The Annals of Statistics
1981, vol. 9, No. 6, 1218–1228

BOOTSTRAPPING REGRESSION MODELS

BY D. A. FREEDMAN[1]

University of California, Berkeley

The regression and correlation models are considered. It is shown that the bootstrap approximation to the distribution of the least squares estimates is valid, and some error bounds are given.

1. Introduction. This paper, a sequel to Bickel and Freedman (1981a), will develop some asymptotic theory for applications of Efron's (1977) bootstrap to regression. Autoregressive models and linear econometric models may be considered elsewhere. Here, only multiple linear regression models will be considered:

$$Y(n) = X(n)\beta + \varepsilon(n). \tag{1.1}$$

In this equation, β is a $p \times 1$ vector of unknown parameters, to be estimated from the data; $Y(n)$ is an $n \times 1$ data vector; $X(n)$ is an $n \times p$ data matrix, of full rank $p \leqq n$; $\varepsilon(n)$ is an $n \times 1$ vector of unobservables. The models differ in the stochastic mechanism supposed to have generated the data. However, in all cases ε is supposed random. Throughout this paper, p is fixed and n is large. The case where p and n are both large will be considered in a future paper, Bickel and Freedman (1981b).

Attention is restricted to the conventional least squares estimate $\hat{\beta}(n)$ for β, given by

$$\hat{\beta}(n) = \{X(n)^T X(n)\}^{-1} X(n)^T Y(n).$$

How close is $\hat{\beta}(n)$ to β? The object of this paper is to compare the bootstrap approximation with the standard asymptotics. Under mild conditions, the bootstrap is valid. However, details depend on the model.

To fix ideas, it may be useful to review two different kinds of models, "regression" and "correlation", and to indicate the results for each. Naturally, the mathematics may apply in other cases as well. In the regression model, X is fixed and the errors are "homoscedastic." In the correlation model, X is random and the errors are in general "heteroscedastic": the conditional distribution of the errors given X depends on X. (As it turns out, only the behavior of the conditional second moment counts.) In order to succeed, the bootstrap simulation must reflect the relevant features of the stochastic model assumed to have generated the data. This point is also discussed by Efron (1977, Section 7), and in the jackknife context by Hinkley (1977).

The regression model. This is appropriate if, for example, the basic source of uncertainty is measurement error. An instance is the weighing designs used in precision calibration work at the National Bureau of Standards. The main assumptions are as follows.

(1.2) The matrix $X(n)$ is not random.

Received August 1980; revised April 1981.

[1] Research partially supported by NSF Grant MCS-80-02535. I worked on this paper while enjoying the hospitality of the ETH, Zurich.

AMS 1980 *subject classifications.* Primary 62E20; secondary 62G05, 62G15.

Key words and phrases. Regression, correlation, least squares, bootstrap, Wasserstein metrics.

(1.3) The components $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ of $\varepsilon(n)$ are independent, with common distribution F having mean 0 and finite variance σ^2; both F and σ^2 are unknown.

The Y-vector is considered to be the observed value of the random vector $X(n)\beta + \varepsilon(n)$. Then $\hat{\beta}(n)$ has mean β and variance-covariance matrix $\sigma^2\{X(n)^TX(n)\}^{-1}$. Suppose

$$\frac{1}{n}X(n)^TX(n) \to V \text{ which is positive definite.} \tag{1.4}$$

Suppose also that the elements of $X(n)$ are uniformly small by comparison with $\sqrt{n}$. Then $\sqrt{n}\{\hat{\beta}(n) - \beta\}$ is asymptotically normal, with mean 0 and variance-covariance matrix σ^2V^{-1}. In particular, the distribution of the pivotal quantity $\{X(n)^TX(n)\}^{1/2}\{\hat{\beta}(n) - \beta\}/\sigma$ 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 $(X^TX)^{1/2}$. "Positive definite" is taken in the strict sense.

The correlation model. This is appropriate if, for example, it is desired to estimate the regression plane for a certain population on the basis of a simple random sample, and to quantify the sampling error in the estimate. Now X must be considered random, and ε may be related to X. The ith row in the data array (X, Y) will be denoted (X_i, Y_i), representing the measurements on the ith subject in the sample; so (X_i, Y_i) is a random $(p + 1)$ dimensional row vector. Potentially, there are infinitely many such vectors; the rows of $X(n)$ consist of the first n of the X_i's and $Y(n)$ consists of the first n of the Y_i's.

(1.5) The vectors (X_i, Y_i) are assumed independent, with common (unknown) distribution μ in R^{p+1}; and $E\{\|(X_i, Y_i)\|^4\} < \infty$, where $\|\cdot\|$ is Euclidean length.

By convention, X_i is a row vector. Let $\Sigma = E(X_i^TX_i)$, the $p \times p$ variance covariance matrix of any row of X. Assume

$$\Sigma \text{ is positive definite.} \tag{1.6}$$

The p-vector β of parameters is defined as the vector which minimizes $E(\|Y_i - X_i\beta\|^2)$; equivalently, $Y_i - X_i\beta$ is orthogonal to X_i:

$$E(X_{ij}\varepsilon_i) = 0 \quad \text{for } j = 1, \cdots, p, \quad \text{where } \varepsilon_i = Y_i - X_i\beta. \tag{1.7}$$

Of course $\varepsilon(n)$ in (1.1) consists of the first n of the ε_i's. Relationship (1.7) entails

$$\beta = \Sigma^{-1}E(X_i^TY_i). \tag{1.8}$$

If the nonnegative definite matrix M is defined by

$$M_{jk} = E(X_{ij}X_{ik}\varepsilon_i^2) \tag{1.9}$$

then the asymptotics in the model can be summarized as follows: $X(n)^T\varepsilon(n)/\sqrt{n}$ is asymptotically normal, with mean 0 and variance-covariance matrix M. Since

$$\sqrt{n}\{\hat{\beta}(n) - \beta\} = \left\{\frac{1}{n}X(n)^TX(n)\right\}^{-1} \cdot \frac{1}{\sqrt{n}}X(n)^T\varepsilon(n) \tag{1.10}$$

and

$$\frac{1}{n}X(n)^TX(n) \to \Sigma \quad \text{a.e.},$$

it follows that $\sqrt{n}\{\hat{\beta}(n) - \beta\}$ is asymptotically normal, with mean 0 and variance-covariance matrix $\Sigma^{-1}M\Sigma^{-1}$.

The Σ in the correlation model plays the role of V in the regression model. However, the asymptotics are quite different, unless the correlation model is "homoscedastic," which can be interpreted mathematically as follows:

$$E(\varepsilon_i^2 \mid X_i) = \sigma^2 \quad \text{a.e., where } \sigma^2 = E(\varepsilon_i^2). \tag{1.11}$$

Then $M = \sigma^2 \Sigma$; as a result, in the homoscedastic case, the correlation model has the same asymptotics as the regression model, Perhaps surprisingly, the condition $E(\varepsilon_i \mid X_i) = 0$ does not seem to be needed, or even $E(\varepsilon_i) = 0$.

Bootstrapping. The object of this paper is to show that if the simulations are done in a manner consistent with the model, the bootstrap will give the same asymptotic results as classical methods. In the regression model, it is appropriate to resample the centered residuals. More specifically, the observable column n-vector $\hat{\varepsilon}(n)$ of residuals is given by $\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta}$. However, $\hat{\mu}_n = (1/n)\sum_{i=1}^n \hat{\varepsilon}_i(n)$ need not vanish, for the column space of X need not include the constant vectors. Let $\hat{F}_n$ be the empirical distribution of $\hat{\varepsilon}(n)$, centered at the mean, so $\hat{F}_n$ puts mass 1/n at $\hat{\varepsilon}_i(n) - \hat{\mu}_n$ and $\int x\, d\hat{F}_n{}^x = 0$. Given $Y(n)$, let $\varepsilon_1^*, \cdots, \varepsilon_n^*$ be conditionally independent, with common distribution $\hat{F}_n$; let $\varepsilon^*(n)$ be the n-vector whose ith component is ε_i^*; and let

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

Informally, ε^* is obtained by resampling the centered residuals. And Y^* is generated from the data, using the regression model with $\hat{\beta}$ as the vector of parameters and $\hat{F}_n$ as the distribution of the disturbance terms ε. Now imagine giving the starred data (X, Y^*) to another statistician, and asking for an estimate of the parameter vector. The least squares estimate is $\hat{\beta}^* = (X^TX)^{-1}X^TY^*$. The bootstrap principle is that the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$, which can be computed directly from the data, approximates the distribution of $\sqrt{n}(\hat{\beta} - \beta)$. As will be shown in Section 2 below, this approximation is likely to be very good, provided n is large and $\sigma^2 p \cdot \text{trace}(X^TX)^{-1}$ is small.

What happens if the residuals are not centered before resampling? Suppose the constant vectors are neither included in nor orthogonal to the column space of X. Then the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ incorporates a bias term which is random (depending on $\varepsilon_1, \cdots, \varepsilon_n$) and which in general has a nondegenerate normal limiting distribution. This is so despite the fact that the empirical distribution of the uncentered residuals converges to F. In short, without centering, the bootstrap will usually fail. Efron (1977) asks about this issue in his Section 7. This completes a sketch of the bootstrap for the regression model; details will be found in Section 2.

Turning now to the correlation model, there is in general some dependence between ε_i and X_i, "heteroscedacity." So it is inappropriate to resample the residuals, for that obliterates the dependence. Instead, it is necessary to resample the vectors. More specifically, let μ_n be the empirical distribution of the (X_i, Y_i) for $i = 1, \cdots, n$. Thus, μ_n is a probability on R^{p+1}, putting mass $1/n$ at each vector (X_i, Y_i). Given $\{X(n), Y(n)\}$, let (X_i^*, Y_i^*) be independent, with common distribution μ_n, for $i = 1, \cdots, m$. Informally, this amounts to taking a resample of size m from the n observed vectors. The technical advantages of letting $m \neq n$ are seen in Bickel and Freedman (1981a). Informally, data from a small sample can be used to judge the likely performance of a larger sample.

Remember that $\hat{\beta}(n)$ minimizes $1/n \sum_{i=1}^n \{Y_i - X_i\hat{\beta}(n)\}^2$. Thus, $\hat{\beta}(n)$ is to μ_n as β is to the true law μ of (X_i, Y_i). Let $\hat{\beta}^*(m)$ be the least squares estimate based on the resample:

$$\hat{\beta}^*(m) = \{X^*(m)^TX^*(m)\}^{-1}X^*(m)^TY^*(m). \tag{1.12}$$

In Section 3, it will be shown that the conditional law of $\sqrt{m}\ \{\hat{\beta}^*(m) - \beta\}$ must be close to the unconditional law of $\sqrt{n}\hat{\beta}\{(n) - \beta\}$, i.e., the bootstrap approximation is valid. Notice that in the correlation model, unlike the regression model, the first m rows of the starred design matrix are random. Notice too that in the correlation model, the residuals should not be centered before resampling: they are already orthogonal to X.

2. The regression model. Assume the regression model, with (1.1-2-3). Let $\Psi_n(F)$ be the distribution of $\sqrt{n}\{\hat{\beta}(n) - \beta\}$, when F is the law of the ε's. So $\Psi_n(F)$ is a probability in R^p. For the next theorem, let G be an alternative law for the ε's: assume G also has mean 0 and finite variance. In applications, G will be the centered empirical distribution of the residuals. Bickel and Freedman (1981a) will be abbreviated B & F, due to frequent citation.

DEFINITION 2.1. Let d_i^p be the Mallows metric for probabilities in R^p, relative to the Euclidean norm $\|\cdot\|$. Thus, if μ and ν are probabilities in R^p, $d_i^p(\mu, \nu)$ is the infimum of $E[\|U - V\|^i]^{1/i}$ over all pairs of random vectors U and V, where U has law μ and V has law ν. Abbreviate d_i for d_i^1. For details, see Section 8 of B & F. Only $i = 1$ or 2 are of present interest.

NOTATION. In the present paper, p is the dimension of a linear space; in Section 8 of B & F, however, p is the index of an L_p space.

THEOREM 2.1. $d_2^p\{\Psi_n(F), \Psi_n(G)\}^2 \leqq n \cdot \text{trace}\{X(n)^T X(n)\}^{-1} \cdot d_2(F, G)^2$.

PROOF. Let $A(n) = \{X(n)^T X(n)\}^{-1} X(n)^T$. Then $\Psi_n(F)$ is the law of $\sqrt{n}A(n)\varepsilon(n)$ where $\varepsilon(n)$ is an n-vector of independent random variables ε_i having common law F. Likewise for G. Now use Lemma 8.9 of B & F, observing that $A(n)A(n)^T = \{X(n)^T X(n)\}^{-1}$. Also see (8.2) of B & F. □

To proceed, let F_n be the empirical distribution function of $\varepsilon_1, \cdots, \varepsilon_n$; let $\tilde{F}_n$ be the empirical distribution of the residuals $\hat{\varepsilon}_1(n), \cdots, \hat{\varepsilon}_n(n)$ from the original regression on n data vectors, and let $\hat{F}_n$ be $\tilde{F}_n$ centered at its mean $\hat{\mu}_n = (1/n)\sum_{i=1}^n \hat{\varepsilon}_i(n)$. Since $\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta}(n)$,

$$\hat{\varepsilon}(n) - \varepsilon(n) = -P(n)\varepsilon(n) \tag{2.1}$$

where $P(n) = X(n)\{X(n)^T X(n)\}^{-1} X(n)^T$ is the projection matrix onto the column space of $X(n)$.

LEMMA 2.1. $E\{d_2(\tilde{F}_n, F_n)^2\} \leqq \sigma^2 p/n$.

PROOF. A routine computation starting from (2.1) shows

$$E\{\|\hat{\varepsilon}(n) - \varepsilon(n)\|^2\} = \sigma^2 p. \tag{2.2}$$

But

$$d_2(\tilde{F}_n, F_n)^2 \leqq \frac{1}{n}\sum_{i=1}^n \{\hat{\varepsilon}_i(n) - \varepsilon_i\}^2 = \frac{1}{n}\|\hat{\varepsilon}(n) - \varepsilon(n)\|^2. \tag{2.3}$$

□

LEMMA 2.2. $E\{d_2(\hat{F}_n, F_n)^2\} \leqq \sigma^2(p + 1)/n$.

PROOF. Two applications of Lemma 8.8 of B & F show that

$$d_2(\hat{F}_n, F_n)^2 = \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right)^2 - \left[\frac{1}{n}\sum_{i=1}^n \{\hat{\varepsilon}_i(n) - \varepsilon_i\}\right]^2 + d_2(\tilde{F}_n, F_n)^2. \tag{2.4}$$

Now use the present Lemma 2.1. □

REMARK. The negative term in (2.4) is a bit disconcerting. However, it is small. To see this, let the $n \times 1$ column vector $v(n)$ be identically 1. Using (2.1),

$$E\left(\left[\frac{1}{n}\sum_{i=1}^n \{\hat{\varepsilon}_i(n) - \varepsilon_i\}\right]^2\right) = \frac{\sigma^2}{n^2}\|P(n)v(n)\|^2 \leqq \frac{\sigma^2}{n}.$$

As will now be shown, these results imply the validity of the bootstrap approximation, in probability, assuming for example (1.4). Recall Ψ_n from the beginning of the section. Now

$$E[d_2^p\{\Psi_n(\hat{F}_n), \Psi_n(F)\}^2] \leqq n \cdot \text{trace}\{X(n)^T X(n)\}^{-1} \cdot E\{d_2(\hat{F}_n, F)^2\} \tag{2.5}$$

and because d_2 is a metric

$$\tfrac{1}{2} d_2(\hat{F}_n, F)^2 \leqq d_2(\hat{F}_n, F_n)^2 + d_2(F_n, F)^2 \tag{2.6}$$

The first term goes to 0 in probability by Lemma 2.2; the second, by Lemma 8.4 of B & F: and $n \cdot \text{trace}\{X(n)^T X(n)\}^{-1} = O(1)$ by (1.4). Of course, condition (1.4) can be weakened appreciably, and p can be allowed to go to infinity slowly: Bickel and Freedman (1981b).

Rather than pursuing this, a theorem for convergence a.e. will be given. Consider $X(n)$ as the first n of an infinite sequence of rows. Likewise, consider the disturbances $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ as the first n of an infinite sequence of independent random variables with common distribution function F. The original regression problem has now been embedded into an infinite sequence of such problems. Allow the resample size m to differ from n. For motivation, consider again the weighing designs used in precision calibration. The error distribution depends in principle only on the apparatus and procedures used, not on the specific weights. Thus, it may be possible to use data from one design to assess the probable accuracy from another.

This simulation will now be spelled out in more detail. Recall that $\hat{\beta}(n)$ is the estimate of β, based on the first n data points. The starred data is generated by the recipe

$$\underset{m\times 1}{Y^*(m)} = \underset{m\times p}{X(m)}\ \underset{p\times 1}{\hat{\beta}(n)} + \underset{m\times 1}{\varepsilon^*(m)}$$

the $\varepsilon_1^* \cdots \varepsilon_m^*$ being independent with common distribution $\hat{F}_n$, the empirical distribution of the residuals from the original data set, but centered at the mean μ_n. Now $\hat{\beta}^*(m)$ is the parameter estimate based on the starred data:

$$\underset{p\times 1}{\hat{\beta}^*(m)} = [\underset{p\times p}{X(m)^T X(m)}]^{-1}\ \underset{p\times m}{X(m)^T}\ \underset{m\times 1}{Y^*(m)}.$$

The starred residuals are

$$\underset{m\times 1}{\hat{\varepsilon}^*(m)} = \underset{m\times 1}{Y^*(m)} - \underset{m\times p}{X(m)}\underset{p\times 1}{\hat{\beta}^*(m)}.$$

The theoretical variance $\sigma^2 = E\{\varepsilon_i^2\}$ is estimated from the n original data vectors by

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2(n) - \mu_n^2, \qquad \text{where} \quad \mu_n = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i(n). \tag{2.7}$$

Likewise, the variance estimate from the m starred vectors is

$$\hat{\sigma}_m^{*2} = \frac{1}{m}\sum_{i=1}^m \hat{\varepsilon}_i^{*2}(m) - \mu_m^{*2}, \qquad \text{where} \quad \mu_m^* = \frac{1}{m}\sum_{i=1}^m \hat{\varepsilon}_i^*(m). \tag{2.8}$$

In principle, the starred data, as well as $\hat{\beta}^*(m)$ and $\hat{\sigma}_m^*$, depend on n; this is suppressed in the notation. The estimates $\hat{\sigma}^2$ are slightly biased, but this is immaterial for present purposes.

The next result is a special case of results in Lai *et al.* (1979), which gives further references. Write $\varepsilon(n)$ for the $n \times 1$ column vector $\varepsilon_1, \cdots, \varepsilon_n$. Likewise, write $\hat{\varepsilon}(n)$ for the $n \times 1$ column vector of residuals from the regression on the first n data points.

LEMMA 2.3. $\dfrac{1}{n} X(n)^T \varepsilon(n) \to 0$ a.e. *and* $\hat{\beta}(n) \to \beta$ a.e.

PROOF. Use Kolmogorov's inequality, along the subsequence of powers of 2. □

LEMMA 2.4. $\dfrac{1}{n} \| \hat{\varepsilon}(n) - \varepsilon(n) \|^2 \to 0$ a.e.

PROOF. As is easily seen from (2.1),

$$\frac{1}{n}\,\|\,\hat{\varepsilon}(n) - \varepsilon(n)\,\|^2 = \left\{\frac{1}{n}\,\varepsilon(n)^T X(n)\right\}\cdot\left\{\frac{1}{n}\,X(n)^T X(n)\right\}^{-1}\cdot\left\{\frac{1}{n}\,X(n)^T\varepsilon(n)\right\}. \tag{2.9}$$

But the first and third factors go to 0 a.e. by Lemma 2.3; the middle factor goes to V^{-1} by assumption (1.4) □

LEMMA 2.5. $d_2(\hat{F}_n, F_n) \to 0$ a.e.

PROOF. Use (2.3), (2.4) and Lemma 2.4. □

LEMMA 2.6. $d_2(\hat{F}_n, F) \to 0$ a.e.

PROOF. Use Lemma 2.5, and Lemma 8.4 of B & F. □

LEMMA 2.7. *Let u_i and v_i be real numbers. Let*

$$\bar{u} = \frac{1}{n}\sum_{i=1}^n u_i \quad \textit{and} \quad s_u^2 = \frac{1}{n}\sum_{i=1}^n (u_i - \bar{u})^2,$$

and likewise for v. Then

$$(s_u - s_v)^2 \leqq \frac{1}{n}\sum_{i=1}^n (u_i - v_i)^2.$$

PROOF. Clearly, $s_u = \|\,u - \bar{u}\,\|/\sqrt{n}$ and likewise for v, so

$$\begin{aligned}(s_u - s_v)^2 &\leqq \frac{1}{n}\,\|\,(u - \bar{u}) - (v - \bar{v})\,\|^2 \\ &= \frac{1}{n}\,[\,\|\,u - v\,\|^2 - (\bar{u} - \bar{v})^2] \\ &\leqq \frac{1}{n}\,\|\,u - v\,\|^2.\end{aligned}$$

□

The next theorem is the a.e. justification of the bootstrap asymptotics. The behavior of the pivot will be considered in more detail in Bickel and Freedman (1981b).

THEOREM 2.2. *Assume the regression model, with* (1.1–4). *Along almost all sample sequences, given $Y_1, \cdots, Y_n$, as m and n tend to ∞,*

a) *the conditional distribution of $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ converges weakly to normal with mean 0 and variance-covariance matrix $\sigma^2 V^{-1}$.*

b) *the conditional distribution of $\hat{\sigma}_m^*$ converges to point mass at σ.*

c) *the conditional distribution of the pivot $\{X(m)^T X(m)\}^{1/2}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}/\hat{\sigma}_m^*$ converges to standard normal in R^p.*

PROOF. Claim (a) is immediate from Theorem 2.1 and Lemma 2.6. Indeed, in the theorem, one puts m for n and $\hat{F}_n$ for G. Now

$$m\cdot\text{trace}[X(m)^T X(m)]^{-1}\cdot d_2(F, \hat{F}_n)^2 = \text{trace}\left[\frac{1}{m}\,X(m)^T X(m)\right]^{-1}\cdot d_2(F, \hat{F}_n)^2 \to 0 \qquad \text{a.e.}$$

because $(1/m)X(m)^T X(m) \to V$ positive definite by assumption (1.4). By construction, $X(n)$ is the first n rows of $X(n+1)$, so (1.4) entails that the elements of $X(n)$ are uniformly $o(\sqrt{n})$, and $X(n)^T\varepsilon(n)/\sqrt{n}$ is asymptotically normal.

Claim (b). Recall $\hat{\sigma}_n$ from (2.7). It will be shown that

$$\hat{\sigma}_n \to \sigma \qquad \text{a.e.} \tag{2.10}$$

To argue this, introduce

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right)^2.$$

Clearly, $\sigma_n \to \sigma$ a.e. In view of Lemmas 2.7 and 2.4,

$$(\hat{\sigma}_n - \sigma_n)^2 \leqq \frac{1}{n}\sum_{i=1}^n \{\hat{\varepsilon}_i(n) - \varepsilon_i\}^2 \to 0 \qquad \text{a.e.}$$

Next, let

$$\sigma_m^{*2} = \frac{1}{m}\sum_{i=1}^m \varepsilon_i^{*2} - \left(\frac{1}{m}\sum_{i=1}^m \varepsilon_i^*\right)^2.$$

Recall from (2.8) that $\hat{\sigma}_m^{*2}$ is the variance of the residuals in the starred regression. Now

$$\begin{aligned} E(|\hat{\sigma}_m^* - \sigma_m^*| \mid Y_1, \cdots, Y_n)^2 &\leqq E\{(\hat{\sigma}_m^* - \sigma_m^*)^2 \mid Y_1, \cdots, Y_n\} \\ &\leqq E\left[\frac{1}{m}\sum_{i=1}^m \{\hat{\varepsilon}_i^*(m) - \varepsilon_i^*\}^2 \mid Y_1, \cdots, Y_n\right] \qquad \text{by Lemma 2.7} \\ &= \hat{\sigma}_n^2 p/m \qquad \text{by (2.2) applied to the starred regression} \\ &\to 0 \qquad \text{a.e. by (2.10).} \end{aligned}$$

What remains is to show that the conditional law of σ_m^{*2} is nearly point mass at σ^2. This follows from the results in Section 8 of B & F. Indeed, condition on $Y_1, \cdots, Y_n$. By Lemma 8.6 of B & F,

$$d_1\left(\frac{1}{m}\sum_{i=1}^m \varepsilon_i^{*2}, \frac{1}{m}\sum_{i=1}^m \varepsilon_i^2\right) \leqq d_1(\varepsilon_i^{*2}, \varepsilon_i^2)$$

(Both sides of the display are random; for the distance computed is between the conditional distribution of the starred quantity and the unconditional distribution of the unstarred quantity.) Now ε_i^* has conditional law $\hat{F}_n$; and ε_i has law F; and $d_2(\hat{F}_n, F) \to 0$ a.e. by Lemma 2.6. So $d_1[\varepsilon_i^{*2}, \varepsilon_i^2] \to 0$ a.e. by Lemma 8.5 of B & F, with $\phi(\varepsilon) = \varepsilon^2$. In short, the conditional law of $1/m\sum_{i=1}^m \varepsilon_i^{*2}$ differs little from the unconditional law of $1/m\sum_{i=1}^m \varepsilon_i^2$, and must therefore concentrate near σ^2. Likewise, the conditional law of $1/m\sum_{i=1}^m \varepsilon_i^*$ concentrates near 0.

Claim (c). This is immediate from (a) and (b). □

To conclude this section, consider the bootstrap when the uncentered residuals are resampled. Let $v(m)$ be a column $(m \times 1)$ vector of 1's. Applying (1.10) to the starred regression,

$$E[\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\} \mid Y_1, \cdots, Y_n] = \left\{\frac{1}{m}X(m)^T X(m)\right\}^{-1} \cdot \frac{1}{m}X(m)^T v(m) \cdot \sqrt{\frac{m}{n}}\,\zeta_n \tag{2.11}$$

where by (2.1)

$$\zeta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\varepsilon}_i(n) = \frac{1}{\sqrt{n}}v(n)^T\{I_{n\times n} - P(n)\}\varepsilon(n).$$

Now ζ_n is scalar, $E(\zeta_n) = 0$ and

$$E(\zeta_n^2) = \sigma^2\frac{1}{n}\|\{I_{n\times n} - P(n)\}v(n)\|^2. \tag{2.12}$$

It is easy to find a sequence $X(n)$ of designs for which (1.4) holds, and $(1/m)X(m)^T v(m)$ converges to a limit L; then the right side of (2.12) converges to $\sigma^2(1 - L^T V^{-1} L)$. Assume L is nonzero, and $L^T V^{-1} L < 1$, i.e., v has a nontrivial projection into the column space of

X, and this projection is substantially shorter than v. If m is of order n, the right side of (2.11) converges to a proper Gaussian limit. If m dominates n, the right side of (2.11) blows up.

3. The correlation model. In this section, the object is to justify the bootstrap in the correlation model, by a straightforward application of the machinery in Section 8 of B & F. The following lemma will be useful. To state it, let μ_n and μ be probabilities on R^{p+1}, for which the fourth power of the Euclidean norm is integrable. A typical point in R^{p+1} will be written (x, y), where $x \in R^p$ is viewed as a row vector, and $y \in R^1$. Assume

$$\Sigma(\mu) = \int x^T x \mu(dx, dy)$$

is positive definite, and let

$$\beta(\mu) = \Sigma(\mu)^{-1} \int x^T y \mu(dx, dy);$$

$$\varepsilon(\mu, x, y) = y - x\beta(\mu).$$

LEMMA 3.1. *If* $d_4^{p+1}(\mu_n, \mu) \to 0$, *then*
a) $\Sigma(\mu_n) \to \Sigma(\mu)$ *and* $\beta(\mu_n) \to \beta(\mu)$,
b) *the* μ_n*-law of* $\varepsilon(\mu_n, x, y)x$ *converges to the* μ*-law of* $\varepsilon(\mu, x, y)x$ *in* d_2^p,
c) *the* μ_n*-law of* $\varepsilon(\mu_n, x, y)^2$ *converges to the* μ*-law of* $\varepsilon(\mu, x, y)^2$ *in* d_1.

PROOF. Claim (a) is immediate from Lemma 8.3c of B & F.

Claim (b). Weak convergence is easy, and then Lemma 8.3a of B & F can be used. Here is a sketch of the argument.

$$\| \varepsilon(\mu_n, x, y)x \|^2 = \varepsilon(\mu_n, x, y)^2 \| x \|^2$$
$$= y^2 \| x \|^2 - 2yx\beta(\mu_n) \| x \|^2 + \beta(\mu_n)^T x^T x \beta(\mu_n) \| x \|^2.$$

Integrate with respect to μ_n, and use claim (a).

Claim (c). First, the μ_n-law of $\varepsilon(\mu_n, x, y)$ converges to the μ-law of $\varepsilon(\mu, x, y)$ in d_2, by the previous argument. Then use Lemma 8.5 of B & F with $\phi(\varepsilon) = \varepsilon^2$. □

Now return to the correlation model described in Section 1. The original n data vectors are (X_i, Y_i) for $i = 1, \cdots, n$; these are independent, with common distribution μ; their empirical distribution is μ_n. Both μ and μ_n are probabilities in R^{p+1}.

LEMMA 3.2. $d_4^{p+1}(\mu_n, \mu) \to 0$ a.e. *as* $n \to \infty$.

PROOF. This is a special case of Lemma 8.4 of B & F; the variables are L_4 by (1.5). □

Turn now to bootstrapping. Given $\{X(n), Y(n)\}$, the resampled vectors (X_i^*, Y_i^*) are independent, with common distribution μ_n, for $i = 1, \cdots, m$. Let $X^*(m)$ be the $m \times p$ matrix whose ith row is X_i^*; and $Y^*(m)$ is the $m \times 1$ column vector of Y_i^*'s. The least squares estimate based on the original data is $\hat{\beta}(n)$; on the starred data, $\hat{\beta}^*(m)$: see (1.12). In the original data, the vector of unobservable disturbances is $\varepsilon(n)$, see (1.7); the observable residuals are

$$\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta}(n). \tag{3.1}$$

In the starred data, the $m \times 1$ column vector of disturbances is ε^*, with

$$\varepsilon_i^* = Y_i^* - X_i^*(m)\hat{\beta}(n). \tag{3.2}$$

The $m \times 1$ column vector of residuals is $\hat{\varepsilon}^*(m)$ with

$$\hat{\varepsilon}_i^*(m) = Y_i^* - X_i^*(m)\hat{\beta}^*(m). \tag{3.3}$$

The next result shows that the asymptotics estimated from the bootstrap are correct. Recall that Σ is the variance-covariance matrix of X_i; and M was defined in (1.9). The dependence of the starred data and $\hat{\beta}^*(m)$ on n is suppressed in the notation.

THEOREM 3.1. *Assume the correlation model, with conditions* (1.5–6). *Along almost all sample sequences, given* (X_i, Y_i) *for* $1 \leqq i \leqq n$, *as m and n go to infinity,*

a) $(1/m)X^*(m)^TX^*(m)$ *converges in conditional probability to* Σ

b) *the conditional law of* $\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}$ *goes weakly to normal with mean* 0 *and variance-covariance matrix* $\Sigma^{-1}M\Sigma^{-1}$.

PROOF. As in (1.10),

$$\sqrt{m}\{\hat{\beta}^*(m) - \hat{\beta}(n)\} = W^*(m)^{-1}Z^*(m), \tag{3.4}$$

where

$$W^*(m) = \frac{1}{m}X^*(m)^TX^*(m) = \frac{1}{m}\sum_{i=1}^m X_i^{*T}X_i^* \tag{3.5}$$

and

$$Z^*(m) = \frac{1}{\sqrt{m}}X^*(m)^T\varepsilon^*(m) = \frac{1}{\sqrt{m}}\sum_{i=1}^m X_i^{*T}\varepsilon_i^*. \tag{3.6}$$

Here, $W^*(m)$ is a $p \times p$ matrix; $Z^*(m)$ is a $p \times 1$ column vector. The corresponding unstarred quantities are

$$W(m) = \frac{1}{m}X(m)^TX(m) = \frac{1}{m}\sum_{i=1}^m X_i^TX_i; \tag{3.7}$$

$$Z(m) = \frac{1}{\sqrt{m}}X(m)^T\varepsilon(m) = \frac{1}{\sqrt{m}}\sum_{i=1}^m X_i^T\varepsilon_i. \tag{3.8}$$

Now $W^*(m)$ is a vector sum in $R^{p\times p}$; condition it on $\{X(n), Y(n)\}$. By Lemma 8.6 of B & F,

$$d_1^{p\times p}\{W^*(m), W(m)\} \leqq d_1^{p\times p}\{X_i^{*T}X_i^*, X_i^TX_i\}.$$

Again, both sides of the display are random variables: for the distance is computed between the conditional distribution of the starred quantity and the unconditional distribution of the unstarred quantity. The right hand side of the display goes to 0 a.e. as $n \to \infty$; this follows from Lemma 3.2, and Lemma 8.5 of B & F; the relevant ϕ is $\phi(x,y) = x^Tx$ from R^{p+1} to $R^{p\times p}$. In other words, the conditional law of $W^*(m)$ is close to the unconditional law of $W(m)$, but the latter concentrates near Σ, the variance-covariance matrix of X_i. This proves:

(3.9) The conditional law of $W^*(m)$ concentrates near Σ.

Likewise, $Z^*(m)$ is a vector sum in R^p. Condition it on $\{X(n), Y(n)\}$ and use Lemma 8.7 of B & F to obtain

$$d_2^p\{Z^*(m), Z(m)\}^2 \leqq d_2^p(X_i^{*T}\varepsilon_i^*, X_i^T\varepsilon_i)^2$$

The right hand side goes to 0 a.e. as $n \to \infty$. Indeed, Lemma 3.2 shows $\mu_n \to \mu$ a.e. in d_4^{p+1}; then use Lemma 3.1b. In other words, the conditional law of $Z^*(m)$ is close to the unconditional law of $Z(m)$, and the latter is essentially multivariate Gaussian, with mean 0 and variance-covariance matrix M defined by (1.9). This proves:

(3.10) The conditional law of $Z^*(m)$ is essentially multivariate normal with mean 0 and variance-covariance matrix M.

To complete the argument, combine (3.9) and (3.10). □

In the correlation model, $(X^TX)^{1/2}(\hat{\beta} - \beta)/\hat{\sigma}$ is not pivotal. However, bootstrapping it may be of interest. The only new issue is $\hat{\sigma}$. As before, let

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i(n)^2.$$

This estimates $\sigma^2 = E(\varepsilon_i^2)$ from the data, as the mean square of the residuals: see (3.1). The corresponding estimate based on the starred data is

$$\hat{\sigma}_m^{*2} = \frac{1}{m}\sum_{i=1}^m \hat{\varepsilon}_i^*(m)^2 \tag{3.11}$$

where the starred residuals are defined in (3.3).

THEOREM 3.2. *Assume the correlation model with conditions* (1.5–6). *Along almost all sample sequences, given* (X_i, Y_i) *for* $1 = 1, \cdots, n$, *as* m *and* n *tend to infinity, the conditional law of* $\hat{\sigma}_m^*$ *converges weakly to point mass at* σ.

PROOF. Using (2.9) on the starred regression,

$$\| \hat{\varepsilon}^*(m) - \varepsilon^*(m) \|^2 = Z^*(m)^T \cdot W^*(m)^{-1} \cdot Z^*(m).$$

Now use (3.9–10) to conclude

(3.12) The conditional law of $\frac{1}{m}\| \hat{\varepsilon}^*(m) - \varepsilon^*(m) \|^2$ concentrates near 0.

For the definition of $\hat{\sigma}_m^*$, see (3.11). Let

$$\sigma_m^{*2} = \frac{1}{m}\sum_{i=1}^m \varepsilon_i^{*2}$$

be the average of the squares of the starred disturbances (as opposed to residuals); see (3.2–3). Now

$$(\hat{\sigma}_m^* - \sigma_m^*)^2 \leqq \frac{1}{m}\sum_{i=1}^m \{\hat{\varepsilon}_i^*(m) - \varepsilon_i^*\}^2$$

so it remains only to show that σ_m^{*2} is nearly σ^2. Condition the ε_i^* on $[X(n), Y(n)]$. In view of Lemma 8.6 of B & F,

$$d_1\left(\frac{1}{m}\sum_{i=1}^m \varepsilon_i^{*2}, \frac{1}{m}\sum_{i=1}^m \varepsilon_i^2\right) \leqq d_i(\varepsilon_i^{*2}, \varepsilon_i^2).$$

But the right hand side tends to 0 a.e. by Lemmas 3.2 and 3.1c. In other words, the conditional law of $(1/m)\sum_{i=1}^m \varepsilon_i^{*2}$ is close to the unconditional law of $(1/m)\sum_{i=1}^m \varepsilon_i^2$. And the latter concentrates near σ^2. □

In particular, as m and n tend to ∞, the conditional law of $\{X^*(m)^TX^*(m)\}^{1/2}\{\hat{\beta}^*(m) - \hat{\beta}(n)\}/\hat{\sigma}_m^*$ converges to the appropriate limit: normal with mean 0 and variance-covariance matrix $\Sigma^{-1/2}M\Sigma^{-1/2}/\sigma^2$. In the homoscedastic case, this is just the $p \times p$ identity matrix $I_{p\times p}$: see (1.11).

What is the role of the 4th moment condition in (1.5)? To secure the conventional asymptotics, the following conditions seem to be needed:

$$E(\| X_i \|^2) < \infty \quad \text{and} \quad E(Y_i^2) < \infty \quad \text{and} \quad E(\| X_i \|^2 \varepsilon_i^2) < \infty.$$

Preliminary calculations suggest that under these minimal conditions, the bootstrap will be valid in probability; convergence a.e. can be secured by requiring $E\{\| X_i \|^{2+\delta}\} < \infty$. Convergence a.e. under the minimal conditions seems to be quite a delicate question.

REFERENCES

BICKEL, P. and FREEDMAN, D. (1981a). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. and FREEDMAN, D. (1981b). More on bootstrapping regression models. Technical report, Statistics Department, University of California, Berkeley.

BILLINGSLEY, P. (1979). *Probability and Measure.* Wiley, New York.

EFRON, B. (1977). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.

HINKLEY, D. (1977). On jackknifing in unbalanced situations. Technometrics, **19** 285–292.

LAI, T., ROBBINS, H., and WEI, C. (1979). Strong consistency of least squares estimates in multiple regression. *J. Multivariate Analysis* **9** 343–361.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720