The Annals of Statistics
1984, Vol. 12, No. 3, 827–842

ON BOOTSTRAPPING TWO-STAGE LEAST-SQUARES ESTIMATES IN STATIONARY LINEAR MODELS

BY D. FREEDMAN[1]

University of California, Berkeley

For models similar to those used in econometric work, under suitable regularity conditions, the bootstrap is shown to give asymptotically valid approximations to the distribution of errors in coefficient estimates.

1. Introduction. The bootstrap is described by Efron (1979, 1982); related papers are Bickel and Freedman (1981, 1983), Freedman (1981), and Shorack (1982). In essence, the bootstrap is a procedure for estimating standard errors by resampling the data in a suitable way, so the model is tested against its own assumptions. The object of this paper is to indicate how the idea might be applied to linear models of the kind used in econometrics, where the technical difficulties include simultaneity, heteroscedasticity, and dynamics. Since the object is purely illustrative, only two theorems will be presented. Section 3 deals with simultaneity, but the model is static; Section 4 allows a dynamic model.

To make the bootstrap appealing, two kinds of evidence are needed:

(i) A showing that the bootstrap gives the right answers with large samples, so it is at least as sound as the conventional asymptotics.

(ii) A showing that in finite samples, the bootstrap actually outperforms the conventional asymptotics.

The present paper focuses on (i). It actually does a bit more, by showing that for large samples the bootstrap will give the right answers even in the presence of heteroscedastic errors, which throw the conventional formulae off. The conditions are appreciably less restrictive than those of White (1982), who assumes normal errors.

With respect to point (ii), there is good empirical evidence in Efron (1979, 1982), or Freedman and Peters (1984a, b); also see Daggett and Freedman (1984). Too, there is some theoretical evidence, in the form of Edgeworth expansions: see Beran (1982), Singh (1981). This paper will not deal with point (ii); it is purely asymptotic.

The balance of this section is intended to give an informal overview of the bootstrap idea, for a dynamic model. In brief, the model has been fitted to data, by some statistical procedure; the residuals are the discrepancies between actual and fitted values. Some stochastic structure was imposed on the theoretical

Received January 1983; revised March 1984.

[1] I would like to thank David Brillinger, Edwin Kuh and Thomas Rothenberg for their help.

AMS 1980 *subject classifications.* Primary 62J05; secondary 62E20.

Key words and phrases. Regression, standard errors, two-stage least squares, bootstrap, linear models.

stochastic disturbance terms, explicitly or implicitly, in the fitting. The key idea is to resample the residuals, preserving this stochastic structure.

Assuming the model and the estimated parameters to be right, the resampling generates "pseudo-data." Now the model can be refitted to the pseudo-data. In this artificial world, the errors in the parameter estimates are directly observable. The Monte Carlo distribution of such errors can be used to approximate the distribution of the unobservable errors in the real parameter estimates. This gives a measure of the statistical uncertainty in the parameter estimates.

A more explicit, but still informal, description of the bootstrap is as follows. Consider a dynamic linear model, of the form

$$\underset{1\times a}{Y_t} = \underset{1\times a}{Y_t}\,\underset{a\times a}{A} + \underset{1\times a}{Y_{t-1}}\,\underset{a\times a}{B} + \underset{1\times b}{X_t}\,\underset{b\times a}{C} + \underset{1\times a}{\varepsilon_t} \tag{1.1}$$

In this equation, A, B, C are coefficient matrices of unknown parameters, to be estimated from the data, subject to identifying restriction; Y_t is the vector of endogenous variables at time t; while X_t is the vector of exogenous variables at time t; and ε_t is the vector of disturbances at time t; identifying restrictions may be imposed on this distribution, especially, given the X's the ε's may be assumed independent and identically distributed (i.i.d.) with mean 0. In the informal discussion which follows, and in Section 2, the X's will be treated as known constants; in the body of the paper, they will be treated as random. To avoid trivial complications, suppose the equations all have intercepts.

Coming back to the model (1.1), data is available for $t = 1, \cdots, n$ and Y_0 is available too. The coefficient matrices are estimated as $\hat{A}$, $\hat{B}$, $\hat{C}$ by some well-defined statistical procedure, like "two-stage least squares," to be discussed in Section 2. Due to the assumed randomness in ε, there is random error in the estimates $\hat{A}$, $\hat{B}$, $\hat{C}$ for A, B, C. How big are these errors? This question can be addressed by the following bootstrap procedure, whose explanation is a bit lengthy. When $\hat{A}$, $\hat{B}$ and $\hat{C}$ are computed, residuals are defined:

$$\hat{\varepsilon}_t = Y_t - Y_t\hat{A} - Y_{t-1}\hat{B} - X_t\hat{C}. \tag{1.2}$$

These are estimates for the true disturbances ε_t in the model (1).

Now consider a model like (1), but where all the ingredients are known:

- Set the coefficients at $\hat{A}$, $\hat{B}$, $\hat{C}$ respectively.
- Make the disturbance terms independent, with common distribution equal to the empirical distribution of $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$.

(The "empirical" distribution puts mass $1/n$ at each of the computed residuals.) The exogenous X's are kept as before, as is Y_0. Using this simulation model, pseudo-data can be generated for periods $t = 1, \cdots, n$. This pseudo-data will be denoted by stars: $Y_1^*, \cdots, Y_n^*$. The construction is iterative: $Y_0^* = Y_0$, and

$$Y_t^* = (Y_{t-1}^*\hat{B} + X_t\hat{C} + \varepsilon_t^*)(I - \hat{A})^{-1}, \tag{1.3}$$

the ε^*'s being independent with the common distribution specified above. This rule applies for $t = 1, \cdots, n$.

Now pretend the pseudo-data (1.3) come from a model like (1.1), with unknown coefficient matrices. Using the given procedures (two-stage least squares), estimate these coefficients; denote the estimates by $\hat{A}^*$, $\hat{B}^*$, $\hat{C}^*$. The distribution of the pseudo-errors $\hat{A}^* - \hat{A}$, and $\hat{B}^* - \hat{B}$, and $\hat{C}^* - \hat{C}$ can be computed, and used to approximate the distribution of the real errors $\hat{A} - A$, and $\hat{B} - B$, and $\hat{C} - C$. This approximation is the bootstrap. It is emphasized that the calculation assumes the validity of the model (1.1). The distribution of the pseudo-errors can be computed, e.g. by Monte Carlo. It is of interest only as an approximation to the distribution of the real errors.

The balance of this paper is organized as follows. Section 2 explains the inference problem addressed by two-stage least squares (2SLS), and the conventional procedure for computing standard errors with 2SLS estimators. Section 3 presents a theorem for "static" models, and Section 4 covers dynamic models.

2. Two-stage least squares. Consider the quantity sold Q_t and price P_t of a commodity in period t. Economists consider that there is a supply curve governing the quantity supplied as a function of price, assumed linear for simplicity:

$$Q = \alpha_0 + \alpha_1 P + \varepsilon_t. \tag{2.1}$$

Here, α_0 and α_1 are parameters governing the market; α_1 is expected to be positive, so supply will increase with price. Likewise, there is a demand curve governing the quantity demanded as a function of price, also taken to be linear:

$$Q = \beta_0 + \beta_1 P + \delta_t. \tag{2.2}$$

Here, β_0 and β_1 are two more parameters, with β_1 expected to be negative, so demand decreases when price increases. The stochastic disturbance terms $(\varepsilon_t, \delta_t)$ are usually held to "represent the effect of omitted variables"; see Pratt and Schlaifer (1983). In this context, it is conventional to assume that the pairs $(\varepsilon_t, \delta_t)$ are i.i.d. in t, with mean 0; but ε_t and δ_t have a general 2×2 covariance matrix.

In period t, the observed "market-clearing" price P_t and quantity Q_t are obtained by solving (2.1)–(2.2) as two linear equations in two unknowns. This is the stochastic model for the data. Thus, Q_t and P_t both depend on ε_t and δ_t. Now α and β are to be estimated. The complication is that ordinary least squares is inconsistent when the disturbances are correlated with the explanatory variables.

To get around this, economists use "instrumental" or "exogenous" variables, which are uncorrelated by assumption with the disturbances. By contrast, in the example, Q_t and P_t are "endogenous": correlated with the disturbances.

By way of illustration, suppose U_t and V_t are exogenous. Multiply (2.1) by U_t or V_t and sum:

$$\begin{aligned} \left(\sum_{t=1}^T U_t Q_t\right) &= \alpha_0 \left(\sum_{t=1}^T U_t\right) + \alpha_1 \left(\sum_{t=1}^T U_t P_t\right) + \xi \\ \left(\sum_{t=1}^T V_t Q_t\right) &= \alpha_0 \left(\sum_{t=1}^T V_t\right) + \alpha_1 \left(\sum_{t=1}^T V_t P_t\right) + \zeta \end{aligned} \tag{2.3}$$

where

$$\xi = \sum_{t=1}^{T} U_t \varepsilon_t \quad \text{and} \quad \zeta = \sum_{t=1}^{T} V_t \delta_t \tag{2.4}$$

should be small, because $E(U_t\varepsilon_t) = E(V_t\delta_t) = 0$. Now drop ξ and ζ from (2.3). Left are two linear simultaneous equations—the analog of the normal equations—for the parameters α_0 and α_1. Solving this pair of equations for α_0 and α_1 gives the "two-stage least squares" estimators; likewise for β_0 and β_1.

To set this up more generally, and to get at the conventional standard errors, it is convenient to use the machinery of generalized least squares. This will now be reviewed very briefly, to fix notation. Consider the model

$$Y = M\beta + \delta, \quad E(\delta) = 0, \quad \text{cov}(\delta) = \Sigma. \tag{2.5}$$

For historical reasons, M is called "the design matrix;" it is usually denoted by X, but that conflicts with present notation. With Σ known, the generalized least squares (gls) estimate is

$$\hat{\beta}_{\text{gls}} = (M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}Y. \tag{2.6}$$

As usual,

$$E(\hat{\beta}_{\text{gls}}) = \beta \tag{2.7}$$

$$\text{cov}(\hat{\beta}_{\text{gls}}) = (M^T\Sigma^{-1}M)^{-1}. \tag{2.8}$$

When Σ is unknown, statisticians routinely use (2.6) and (2.8) with Σ replaced by some estimate $\hat{\Sigma}$. Iterative procedures are often used, as follows. Let $\hat{\beta}^{(0)}$ be some initial estimate for β, typically from a preliminary ordinary least squares (ols) fit. There are residuals $\hat{e}^{(0)} = Y - M\hat{\beta}^{(0)}$. Suppose the procedure has been defined through stage k, with residuals

$$\hat{e}^{(k)} = Y - M\hat{\beta}^{(k)}_{\text{gls}}. \tag{2.9}$$

Let $\hat{\Sigma}_k$ be an estimator for Σ, based on $\hat{e}^{(k)}$. Then

$$\hat{\beta}^{(k+1)}_{\text{gls}} = (M^T\hat{\Sigma}_k^{-1}M)^{-1}M^T\hat{\Sigma}_k^{-1}Y. \tag{2.10}$$

This procedure can be continued for a fixed number of steps, or until $\hat{\beta}^{(k)}_{\text{gls}}$ settles down: a convexity argument shows that $\hat{\beta}^{(k)}_{\text{gls}}$ converges to the maximum likelihood estimate for β, assuming ε is independent of M and multivariate Gaussian with mean 0.

The covariance matrix for $\hat{\beta}^{(k+1)}_{\text{gls}}$ is usually estimated from (2.8), with $\hat{\Sigma}_k$ put in for Σ:

$$\widehat{\text{cov}}^{(k+1)} = (M^T\hat{\Sigma}_k^{-1}M)^{-1}. \tag{2.11}$$

This may be legitimate, asymptotically. In finite-sample situations, all depends on whether $\hat{\Sigma}_k$ is a good estimate for Σ or not. If $\hat{\Sigma}_k$ is a poor estimate for Σ, the standard errors estimated from (2.11) may prove to be unduly optimistic, and approximate gls estimators are often used when there is too little data to offer any hope of estimating Σ with reasonable accuracy: an example is given in

Freedman and Peters (1984a). In such circumstances, the bootstrap is a useful diagnostic, and in many cases it gives a more realistic estimate of the standard errors.

To ease notation, $\hat{\beta}^{(k)}_{\text{gls}}$ will be referred to as the (gls, k)-estimator. This paper only consider the (gls, 1) estimator, which in many situations has full asymptotic efficiency; see Cox and Hinkley (1974, page 308). In some examples, further iteration seems to make the coefficient estimates better, but also exaggerates the optimism of the standard error estimates. In other examples, iteration actually makes the coefficient estimators worse. The effects of additional iteration are considered in Peters (1983).

The next object is to review two-stage least squares (2SLS). The present exposition is self-contained but terse. For a fuller account, see Theil (1971).

We return to the model (1.1). We suppose for the balance of this section that exogenous X's are nonrandom. Multiply (1.1) by X_t^T and sum:

$$\underset{b\times a}{R} = \underset{b\times a}{R}\,\underset{a\times a}{A} + \underset{b\times a}{S}\,\underset{a\times a}{B} + \underset{b\times b}{T}\,\underset{b\times a}{C} + \underset{b\times a}{\Delta} \tag{2.12}$$

where

$$R = \textstyle\sum_{t=1}^n X_t^T Y_t, \quad S = \sum_{t=1}^n X_t^T Y_{t-1},$$
$$T = \textstyle\sum_{t=1}^n X_t^T X_t, \quad \Delta = \sum_{t=1}^n X_t^T \varepsilon_t. \tag{2.13}$$

Notice that the jth column of (2.12) corresponds to the jth equation in (1.1).

In applications, $[A, B, C]$ is constrained to fall in some linear space of dimension at most ab: then A, B, C can be estimated from (2.12) by some variant of constrained least squares. (Without constraints, the parameters are not estimable, since there are only ab equations.) Notice that T is constant (nonrandom) since X is. It is conventional to treat R and S on the right side of (2.12) as constant. This may be legitimate asymptotically, but is false in any finite sample. Moreover, R and S are correlated with Δ, and this is the source of "small-sample bias" in 2SLS; see Daggett and Freedman (1984) for a bootstrap investigation of the bias.

The matrix of errors Δ on the right-hand side of (2.12) has covariance structure, so generalized least squares is the procedure of choice. To make contact with the standard format of (2.5), we stack the columns in (2.12): column #1 on top of column #2, $\cdots$, on top of column #q. In the stack, information corresponding to the first equation comes first, information about the last equation comes last.

The parameter vector β in (2.5) is obtained by stacking A, B and C: column #1 of A, followed by column #1 of B, followed by column #1 of C, $\cdots$, followed by column #q of A, followed by column #q of B, followed by column #q of C. The design matrix is obtained by writing R, S and T down the diagonal, and padding with zeros.

The left-hand side Y vector in (2.5) consists of the stacked R matrix; the error δ vector consists of the stacked Δ matrix. The full system of equations (2.12) is layed out in stacked form below, with R_j being the jth column of the matrix R,

and likewise for the other matrices.

$$(2.14)\qquad \begin{bmatrix} R_1 \\ \cdot \\ \cdot \\ \cdot \\ R_a \end{bmatrix} = \begin{bmatrix} R & S & T & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & R & S & T & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & R & S & T \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \\ C_1 \\ \vdots \\ A_a \\ B_a \\ C_a \end{bmatrix} + \begin{bmatrix} \Delta_1 \\ \cdot \\ \cdot \\ \cdot \\ \Delta_a \end{bmatrix}.$$

At this point, the design matrix is highly singular. Usually the elements of β known a priori to vanish are suppressed, and the design matrix is adjusted accordingly by deleting the corresponding columns. An alternative approach is to use generalized inverses: see Chapter 6 of Theil (1971). The covariance matrix of the error vector (the stacked Δ matrix) is the Kronecker product

$$(2.15)\qquad \Sigma = \Gamma \otimes T = \begin{bmatrix} \Gamma_{11}T & \Gamma_{12}T & \cdots & \Gamma_{1a}T \\ \Gamma_{21}T & \Gamma_{22}T & \cdots & \Gamma_{2a}T \\ \vdots & \vdots & & \vdots \\ \Gamma_{a1}T & \Gamma_{a2}T & \cdots & \Gamma_{aa}T \end{bmatrix}$$

where Γ is the $a \times a$ covariance matrix of the error vector ε in (1.1).

We can now give a brief description of two-stage least squares, focusing on the connection with generalized least squares, and sharpening the idea that 2SLS is a "single-equation" method. Consider each column of (2.12) in isolation. Take column j, corresponding to equation j in system (1.1):

$$(2.16)\qquad \underset{b\times 1}{R_j} = \underset{b\times a}{R}\ \underset{a\times 1}{A_j} + \underset{b\times a}{S}\ \underset{a\times 1}{B_j} + \underset{b\times b}{T}\ \underset{b\times 1}{C_j} + \underset{b\times 1}{\Delta_j}$$

The 2SLS procedure amounts to estimating (2.16) by gls, treating R and S on the right as constant. The constraints specific to the jth equation would be imposed, but not the cross-equation constraints. The covariance matrix of Δ_j is required. Plainly, cov $\Delta_j = \Gamma_{jj}T$, where T was defined in (2.13) and is computable from the data; Γ_{jj} is unknown, but enters only as a constant of proportionality, and its value is immaterial. With large enough samples, this procedure is preferable to ols, because it takes account of the correlation between ε and Y on the right side of (1.1): this correlation would make ols inconsistent.

Let $\hat{A}_{II}$, $\hat{B}_{II}$, $\hat{C}_{II}$ denote the 2SLS estimators. To estimate their covariances, let

$$(2.17a)\qquad \hat{\varepsilon}_t = Y_t - Y_t\hat{A}_{II} - Y_t\hat{B}_{II} - X_t\hat{C}_{II}$$

$$(2.17b)\qquad \hat{\Gamma} = (1/n)\ \textstyle\sum_{t=1}^{n} \hat{\varepsilon}_t^T\hat{\varepsilon}_t.$$

The $\hat{\varepsilon}$'s are the residuals, and $\hat{\Gamma}$ is an empirical covariance matrix which estimates Γ in (2.15). And the covariance matrix of $[\hat{A}_{II}, \hat{B}_{II}, \hat{C}_{II}]$ can be estimated by the gls formula as $(M^T\hat{\Sigma}^{-1}M)^{-1}$, where M denotes the relevant design matrix, and Σ = cov $\Delta_j = \Gamma_{jj}T$ is estimated as $\hat{\Sigma} = \hat{\Gamma}_{jj}$T. It is conventional, for the purpose of

estimating $\text{cov}[\hat{A}_{II}, \hat{B}_{II}, \hat{C}_{II}]$ only, to inflate $\hat{\Gamma}_{jj}$ by $n/(n-r)$, where r is the number of variables actually coming into the jth equation.

3. Simultaneity without dynamics. The model to be discussed in this section is relevant to cross-sectional data, where the problem is to estimate population parameters from a sample. Only a single-equation estimation procedure will be considered, namely, two-stage least squares. Consider observable random vectors Y, U, V; and a coefficient matrix A. The model assumed is

$$\begin{array}{ccccccc} Y & = & U & A & + & \varepsilon \\ 1\times 1 & & 1\times p & p\times 1 & & 1\times 1 \end{array} \tag{3.1}$$

with an $r \times 1$ vector V of instrumental variables orthogonal to ε, in the sense $E\{V\varepsilon\} = 0$; where E is mathematical expectation and $V\varepsilon$ is the matrix product of V and ε, of dimension $r \times 1$. We do not assume that $E\{U^T\varepsilon\} = 0$, or even $E\{\varepsilon\} = 0$. Also, we do not assume that $E\{\varepsilon \mid V\} = 0$ or $E\{\varepsilon^2 \mid V\}$ is constant. Thus, heteroscedastic errors are allowed.

We view (3.1) as one equation in a system. Ordinarily, some components of U would be endogenous (correlated with ε) and others exogenous (uncorrelated with ε); the exogenous ones would turn up among the instruments V, as would exogenous variables from other equations in the system.

Write $|\ |$ for Euclidean norm. Now (Y, U, V) is a random vector of dimension $1 + p + r$. Ordinarily, this vector would be assumed L_2. For the bootstrap to succeed, however, a bit more is needed, and L_4 is convenient: $E\{|(Y, U, V)|^4\} < \infty$.

Before proceeding to the bootstrap, it will be helpful to review the standard theory in the present setting. Let

$$\begin{array}{ccccccc} Q & = & E\{VY\}, & R & = & E\{VU\}, & \text{and} \quad S = E\{VV^T\} \\ r\times 1 & & & r\times p & & & r\times r \end{array}. \tag{3.2}$$

Multiply (3.1) on the left by V and take expectations, using the assumed orthogonality:

$$Q = RA. \tag{3.3}$$

Assume that the system is identified:

$$r \geq p, \quad R \text{ has full rank } p, \quad \text{and } S \text{ is invertible.} \tag{3.4}$$

Thus, Q is in the range of R, by (3.3); and A is the unique p-vector satisfying (3.3), by (3.4).

So far, we have a probability structure but no data; the data are modeled as a sample of size n from this structure. More particularly, let $(Y_i, U_i, V_i, \varepsilon_i)$ be independent, and distributed as (Y, U, V, ε). In particular, V_i is orthogonal to ε_i in the sense $E\{V_i\varepsilon_i\} = 0$; and $Y_i = U_iA + \varepsilon_i$.

These data are used to estimate A by 2SLS, as follows. Let

$$\begin{aligned} Q_n &= (1/n)\textstyle\sum_{i=1}^n V_iY_i, \quad R_n = (1/n)\sum_{i=1}^n V_iU_i, \\ S_n &= (1/n)\textstyle\sum_{i=1}^n V_iV_i^T, \quad \Delta_n = (1/n)\sum_{i=1}^n V_i\varepsilon_i. \end{aligned} \tag{3.5}$$

We have

$$\underset{r\times 1}{Q_n} = \underset{r\times p}{R_n}\ \underset{p\times 1}{A} + \underset{r\times 1}{\Delta_n}. \tag{3.6}$$

Now A can be estimated from (3.6) by regression, taking into account that the components of Δ_n are correlated. In the conventional homoscedastic case, the variance-covariance matrix of Δ_n would be estimated as proportional to S_n, so

$$\hat{A}_n = (R_n^T S_n^{-1} R_n)^{-1} R_n^T S_n^{-1} Q_n. \tag{3.7}$$

This is the conventional two-stage least squares estimator.

Some algebraic manipulation gives

$$\sqrt{n}(\hat{A}_n - A) = (R_n^T S_n^{-1} R_n)^{-1} R_n^T S_n^{-1} (\sqrt{n}\Delta_n). \tag{3.8}$$

Now $\sqrt{n}\Delta_n$ satisfies the central limit theorem in r-dimensional space; the other factors on the right side of (3.7) can be treated as constants; since $Q_n \to Q$, $R_n \to R$, $S_n \to S$ by the law of large numbers: the usual asymptotics follow.

The estimation procedure is efficient only in the homoscedastic case; likewise, the conventional formulae for standard errors assume homoscedasticity. But the analysis which follows is valid whether the errors are homoscedastic or not. A referee asks about the alternative of modeling and estimating the heteroscedasticity. The simulations in Freedman and Peters (1984a) make one pessimistic about this approach.

Let $\hat{\varepsilon}_i(n)$ be the residual from the fit:

$$\hat{\varepsilon}_i(n) = Y_i - U_i \hat{A}_n. \tag{3.9}$$

As data, the residuals will not in general be exactly orthogonal to the instruments, i.e., in general $(1/n)\sum_{i=1}^n V_i \hat{\varepsilon}_i(n) \neq 0$. Let $\tilde{\varepsilon}(n)$ be the part of the residual vector orthogonal to the vector of instruments:

$$\tilde{\varepsilon}_i(n) = \hat{\varepsilon}_i(n) - \hat{b}_n^T V_i = \hat{\varepsilon}_i(n) - V_i^T \hat{b}_n \tag{3.10}$$

where the $r \times 1$ vector $\hat{b}_n$ is defined as follows:

$$\hat{b}_n = S_n^{-1}\ (1/n)\ \textstyle\sum_{i=1}^n V_i \hat{\varepsilon}_i(n) = S_n^{-1}[Q_n - R_n \hat{A}_n]. \tag{3.11}$$

Coming now to the bootstrap, given the data, let $\tilde{\mu}_n$ be the empirical distribution of $(U_i, V_i, \tilde{\varepsilon}_i(n))$ for $i = 1, \cdots, n$. Thus, $\tilde{\mu}_n$ is an atomic probability measure in $(p + r + 1)$-dimensional Euclidean space; it assigns measure $(1/n)$ to each of the n points $(U_i, V_i, \tilde{\varepsilon}_i(n))$.

It is now time to resample the data. Given (Y_i, U_i, V_i) for $i = 1, \cdots, n$, let $(U_j^*, V_j^*, \varepsilon_j^*)$ be conditionally independent for $j = 1, \cdots, n$, with common distribution $\tilde{\mu}_n$; let $Y_j^* = U_j^* \hat{A}_n + \varepsilon_j^*$. Resampling the data this way preserves any relationship there may be between instruments and disturbances.

Now imagine giving the starred data to another investigator, to calculate the two-stage least squares estimates: the results will be

$$\begin{aligned} &Q_n^* = (1/n)\textstyle\sum_{j=1}^n V_j^* Y_j^*, \quad R_n^* = (1/n)\sum_{j=1}^n V_j^* U_j^*, \\ &S_n^* = (1/n)\textstyle\sum_{j=1}^n V_j^* V_j^{*T}, \quad \Delta_n^* = (1/n)\sum_{j=1}^n V_j^* \varepsilon_j^* \end{aligned} \tag{3.12a}$$

$$\begin{aligned}\hat{A}_n^* &= (R_n^{*T}S_n^{*-1}R_n^*)^{-1}R_n^{*T}S_n^{*-1}Q_n^* \\ &= \hat{A}_n + (R_n^{*T}S_n^{*-1}R_n^*)^{-1}R_n^{*T}S_n^{*-1}\Delta_n^*.\end{aligned} \tag{3.12b}$$

The bootstrap principle is that the error structure of the starred estimates mimics that in the original estimates, as the following theorem shows.

THEOREM 3.1. *Along almost all sample sequences, as* $n \to \infty$,
(a) $Q_n^* \to Q$ *and* $R_n^* \to R$ *and* $S_n^* \to S$ *in conditional probability*;
(b) *the conditional law of* $\sqrt{n}\Delta_n^*$ *and the unconditional law of* $\sqrt{n}\Delta_n$ *converge to the same limit.*

In particular, by a variant on Slutzky's lemma, the conditional law of $\sqrt{n}(\hat{A}_n^* - \hat{A}_n)$ and the unconditional law of $\sqrt{n}(\hat{A}_n - A)$ have the same limit too. The theorem will be proved in Section 5 below. For extensions, see the discussion at the end of Section 4.

4. A dynamic model. The model to be discussed in this section is relevant to a single realization of a multivariate time series. The discussion parallels that in Section 3, but there are a few annoying complications. Only single-equation methods will be considered, but all equations must be specified and estimated, so the bootstrap dynamics will match the original dynamics. Consider observable random vectors Y_t and X_t in each time period t, where t runs through the integers; and coefficient matrices A, B, C. The model assumed is

$$\begin{array}{ccccccccccc} Y_t & = & Y_t & A & + & Y_{t-1} & B & + & X_t & C & + & \varepsilon_t \\ 1\times a & & 1\times a & a\times a & & 1\times a & a\times a & & 1\times b & b\times a & & 1\times a \end{array} \tag{4.1}$$

$$(X_t, \varepsilon_t) \quad \text{are i.i.d. and} \quad L_4 \quad \text{for} \quad t = 0, \pm 1, \pm 2, \cdots \tag{4.2}$$

$$X_t \quad \text{is orthogonal to} \quad \varepsilon_t. \tag{4.3}$$

We assume the first component of X_t is 1, so $E\{\varepsilon_t\} = 0$. Identifying restrictions are imposed:

$$\begin{aligned} A_{jk} = 0 \quad \text{for} \quad jk \in N_A, \qquad B_{jk} = 0 \quad \text{for} \quad jk \in N_B, \\ C_{jk} = 0 \quad \text{for} \quad jk \in N_C. \end{aligned} \tag{4.4}$$

Here, the N's are finite sets, $jj \in N_A$ for all j, and $1k \in N_C$ for no k, so all equations have intercepts.

Let I be the identity matrix. Assume further:

$$I - A \quad \text{is invertible;} \tag{4.5}$$

$$r = \|B(I - A)\| < 1. \tag{4.6}$$

Here, $\|\cdot\|$ is the operator norm: $\|D\| = \sup\{|yD|: |y| \le 1\}$. Clearly, $\|D\| \le |D|$, where $|D| = (\sum_{jk} D_{jk}^2)^{1/2}$. It is assumed

$$\{Y_t\} \quad \text{is stationary} \tag{4.7}$$

so that

$$Y_t = \sum_{s=0}^{\infty} \xi_{t,s} \tag{4.8}$$

where

$$\xi_{t,s} = (\varepsilon_{t-s} + X_{t-s}C)(I - A)^{-1}[B(I - A)^{-1}]^s.$$

Consider for example the first equation in the system. Let y_t be the first component of Y_t, and δ_t the first component of ε_t. Rewrite the first equation of the system in notation like that of (3.1):

$$\begin{array}{ccccccc} Y_t & = & U_t & \alpha & + & \delta_t, & \delta_t \perp V_t \\ 1 \times 1 & & 1 \times p & p \times 1 & & 1 \times 1 & r \times 1 \end{array} \tag{4.9}$$

where U_t consists of the components of Y_t which are really in the equation, followed by the relevant components of Y_{t-1}, followed by those of X_t; while the vector α consists of the elements from the first columns of the matrices A, B, C respectively which are not constrained to vanish in (4.4). This procedure drops the variables defined as irrelevant by the constraints. Likewise, V_t consists of Y_{t-1}^T followed by X_t: thus, $r = a + b$. Verify that $V_t \perp \delta_t$: the condition $E\{\delta_t\} = 0$ will be used here. Verify too that $(y_t, U_t, V_t, \delta_t)$ is stationary and ergodic.

As before, let

$$Q = E\{V_t y_t\}, \quad R = E\{V_t U_t\}, \quad S = E\{V_t V_t^T\}. \tag{4.10}$$

These do not depend on t, due to the assumed stationarity. Multiply (4.9) on the left by V_t and take expectations:

$$Q = R\alpha. \tag{4.11}$$

Assume the system is identified:

$$r \geq p, \quad R \text{ has full rank } p, \quad \text{and } S \text{ is invertible.} \tag{4.12}$$

Suppose observations on all the time series are available for periods $t = 1, \cdots, n$, and y_0 is available too. These data can be used exactly as before to estimate the coefficients by instrumental-variables regression. As in (3.5), let

$$\begin{aligned} Q_n &= (1/n) \sum_{t=1}^n V_t y_t, \quad & R_n &= (1/n) \sum_{t=1}^n V_t U_t, \\ S_n &= (1/n) \sum_{t=1}^n V_t V_t^T, \quad & \Delta_n &= (1/n) \sum_{t=1}^n V_t \delta_t. \end{aligned} \tag{4.13}$$

Again, $Q_n \to Q$ and $R_n \to R$ and $S_n \to S$, for instance, by the ergodic theorem. As before, the estimator to be bootstrapped is

$$\hat{\alpha}_n = (R_n^T S_n^{-1} R_n)^{-1} R_n^T S_n^{-1} Q_n = \alpha + (R_n^T S_n^{-1} R_n)^{-1} R_n^T S_n^{-1} \Delta_n. \tag{4.14}$$

Now for the analogs of (3.9)–(3.11). Let $\hat{\delta}_t(n)$ be the residual from the fit:

$$\hat{\delta}_t(n) = y_t - U_t \hat{\alpha}_n. \tag{4.15}$$

Let $\tilde{\delta}_t(n)$ be the part of the residual vector orthogonal to the vector of instruments:

$$\tilde{\delta}_t(n) = \hat{\delta}_t(n) - \hat{b}_n^T V_t \tag{4.16}$$

where

$$\hat{b}_n = S_n^{-1}(1/n)\sum_{t=1}^{n} V_t\hat{\delta}_t(n) = S_n^{-1}[Q_n - R_n\hat{\alpha}_n].$$

The foregoing can be carried out separately for each equation in the system. The constraints may vary from equation to equation, according to (4.4), but no constraints are imposed across equations. Let $\hat{A}_n$, $\hat{B}_n$ and $\hat{C}_n$ be the resulting estimates for the coefficient matrices, and $\tilde{\varepsilon}_t(n)$ the vector of the $\hat{\delta}_t(n)$ from the various equations strung together. Given the data: let $\tilde{\mu}_n$ be the empirical distribution of $(X_t, \tilde{\varepsilon}_t(n))$ for $t = 1, \cdots, n$; let (X_s^*, ε_s^*) be independent, with common distribution $\tilde{\mu}_n$, for $s = 0, \pm 1, \cdots$; let

$$Y_t^* = \sum_{s=0}^{\infty}[\varepsilon_{t-s}^* + X_{t-s}^*\hat{C}_n](I - \hat{A}_n)^{-1}[\hat{B}_n(I - \hat{A}_n)^{-1}]^s. \tag{4.17}$$

Then the starred data will satisfy the model

$$Y_t^* = Y_t^*\hat{A}_n + Y_{t-1}^*\hat{B}_n + X_t^*\hat{C}_n + \varepsilon_t^* \tag{4.18}$$

with X_t^* orthogonal to ε_t^*. The starred instrumental-variables estimate for the first equation in the system is obtained as follows, with y^*, U^*, V^* and δ^* built up from X^* and ε^* just as y, U, V and δ were from X and ε:

$$\begin{aligned} Q_n^* &= (1/n)\sum_{t=1}^{n} V_t^* y_t^*, \quad & R_n^* &= (1/n)\sum_{t=1}^{n} V_t^* U_t^*, \\ S_n^* &= (1/n)\sum_{t=1}^{n} V_t^* V_t^{*T}, \quad & \Delta_n^* &= (1/n)\sum_{t=1}^{n} V_t^*\delta_t^* \end{aligned} \tag{4.19}$$

$$\begin{aligned} \hat{\alpha}_n^* &= (R_n^{*T}S_n^{*-1}R_n^*)^{-1}R_n^{*T}S_n^{*-1}Q_n^* \\ &= \hat{\alpha}_n + (R_n^{*T}S_n^{*-1}R_n^*)^{-1}R_n^{*T}S_n^{*-1}\Delta_n^*. \end{aligned} \tag{4.20}$$

The bootstrap principle is stated in the next theorem; the proof is deferred to Section 6.

THEOREM 4.1. *Along almost all sample sequences, as $n \to \infty$, conditionally on the data:*

(a) *$Q_n^* \to Q$ and $R_n^* \to R$ and $S_n^* \to S$ in conditional probability.*
(b) *The conditional law of $\sqrt{n}\Delta_n^*$ has the same limit as the unconditional law of $\sqrt{n}\Delta_n$.*

The stationarity condition is easy to relax, because the effect of Y_0 dwindles exponentially fast. The independence condition (3.2) is quite strong; it may be replaced by the assumption of an autoregressive structure, which too is estimated from the data. Formally, the theorem only covers the joint distribution of estimates for one equation; the extension to the whole system is done in Section 6.

In many situations, the X_t are treated like constants. In effect, this assumes "homoscedasticity:" given X_t, the ε_t are independent and identically distributed, with mean 0 and finite variance. Then, it is appropriate to resample residuals (after orthogonalization). However, convergence must be assumed for $(1/n)\sum_{t=1}^{n} X_t^T X_{t-s}$. Three-stage least squares, with arbitrary linear constraints

within and across equations, may be bootstrapped too, although we have not checked the details completely. The variance-covariance matrix of the errors is to be estimated by the empirical variance-covariance matrix of the residuals from a 2SLS fit. More ambitious iterative procedures can be used too.

5. Technical details: Theorem 3.1. The argument is in terms of the "Mallows metrics" discussed in Section 8 of Bickel and Freedman (1981); hereafter, "B&F." If R^j is j-dimensional space equipped with the Euclidean norm $|\cdot|$, and $\alpha \geq 1$, then $d_\alpha^j(\mu, \nu)$ is the "distance" between probabilities μ and ν in R^j, defined as the infimum of $E\{|\xi - \zeta|^\alpha\}^{1/\alpha}$ over all pairs of random j-vectors ξ and ζ, where ξ has law μ and ζ has law ν. In effect, the infimum is over all possible dependencies between ξ and ζ.

LEMMA 5.1. *Let ν_n, ν be probabilities in R^j. Let $\alpha \geq 1$, and suppose the Mallows metric $d_\alpha^j(\nu_n, \nu) \to 0$. Let M_n be a linear map from R^j to R^k, also equipped with the Euclidean norm. Suppose $M_n \to M$. Then $d_\alpha^k(\nu_n M_n^{-1}, \nu M^{-1}) \to 0$.*

PROOF. Construct U_n and U with distributions ν_n and ν respectively, and $E[|U_n - U|^\alpha]^{1/\alpha} = d_\alpha^j(\nu_n, \nu)$. See Lemma 8.1 of B&F. Recall that $\|\ \|$ is the operator norm, so e.g. $|M_n u| \leq \|M_n\| \cdot |u|$. Then

$$\begin{aligned} d_\alpha^k(\nu_n M_n^{-1}, \nu M^{-1}) &\leq E\{|M_n U_n - MU|^\alpha\}^{1/\alpha} \\ &\leq E\{|M_n(U_n - U)|^\alpha\}^{1/\alpha} + E\{|(M_n - M)U|^\alpha\}^{1/\alpha} \\ &\leq \|M_n\| \cdot E\{|U_n - U|^\alpha\}^{1/\alpha} \\ &\quad + \|M_n - M\| \cdot E\{|U|^\alpha\}^{1/\alpha} \to 0. \quad \square \end{aligned}$$

LEMMA 5.2. *Let μ_n be the empirical distribution of (Y_i, U_i, V_i) for $1 \leq i \leq n$. Let μ be the common theoretical distribution of (Y_i, U_i, V_i). Then $d_4^{1+p+r}(\mu_n, \mu) \to 0$ a.e.*

PROOF. This is Lemma 8.4 of B&F. $\square$

LEMMA 5.3. *Let $\tilde{\mu}_n$ be the empirical distribution of $(U_i, V_i, \tilde{\varepsilon}_i(n))$ for $1 \leq i \leq n$. Let $\tilde{\mu}$ be the common theoretical distribution of $(U_i, V_i, \varepsilon_i)$. Then $d_4^{p+r+1}(\tilde{\mu}_n, \tilde{\mu}) \to 0$ a.e.*

PROOF. This follows from Lemmas 5.1–5.2, because $\tilde{\mu}_n$ is the image of μ_n under the linear mapping L_n:

$$L_n(y, u, v) = (u, v, y - u\hat{A}_n - \hat{b}_n^T v)$$

and $\hat{A}_n \to A$, $\hat{b}_n \to 0$ a.e. So, $\tilde{\mu}_n$ tends to the image of μ under the linear mapping L:

$$L(y, u, v) = (u, v, y - uA).$$

This is $\tilde{\mu}$. $\square$

Claim (a) of the theorem follows from Lemma 8.6 of B&F. Likewise, claim (b) follows from Lemma 8.7 of B&F: indeed, $V_j^* \varepsilon_j^*$ has mean 0 due to the orthogonalization, and its conditional law is close in d_2^r to the unconditional law of $V_j \varepsilon_j$, by Lemma 8.5 of B&F and the present Lemma 5.3. This is where L_4 is needed. Claim (b) of the theorem can be sharpened; a.e., as $n \to \infty$, given the data, the d_2^r-distance between the conditional law of $\sqrt{n}\Delta_n^*$ and the unconditional law of $\sqrt{n}\Delta_n$ tends to 0.

6. Technical details; Theorem 4.1. The random vectors X_t and ε_t are defined on some probability triple $(\Omega, \mathcal{F}, P)$, with ω a typical element of Ω. Thus, $X_t(\omega)$ is a $1 \times b$-vector in Euclidean space. Let $\mu_n(\omega)$ be the empirical distribution of the 3-tuples $[X_t(\omega), Y_t(\omega), Y_{t-1}(\omega)]$ for $t = 1, \cdots, n$; so $\mu_n(\omega)$ is an atomic probability in R^k, with $k = b + 2a$. Let μ be the theoretical distribution of (X_t, Y_t, Y_{t-1}), for any particular t. This too is a probability in R^k.

LEMMA 6.1. *Assume conditions* (4.1)–(4.8) *only. Then* $d_4^k(\mu_n, \mu) \to 0$ *a.e.*

PROOF. This follows from the ergodic theorem, and Lemma 8.3 of Bickel and Freedman (1981), referenced hereafter as B&F. □

Recall that the $\tilde{\varepsilon}_t(n)$ are the residuals from the fitting, made orthogonal to the X_t's. Let S be the set of pairs (x, z), where x is $1 \times b$ and z is $1 \times a$. So S is a Euclidean space, of dimension $b + a$. Equip S with the Euclidean norm $|\cdot|$. Let $\tilde{\mu}$ be the distribution of (X_t, ε_t). So $\tilde{\mu}$ is a probability in S, satisfying

$$\int_S |(x, z)|^4 \tilde{\mu}(dx, dz) < \infty \tag{6.1}$$

$$\int_S x^T z \tilde{\mu}(dx, dz) = 0 \tag{6.2}$$

by (4.2) and (4.3) respectively. Let $\tilde{\mu}_n$ be the empirical distribution of $(X_t, \tilde{\varepsilon}_t(n))$ for $t = 1, \cdots, n$. So $\tilde{\mu}_n$ is an atomic probability in S, satisfying (6.2) by construction. Let $\mathcal{S}$ be the set of probabilities in S satisfying (6.1)–(6.2), so $\tilde{\mu}_n \in \mathcal{S}$ and $\tilde{\mu} \in \mathcal{S}$.

LEMMA 6.2. *Assume* (4.1)–(4.16). *Then* $d_4^{b+a}(\tilde{\mu}_n, \tilde{\mu}) \to 0$ *a.e.*

PROOF. This is immediate from Lemmas 5.1 and 6.1, because $\hat{A}_n \to A$, $\hat{B}_n \to B$, and $\hat{C}_n \to C$ a.e.; and $\hat{b}_n \to 0$ a.e. for each equation in the system: see (4.16). □

Turn to claim (a) of the theorem. We focus on Q_n^*, the argument for R_n^* and S_n^* being similar. Notice that Q_n^* consists of some elements of $(1/n) \sum_{t=1}^n X_t^{*T} Y_t^*$, and some of $(1/n) \sum_{t=1}^n Y_{t-1}^{*T} Y_t^*$. We focus on the first group, the second being similar. If ζ_s are independent random vectors, and $|\ |$ is the

Euclidean norm, then of course

$$E\{|\sum_s \zeta_s|^2\} = \sum_s E\{|\zeta_s - E(\zeta_s)|^2\} + |\sum_s E\{\zeta_s\}|^2$$
$$\leq \sum_s E\{|\zeta_s|^2\} + |\sum_s E\{\zeta_s\}|^2. \tag{6.3}$$

If M is a matrix, recall $|M| = (\Sigma_{ij} M_{ij}^2)^{1/2}$, so $\|M\| \leq |M|$, where $\|M\|$ is the L_2 operator norm of M. If v is $r \times 1$ and x is $1 \times q$, confirm that $|vx| = |v| \cdot |x|$. Let $\psi_n(\mu)$ be the law of $(1/n)\sum_{t=1}^n X_t^T Y_t$, when (X_t, ε_t) has law μ. Metrize ψ's by $d_1^{b\times a}$ and μ's by d_2^{b+a}. Recall S and $\mathscr{S}$ from (6.1)–(6.2).

LEMMA 6.3. *The $\psi_n(\mu)$ are equiuniformly continuous functions of μ on the "ball"*

$$\left\{\mu: \mu \in \mathscr{S} \text{ and } \int_S |(x, z)|^2 \mu(dx, dz) \leq c^2 < \infty\right\}.$$

PROOF. Fix μ and μ' in the ball. We must estimate $d_1^{b\times a}[\psi_n(\mu), \psi_n(\mu')]$. To this end, construct independent, identically distributed 4-tuples

$$(X_t, \varepsilon_t, X_t'\varepsilon_t'): t = 0, \pm 1, \cdots$$

such that (X_t, ε_t) has law μ, (X_t', ε_t') has law μ', and

$$E\{|X_t - X_t'|^2 + |\varepsilon_t - \varepsilon_t'|^2\} = d_2^{b+a}(\mu, \mu')^2. \tag{6.4}$$

See Lemma 8.1 of B&F. Build Y_t in terms of $\xi_{t,s}$ from the $(X_{t-s}, \varepsilon_{t-s})$ and Y_t' from $\xi_{t,s}'$ in terms of $(X_{t-s}', \varepsilon_{t-s}')$, as in (4.8)–(4.9). Then

$$d_1^{b\times a}[\psi_n(\mu), \psi_n(\mu')] \leq E\{|(1/n)\sum_{t=1}^n (X_t^T Y_t - X_t'^T Y_t')|\}$$
$$\leq E\{|X_t^T Y_t - X_t'^T Y_t'|\}$$
$$\leq E\{|X_t^T(Y_t - Y_t')|\} + E\{|(X_t - X_t')^T Y_t'|\}$$
$$= E\{|X_t| \cdot |Y_t - Y_t'|\} + E\{|X_t - X_t'| \cdot |Y_t'|\}.$$

Only the first term will be estimated, the second being easier. By the Cauchy-Schwarz inequality,

$$E\{|X_t| \cdot |Y_t - Y_t'|\}^2 \leq E\{|X_t|^2\} E\{|Y_t - Y_t'|^2\}$$
$$\leq c^2 \cdot E[|Y_t - Y_t'|^2].$$

Recall $\xi_{t,s}$ from (4.8); and $\xi_{t,s}'$ is defined analogously, in terms of ξ_{t-s}' and X_{t-s}'. Now $\xi_{t,s} - \xi_{t,s}'$ are independent for $s = 0, 1, \cdots$; equation (6.3) applies, and shows

$$E\{|Y_t - Y_t'|^2\} \leq E\{|\delta_0|^2\} \frac{1}{1 - r^2} + |E\{\delta_0\}|^2 \frac{1}{(1 - r)^2}$$

where $r < 1$ by (4.6) and

$$\delta_0 = [(\varepsilon_0 - \varepsilon_0') + (X_0 - X_0')C](I - A)^{-1}.$$

This is small if μ and μ' are close in d_2: see (6.4). □

Claim (a) of Theorem 4.1 now follows in respect of $(1/n)\sum_{t=1}^n X_t^{*T}Y_t^*$. Indeed, by Lemma 6.3 the conditional law of $(1/n)\sum_{t=1}^n X_t^{*T}Y_t^*$ given the data differs little in the sense of $d_1^{b\times a}$ from the unconditional law of $(1/n)\sum_{t=1}^n X_t^T Y_t$, because $\tilde{\mu}_n$ differs little in the sense of d_2^{b+a} from $\tilde{\mu}$, by Lemma 6.2. The balance of the argument is omitted.

Turn to claim (b) of the theorem. Notice that $\sqrt{n}\Delta_n^*$ consists of some elements of $(1/\sqrt{n})\sum_{t=1}^n X_t^{*T}\varepsilon_t^*$, followed by $(1/\sqrt{n})\sum_{t=1}^n Y_{t-1}^{*T}\varepsilon_t^*$. The first group can be handled by Lemma 6.2, and Lemma 8.7 of B&F. For the second group, new arguments are needed, and these will now be given.

Recall S and $\mathscr{S}$ from (6.1)–(6.2); each equation has an intercept, so for $\mu \in \mathscr{S}$, we may also assume

$$\int_S z\mu(dx, dz) = 0. \tag{6.5}$$

Let $\Phi_n(\mu)$ be the law of $(1/\sqrt{n})\sum_{t=1}^n Y_{t-1}^T\varepsilon_t$, when (X_t, ε_t) has law μ. Metrize Φ's by $d_2^{a\times a}$, and μ's by d_4^{b+a}.

LEMMA 6.4. *The $\Phi_n(\cdot)$ are equiuniformly continuous functions of μ on the "ball"*

$$\left\{\mu\colon \mu \in \mathscr{S} \text{ and } \int_S |(x, z)|^4\mu(dx, dz) \le c^4 < \infty\right\}.$$

PROOF. Lemma 8.7 of B&F does not apply directly, because the $Y_{t-1}^T\varepsilon_t$ are not independent. However, the argument for Lemma 6.3 can be pushed through. Fix μ and μ' in the ball. Modify the construction so that

$$E\{[|X_t - X_t'|^2 + |\varepsilon_t - \varepsilon_t'|^2]^2\} \le d_4^{b+a}(\mu, \mu')^4. \tag{6.6}$$

Then

$$d_2^{a\times a}[\Phi_n(\mu), \Phi_n(\mu')]^2 \le E\{|(1/\sqrt{n})\sum_{t=1}^n (Y_{t-1}^T\varepsilon_t - Y_{t-1}'^T\varepsilon_t')|^2\}. \tag{6.7}$$

The terms are stationary and pairwise orthogonal because $E\{\varepsilon_t\} = E\{\varepsilon_t'\} = 0$. So the right side of the inequality (6.7) is

$$E\{|Y_0^T\varepsilon_1 - Y_0'^T\varepsilon_0'|^2\}. \tag{6.8}$$

Now

$$Y_0^T\varepsilon_1 - Y_0'^T\varepsilon_1' = Y_0^T(\varepsilon_1 - \varepsilon_1') + (Y_0^T - Y_0'^T)\varepsilon_1',$$

so the expression (6.8) is bounded above by

$$2E\{|Y_0^T(\varepsilon_1 - \varepsilon_1')|^2\} + 2E\{|(Y_0^T - Y_0'^T)\varepsilon_1'|^2\}$$
$$= 2E\{|Y_0|^2\cdot|\varepsilon_1 - \varepsilon_1'|^2\} + 2E\{|Y_0 - Y_0'|^2\cdot|\varepsilon_1'|^2\}.$$

Only the second term will be estimated. By the Cauchy-Schwarz inequality,

$$E\{|Y_0 - Y_0'|^2\cdot|\varepsilon_1'|^2\}^2 \le c^4\cdot E\{|Y_0 - Y_0'|^4\}.$$

Abbreviate

$$\xi_s = [(\varepsilon_{-s} - \varepsilon'_{-s}) + (X_{-s} - X'_{-s})C](I - A)^{-1}$$

$$|\xi_s| = \zeta_s \quad \text{and} \quad \bar{B} = B(I - A)^{-1}.$$

Now $r = \|\bar{B}\| < 1$ by (4.6) and $Y_0 - Y'_0 = \sum_{s=0}^{\infty} \xi_s \bar{B}^s$ by (4.8) so

$$|Y_0 - Y'_0|^4 \leq (\sum_{s=1}^{\infty} \zeta_s r^s)^4 = \sum_{ijk\ell} \zeta_i \zeta_j \zeta_k \zeta_\ell r^{i+j+k+\ell}.$$

Therefore $E\{|Y_0 - Y'_0|^4\} \leq \Delta/(1-r)^4$, where $\Delta = \max_{ijk\ell} E\{\zeta_i \zeta_j \zeta_k \zeta_\ell\}$ is small if μ and μ' are close, by (6.6). □

REFERENCES

BERAN, R. (1982). Estimated sampling distributions: the bootstrap and competitors. *Ann. Statist.* **10** 212–225.

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

BICKEL, P. and FREEDMAN, D. (1983). Bootstrapping regression models with a large number of parameters. In *The Lehmann Festschrift*, ed. by P. Bickel, J. L. Hodges, and K. Doksum. Wadsworth, Belmont, California.

BROWN, T. M. (1954). Standard errors of forecast of a complete econometric model. *Econometrica* **22** 178–192.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman, London.

DAGGETT, R. and FREEDMAN, D. A. (1984). Econometrics and the law: a case study in the proof of antitrust damages. Technical report no. 23, Department of Statistics, University of California, Berkeley. To appear in the Neyman-Kiefer volume, ed. by L. Le Cam, pub. by Wadsworth, Belmont, California.

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

EFRON, B. (1982). The jackknife, the bootstrap and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics. Monograph 38. Society for Industrial and Applied Mathematics, Philadelphia.

FAIR, R. (1979). An analysis of the accuracy of four macro-econometric models. *J. Political Econ.* **87** 701–718.

FAIR, R. (1980). Estimating the expected predictive accuracy of econometric models. *Internat. Econ. Rev.* **21** 355–378.

FREEDMAN, D. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

FREEDMAN, D. and PETERS, S. (1984a). Bootstrapping a regression equation: some empirical results. *J. Amer. Statist. Assoc.* **79** 97–106.

FREEDMAN, D. and PETERS, S. (1984b). Bootstrapping an econometric model: some empirical results. *J. Bus. Econ. Statist.* **2** 150–158.

GOLDBERGER, A., NAGAR, A. L. and ODEH, H. S. (1961). The covariance matrices of reduced-form coefficients and of forecasts for a structural econometric model. *Econometrica* **29** 556–573.

PRATT, J. and SCHLAIFER, R. (1984). On estimating structure. *J. Amer. Statist. Assoc.* **79** 9–21.

SHORACK, G. (1982). Bootstrapping robust regression. *Comm. Statist. (A)* **11** 961–972.

SINGH, K. (1981). On asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

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

WHITE, H. (1982). Instrumental variables regression with independent observations. *Econometrica* **50** 483–500.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720