SUBSAMPLING FOR ECONOMETRIC MODELS

BY

DIMITRIS N. POLITIS and JOSEPH P. ROMANO

TECHNICAL REPORT NO. 484

AUGUST 1995

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS95–04379

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Subsampling for Econometric Models

By

Dimitris N. Politis and Joseph P. Romano

Technical Report No. 484

August 1995

Prepared Under the Auspices

Of

National Science Foundation Grant DMS95-04379

Department of Statistics

Stanford University

Stanford, California
Subsampling for Econometric Models

_Dimitris N. Politis and Joseph P. Romano_

Department of Statistics, Purdue University

and Department of Statistics, Stanford University

June 1995
1. INTRODUCTION

The survey article, *Bootstrapping Time Series Models*, by Li and Maddala is a timely and thoughtful account of bootstrap and other resampling methods that are finding increasing usefulness in econometrics. The purpose of the present article is to present an approach not discussed by Li and Maddala that has certain virtues. Specifically, a general theory for the construction of hypothesis tests and confidence intervals or regions is presented. The basic idea is to approximate the sampling distribution of a statistic based on the values of the statistic computed over smaller subsets of the data. For example, in the case where the data are \( n \) observations which are independent and identically distributed, a statistic is computed based on the entire data set and is recomputed over all \( \binom{n}{b} \) data sets of size \( b \). Implicit is the notion of a statistic sequence, so that the statistic is defined for samples of size \( n \) and \( b \). These recomputed values of the statistic are suitably normalized to approximate the true sampling distribution.

The approach presented here is perhaps the most general theory for the construction of first order asymptotically valid confidence regions. That is, it will be seen that, under very weak assumptions on \( b \), the method is valid whenever the original statistic, suitably normalized, has a limit distribution under the true model. No knowledge, such as normality, of the limit distribution is required. Other methods, such as Efron’s (1979) bootstrap, require that the distribution of the statistic is somehow locally smooth as a function of the unknown model. Indeed, the method here is applicable even in the several known situations which represent counterexamples to the bootstrap. To appreciate why our method behaves well under such weak assumptions, note that each subset of size \( b \) (taken without replacement from the original data) is indeed a sample of size \( b \) from the true model. Hence, it should be intuitively clear that one can at least approximate the sampling distribution of the (normalized) statistic based on a sample of size \( b \). But, under the weak convergence hypothesis, the sampling distributions based on samples of size \( b \) and \( n \) should be close.

The method has a clear extension to the context of a stationary time series or, more generally, a homogeneous random field. The only difference is that the statistic is computed over a smaller number of subsets of the data that retain the dependence structure of the observations. For example,
if \( X_1, \ldots, X_n \) represent \( n \) observations from some stationary time series, the statistic is recomputed over the \( n - b + 1 \) subsets of size \( b \) of the form \( \{X_i, X_{i+1}, \ldots, X_{i+b-1}\} \).

The use of subsample values to approximate the variance of a statistic is well-known. In the i.i.d. case, the Quenouille-Tukey jackknife estimates of bias and variance based on computing a statistic over all subsamples of size \( n - 1 \) has been well-studied and is closely related to the mean and variance of our estimated sampling distribution with \( b = n - 1 \). In fact, the jackknife was originally conceived by Quenouille (1949) in the context of bias reduction of a correlation estimate based on an observed stretch of a time series. Our approach is to actually use the pseudo-values to approximate an entire sampling distribution. The result is a method that is first-order correct under minimal assumptions.

The i.i.d. case is reviewed in Section 2 and the stationary time series case is presented in Section 3. The viability of our method in the context of time series is particularly important because the distribution theory of many estimators is quite complicated. The approach presented here is mathematically quite simple, and the hope is that these ideas can be applied quite fruitfully to very complex models. Future work will address the higher order accuracy of the subsampling approach, the choice of subsample size, and comparisons with other methods.
2. SUBSAMPLING IN THE I.I.D. CASE

2.1 The Basic Theorem.

Suppose $X_1, \ldots, X_n$ is a sample of $n$ independent and identically distributed random variables taking values in an arbitrary sample space. The probability distribution generating the observations is denoted $P$. The goal is to construct a confidence region for some parameter $\theta(P)$. For now, assume $\theta$ is real-valued, but this can be considerably generalized to allow for the construction of confidence regions for multivariate parameters or uniform confidence bands for functions.

An estimator $T_n = T_n(X_1, \ldots, X_n)$ of $\theta(P)$ is given. It is desired to estimate or approximate the true sampling distribution of $T_n$ in order to make inferences about $\theta(P)$. Nothing is assumed about the form of the estimator, though it is natural in the i.i.d. context to assume $T_n$ is symmetric in its arguments. Let $J_n(P)$ to be the sampling distribution of $\tau_n(T_n - \theta(P))$ based on a sample of size $n$ from $P$. Also define the corresponding cumulative distribution function:

$$J_n(x, P) = \text{Prob}_P\{\tau_n[T_n(X_1, \ldots, X_n) - \theta(P)] \leq x\}.$$  

The basic idea is that, if $J_n(P)$ were known, a confidence interval for $\theta(P)$ could be constructed by the usual pivotal technique. The approach can be stated more generally for studentized roots as well. In any case, the bootstrap approach is to estimate $J_n(P)$ by $J_n(\hat{Q}_n)$ for some appropriate estimate $\hat{Q}_n$ of $P$. Hence, the bootstrap requires that $J_n(\cdot)$ be somehow smooth in the argument $P$. Essentially, the only assumption that we will need to construct asymptotically valid confidence intervals for $\theta(P)$ is that $J_n(P)$ converges weakly to a limit law $J(P)$ as $n \to \infty$.

To describe the method studied in this section, let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Only a very weak assumption on $b$ will be required. In typical situations, it will be assumed that $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Now, let $S_{n,i}$ be equal to the statistic $T_b$ evaluated at the data set $Y_i$. The approximation to $J_n(x, P)$ we study is defined by

$$L_n(x) = N_n^{-1} \sum_{i=1}^{N_n} 1\{\tau_b(S_{n,i} - T_n) \leq x\}.$$  \hspace{1cm} (2.1)

The motivation behind the method is the following. For any $i$, $Y_i$ is a random sample of size $b$ from $P$. Hence, the exact distribution of $\tau_b(S_{n,i} - \theta(P))$ is $J_b(P)$. The empirical distribution of the
$N_n$ values of $\tau_b(S_n,i - \theta(P))$ should then serve as a good approximation to $J_n(P)$. Of course, $\theta(P)$ is unknown, so we replace $\theta(P)$ by $T_n$, which is asymptotically permissible because $\tau_b(T_n - \theta(P))$ is of order $\tau_b/\tau_n \to 0$.

**Theorem 1.** Assume $J_n(P)$ converges weakly to a limit law $J(P)$. Also assume $\tau_b/\tau_n \to 0$, $b \to \infty$ and $b/n \to 0$ as $n \to \infty$. Let $x$ be a continuity point of $J(\cdot, P)$.

(i) Then, $L_n(x) \to J(x, P)$ in probability.

(ii) If $J(\cdot, P)$ is continuous, then

$$
\sup_x |L_n(x) - J_n(x, P)| \to 0 \quad (2.2)
$$

in probability.

(iii). Let $c_n(1 - \alpha) = \inf \{x : L_n(x) \geq 1 - \alpha\}$. Correspondingly, define $c(1 - \alpha, P) = \inf \{x : J(x, P) \geq 1 - \alpha\}$. If $J(\cdot, P)$ is continuous at $c(1 - \alpha, P)$, then

$$
\text{Prob}_P\{\tau_n[T_n - \theta(P)] \leq c_n(1 - \alpha)\} \to 1 - \alpha \quad (2.3)
$$

as $n \to \infty$. Thus, the asymptotic coverage probability under $P$ of the interval $[T_n - \tau_n^{-1}c_n(1 - \alpha), \infty)$ is the nominal level $1 - \alpha$.

(iv). Assume, for every $d > 0$, $\sum_n \exp\{-d[n/b]\} < \infty$ and $\tau_b(T_n - \theta(P)) \to 0$ almost surely. Then, the convergences in (i) and (ii) hold with probability one.

The theorem is proved in Politis and Romano (1994).

2.2. Comparison With The Bootstrap.

The usual bootstrap approximation to $J_n(x, P)$ is $J_n(x, \hat{Q}_n)$, where $\hat{Q}_n$ is some estimate of $P$. In many (but not all) i.i.d. situations, $\hat{Q}_n$ is taken to be the empirical distribution of the sample $X_1, \ldots, X_n$. The analogous results to (2.2) and (2.3) with $L_n(x)$ replaced by $J_n(x, \hat{Q}_n)$ have been proved in many situations; see Bickel and Freedman (1981) and Beran (1984). In fact, dozens of other papers exist whose sole purpose is to prove such results in very specific situations. Our theorem immediately applies very generally with no further work.
To elaborate a little further, analogous bootstrap limit results are typically proved in the following manner. For some choice of metric (or pseudo-metric) $d$ on the space of probability measures, it must be known that $d(P_n, P) \to 0$ implies $J_n(P_n)$ converges weakly to $J(P)$. That is, our pointwise weak convergence assumption that $J_n(P)$ converges weakly to $J(P)$ for fixed $P$ must be strengthened so that the convergence of $J_n(P)$ to $J(P)$ is suitably locally uniform in $P$. In addition, the estimator $\hat{Q}_n$ must then be known to satisfy $d(\hat{Q}_n, P) \to 0$ almost surely or in probability under $P$. In contrast, no such local uniform convergence hypothesis is required in Theorem 2.1. In the known counterexamples to the bootstrap, it is precisely a certain lack of uniformity in convergence which leads to failure of the bootstrap.

In some special cases, it has been realized that a sample size trick can often remedy the inconsistency of the bootstrap. To describe how, focus on the case where $\hat{Q}_n$ is the empirical measure, denoted $\hat{P}_n$. Rather than approximating $J_n(P)$ by $J_n(\hat{P}_n)$, the suggestion is to approximate $J_n(P)$ by $J_b(\hat{P}_n)$ for some $b$ which usually satisfies $b/n \to 0$ and $b \to \infty$. The resulting estimator $J_b(x, \hat{P}_n)$ is obviously quite similar to our $L_n(x)$ given in (2.1). In words, $J_b(x, \hat{P}_n)$ is the bootstrap approximation defined by the distribution (conditional on the original data) of $\tau_b[T_b(X_1^*, \ldots, X_b^*) - T_n]$, where $X_1^*, \ldots, X_b^*$ are chosen with replacement from $X_1, \ldots, X_n$. In contrast, $L_n(x)$ is the distribution (conditional on the data) of $\tau_b[T_b(Y_1^*, \ldots, Y_b^*) - T_n]$, where $Y_1^*, \ldots, S_b^*$ are chosen without replacement from $X_1, \ldots, X_n$. Clearly, these two approaches must be similar if $b$ is so small that sampling with and without replacement are essentially the same. Indeed, if one resamples $b$ numbers (or indices) from the set $\{1, \ldots, n\}$, then the chance that none of the indices is duplicated is $\prod_{i=1}^{b-1}(1 - \frac{i}{n})$. This probability tends to 0 if $b^2/n \to 0$. (To see why, take logs and do a Taylor expansion analysis.) Hence, the following is true.

**Corollary 2.1.** Under the further assumption that $b^2/n \to 0$, parts (i)–(iii) of Theorem 2.1 remain valid if $L_n(x)$ is replaced by the bootstrap approximation $J_b(x, \hat{P}_n)$.

3. **STATIONARY TIME SERIES**

To fix ideas, suppose the sample $X_1, \ldots, X_n$ is known or suspected to exhibit serial dependence, and that it can generally be modeled as a stationary time series. Assume that the parameter of interest $\theta$ is the common mean $EX_i$, and the statistic $T_n$ is the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. If the serial dependence is weak enough such that $\sum |R(k)| < \infty$, where $R(k) = Cov(X_1, X_{1+k})$, then (under regularity conditions, cf. Brockwell and Davis (1991)) $\bar{X}_n$ is asymptotically normal, i.e., $\sqrt{n}(\bar{X}_n - \theta)$ has the limiting normal $N(0, \sigma^2_{\infty})$ distribution, where $\sigma^2_{\infty} = Var(X_1) + 2 \sum_{i=1}^{\infty} R(k)$. Note that $\sigma^2_{\infty} = \lim_{n \to \infty} \sigma^2_n$, where $\sigma^2_n = Var(\sqrt{n}\bar{X}_n) = Var(X_1) + 2 \sum_{i=1}^{n} (1 - |i|/n) R(k)$.

Now to construct confidence intervals for $\theta$ using the previously mentioned Central Limit Theorem, a consistent estimate of $\sigma^2_{\infty}$ is required. A straightforward approach would be to estimate the covariances $R(k)$, $k = 1, \ldots, n$, and plug them in the formula for $\sigma^2_n$, since $\sigma^2_n \to \sigma^2_{\infty}$. However, this naive procedure is not consistent, because the estimates of $R(k)$ for $k$ close to $n$ are highly inaccurate; indeed, $\sigma^2_{\infty}$ is just a constant multiple of the spectral density of the time series evaluated at point zero, this is a well known difficulty in the literature concerning spectral estimation (cf., for example, Priestley (1981)). Apparently, based on a sample of size $n$ we could only accurately estimate $R(k)$, for $k = 1, \ldots, b$, where $b << n$. It then follows that we can only hope to estimate well $\sigma^2_b$, and not $\sigma^2_n$, where $\sigma^2_b = Var(X_1) + 2 \sum_{i=1}^{b} (1 - |i|/b) R(k)$. But there is a natural way to estimate $\sigma^2_b$ from $X_1, \ldots, X_n$, namely to look at the sample variability of $\frac{1}{\sqrt{b}} \sum_{t=i}^{i+b-1} X_t$, for $i = 1, \ldots, n - b + 1$. This is equivalent to considering the ‘sample variance’ estimator

$$\hat{\sigma}^2_b = \frac{1}{n - b + 1} \sum_{i=1}^{n-b+1} \left( \frac{1}{\sqrt{b}} \sum_{t=i}^{i+b-1} X_t - \sqrt{b}\bar{X}_n \right)^2$$

that was studied in Carlstein (1986) and in Politis and Romano (1993). This proposed estimator is consistent, under some moment and mixing conditions, essentially because both $\sigma^2_b$ and $\sigma^2_n$ converge asymptotically to $\sigma^2_{\infty}$, if both $b$ and $n$ are assumed to tend to infinity.

In the same light, one could look at the more general problem of estimating the distribution of $\sqrt{n}(\bar{X}_n - \theta)$. But this can be done by looking at the sample variability of $\frac{1}{\sqrt{b}} \sum_{t=i}^{i+b-1} X_t$, for $i = 1, \ldots, n - b + 1$, and defining the corresponding ‘empirical’ distribution

$$L_n(x) = (n - b + 1)^{-1} \sum_{i=1}^{n-b+1} 1\{\sqrt{b}(b^{-1} \sum_{t=i}^{i+b-1} X_t - \bar{X}_n) \leq x\}$$

as an approximation to the sampling distribution of $\sqrt{n}(\bar{X}_n - \theta)$. Here the underlying principle is that both $\sqrt{b}(\bar{X}_b - \theta)$ and $\sqrt{n}(\bar{X}_n - \theta)$ have the same asymptotic distribution, (which just happens
to be the normal $N(0, \sigma^2_{\infty})$ distribution), where of course $\bar{X}_b = b^{-1} \sum_{i=1}^{b} X_i$. Although variance estimation is intimately linked with the assumption of asymptotic normality, this more general idea of directly approximating the sampling distribution would work in a variety of different situations, including cases where asymptotic normality does not hold, where the rate of convergence is not $\sqrt{n}$, or where variance estimation is not consistent.

3.2. The General Stationary Case.

Suppose $X_1, \ldots, X_n$ is an observed stretch of a time series. For simplicity, we’ll assume the time series is stationary, but this can be weakened. The probability measure generating the whole infinite sequence will be denoted $P$ and interest focuses on a real-valued parameter $\theta(P)$. (The case where $\theta(P)$ may be a function is considered in Romano, Politis, and You (1993), where an asymptotically valid approach for constructing a uniform confidence band for the spectrum is given.) As before, suppose $T_n = T_n(X_1, \ldots, X_n)$ is the statistic of interest, and let $J_n(P)$ be the distribution of $\tau_n[T_n - \theta(P)]$. Let $S_{n,i}$ be the statistic $T_b$ evaluated at the subseries $X_i, \ldots, X_{i+b-1}$. The subsampling approximation to the true sampling distribution of $\tau_n[T_n - \theta(P)]$ is given by the ‘empirical’ distribution

$$L_n(x) = q^{-1} \sum_{i=1}^{q} 1\{\tau_b(S_{n,i} - T_n) \leq x\}$$

and $q = n - b + 1$. In order to prove anything in the dependent data case, some sort of weak dependence condition must be in force. Before writing down limit results, we recall some standard notation. For a stationary time series $X = \{X_n, n \in \mathbb{Z}^+\}$, define Rosenblatt’s $\alpha$-mixing coefficient by $\alpha_X(j) = \sup_{A,B} |P(AB) - P(A)P(B)|$, where events $A$ and $B$ vary in the $\sigma$-fields generated by $\{X_n, n \leq k\}$ and $\{X_n, n \geq j + k\}$, respectively for any $k \geq 1$. The sequence $X$ is said to be $\alpha$-mixing if $\alpha_X(j) \to 0$ as $j \to \infty$. random fields.

We now describe how subsampling can be used in the context of hypothesis testing. Suppose the parameter of interest is $\theta(P)$ and we are interested in testing the null hypothesis $\theta(P) = \theta_0$ versus the alternative $\theta(P) > \theta_0$. The two-sided case can be considered in an analogous manner.

**Theorem 2.** Assume the underlying stationary process generating the data is a stationary strong mixing sequence. Suppose $J_n(P)$ converges weakly to a continuous limit law $J(P)$. As before, let

- 7 -
\[ c_n(1 - \alpha) = \inf\{x : L_n(x) \geq 1 - \alpha\} \]. Then, under the same assumptions for \( \tau_n \) and \( b \) in Theorem 1, the results (i)-(iii) continue to hold with \( L_n \) defined by (3.1). Thus, if \( \theta(P) = \theta_0 \), the test the rejects if \( \theta_0 < T_n - \tau_n^{-1}c_n(1 - \alpha) \) has asymptotic rejection probability \( \alpha \). Alternatively, if \( \theta(P) = \theta_1 > \theta_0 \), then the same test has rejection probability (or power at \( P \)) tending to one.

**Proof.** We provide a sketch, and refer the reader to Politis and Romano (1995) for similar arguments. By assumption, \( J_n(x, P) \rightarrow J(x, P) \). We need to show \( L_n(x) \rightarrow J(x, P) \) in probability. To do this, argue that it suffices to show

\[
U_n(x) = q^{-1} \sum_{i=1}^{q} q_1 \{ \tau_b(S_n,i - \theta(P)) \leq x \}
\]

 tends to \( J(x, P) \) in probability. (This is not hard since \( \tau_b(T_n - \theta(P)) \) tends to 0 in probability.) But, \( U_n(x) \) has mean \( J_b(x, P) \rightarrow J(x, P) \). Thus, it suffices to show the variance of \( U_n \) tends to zero. But, noting that the \( S_{n,i} \) form a stationary sequence themselves, we can write

\[
\text{var}[U_n(x)] = \frac{1}{q} \text{var}(Y_{n,1}) + \frac{2}{q} \sum_{i=1}^{q} (1 - \frac{i}{q}) \text{cov}(Y_{n,1}, Y_{n,1+i}),
\]

 where \( Y_{n,i} = 1\{\tau_b[S_{n,i} - \theta(P)] \leq x\} \). Bound \( \text{cov}(Y_{n,1}) \) by one, and for \( i = 1, \ldots, b - 1 \), bound \( \text{cov}(Y_{n,1}, Y_{n,1+i}) \) by one. So,

\[
\text{var}[U_n(x)] \leq (2b + 1)/N + \frac{2}{N} \sum_{i=b}^{N} |\text{cov}(Y_{n,1}, Y_{n,1+i})|.
\]

 But, \( Y_{n,1} \) is a function of \( \{X_1, \ldots, X_b\} \) and \( Y_{n,1+i} \) is a function of \( \{X_{1+i}, \ldots, X_{i+3}\} \), so that \( Y_{n,1} \) and \( Y_{n,1+i} \) are determined by \( X_j \)'s separated by \( 1 + i - b \). So, by the well-known mixing inequality (c.f. Ibragimov (1962)),

\[
|\text{cov}(Y_{n,1}, Y_{n,1+i})| \leq 4\alpha_X(1 + i - b),
\]

 for \( i = b, \ldots, N \). Hence,

\[
\text{var}[U_n(x)] \leq (2b + 1)/N + \frac{8}{N} \sum_{i=b}^{N} \alpha_X(1 + i - b).
\]

\[
\leq (2b + 1)/N + \frac{8}{N} \sum_{j=1}^{N} \alpha_X(j).
\]
The assumptions on \( b \) imply \( (2b+1)/N \to 0 \) as \( n \to \infty \). Strong mixing implies \( N^{-1} \sum_{j=1}^{N} \alpha_X(j) \to 0 \). Thus, \( U_n(x) \to J(x, P) \) in probability, if \( x \) is a continuity point of \( J(\cdot, P) \). By Polya's theorem, it follows that \( U_n(\cdot) \), and hence \( L_n(\cdot) \) converges to \( J(\cdot, P) \) uniformly. If \( \theta(P) = \theta_0 \), then the rejection probability of the test is

\[
P\{\theta_0 < T_n - \tau_n^{-1}c_n(1 - \alpha)\} = P\{\tau_n(T_n - \theta_0) > c_n(1 - \alpha)\},
\]

which tends to \( \alpha \) by application of Slutsky's theorem. If \( \theta(P) = \theta_1 > \theta_0 \), then this rejection probability becomes

\[
P\{\tau_n(T_n - \theta_1) > c_n(1 - \alpha) - \tau_n(\theta_1 - \theta_0)\}.
\]

But, \( c_n(1 - \alpha) \) is bounded in probability, \( \tau_n(T_n - \theta_1) \) has a limit law, and \( \tau_n(\theta_1 - \theta_0) \to \infty \) implies this probability tends to one.

The result is proved in the more general context of homogeneous random fields in Politis and Romano (1994). Generalizations to the nonstationary case are considered in Politis, Romano and Wolf (1995). The weak convergence hypothesis may be weakened to allow for \( \tau_n \) to be unknown. Some work in this direction is presented in Bertrand, Politis and Romano (1995).
References


