{1}{n^2}\Sigma_{i,j}[\Pi_{k=1}^n \psi(t(\Gamma_{ik}+\Gamma_{jk})) - \Pi_{k=1}^n \psi(t\Gamma_{ik})\psi(t\Gamma_{jk})] . \tag{A.6}$$

The two terms in each summand are bounded in absolute value by $\psi^n(2t)$ since $|\Gamma_{ik}| \le 1$ for all i, k. So, if $|t| < \frac{a}{2}$, we can apply claim (1) and the inequality $|e^a-e^b| \le \max(e^{|a|},e^{|b|})|b-a|$ to bound the right-hand side of (A.6) by a constant depending only on $\Sigma_{t=2}^\infty \frac{|\kappa_r|}{r!}|t|^r$ (but not n) times

$$\frac{1}{n^2}\Sigma_{i,j}\left|\Sigma_{k=1}^n[\log\psi(t(\Gamma_{ik}+\Gamma_{jk})) - \log\psi(t\Gamma_{ik}) - \log\psi(t\Gamma_{jk})]\right| .$$

Arguing as for (A.4) we can bound this expression by

$$\frac{1}{n^2}\Sigma_{i,j}\left|\Sigma_k \Sigma_{r=2}^\infty \frac{\kappa_r}{r!}t^r[(\Gamma_{ik}+\Gamma_{jk})^r - \Gamma_{ik}^r - \Gamma_{jk}^r]\right| .$$

Recalling that $\Sigma_k \Gamma_{ik}\Gamma_{jk} = \Gamma_{ij}$ we can bound this expression by

$$t^2\kappa_2 n^{-2} \Sigma_{i,j}|\Gamma_{ij}| + \Sigma_{r=3}^{\infty}|t|^r \frac{|\kappa_r|}{r!} \Sigma_{\ell=1}^{r-1} \binom{t}{\ell} n^{-2} \Sigma_{i,j,k}|\Gamma_{ik}|^{\ell}|\Gamma_{jk}|^{r-\ell}.$$

Since $\min(\ell, r-\ell) \geq 1$, $\max(\ell, r-\ell) \geq 2$ we can apply Lemma A.1 to get the bound

$$n^{-1/2}\left(t^2\kappa_2 + \Sigma_{r=3}^{\infty} \frac{|\kappa_r|}{r!}|t|^r 2^r\right) = O(n^{-1/2})$$

if $|t| < \frac{a}{2}$. So, if $|t| < \frac{a}{2}$

$$E(\hat{\psi}_n(t)-\psi_n(t))^2 = O(n^{-1/2}) \text{ and } \hat{\psi}_n(t) - \psi_n(t) \xrightarrow{p} 0. \quad (A.7)$$

Now, for any pair s, t, $|s|$, $|t| \leq b < \frac{a}{2}$

$$|\hat{\psi}_n(t) - \psi_n(t) - \hat{\psi}_n(s) + \psi_n(t)| \leq \sup\{|\hat{\psi}_n'(s)| + |\psi_n'(s)|: |s| \leq b\}|t-s| .$$

Since ψ_n, $\hat{\psi}_n$ are Laplace transforms, the first factor above is

$$O_p\left(\int \exp\{(b+\delta)|x|\} d\hat{F}_n(x) + \int \exp\{(b+\delta)|x| dF_n(x)\right)$$

for any $\delta > 0$. By (1) of Lemma 3.1 again we conclude that $\hat{\psi}_n - \psi_n$ satisfies a uniform Lipschitz condition on any closed subinterval of $[-a/2, a/2]$. In view of (A.7), part (2) of the lemma follows.

REFERENCES

Albers, W., Bickel, P.J., and van Zwet, W.R. (1976), "Asymptotic Expansions for the Power of Distribution Free Tests in the One Sample Problem." *Annals of Statistics*, 4, 108-156.

Bickel, P. and Freedman, D. (1981), "Some Asymptotic Theory for the Bootstrap." *Annals of Statistics*, 9, 1196-1217.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statictics*, 7, 1-26.

Freedman, D. (1981), "Bootstrapping Regression Models." *Annals of Statistics*, 9, 1218-1228.

Freedman, D. and Lane, D. (1981), "The Empirical Distribution of the Fourier Coefficients of Independent, Identically Distributed Long-Tailed Random Variables." *Z. Wahrscheinlichkeitstheorie* 58, 21-39.

Huber, P.J. (1973), "Robust Regression." *Annals of Statistics*, 1, 799-821.

Festschrift, 1983 (P.J. Bickel, K. Doksum, and J.L. Hodges, Jr., eds). Wadsworth Press, Belmont.