

ANALYSIS OF TIDAL DATA VIA THE BLOCKWISE BOOTSTRAP

Michael Sherman and F. Michael Speed, Jr. and F. Michael Speed *
Texas A&M University

Summary:

We analyze tidal data from Port Mansfield, TX using Kunsch's (1989) blockwise bootstrap in the regression setting. In particular, we estimate the variability of parameter estimates in a harmonic analysis via block subsampling of residuals from a least squares fit. We see that naive least squares variance estimates can be either too large or too small depending on the strength of correlation and the design matrix. We argue that the block bootstrap is a simple, omnibus method of accounting for correlation in a regression model with correlated errors.

Keywords: bootstrap, resampling, subsampling, correlation, harmonic analysis

* Michael Sherman is Assistant Professor, F. Michael Speed, Jr. is Visiting Assistant Professor, and F. Michael Speed is Associate Professor in the Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

1: Introduction

Efron's bootstrap (1982) has become a popular tool for analyzing the distribution of complicated statistics in a wide variety of settings. In his paper justifying use of the bootstrap in the independent data setting, Singh (1981) pointed out that the bootstrap could not work for dependent (correlated) data without modification. Since this time there has been a considerable amount of theoretical research into model free subsampling methods that account for temporal and/or spatial dependence.

There are two main approaches to subsampling in the dependent data setting: model based and model free. In the former the dependence structure is modeled in terms of a few unknown parameters and independent errors (see, e.g., Bose (1988) who estimates the distribution of the least squares estimator of autoregressive parameters). In the latter the observed series is divided into "blocks", and these blocks are used to capture the dependence in the original series. In the sequel we focus on model free methods. Seminal works in this area are, e.g., Carlstein (1986) and Künsch (1989) who introduced "subseries" and the "blockwise bootstrap", respectively. Although these methods have been justified by their asymptotic properties (e.g., consistency, asymptotic normality) they seem to have not been often applied in data analyses.

In this paper, our basic goal is to obtain valid inferences for model parameters in regression in the presence of correlated errors without having to specify the correlation structure. Our analysis of tidal data at Port Mansfield, TX, in the Gulf Coast U.S.A. is a natural setting in which to employ model free block resampling due to the large size of the data set and the approximate stationarity of the residuals from the least squares fit. An alternative approach to this problem is to model the correlation structure in the residuals, and then use the estimated correlations in a generalized least squares (GLS) fit. It is important to note the assumptions in this approach. First, the correct class of models describing the correlation in the residuals must be known, e.g., the autoregressive class of models, $ARMA(p, q)$. Secondly, in this case the correct order of the autoregressive component, p , and of the moving average component, q , needs to be known. Thirdly, these $p + q$ parameters must be estimated.

The block bootstrap captures the dependence in the series of residuals without the need to know or guess the correlation structure. One apparent drawback to this approach is that the ordinary least squares (OLS) estimates which we will employ, are known to be generally inefficient relative to GLS. It turns out that for harmonic regression with stationary errors OLS is fully efficient relative to GLS, as noted by Priestly (1981), Chapter 7. We note that our procedure is quite easy to implement in matrix based software like Splus or Gauss.

2: Correlated Errors in Regression

Consider the simple linear regression model:

$$Y_t = \alpha + \beta x_t + \epsilon_t, \quad t = 1, \dots, T,$$

where Y_t is the response at time t , x_t is a covariate believed to be related to the response, and ϵ_t are possibly correlated errors. The model parameters α and β are to be estimated via the usual method of OLS. A common belief is that ignoring positive correlation between the errors typically leads to underestimation of the variance of the estimated regression parameters (see, e.g., Neter, Kutner, Nachsteim, and Wasserman (1996), Chapter 12). It is not difficult to see that this conclusion depends crucially on the form of the design points x_t . To see this, note that if $\hat{\beta}$ is the least squares estimator of the slope parameter β then:

$$Var(\hat{\beta}) = \frac{\sum_{t=1}^T \sum_{u=1}^T (x_t - \bar{x})(x_u - \bar{x}) E(\epsilon_t \epsilon_u)}{\{\sum_{t=1}^T (x_t - \bar{x})^2\}^2}.$$

From this expression we see that there are two contributions to the variability of the slope estimate: 1) the correlations between errors and 2) the structure of the design. To see the effect of the design consider the variability of the slope estimate under the AR(1) correlation structure given by $\epsilon_t = \rho \epsilon_{t-1} + \eta_t$, where $\{\eta_t\}$ are independent and identically distributed disturbances with $E(\eta_i) = 0$ and $Var(\eta_i) = 1$. In this case $E(\epsilon_t \epsilon_u) = (1 - \rho^2)^{-1} \rho^{|t-u|}$, and we compare the expected variance estimate obtained by ignoring the correlation with the correct variance in the following two designs:

Design 1:

$$x_i = i, \quad i = 1, \dots, T,$$

Design 2:

$$x_i = \begin{cases} (i+1)/2, & \text{if } i \text{ is odd;} \\ T - (i/2) + 1 & \text{if } i \text{ is even.} \end{cases}$$

For example, when $T = 10$ and $\rho = .5$, we have $Var(\hat{\beta}) = .0306$ and $.00309$ in Designs 1 and 2, respectively, but in both cases $Var_{ind}(\hat{\beta}) := Var(\epsilon_i) / \sum_{t=1}^T (x_t - \bar{x})^2 = .01616$. Thus, we see that the usual OLS variance estimates which ignore correlation can either under or over estimate the correct variance of parameter estimates depending on the structure of the design. This principle holds more generally and will explain variance estimates of parameters in the harmonic analysis of the Port Mansfield tidal data given in Section 4.

3: Harmonic Analysis and a Description of the Block Bootstrap

We now describe the model for a harmonic analysis of tidal data. Let Y_t denote the tide level at time $t, t = 1, \dots, T$. The standard harmonic model is given in Hartley (1949):

$$Y_t = a_0 + \sum_{i=1}^m a_i \cos(it\gamma) + b_i \sin(it\gamma) + \epsilon_t, \quad (1)$$

where $\gamma = 2\pi/T$ and a_0, a_i and b_i are the unknown amplitudes (regression parameters) of the model to be estimated. It has been traditionally assumed that the errors, ϵ_i , are independently and identically distributed variates. Pictured in Figure 1 are the (first 200) residuals from the OLS fit to the Port Mansfield data plotted as a function of time, t . We see that the residuals are heavily correlated and thus, using the results of Section 2, the ordinary least squares estimates of variance cannot be trusted. The effects of ignoring this correlation in the Port Mansfield data will be shown in Section 4.

The block bootstrap algorithm depends on a block length, l . Hueristically, this l should be the shortest lag at which correlations become “negligible”. We give further guidance into this choice and discuss the significance of this choice in Section 4. We now give the block bootstrap procedure for estimating the variance and the distribution of parameter estimates (borrowing notation from Götze and Künsch (1996)). Efficient SPLUS code to implement the procedure for general regression models has been developed and is available from the first author upon request.

The Block Bootstrap Algorithm in Harmonic Analysis

- 1) Let $\hat{\epsilon}_t, t = 1, \dots, T$ be the residuals from the model fit:

$$\hat{\epsilon}_t = Y_t - \left\{ \hat{a}_0 + \sum_{i=1}^m \hat{a}_i \cos(it\gamma) + \hat{b}_i \sin(it\gamma) \right\},$$

where $\hat{a}_0, \hat{a}_i, \hat{b}_i$ are the OLS parameter estimates.

- 2) Now assuming that $T = bl$ with b and l integers: Let N_1, \dots, N_b denote b uniform draws with replacement from the integers $\{0, 1, \dots, T - l\}$. These represent the starting point for each block of length l . A block bootstrap resample of residuals, $(\hat{\epsilon}_1^*, \dots, \hat{\epsilon}_T^*)$, is defined by:

$$\hat{\epsilon}_{(j-1)l+t}^* = \hat{\epsilon}_{N_j+t}, \quad (1 \leq j \leq b, 1 \leq t \leq l).$$

- 3) The bootstrapped responses, Y_t^* , are then generated from the estimated model with residuals $\hat{\epsilon}_t^*$ and the original covariates:

$$Y_t^* = \hat{a}_0 + \sum_{i=1}^m \hat{a}_i \cos(it\gamma) + \hat{b}_i \sin(it\gamma) + \hat{\epsilon}_t^*.$$

4) From the resampled responses, Y_t^* , and original covariates we fit model (1) and obtain new parameter estimates, \hat{a}_0^* , \hat{a}_i^* , \hat{b}_i^* , $i = 1, \dots, m$.

5) Repeating steps 2) through 4) a large number, B , of times one obtains B bootstrap replicates from which features of the distribution of the parameter estimates can be estimated. In particular, the bootstrap variance estimates are simply the sample variance of the B computed values for each parameter.

Notes on the Algorithm:

a) The algorithm is generalizable to any regression model of the form $Y = X\beta + \epsilon$ with stationary errors using the natural modification.

b) By resampling blocks of residuals we automatically retain the approximately correct dependence structure in a block without having to know (or guess) a model for the correlation in the residuals. Although the resulting inferences will not be, in general, optimal, if large numbers of regressions need to be carried out the block bootstrap obviates the need for extensive modelling (this was the situation here, as several harmonic analyses were carried out on different sections of coast). Even for a single regression analysis the block bootstrap avoids the need to rely on possibly incorrect modelling assumptions concerning the joint distribution of the residuals.

c) If $T \neq bl$ for convenient (reasonable) integers b and l then letting $T' = bl$, one can simply take one block of length $T - T'$ at the end as suggested by Hall, Horowitz, and Jing (1995).

d) Choice of block length, l , can be a delicate issue. The basic principle is that the block length should be relatively longer when the correlations are strong in the residual series. This is because long blocks are needed to capture the full extent of the correlation. On the other hand, long blocks decrease the number of available blocks thus increasing the variability in the procedure. Carlstein (1986) proposed a simple method for choosing block length that is model dependent while Hall, Horowitz, and Jing (1995) have recently proposed a more model free method for choosing block length. The latter, however, does not seem reliable for strong correlations, like those in the Port Mansfield residuals. For this reason, and due to the simplicity of Carlstein's approach we will use a variant of his method in our analysis in Section 4.

4: Implementation for the Tidal Data and Results

We employ the block bootstrap methodology in analyzing annual tidal data from Port Mansfield, TX, collected in 1993. The arguments in the $\cos(\cdot)$ and $\sin(\cdot)$ functions (see equation (1)) in this particular model are given by $(\pi t S_i)/180$ where S_i denotes the speed of the i^{th} constituent, $i = 1, \dots, m$. The tidal level is measured hourly and thus we have $T = 24 * 365 = 8760$ total observations. The number of terms in the harmonic analysis is $m = 37$ and thus (with the intercept) there are 75 total parameters to be estimated. This model was dictated by the National Oceanographic

and Atmospheric Administration (NOAA).

In order to implement the block bootstrap algorithm we need to choose a block length, l . In order to avoid the problem discussed in Note c) of Section 3 we restrict attention to four candidate values of block length l : $l = 10$, $l = 40$, $l = 219$, $l = 438$ which yield $b = 876$, $b = 219$, $b = 40$, $b = 20$ blocks, respectively. This ensures that the total series length is a constant multiple of the block length, l . In order to choose a block length from these candidates we do the following. If we assume temporarily that the residuals come from an AR(1) sequence, then it can be shown using arguments analogous to Carlstein (1986) that the optimal choice of block length is $l = ((6^{1/2}\rho)/(1 - \rho^2))^{2/3}T^{1/3}$ where ρ is the AR(1) parameter (Carlstein gives a slightly different value for his “subseries” procedure). Plugging in a least squares estimate of ρ into the last expression gives a reasonable value of l . We then choose the one of four candidate block lengths that is closest to this value. For the Port Mansfield data the estimated AR(1) parameter is $\hat{\rho} = .9923$ giving an estimated block length of 604. Thus, we choose $l = 438$ for this data set.

We give in Table 1 (selected) variance estimates given by OLS and the block bootstrap procedure as well as the t-ratio for testing if the parameter is zero derived from each variance estimate. Each block bootstrap variance estimate was calculated from $B = 2000$ bootstrap replicates.

Table 1: Variance Estimates from OLS and Block Bootstrap (Port Mansfield Data)

Speed	Par. Est.	i.i.d. Var	t_{ind}	Boot Var.	t_{BB}
.04107	-.03946	.0000183	9.22	.00345	.672
1.0980	-.01216	.0000182	2.95	.00080	.430
30.0000	-.00647	.0000182	1.52	.00000193	4.66

We see that for the low speeds (long periods) the variance estimates which assume independence are much too small and that for the high speeds they are too big. The reason for this can be seen by considering the results in Section 2. For low speeds (long period), neighboring x_i 's are positively correlated which leads to an underestimation by the variance estimate which assumed independence. This effect diminishes and is eventually reversed as the speed increases (and the period decreases). In the former case the incorrect analysis finds an insignificant covariate significant and in the latter it finds a significant predictor to be insignificant. We see that ignoring the correlation in this example gives entirely incorrect inferences. Our results are in basic agreement with a completely parametric result that entailed extensive modelling of the errors under the assumption that the errors are from the ARMA(p,q) class of models. The validity of our procedure, however, rests only on the assumption of stationary errors, of which the ARMA(p,q) class is a special case.

In regards to identifying the proper distribution of the parameter estimates by using the 2000 bootstrap replicates we were able to calculate percentile confidence intervals for the parameters by

identifying the 5th and 95th percentiles of the bootstrap replicates as suggested by Efron (1982). There was essentially no difference when the finer Bias Corrected and Accelerated versions (Efron (1987)) were used. The percentile intervals, thus, agree closely with the usual intervals based on an assumed asymptotic normal distribution that used the bootstrap estimate of standard error. Practically, that the usual intervals are reasonable can be seen by examining the bootstrap histograms of the parameter estimates and checking for approximate normality.

4.1: A Simulation

The analysis of the Port Mansfield data indicates that the block bootstrap performs better than OLS in the presence of correlated residuals. In order to obtain a feel for how well the block bootstrapping performs in a case where the target variances are known we performed the following simulation. We generated $T = 300$ observations from the model: $Y_t = \alpha + (t/T)\beta_1 + (t/T)^2\beta_2 + \epsilon_t$, $t = 1, \dots, T$, where the errors are AR(1): $\epsilon_i = \rho\epsilon_{i-1} + \eta_i$, $E(\eta_i) = 0$, $Var(\eta_i) = 1$, and $\rho = .5$. For estimating the variance of the sample mean, under this correlation structure, the correct variances of parameter estimates are approximately $(1+\rho)/(1-\rho) = 3$ times larger than the estimates derived under independence. For each of 250 simulated data sets with $T = 300$, $l = 10$, $\alpha = \beta_1 = \beta_2 = 1$, we computed the bootstrap estimates of variance using the algorithm given in Section 3. In Table 2 we give the expected value of the OLS variance estimator, and the mean of the 250 bootstrap variance estimates with an associated estimated standard error of the variance estimator computed from the 250 independent simulations.

Table 2: Variance Estimates from OLS and Block Bootstrap (Simulation)

	α	β_1	β_2
True	0.1189	2.520	2.347
OLS variance estimates	0.0405	.8595	.8000
Bootstrap variances:			
Mean	0.0960	2.0384	1.897
(Standard Error)	.0015	.248	.216

We see that the block bootstrap estimates perform much better than the OLS variance estimates which ignore the correlation. The block bootstrap has a negative bias as is predicted by the theory (Künsch (1989), Theorems 3.2 and 3.4). For each of the three parameters the average bootstrap estimate of variance is approximately 80% of the correct and thus the estimated standard error is approximately 90% of the true standard error. This is fairly good performance given the modest length of the series and the omnibus nature of the estimator.

5: Summary

Using the block bootstrap in the context of tidal analysis, we have been able to calculate estimated standard errors of parameter estimates which account for dependence in a model free way. The bootstrap methodology also gives an estimate of the distribution of parameter estimates and thus can be used to construct confidence intervals. For the tidal data, there is every indication that the distribution is approximately normal and thus the bootstrap confidence intervals closely agree with the usual intervals based on asymptotic normality which employ the bootstrap estimates of standard error. We conclude that the block bootstrap is a reliable and omnibus method for drawing inferences regarding regression parameters in the presence of correlated errors.

REFERENCES

- Bose, A. (1988). Edgeworth Correction by Bootstrap in Autoregressions, *Annals of Statistics*, 16, 1709-1722.
- Carlstein, E. (1986). The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence, *Annals of Statistics*, 14, 1171-1179.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, Philadelphia.
- Efron, B. (1987). Better Bootstrap Confidence Intervals, *Journal of the American Statistical Assoc.*, 82, 171-185.
- Götze, F. and Künsch, H.R. (1996). Second Order Correctness of the Blockwise Bootstrap for Stationary Observations, *Annals of Statistics*, to appear.
- Hall, P., Horowitz, J.L., and Jing, B.Y. (1995). On Blocking Rules for the Bootstrap with Dependent Data, *Biometrika*, 3, 561-574.
- Hartley, H.O. (1949). Tests of Significance in Harmonic Analysis, *Biometrika*, 36, 194-201.
- Künsch, H. (1989). The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics*, 17, 1217-1241.
- Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models*, Irwin, Chicago.
- Priestly, M.B. (1981). *Spectral Analysis and Time Series*, Academic Press, London.
- Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap, *Annals of Statistics*, 9, 1187-1195.