Resampling: Theory and Applications

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Jaideep Mahajan

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2020

Supervisor: Prof. Arup Bose, ISI-Kolkata

Certificate

This is to certify that this dissertation entitled Resampling: Theory and Applications towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Jaideep Mahajan at Indian Institute of Science Education and Research under the supervision of Prof. Arup Bose, ISI-Kolkata, Professor, Department of Mathematics , during the academic year 2019-2020.

Arup Bose

(ARUP BOSE)

Prof. Arup Bose, ISI-Kolkata

Committee:

Prof. Arup Bose, ISI-Kolkata

Prof. Anindya Goswami

This thesis is dedicated to my parents and sister for their endless love and support.

Declaration

I hereby declare that the matter embodied in the report entitled Resampling: Theory and Applications are the results of the work carried out by me at the Department of Mathematics, Indian Institute of Science Education and Research, Pune, under the supervision of Prof. Arup Bose, ISI-Kolkata and the same has not been submitted elsewhere for any other degree.

Jaideep Mahajan

Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Arup Bose for his constant guidance throughout this project. I am extremely thankful to Prof. Uttara Naik-Nimbalkar and Sir Bhas Bhamre for building the preliminaries required for this project.
My completion of this project could not have been completed without the constant support and care of my friends Akriti, Anirudhha, Ashwin, Devesh, Kaarthik, Kush, Madhav, Prateek, Radhika, Rahul, Rajdeep, Saswata, Saurav, Shubhangi, Vishnu. My heartfelt thanks.

Abstract

Let $X = (X_1, \ldots, X_n)$ be the random sample obtained from some unknown distribution F. Let $\theta = T(F)$ be the parameter of interest estimated by $\hat{\theta} = S(X)$. We discuss bias and variance estimators for $\hat{\theta}$ obtained by resampling methods such as jackknife and bootstrap. Along these lines, we try to extrapolate these methods originally introduced for independent data set to specific models such as linear model and Dependent data set. More specifically, we consider least square estimates for the linear model and apply resampling methods for the bias and the variance estimates for the same. Time series model has been looked at in dependent data sequence and these resampling procedures have been modified to produce a consistent estimates for the statistic of interest.

Contents

Abstract **xi**

1 Resampling Methods **3**

1.1 Standard Errors and Bias 3

1.2 Bootstrap Distribution Estimator 7

1.3 Confidence Intervals 9

2 Linear Models **19**

2.1 Introduction 19

2.2 Variance and Bias Estimation 20

2.3 Prediction 25

2.4 Model selection 27

3 Dependent Data Set **29**

3.1 Introduction 29

3.2 m-dependent data 29

3.3 Bootstrap Based on lID Innovations 31

3.4 Moving Block Bootstrap 32

3.5 Non-overlapping Block Bootstrap 33

3.6 Time series 34
3.7 Simulations 36

4 Conclusion **41**

Introduction

Statistical analysis is often concerned with making inferences about accuracy and distributions of statistics. While inferring, traditional approach involves dealing with underlying distribution of the data. This distribution is itself hard to estimate or sometimes statistic is too complicated to use this theoretical approach. Resampling methods resamples from the obtained data and infer from the resampled data to estimate this underlying distribution and make estimations about accuracy of the statistic of interest. Widely used resampling methods are jackknife approach and bootstrap approach. Though these methods require intensive computer operations, technology has made it easier to use these approaches.

In 1949, Quenouille introduced jackknife method. This method involves deleting one data point from the original data set and making the inference based on this new data set to estimate the properties of the statistic of interest. This method can be used to calculate standard error and bias of the statistic based on i.i.d data set. Rather than considering, delete-1-point subsets of the sample in jackknife, Hastigon (1969), thought of using all the $2^n - 1$ nonempty subset of the data to improve the accuracy. But with the increased subsets, more computer power was required to use this method. Over next several years many more statistical methods were introduced which had broader applications. One of this methods is bootstrap which was introduced by Efron. This method involves bootstrap samples obtained randomly from the original sample. Accuracy of the statistic T_n as we have explained earlier depends upon its distribution. Bootstrap can be used to estimate this distribution. Based on this estimated distribution, various other accuracy results such as standard error and bias can be obtained for the statistic T_n. Other applications of bootstrap consist of estimating confidence interval of unknown parameter θ. This approach was introduced by Efron in 1982. We will discuss all these approaches for i.i.d. case in details in **Chapter 1**. We will also try to provide intuitive idea of why these approaches are good and also mention asymptotic

properties when needed.

Other than the independent data set, two of the main applications of resampling methods are linear models and dependent data set. In 1974, Miller started working with the applications of jackknife method on linear models. He later showed the inconsistency of bias estimator for least square estimate. To compensate this lack of consistency, Hinkley(1977), and Wu(1986) worked with weighted jackknife estimators and showed their consistency for variance and bias estimates for least square estimates under certain conditions. Bootstrap methods using the structure of linear models were introduced by Efron in 1979. These estimates produced by jackknife and bootstrap method, seem to be robust in the following sense. LSE estimates are based on the equal variance assumption. But even if this assumptions are violated these resampling procedures produces fairly consistent estimates. Further study regarding this robustness were studied by many statisticians (Hinkley, 1977; Wu,1986; Shao and Wu, 1987; Shao, 1988a; Liu and Singh, 1992a). Another important application concerning linear models is the prediction problem. **Chapter 2** is entirely focused on the general linear model and the prediction problem. Proofs of the asymptotic results are skipped, but reader can refer to Shao (1988a, On resampling methods for variance and bias estimation in linear models) and Shao(1988c, Bootstrap variance and bias estimation in linear models).

In 1981, Singh gave an example for the inadequacy of Efron's bootstrap method if there is any dependency in the data. Since then many have tried to modify bootstrap methods for dependent data set. **Chapter 3** describes some of these methods. We also discuss why Efron's bootstrap does not work for the dependent data by giving an example of m-dependent data set. Last part of chapter-3 introduces bootstrap method for autoregressive model for time series. Model structure of this particular time series is innovatively used to produce consistent bootstrap estimates. We provide result based on higher properties of this procedure studied by Bose (1988, Edgeworth correction by bootstrap in autoregressions). We have also tried working with some simulations to provide practical insight of the bootstrap method for the time series model.

Chapter 1

Resampling Methods

1.1 Standard Errors and Bias

Accuracy of the statistic can be evaluated by standard error, Bias, confidence interval etc. For example, mean of the population is generally estimated by sample average. Standard error of the sample average has closed form, which might or might not be available for other statistic. Resampling techniques such as Jackknife and Bootstrap can be used to estimate the variance or Bias of the statistic of interest without going into theoretic calculations involving underlying distribution.

1.1.1 Jackknife approach

Suppose we have sample $X = (X_1, X_2, \ldots, X_n)$ drawn from distribution F. Let θ be the parameter of interest. Suppose we estimate this θ with say $\hat{\theta} = S(X)$ [Some function of data X]. We wish to estimate the Standard error and the bias of $\hat{\theta}$.
Jackknife method works with leave one out samples. Consider $X_{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1} \ldots, X_n)$ for $i = 1, \ldots, n$. $X_{(i)}$ is known as i-th Jackknife sample which consists of all the sample points but the i-th one. Jackknife analogue of $\hat{\theta}$ is given by $\hat{\theta}_{(i)} = S(X_{(i)})$, for $i = 1, \ldots, n$.

Jackknife estimates for the Standard error and the bias of $\hat{\theta}$ are given by,

$$\widehat{\text{Se}}_{jack} = \left[\frac{n-1}{n}\sum_{i=1}^{n}\{\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\}^2\right]^{1/2}$$

where $\hat{\theta}_{(.)} = \sum_{i=1}^{n}\hat{\theta}_{(i)}/n$, and

$$\widehat{\text{Bias}}_{jack} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta})$$

respectively.

Jackknife estimate for the standard error looks like the standard deviation of the n jackknife analogues of $\hat{\theta}$ if we replace $(n-1)/n$ by $1/n$. In jackknife approach, for calculating jackknife replica of $\hat{\theta}$, only one point X_i is excluded from the original sample X. Thus jackknife replicas are usually numerically close to $\hat{\theta}$. As, $(n-1)/n$ is bigger than $1/n$, intuitively speaking, it compensates for this small difference between $\hat{\theta}$ and $\hat{\theta}_{(i)}$. Similar intuition can be given for the term $n-1$ in the jackknife estimate for the bias. To get the exact form of leading terms, consider $\hat{\theta} = \bar{X}$. Jackknife estimate for the standard error for this statistics turns out to be,

$$\widehat{\text{Se}}_{jack} = \Big[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2)}{n(n-1)}\Big]^{1/2}$$

To make this estimate approximately equal to the standard unbiased estimate of the sample mean, $\left[\frac{n-1}{n}\right]^{1/2}$ needs to be multiplied. Similarly, consider Standard error as a statistics and try to find jackknife estimate for its bias. In this case, leading term turns out to be $(n-1)$. Hence it is arbitrary convention to use these factors in the jackknife estimates for standard error and bias.

1.1.2 Bootstrap Approach

Let $X = (X_1, X_2, ..., X_n)$ be the random sample chosen from some unknown distribution F. Let $\hat{F}$ be the corresponding empirical distribution which puts mass $1/n$ on each of the X_i and $\hat{\theta} = S(X)$ be the statistics of interest.

Bootstrap sample X^* is defined as the random sample of size n drawn from the empirical distribution. Intuitively, it can be seen as a random sample of size n drawn from the data points $X_1 = x_1, \ldots, X_n = x_n$ with replacement. Bootstrap replication of $\hat{\theta}$ is defined as

$\hat{\theta^*} = S(X^*)$. Ideal Bootstrap estimate of standard error of $\hat{\theta}$ is defined as

$$Se_{\hat{F}}(\hat{\theta^*}) = \left[E_*(\hat{\theta^*} - E_*(\hat{\theta^*}))^2\right]^{1/2}$$

Where E_* denotes that the expectation needed to calculate standard error are taken with respect to the empirical distribution $\hat{F}$.
For most of the statistics other than the sample mean, there is no closed form for the ideal bootstrap estimate. Bootstrap algorithm explained below tries to approximate the ideal bootstrap estimate for the standard error.

Bootstrap Algorithm:

- Select B independent bootstrap samples $X^{*1}, \ldots, X^{*B}$ from the empirical distribution as stated earlier. For good estimate $B = 25$ to 100 is sufficient.
- Calculate corresponding bootstrap replication: $\hat{\theta}^*(b) = S(X^{*b})$ for each $b = 1, \ldots, B$.
- Then estimate the standard error of $\hat{\theta}$ by following ,

$$\widehat{SE}_B = \{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2/(B-1)\}^{1/2}$$

  Where $\hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$

As B goes to infinity, Law of Large Numbers suggests that $\widehat{SE}_B$ will be exactly equal to $Se_{\hat{F}}(\hat{\theta^*})$.

Bootstrap can also be used to approximate the Bias of any estimator $\hat{\theta} = S(X)$ of $\theta = t(F)$. Bias of $\hat{\theta} = S(X)$ is defined as

$$Bias(\hat{\theta}) = E(S(X)) - t(F)$$

Let $t(\hat{F})$ be the Plug-in estimate for θ. The bootstrap estimate for the bias for $\hat{\theta}$ is defined as,

$$Bias_{\hat{F}} = E_{\hat{F}}(S(X)) - t(\hat{F})$$

Thus, $Bias_{\hat{F}}$ is just a plug-in estimate for $Bias(\hat{\theta})$. Note that $\hat{F}$ denotes that calculation required for expectation of $S(X)$ are taken with respect to emperical distribution.

Similar to the standard error, there does not exist closed form for this bootstrap bias estimate for most of the complicated statistics. Hence, Bootstrap algorithm approximates the ideal bootstrap estimate for bias.

Bootstrap Algorithm:

- Generate B bootstrap samples namely $X^{*1}, \ldots, X^{*B}$ from emperical distribution corresponding to sample X.
- Then corresponding to each bootstrap sample X^{*b}, evaluate bootstrap replication of the statistic $\hat{\theta}$, which we will denote as , $\hat{\theta}^*(b)$
- Approximation to the ideal bootstrap estimate for bias of $\hat{\theta}$ is given by ,

$$\widehat{\text{Bias}_B} = \hat{\theta}^*(.) - t(\hat{F})$$

where $\hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$.

As B goes to infinity, Law of Large Numbers suggests that $\widehat{\text{Bias}}$ will be exactly equal to $Bias_{\hat{F}}$.

When $\hat{\theta} = S(X) = t(\hat{F})$ i.e. the plug in estimate of $\theta = t(F)$, the bias can be approximated by a better estimate. Let $X^* = (X_1^*, \ldots, X_n^*)$ be the bootstrap sample same as earlier. Let,

$$P_k^* = \#\{X_i^* = X_k\}/n$$

P_k can be thought of as the number in how much proportion X_k appears in the bootstrap sample X^*. Then Resampling vector P^* is defined as,

$$P^* = (P_1^*, \ldots, P_n^*)$$

It is easy to see that sum of all components of P adds upto 1. Let $P^0 = (1/n, \ldots, 1/n)$. Hence P_0 corresponds to the original sample X(Basically bootstrap sample in this case is the original sample itself).
As we are dealing with the specific case of $\hat{\theta}$ being $t(\hat{F})$, the bootstrap replication $\hat{\theta}^* = S(X^*)$ then can be thought of a function of P. Let $\hat{\theta}^* = T(P^*)$. Here, $T(P^0)$ will be exactly equal to $\hat{\theta}$.

Let $X_1^*, \ldots, X_B^*$ be the B bootstrap samples drawn from original sample X. Corresponding to each bootstrap sample X_b^* we have resampling vector P_b^*. Define $\bar{P}^*$ as

$$\bar{P}^* = \sum_{b=1}^{B} P_b^*/B$$

Then the better bootstrap estimate of the Bias denoted by $\overline{Bias_B}$ can be given by

$$\overline{Bias_B} = \hat{\theta}^*(.) - T(\bar{P}^*)$$

Where $\hat{\theta}^*(.)$ is the mean of bootstrap replications corresponding to X_b^*. Note that the bootstrap estimate we defined earlier is,

$$\widehat{\text{Bias}_B} = \hat{\theta}^*(.) - T(P^0)$$

It seems that both $\widehat{\text{Bias}_B}$ and $\overline{Bias_B}$ converges to $Bias_{\hat{F}}$. But convergence rate of $\overline{Bias_B}$ is much faster than $\widehat{\text{Bias}_B}$. Hence $\overline{Bias_B}$ is a better bootstrap estimate than $\widehat{\text{Bias}_B}$ in this particular case of plug in estimate.

1.2 Bootstrap Distribution Estimator

Let $X_1, \ldots, X_n$ be the random variables from some model P_n. And our statistic of interest is $T_n = T_n(X_1, \ldots, X_n, P_n)$ which is dependent upon these variables and model P_n. Let H_{P_n} be the sampling distribution of T_n with respect to the model P_n. That means

$$H_{P_n}(x) = P(T_n \leq x | P_n)$$

Sometimes model P_n is determined by distribution F and in that case we use $H_{n,F}$ or H_n instead of H_{P_n} as our notation for the sampling distribution. Suppose based on the data $X_1, \ldots, X_n$ we estimate P_n. Let $\hat{P}_n$ be this estimator(In general this is going to be emperical distribution). Then, $X_1^*, \ldots, X_n^*$ will be bootstrap sample drawn randomly from $\hat{P}_n$. And $T_n^* = T_n(X_1^*, \ldots, X_n^*, \hat{P}_n)$ will be the corresponding bootstrap replication of T_n. Let P_* denote the conditional probability $P(.|\hat{P}_n)$, then the estimator for the sampling distribution

of T_n obtained by bootstrap approach is given by

$$H_{Boot}(x) = P_*(T_n^* \leq x)$$

Definition 1.2.1. *Suppose ρ is any metric on $\mathbb{F} = \{$all distribution on $\mathbb{R}\}$. Then H_{Boot} is said to be weakly ρ-consistent if $\rho(H_{Boot}, H_{P_n}) \to_p 0$ as $n \to \infty$. And H_{Boot} is said to be strongly ρ-consistent if $\rho(H_{Boot}, H_{P_n}) \to_{a.s.} 0$ as $n \to \infty$.*

Metric ρ_∞ which is sup-norm metric is often used in such consistency problems. One of the techniques that can be used in proving consistencies of bootstrap distributions has been discussed later in this section. One way to prove consistency is to show that limiting distributions of T_n and T_n^* are the same. Berry-Esseen inequality is one way to prove this. Berry-Esseen inequality has following form: Suppose H is some distribution on $\mathbb{R}$ and

$$\rho_\infty(H_{P_n}, H) \leq cB_n(P_n)$$

here c is some constant independent of n and B_n is a function on P_n which satisfies $B_n(P_n) \to 0$ as $n \to \infty$. Now if estimator $\hat{P}_n$ of P_n satisfies this inequality with distribution $\hat{H}$ i.e.

$$\rho_\infty(H_{Boot}, \hat{H}) \leq cB_n(\hat{P}_n)$$

Then Singh(1981) used following technique to show the consistency of bootstrap distribution. Note that

$$\rho_\infty(H_{Boot}, H_{P_n}) \leq \rho_\infty(H_{Boot}, \hat{H}) + \rho_\infty(\hat{H}, H) + \rho_\infty((H_{P_n}, H)$$

So consistency of H_{Boot} follows if we show that $B_n(\hat{P}_n) \to_{a.s} 0$ and $\rho_\infty(\hat{H}, H) \to_{a.s} 0$. We will illustrate this technique by considering statistic $T_n = \sqrt{n}(\bar{X}_n - \mu)$, where $\bar{X}$ is sample mean and $\mu = E(X_1)$. Then bootstrap analog of T_n is $T_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X})$, where $\bar{X}_n^*$ is the mean of bootstrap sample obtained from emperical distribution based on $X_1, \ldots, X_n$. Let $Var(X_1) = \sigma^2$. If we apply Berry-Esseen's theorem on i.i.d $X_1^* \ldots, X_n^*$, we get,

$$\underset{x}{Sup}\, |P_*(T_n^*/\hat{S}_n \leq x) - \phi(x)| \leq 2.75\hat{\Delta}_n$$

Where $\hat{S}_n = E_*(X_1^* - \bar{X}_n)^2$ (E_* denotes expectation taken with respect to emperical distribution on $X_1, \ldots, X_n$), $\hat{\Delta}_n = E_*(|X_1^* - \bar{X}_n|)^3/({\hat{S}_n}^3\sqrt{n})$ and $\phi(x)$ denotes distribution function

of standard normal distribution. Hence,

$$\underset{x}{Sup}\,|P_*(T_n^*/\hat{S}_n \leq x) - \phi(x)| \leq 2.75n^{-3/2}\hat{S}_n{}^{-3}\sum_{i=1}^{n}(|X_i - \bar{X}_n|^3)$$

This can be rewritten as,

$$\underset{x}{Sup}\,|H_{Boot}(x) - \phi(x/\hat{S}_n)| \leq 2.75n^{-3/2}\hat{S}_n{}^{-3}\sum_{i=1}^{n}(|X_i - \bar{X}_n|^3)$$

By Marcinkiewicz strong law of large numbers we get that,

$$n^{-3/2}\sum_{i=1}^{n}(|X_i - -\bar{X}_n|^3) \rightarrow_{a.s} 0$$

Now by SLLN, $\hat{S}_n{}^2 \rightarrow_{a.s} \sigma^2$. Hence, we have

$$\underset{x}{Sup}\,|H_{Boot}(x) - \phi(x/\hat{S}_n)| \rightarrow_{a.s} 0$$

Also note that, $\rho_\infty(\phi(x/\hat{S}_n), \phi(x/\sigma)) \rightarrow_{a.s} 0$ and by cental limit theorem, $\rho_\infty(H_n, \phi(x/\sigma)) \rightarrow_{a.s} 0$. Thus, we finally have

$$\rho_\infty(H_{Boot}, H_n) \rightarrow_{a.s} 0$$

1.3 Confidence Intervals

Construction of the confidence interval and the hypothesis testing are two other major parts of statistical inference. Here we will give brief introduction to confidence intervals.
Let $X_1, \ldots, X_n$ be the i.i.d. random variables from some distribution F. Suppose I_n is the function of these random variables (i.e. some subset of R) and if it follows that

$$P(\theta \in I_n) = \alpha$$

then I_n is said to be the confidence set with confidence coefficient α. If I_n turns out to be an interval then it is known as the confidence interval. Statistical inference mainly focuses on One sided intervals $(-\infty, \bar{\theta}]$ and $[\underline{\theta}, \infty)$ or two sided interval $[\underline{\theta}, \bar{\theta}]$. In this context, $\bar{\theta}$ will be

referred as the upper bound and $\underline{\theta}$ as the lower bound. Since there is structural similarities between these intervals, we will focus on the lower confidence bound. If the Desired level $1-\alpha$ or $1-2\alpha$ matches with the probability of θ being in the confidence interval then that interval is known as the exact confidence interval. In standard approach of constructing confidence interval we consider pivotal quantity, say R_n, whose distribution (say G_n) does not depend on F. But it is usually hard to estimate the distribution of R_n and is usually estimated by some distribution. In bootstrap approach we try to estimate the distribution G_n. Let G_{Boot} be the estimation of G_n. But as this is simply an approximation coverage probability of the resultant confidence interval would be different than the desired level. Bootstrap confidence interval seems to be easy to construct and they do have higher accuracy in the sense of coverage probability. In this section we will discuss different bootstrap approach to construct confidence interval. As stated earlier we will stick to the lower confidence bound.

1.3.1 The bootstrap-t interval

Let $X_1^*, \ldots, X_n^*$ be the bootstrap sample from emperical distribution $\hat{F}$ on random sample $X_1, \ldots, X_n$ drawn from F. In this bootstrap approach, we look at pivotal quantity $R_n = \frac{\hat{\theta_n} - \theta}{\hat{\sigma}_n}$. Here, $\hat{\theta}_n$ is estimator of θ and $\hat{\sigma}_n^2$ is variance estimator of $\hat{\theta}_n$. Distribution of R_n (i.e. G_n) is unknown in most of the cases hence we approximate it by bootstrap estimator G_{Boot}. G_{Boot} is defined as follows,

$$G_{Boot}(x) = P_*(\frac{\hat{\theta^*}_n - \hat{\theta}_n}{\hat{\sigma}_n^*} \leq x)$$

$\hat{\theta^*}_n$ and $\hat{\sigma}_n^*$ are bootstrap replications of $\hat{\theta}_n$ and $\hat{\sigma}_n$, respectively. Definition of P_* remains same as in previous section. In this case, if we replace G_n in the confidence interval by G_{Boot}, we get following lower confidence bound

$$\underline{\theta}_{BT} = \hat{\theta}_n - \hat{\sigma}_n G_{Boot}^{-1}(1-\alpha)$$

This will be known as Bootstrap-t lower confidence bound.
The G_{Boot} defined earlier can be used if we are able to find the distribution of $R_n^* = \frac{\hat{\theta}_n^* - \hat{\theta_n}}{\hat{\sigma}_n^*}$. But in practice, we will have to approximate it based on data at hand. Based on independent bootstrap samples $X_1^*, \ldots, X_B^*$ drawn from $\hat{F}$, we will have bootstrap replications for $R^*{}_n$,

say ${R^*}_n(b)$. Then αth percentile for G_{Boot} can be approximated by $\hat{t}^{(\alpha)}$ such that,

$$\#\{{R^*}_n(b) \leq \hat{t}^{(\alpha)}\}/B = \alpha$$

That is, if 1000 bootstrap samples have been generated then 50th value of ordered list of ${R^*}_n(b)$ is the estimate for 0.05th quantile.If $B \cdot \alpha$ is not an integer then estimate the αth quantile is defined by kth largest values of ${R^*}_n(b)$, respectively. Here, $k = [(B+1)\alpha]$ where [ ] denotes the greatest integer function.

Disadvantages:

1. Note that denominator of ${R^*}_n(b)$ is standard error of $\hat{\theta}^*$ for the b-th bootstrap sample. Which means if the statistic is too complicated to have a closed form for standard error, we will have to use another sequence of bootstrap on bootstrap sample B to calculate standard error. This is computationally expensive.

2. Sometimes it is easier to transform the statistic to calculate confidence interval and then transform it back to get the interval for the original statistic. Bootstrap t-method highly depends on which scale confidence interval is being constructed. In other words it is not transformation respecting.

1.3.2 Percentile Intervals

Rather than looking at the distribution of some pivotal quantity, this apporaoch directly looks at the distribution of $\hat{\theta}_n$. Distribution of $\hat{\theta}_n$ is estimated by distribution of its bootstrap analog.

$$K_{Boot}(x) = P_*(\hat{\theta}^*_n \leq x)$$

Based on this estimate, bootstrap percentile method provides following lower bound:

$$\underline{\theta}_{BP} = K^{-1}_{Boot}(\alpha)$$

Unlike bootstrap-t interval approach this approach is invariant with respect to any reparametrization.

In practice, approximation to this estimate can be done as follows: Draw B independent bootstrap samples $x^{*1}, \ldots, x^{*B}$. Corresponding bootstrap replications for $\hat{\theta}$ will be $\hat{\theta}^*(b)$. Then we approximate $K^{-1}_{Boot}(\alpha)$ by the $B.\alpha$the value of ordered list of above bootstrap replications. For example, if $B = 1000$ then $K^{-1}_{Boot}(0.05)$ is approximated by 50th value of ordered list of bootstrap replications. If $B \cdot \alpha$ is not an integer then in that case approach is similar to that of previous section.

Suppose that there exist an increasing function $\phi_n(x)$ such that

$$P(\hat{\phi}_n - \phi_n(\theta) \leq x) = \psi(x) \tag{1.1}$$

is being satisfied by all distributions F including $\hat{F}$, where $\hat{\phi}_n = \phi_n(\hat{\theta})$ and ψ is some increasing, continuous and symmetric distribution. If we know the transformation ϕ_n and distribution ψ, then exact lower confidence bound (say $\underline{\theta}_{Exact}$) for θ can be given as

$$\underline{\theta}_{Exact} = \phi_n^{-1}(\hat{\phi} + z_\alpha) \tag{1.2}$$

Here $z_\alpha = \psi^{-1}(\alpha)$.
Let $w_n = \phi_n(\underline{\theta}_{BP}) + \hat{\psi}$. As stated earlier if we assume (1.1) holds in the case of $\hat{F}$ and as ψ is increasing, we get

$$\psi(w_n) = \psi(\phi_n(\underline{\theta}_{BP}) + \hat{\psi}) = P_*(\hat{\theta}^*_n \leq \underline{\theta}_{BP})$$

Now according to definition of $\underline{\theta}_{BP}$,

$$P_*(\hat{\theta}^*_n \leq \underline{\theta}_{BP}) = \alpha$$

Hence, $w_n = \psi^{-1}(\alpha) = z_\alpha$. Using this we can rewrite $\underline{\theta}_{BP}$ as

$$\underline{\theta}_{BP} = \phi_n^{-1}(\hat{\phi} + z_\alpha) = \underline{\theta}_{Exact}$$

This shows that under the assumption (1.1) $\underline{\theta}_{BP} = \underline{\theta}_{Exact}$. So even if $\hat{\phi}$ and ψ are unknown we can still find exact lower confidence bound using $\underline{\theta}_{BP}$. Sometimes assumption (1.1) holds if n is very large. So for large n, $\underline{\theta}_{BP}$ will be a good approximation to exact lower confidence bound. Frequently, $\hat{\phi}_n$ is not a linear function and hence the bias of $\hat{\phi}_n - \phi(\theta)$ does not vanish rapidly as n becomes larger and larger. So approximation (1.1) will be good enough only if n is very large. Efron suggested some relaxation on assumption (1.1), which lead to

the development of the following two bootstrap methods.

1.3.3 The bootstrap bias-corrected percentile

Discussion of Bias not converging fast enough to 0 suggests that we should add some bias term in the assumption (1.1). Consider the example of bivariate normal distribution F. Let θ be the correlation coefficient and $\hat{\theta}_n$ be sample based correlation coefficient. For this setup, $\sqrt{n}(\tanh^{-1}(\hat{\theta}_n) - \tanh^{-1}(\theta)) - \theta/(2\sqrt{n})$ seems to be better approximation to the standard normal distribution than $\sqrt{n}(\tanh^{-1}(\hat{\theta}_n) - \tanh^{-1}(\theta)$. In the view of such scenarios, Efron suggested more general assumption than that of (1.1):

$$P(\hat{\phi}_n - \phi_n(\theta) + z_0 \leq x) = \psi(x) \tag{1.3}$$

Where notations $\hat{\phi}_n$, ϕ, ψ mean exactly the same as that of in the previous subsection. Considering this assumption, $\underline{\theta}_{Exact}$ can be written as,

$$\underline{\theta}_{Exact} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha + z_0)$$

As, (1.3) is valid for $\hat{F}$ also we get,

$$K_{Boot}(\hat{\theta}_n) = P_*(\hat{\theta}_n^* \leq \hat{\theta}_n) = P_*(\hat{\phi}_n^* - hat\phi_n + z_o \leq z_0) = \psi(z_0)$$

Hence,

$$\psi^{-1}(K_{Boot}(\hat{\theta}_n)) = z_0$$

As ψ is symmetric and due to the form of the exact lower bound based on (1.3) we get,

$$\begin{aligned}
1 - \alpha &= \psi(-z_\alpha) \\
&= \psi(\hat{\phi}_n - \phi_n(\underline{\theta}_{Exact}) + z_0) \\
&= P_*(\hat{\phi}_n^* - \hat{\phi}_n \leq \hat{\phi}_n - \phi_n(\underline{\theta}_{Exact})) \\
&= P_*(\hat{\phi}_n^* \leq 2\hat{\phi}_n - \phi_n(\underline{\theta}_{Exact})) \\
&= P_*(\hat{\phi}_n^* \leq \hat{\phi}_n - z_\alpha - z_0) \\
&= P_*(\hat{\phi}_n^* \leq \hat{\phi}_n^{-1}(\hat{\phi}_n - z_\alpha - z_0)) \\
&= K_{Boot}(\hat{\phi}_n^{-1}(\hat{\phi}_n - z_\alpha - z_0))
\end{aligned}$$

This implies,

$$K_{Boot}^{-1}(1-\alpha) = \hat{\phi}_n^{-1}(\hat{\phi}_n - z_\alpha - z_0)$$

As this is true for any α and also ψ is symmetric function, for any $0 < x < 1$ we get,

$$K_{Boot}^{-1}(x) = \hat{\phi}_n^{-1}(\hat{\phi}_n + \psi^{-1}(x) - z_0)$$

By the expression of $\underline{\theta}_{Exact}$ and above equation of $K_{Boot}^{-1}(x)$ we finally get,

$$\underline{\theta}_{Exact} = K_{Boot}^{-1}(\psi(z_\alpha + 2z_0)) = K_{Boot}^{-1}(\psi(z_\alpha + 2\psi^{-1}(K_{Boot}(\hat{\theta}_n))))$$

Hence if ψ is known, then the bootstrap bias-corrected percentile lower confidence bound for θ is given by,

$$\underline{\theta}_{BC} = K_{Boot}^{-1}(\psi(z_\alpha + 2z_0)) = K_{Boot}^{-1}(\psi(z_\alpha + 2\psi^{-1}(K_{Boot}(\hat{\theta}_n))))$$

As bias term is included in the assumption (1.3), this should improve bootstrap percentile approach. If (1.3) is exact then $\underline{\theta}_{BC}$ is exact lower confidence bound. But if (1.3) holds for large n , then $\underline{\theta}_{BC}$ improves as n gets larger. Many examples still suggest that assumption (1.3) is not satisfied nicely , hence more comprehensive assumption was taken into account, which is explained in the subsequent section.

1.3.4 The bootstrap accelerated bias-corrected percentile

Schenker (1985) considered the example of normal distribution(denote it by F), where the parameter of interest was variance. He shown that even if distribution of sample mean was transformed into that of a normal distribution, transformed variable's variance is still dependent on F. This shows that it is hard to obtain stabilization normally and variance wise. Taking account of this example, Efron(1987) generalized assumption (1.3) even further. More general assumption is,

$$P(\frac{\hat{\phi}_n - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x) = \psi(x) \tag{1.4}$$

Here, notations are similar to that of previous section. The extra parameter a depends on the distribution F and n, and is known as acceleration constant as it measures the change

of standard deviation of $\hat{\phi}_n$ with respect to $\phi_n(\theta)$.

In case when ϕ, ψ, z_0 and a are known we can determine exact lower confidence bound for θ. It turns out to be,

$$\underline{\theta}_{Exact} = \phi_n^{-1}(\hat{\phi}_n + (z_\alpha + z_0)(1 + a\hat{\phi}_n)/[1 - a(z_\alpha + z_0)])$$

Similar to previous section we will show that $\underline{\theta}_{Exact}$ is equal to some percentile of K_{Boot}. We can note that calculation upto the part where we showed $\psi^{-1}(K_{Boot}(\hat{\theta}_n)) = z_0$ still holds. However change in assumption gives us

$$K_{Boot}^{-1}(x) = \hat{\phi}_n^{-1}(\hat{\phi}_n + [\psi^{-1}(x) - z_0](1 + a\hat{\phi}_n))$$

For $0 < x < 1$. Now let $x = \psi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)]))$ we get,

$$K_{Boot}^{-1}(\psi(z_0+(z_\alpha+z_0)/[1-a(z_\alpha+z_0)]))) = \phi_n^{-1}(\hat{\phi}_n+(z_\alpha+z_0)(1+a\hat{\phi}_n)/[1-a(z_\alpha+z_0)]) = \underline{\theta}_{Exact}$$

Hence if ψ and a are known , the lower confidence bound for θ obtained based on assumption (1.4) is given by

$$\underline{\theta}_{BC_a} = K_{Boot}^{-1}(\psi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)])))$$

This is known as the bootstrap accelerated bias-corrected percentile lower confidence bound. If a is unknown then we can estimate it by $\hat{a}$, then the resulting confidence bound is obtained by substituting $\hat{a}$ instead of a in above equation.

$$\underline{\theta}_{BC_a} = \underline{\theta}_{BC_a}(\hat{a})$$

But it should be known that parameter a is hard to estimate in some cases.

1.3.5 Consistency of Bootstrap confidence intervals

In bootstrap approach we approximate distribution of the pivotal quantity with its bootstrap distribution. Also, latter techniques are based upon approximation of the distribution of the transformation of a parameter. hence we would like to know that whether coverage probability converges to the desired level as $n \to \infty$. Consistency of confidence sets are defined in the following way.

Definition 1.3.1. *A confidence set I_n for θ with desired level $1-\alpha$ is said to be consistent if,*

$$P(\theta \in I_n) \to 1-\alpha$$

as $n \to \infty$.

Let H_n be the distribution of $n^t(\hat{\theta}_n - \theta)$, where t is a constant and G_n be the distribution of pivotal quantity $\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}$. Let H_{Boot} and G_{Boot} be their bootstrap estimators, respectively. Following theorem shows the consistency of bootstrap lower confidence bound.

Theorem 1.3.1. *1. If G_{Boot} is consistent in the sense of(Definition 1.2.1), then θ_{BT} is consistent.*

2. Suppose H_{Boot} is consistent and $\lim_{n\to\infty} \rho_\infty(H, H_n) = 0$ for a strictly increasing, continuous and symmetric H. Then $\underline{\theta}_{BP}$, $\underline{\theta}_{BC}$, and $\underline{\theta}_{BC_a}$ are consistent lower confidence bound.

Proof. 1. As G_{Boot} is consistent, $\rho_\infty(G_{Boot}, G_n) \to 0$ as $n \to \infty$. Also note that functional $T(G) = G^{-1}(1-\alpha)$ is continuous with respect to ρ_∞. This proves the consistency of θ_{BT}.

2. By definition of $\underline{\theta}_{BP}$,

$$\begin{aligned} P(\underline{\theta}_{BP} \le \theta) &= P(\alpha \le K_{Boot}(\theta)) \\ &= P(\alpha \le P_*(\hat{\theta}_n^* - \hat{\theta}_n \le \theta - \hat{\theta}_n)) \\ &= P(\alpha \le H_{Boot}(n^t(\theta - \hat{\theta}_n))) \\ &= P(n^t(\theta - \hat{\theta}_n) \ge H_{Boot}^{-1}(\alpha)) \end{aligned} \tag{1.5}$$

Now as H_{Boot} is consistent and $\lim_{n\to\infty} \rho_\infty(H, H_n) = 0$ we get,

$$\begin{aligned} P(\underline{\theta}_{BP} \le \theta) &= P(n^t(\hat{\theta}_n - \theta) \le -H^{-1}(\alpha)) + o(1) \\ &= H(-H^{-1}(\alpha)) + o(1) \end{aligned} \tag{1.6}$$

As H is a symmetric function, we have $H(-H^{-1}(\alpha)) = 1-\alpha$. Hence, $P(\underline{\theta}_{BP} \le \theta) \to 1-\alpha$ as $n \to \infty$.

Note that, in case of $\underline{\theta}_{BC}$, and $\underline{\theta}_{BC_a}$ we had,

$$z_0 = \psi^{-1}(K_{Boot}(\hat{\theta}_n)) = \psi^{-1}(P_*(\hat{\theta}_n^* \leq \hat{\theta}_n)) = \psi^{-1}(P_*(\hat{\theta}_n^* - \hat{\theta}_n \leq 0)) = \psi^{-1}(H_{Boot}(0))$$

Consistency of H_{Boot} and $\lim_{n\to\infty} \rho_\infty(H, H_n) = 0$ gives us,

$$z_0 \to_p \psi^{-1}(H(0)) = 0$$

This implies that asymptotically we will have $\psi(z_\alpha)$ instead of α in above equations in case of $\underline{\theta}_{BC}$, and $\underline{\theta}_{BC_a}$. But $\psi(z_\alpha) = \alpha$. Hence all of the above equation goes through in these cases also. Hence we have consistency for $\underline{\theta}_{BP}$, $\underline{\theta}_{BC}$, and $\underline{\theta}_{BC_a}$.

□

This theorem basically implies that all the bootstrap methods produces consistent lower confidence bounds. Hence larger the n closer we will get to the desired level of confidence set.

Chapter 2

Linear Models

2.1 Introduction

Bootstrap methods can be applied to various problems. It is kind of model dependent method in terms of generation of the bootstrap data which depends on specific model we are dealing with. Generally bootstrap approaches can be applied to any model P where it can be estimated by say $\hat{P}$. Performance of these approaches are dependent upon how good estimate $\hat{P}$ is and how the bootstrap data has been generated.
In this chapter, we will be dealing with the application of the bootstrap and the jackknife to linear models. General linear model is given by:

$$y_i = x_i'\beta + \epsilon_i \text{ for } i = 1, \ldots, n. \tag{2.1}$$

Here y_i is i-th response, x_i is a p-vectors of variables associated with response variable, a p-vector of unknown parameters is denoted by β, and ϵ_i are random errors. In deterministic model, ϵ_i are independent and have mean , variance 0 and σ^2 respectively. While in correlation model, (y_i, x_i) are assumed to be independent and identically distributed with $E(y_i|x_i) = x_i'\beta$ and finite second order moment. We denote $E(y_i^2|x_i)$ by σ_i^2. For simplicity, from now onwards we will denote this conditional expectation and conditional variance by E and Var respectively for the rest of the chapter.

Most of the time it is assumed that ϵ_i have same variance, $\sigma_i^2 = \sigma^2$. Under this assumption,

usually β is estimated by least square estimate $\hat{\beta}_{LS}$ given by,

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y \tag{2.2}$$

Where X is $n \times p$ matrix with x_i as its i-th row and Y is $n \times p$ matrix with y_i as i-th row. It is assumed that X has rank p for simplicity and which assures the existence of inverse of the matrix $X'X$. it is easy to note that $\hat{\beta}_{LS}$ is an unbiased estimator as $E(\hat{\beta}_{LS}) = \beta$.
If $\sigma_i^2 = \sigma^2$, then

$$Var(\hat{\beta}_{LS}) = Var((X'X)^{-1}X'Y) = (X'X)^{-1}X'Var(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{2.3}$$

Unbiased estimator for this variance is usually given by,

$$(\frac{1}{n-p}\sum_{i=1}^{n} r_i^2)(X'X)^{-1}$$

where i-th residual r_i is given by $(y_i - x_i'\hat{\beta}_{LS})$. But like most of the problems, variance of other estimates of β does not have closed form and hence we approximate them by asymptotic approaches.

2.2 Variance and Bias Estimation

Let $\theta = g(\beta)$ and is estimated by $\hat{\theta} = g(\hat{\beta})$, where $\hat{\beta}$ is an estimator of β. In this section, estimators for variance and bias of $\hat{\theta}$ using the bootstrap and the jackknife approach have been discussed. For simplicity we will stick to the least square estimate of β.

2.2.1 Jackknife Approach

Miller extended the jackknife method for the estimation of bias and variance of the statistic $\hat{\theta}_{LS} = g(\hat{\beta}_{LS})$. Extension is as follows: Suppose we have data (y_i, x_i), $i = 1, \ldots, n$. Then delete the i-th pair. Now consider the rest of the data and try to find $\hat{\beta}_{LS}$ based on remaining data points. Let us denote least square estimate we get after deleting i-th pair by $\hat{\beta}_{LS,i}$. Then jackknife replication of $\hat{\theta}$ would be $\hat{\theta}_{LS,i} = g(\hat{\beta}_{LS,i})$. Then bias and variance estimates are

given by:

$$b_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{LS,i} - \hat{\theta}_{LS}) \tag{2.4}$$

And

$$v_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{LS,i} - \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{LS,i})^2 \tag{2.5}$$

respectively. Note that $\hat{\beta}_{LS,i}$ can be written as,

$$\hat{\beta}_{LS,i} = (X'X - x_i x_i')^{-1}(X'Y - x_i y_i)$$

For vector c and matrix A we can show that inverse of $(A - cc')$ is

$$(A - cc')^{-1} = (A^{-1} + \frac{A^{-1}cc'A^{-1}}{1 - c'A^{-1}c})$$

Using this, we get,

$$\begin{aligned}
\hat{\beta}_{LS,i} &= \left((X'X)^{-1} + \frac{X'X)^{-1}x_i x_i' X'X)^{-1}}{1 - x_i' X'X)^{-1}x_i}\right)(X'Y - x_i y_i) \\
&= \hat{\beta}_{LS} - \left[z_i y_i - \frac{z_i x_i' \hat{\beta}_{LS}}{1 - h_i} + \frac{z_i h_i y_i}{1 - h_i}\right] \\
&= \hat{\beta}_{LS} - \frac{z_i r_i}{1 - h_i}
\end{aligned}$$

where $h_i = x_i'(X'X)^{-1}x_i$, $z_i = (X'X)^{-1}x_i$ and $r_i = y_i - x_i\hat{\beta}_{LS}$. In case when, $\theta = g(\beta) = c'\beta$ then we have $\hat{\theta}_{LS} = g(\hat{\beta}_{LS}) = c'\hat{\beta}_{LS}$. So,

$$b_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}\frac{c'z_i r_i}{1 - h_i}$$

Note that $\sum_{i=1}^{n} z_i r_i = 0$. So we can write b_{Jack} as,

$$b_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\frac{c'z_i r_i}{1 - h_i} - c'z_i r_i) = \frac{n-1}{n}\sum_{i=1}^{n}\frac{c'z_i h_i r_i}{1 - h_i}$$

We already know that $\hat{\beta}_{LS}$ is an unbiased estimator of β. So we expect $\hat{\theta}_{LS}$ to be unbiased too. But from above equation we can deduce that b_{Jack} is not 0 if h_i's are not same for all i.

And hence it may not converge to 0 rapidly. Similarly, for the same $\theta = g(\beta) = c'\beta$, we have

$$v_{Jack} = \frac{n-1}{n}\sum_{i=1}^{n}\frac{(c'z_i)^2 r_i^2}{(1-h_i)^2} - \frac{n-1}{n^2}(\sum_{i=1}^{n}\frac{c'z_i r_i}{1-h_i})^2$$

Miller(1974) proved that under weak conditions, v_{Jack} is a consistent estimator for variance. He also showed that b_{Jack} may be inconsistent bias estimator (Which intuitively matches with our discussion of h_i not being same for all i).

Inconsistency of jackknife bias estimator can be explained as follows. Under linear model, we assumed that $E(y_i) = x_i'\beta$. Which may not be same for all i. This makes our linear model kind of unbalanced in nature. Considering this unbalanced nature other jackknife estimators were proposed. These are known as weighted jackknife estimators. We will list some of them here.
Hinkley defined pseudovalues as follows:

$$\tilde{\theta}_i = \hat{\theta}_{LS} + n(1-h_i)(\hat{\theta}_{LS} - \hat{\theta}_{LS,i}) \tag{2.6}$$

Based on these weighted pseudovalues, he gave following weighted jackknife variance and bias estimator.

$$v_{HJack} = \frac{1}{n(n-p)}\sum_{i=1}^{n}(\tilde{\theta}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i)^2$$

and

$$b_{HJack} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{LS} - \tilde{\theta}_i)$$

If we put value of $\tilde{\theta}_i$ in b_{HJack} it can be easily shown that $b_{HJack} = 0$. Hence this b_{HJack} is correct bias estimator.
Wu(1986) gave different weighted jackknife estimators for bias and variance. They are as follows:

$$v_{WJack} = \sum_{i=1}^{n}(1-h_i)(\hat{\theta}_{LS,i} - \hat{\theta}_{LS})^2$$

and,

$$b_{WJack} = \sum_{i=1}^{n}(1-h_i)(\hat{\theta}_{LS,i} - \hat{\theta}_{LS}).$$

All these variance estimators are consistent in the following sense: Let $\theta = g(\beta)$ and is

estimated by $\hat{\theta}_{LS} = g(\hat{\beta}_{LS})$. Moreover let us assume that g is continuously differentiable and $\nabla g(\beta)$ is not zero at β. Then

Theorem 2.2.1. *If* $X'X \to \infty$, $h_{max} = \underset{i \le n}{max}\ h_i \to 0$, *and* $lim\ \underset{n}{Sup}\ \underset{i \le n}{max} E|\epsilon_i|^{(2+\delta)} < \infty$ *for some* $\delta > 0$, *then*

$$v/\nabla g(\beta)' Var(\hat{\beta}_{LS}) \nabla g(\beta) \to_p 1$$

Here v denotes either jackknife estimator or both weighted jackknife estimator for $g(\hat{\beta}_{LS})$.

2.2.2 Bootstrap Approach

As we have seen earlier, bootstrap variance and bias are defined as $Var_*(\hat{\theta}^*)$ and $E_*(\hat{\theta}^*) - \hat{\theta}$, respectively, where $\hat{\theta}^*$ is bootstrap replica of $\hat{\theta}$. Here we will consider $\hat{\theta} = \hat{\theta}_{LS}$. For the linear model, we have to apply bootstrap method in a different way. There are three types of approaches in the literature:

Residual Based Bootstrap

In 1979, Efron proposed idea of residual based bootstrap (RB). General idea in bootstrap , as we described earlier, is to estimate the model by some parameter. Then estimate the parameter by some estimator based on data at hand to generate bootstrap data set. In this case when x_i are deterministic and errors are i.i.d., linear model can be identified with two parameters: β and unknown distribution F_ϵ (Distribution of ϵ_i). As β is estimated by least square estimate $\hat{\beta}_{LS}$, we only need to estimate F_ϵ. Let $r_i = y_i - x_i'\hat{\beta}_{LS}$ be the i-th residual. F_ϵ can be estimated by emperical distribution with respect to these centered residuals i.e. putting n^{-1} mass on each $r_i - \bar{r}$, where $\bar{r} = \sum_{i=1}^n r_i/n$. Let us denote this emperical distibution by $F_{\hat{\epsilon}}$. Hence for general linear model, (β, F_ϵ) is estimated by $(\hat{\beta}_{LS}, F_{\hat{\epsilon}})$. An easy way to generate bootstrap data from this estimated model is as follows. Draw i.i.d. random variables $\epsilon_1^*, \ldots, \epsilon_n^*$ from $F_{\hat{\epsilon}}$) . Let $y_i^* = x_i'\hat{\beta}_{LS} + \epsilon_i^*$ for $i = 1, \ldots, n$. Bootstrap data obtained is thus $(y_1^*, x_1'), \ldots, (y_n^*, x_n')$. Let $\hat{\beta}_{LS}^*$ be the least square estimator obtained using above bootstrap data. Thus,

$$\hat{\beta}_{LS}^* = (X'X)^{-1}X'Y^*$$

Where $Y^* = (y_1^*, \ldots, y_n^*)'$. Bootstrap replication of $\theta = g(\beta)$ will be given by $\hat{\theta}_{LS}^* = g(\hat{\beta}_{LS}^*)$. Note that, $E_*(y_*) = x_i'\hat{\beta}_{LS}$ as $E_*(\epsilon_i^*) = 0$. This implies,

$$E_*(\hat{\beta}_{LS}^*) = (X'X)^{-1}X'X\hat{\beta}_{LS} = \hat{\beta}_{LS}$$

and

$$Var_*(\hat{\beta}_{LS}^*) = (X'X)^{-1}X'Var_*(Y)X(X'X)^{-1} = [\frac{1}{n}\sum_{i=1}^{n}(r_i - \bar{r})^2](X'X)^{-1}$$

Here E_* denotes expectation with respect to emperical distribution $F_{\hat{\epsilon}}$. Shao(1988c) showed that the residual bootstrap procedure generates consistent estimator for bias and variance of $\hat{\beta}_{LS}$ and in turn, if g is smooth then some moment condition ensures consistency for $\hat{\theta}_{LS}$ also.

Paired Bootstrap

In case when x_is are random and data points (y_i, x_i') are identically distributed and independent, we can recognize the linear model by joint distribution of these data points. An obvious estimate to this model would be , emperical distribution based on $(y_1, x_1'), \ldots, (y_n, x_n')$. We can thus generate bootstrap data set from drawing random points from this emperical distribution. This bootstrap method is known as paired bootstrap(PB). Application of this bootstrap method is not confined to least square estimates.
In case of least square estimate $\hat{\beta}_{LS}$, notice that matrix X^*(replica of X) won't have rank p. Thus we have to derive the bootstrap replication of $\hat{\beta}_{LS}$ using the generalized inverse of the matrix $X^{*\prime}X^*$. Bootstrap replication of $\hat{\beta}_{LS}$ is defined as

$$\hat{\beta}_{LS}^* = \begin{cases} (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*, & \text{if } \lambda^* \geq \lambda/2 \\ \hat{\beta}_{LS}, & \text{otherwise} \end{cases} \tag{2.7}$$

where λ^* and λ are two smallest eigenvalues of $(X^{*\prime}X^*)$ and $(X'X)$, respectively. It turns out that approximately the expectation of $\hat{\beta}_{LS}^*$ is $\hat{\beta}_{LS}$ and variance of the same is same as Hinley's weighted variance estimator. Hence, under some regularity conditions, paired bootstrap produces consistent estimators for variance and bias of $\hat{\theta}_{LS}$.

External Bootstrap

Wu(1986) proposed this method specifically for the least square estimates. e_i for $i = 1, \ldots, n$ be the i.i.d random variables drawn from some distribution with expectation 0 and variance 1. Then he generated bootstrap data set by

$$y_i^* = x_i'\hat{\beta}_{LS} + \frac{|r_i|}{\sqrt{1-h_i}}e_i^*$$

here r_i and h_i hold same meaning as that of in jackknife section. Hence the bootstrap analog based on this bootstrap data set can be given by $(X'X)^{-1}X'Y^*$, where $Y^* = (y_1^*, \ldots, y_n^*)'$. It turns out that the variance estimator based on this method is same as that of Wu's weighted jackknife variance estimator. Thus external bootstrap also seems to produce consistent variance and bias estimator.

2.3 Prediction

General linear model can be applied to the prediction of future values y_f based on given value of x_f. Bootstrap method can be used to estimate the prediction error in the point prediction problem. For model (2.1) with i.i.d error terms, point prediction is given by

$$\hat{y}_f = x_f'\hat{\beta}_{LS}$$

In case of this point prediction, Mean squared prediction error is defined as,

$$mse(x_f) = E(y_f - \hat{y}_f)^2$$

Here expectation is taken over joint distribution of y_f and $\hat{y}_f$. Note that $mse(x_f)$ can also be written as $Var(y_f - \hat{y}_f)$. Now in general, y_f and $y_1, \ldots, y_n$ are independent. So y_f and $\hat{y}_f$ are independent. Thus we can write this mean squared prediction error as,

$$mse(x_f) = Var(y_f) + Var(\hat{y}_f) = \sigma^2 + x_f'Var(\hat{\beta}_{LS})x_f = \sigma^2 + \sigma^2 x_f'(X'X)^{-1}x_f$$

Residual bootstrap method can be applied to estimate this mean squared prediction error. Recall that $\hat{F}_\epsilon$ denotes the emperical distribution with respect to adjusted residuals. Let

$\epsilon_1^*, \ldots, \epsilon_n^*$ and ϵ_f^* be the i.i.d random variables drawn from the distribution $\hat{F}_\epsilon$. Bootstrap replication for the least square estimate be the same as that of in section (2.2)(Residual Bootstrap) and define the future value as $y_f^* = x_f'\hat{\beta}_{LS} + \epsilon_f^*$. Bootstrap prediction of this future value is given by $\hat{y}_f^* = x_f'\hat{\beta}_{LS}^*$. Then bootstrap estimate for $mse(x_f)$ is given by

$$\widehat{mse}_{Boot}(x_f) = E_*(y_f^* - \hat{y}_f^*)$$

As ϵ_f^* and all the other bootstrap errors are independent we have,

$$\begin{aligned} \widehat{mse}_{Boot}(x_f) &= Var_*(y_f^*) + Var_*(\hat{y}_f^*) \\ &= Var_*(y_f^*) + Var_*(x_f'\hat{\beta}_{LS}^*) \\ &= \hat{\sigma}^2 + \hat{\sigma}^2 x_f'(X'X)^{-1}x_f \end{aligned} \tag{2.8}$$

Here $\hat{\sigma}^2$ is similarly defined as that in residual bootstrap section. Notice that, $\hat{\sigma}^2$ is unbiased estimator for σ^2. Which implies $\widehat{mse}_{Boot}(x_f)$ is an unbiased estimator of $mse(x_f)$.
Suppose instead of one value we have to estimate mean squared error of prediction for set of values χ of x_f.And let ν be the probability measure on χ. In that case average mean squared error for predicting these values is

$$\int mse(x_f)\, \mathrm{d}\nu(x_f) = \sigma^2 + \sigma^2 \int x_f'(X'X)^{-1}x_f\, \mathrm{d}\nu(x_f)$$

Also, As being a scalar, $x_f'(X'X)^{-1}x_f = tr(x_f'(X'X)^{-1}x_f)$, where $tr(A)$ denotes trace of the matrix A. Hence we have,

$$E(x_f'(X'X)^{-1}x_f) = E(tr(x_f'(X'X)^{-1}x_f)) = E(tr((X'X)^{-1}x_f x_f')) = tr(E((X'X)^{-1}x_f x_f'))$$

Thus,

$$\int mse(x_f)\mathrm{d}\,\nu(x_f) = \sigma^2 + \sigma^2 tr((X'X)^{-1} \int x_f x_f'\, \mathrm{d}\nu(x_f)) = \sigma^2 + \sigma^2 tr((X'X)^{-1}\Sigma_x)$$

Here $\Sigma_x = \int x_f x_f'\, \mathrm{d}\nu(x_f)$.
It is easy to see that bootstrap estimator for average prediction error is,

$$\int \widehat{mse(x_f)}_{Boot}\mathrm{d}\,\nu(x_f) = \hat{\sigma}^2 + \hat{\sigma}^2 tr((X'X)^{-1}\Sigma_x)$$

Lets see an example. Suppose the set of values we want to predict are $x_1, \ldots, x_n$ itself. Then we provide probability measure on this set by putting mass n^{-1} on each one of them. Note that, $(X'X)^{-1} = \sum_{i=1}^{n} x_i x_i'$ and $\Sigma_x = \frac{\sum_{i=1}^{n} x_i x_i'}{n}$. Then in this case average prediction error and its bootstrap estimator are respectively given by,

$$\sigma^2 + \sigma^2 p/n \text{ and } \hat{\sigma}^2 + \hat{\sigma}^2 p/n$$

2.4 Model selection

In model (2.1), not every component x is related to the response value y. For example components those components of x won't be important for the given response if their corresponding components β_i in true parameter β are zero. In this case it is better to remodel the general linear model. Let α be any subset of set $\{1, \ldots, p\}$. Suppose $x_{i\alpha}$ and β_α be sub-vectors of x and β which considers components corresponding to the set α. Hence more dense model for the linear problem can be given by,

$$y_i = x_{i\alpha}'\beta_\alpha + \epsilon_i, \quad i = 1, \ldots, n \tag{2.9}$$

The ideal model in this case would be the model corresponding to α_0 , where α_0 is the set containing all nonzero components of the true parameter β. But in our case β is unknown, and in turn α_0 is unknown. Thus optimal model based on the data $(y_1, x_1'), \ldots, (y_n, x_n')$ should be estimated.

Let S be the class of all models that can be selected. Without any loss of generality model corresponding to (2.1) is assumed to be in S. For model $\alpha \in S$, size of the model is given be size of the set α. In this section, we will consider errors ϵ_i are i.i.d., and with mean and variance to be 0 and 1, respectively. Corresponding to model the (2.10), least square estimate of β_α is given by

$$\hat{\beta_\alpha} = (X_\alpha' X_\alpha)^{-1} X_\alpha' Y$$

Where $X_\alpha = (x_{1\alpha}, \ldots, x_{n\alpha})'$. Hence the prediction for the future values based on the least square estimate is given by,

$$\hat{y}_{f\alpha} = x_{f\alpha}' \hat{\beta}_\alpha$$

The model α can be judged based on the mean prediction error it is producing. Mean square prediction error corresponding to model α is given by,

$$mse(x_f, \alpha) = E(y_f - \hat{y}_{f\alpha})^2$$

As, $E(y_f - \hat{y}_{f\alpha})^2 = Var(y_f - \hat{y}_{f\alpha}) + [E(y_f - \hat{y}_{f\alpha})]^2$, mean squared error can be rewritten as,

$$mse(x_f, \alpha) = \sigma^2 + \sigma^2 x_f'(X_\alpha' X_\alpha)^{-1} x_f + [x_f'\beta - x_f'(X_\alpha' X_\alpha)^{-1} X\beta]^2$$

So, if we know the $mse(x_f, \alpha)$, we can obtain optimal model by minimizing $mse(x_f, \alpha)$ over all $\alpha \in S$. But we don't know the mean prediction error corresponding to α, hence we will first have to estimate it and then proceed with the minimization. We will discuss in short about the model selection method known as cross-validation, which estimates $mse(x_f, \alpha)$

Cross-Validation

This method was based on jackknife approach. Let $\hat{\beta}_{\alpha,i}$ be the least square estimate based on the data without the i-th point (y_i, x_i'). We can write $\hat{\beta}_{\alpha,i}$ as

$$\hat{\beta}_{\alpha,i} = (\sum_{j \neq i} x_{\alpha,j} x_{\alpha,j}')\sum_{j \neq i} x_{\alpha,j} y_j$$

Optimal model α can also be estimated using mse based on set $\chi = \{x_1, \ldots, x_n\}$. In this scenario, average mse can be given as,

$$\overline{mse(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} mse(x_i, \alpha) \tag{2.10}$$

Considering this mean squared error and since y_i and $\hat{\beta}_{\alpha,i}$ are independent to each other, cross-validation gives following estimate for $\overline{mse(\alpha)}$,

$$\widehat{\overline{mse(\alpha)}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_{\alpha,i}'\hat{\beta}_{\alpha,i})^2$$

Then optimal model α can be determined by finding minimizer of estimate $\widehat{\overline{mse(\alpha)}}$ over $\alpha \in S$.

Chapter 3

Dependent Data Set

3.1 Introduction

In the previous chapters, we looked at problems that had data itself independent or had some kind of structure where observation from different units were independent to each other. We applied bootstrap to this independent clusters to get variance and other estimates. In this chapter we will take a look at the dependent data set such as m-dependent data and time series. For a generic dependent data, if we apply normal bootstrap approach it fails to capture the essence of dependency between the data. Thus, We will look at the common modification to bootstrap methods and other approaches that have been described in literatures to these kind of dependent data sets.

3.2 m-dependent data

Definition 3.2.1 (Stationary sequence)**.** *Sequence of r.v.* $\{X_i : i = \pm 0, \pm 1, \ldots\}$ *is said to be strongly stationary sequence if for integers* $p > 0$ *and* q, $\{X_1, \ldots, X_p\}$ *and* $\{X_{1+q}, \ldots, X_{p+q}\}$ *have same distributions.*

All the random variable sequences are assumed to be strongly stationary for this section.

Definition 3.2.2 (m-dependent data)**.** *Sequence of r.v.* $\{X_i : i = \pm 0, \pm 1, \ldots\}$ *is called m-dependent if there exist a non-negative integer m such that sequences* $\{\ldots, X_{t-1}, X_t\}$ *and* $\{X_{t+m+1}, X_{t+m+2} \ldots\}$ *are independent for all integers t..*

3.2.1 Failure of i.i.d Boostrap in case of m-dependent sequence

Let $\{X_1, X_2, \ldots, \}$ be the m-dependent sequence. Moreover assume, $E(X_1) = \mu$ and $E(X_1^2) < \infty$. Also, we define,

$$\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{i+1})$$

If, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\sigma_m^2 \in (0, \infty)$, then by central limit theorem for m-dependent variables we know that

$$\sqrt{n}(\bar{X} - \mu) \to^d N(0, \sigma_m^2)$$

For simplicity, let us denote $\sqrt{n}(\bar{X} - \mu)$ by T_n. As we have seen already T_n converges to normal distribution with specific variance in distribution. But, let us try to find the bootstrap distribution of T_n using i.i.d bootstrap approach and see where it converges.
Let us draw bootstrap sample from empirical distribution with respect to $X_1, \ldots, X_n$. Now, in this case, bootstrap analog for T_n would be $T_n^* = \sqrt{n}(\bar{X_n}^* - \bar{X_n})$ where $\bar{X_n}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$. Following theorem suggest that distribution of T_n^* conditioned upon $X_1, \ldots, X_n$ will converge to the normal distribution with different mean.

Theorem 3.2.1. *Let* $\{X_1, X_2, \ldots, \}$ *be the stationary m-dependent sequence of random variables such that* $E(X_1) = \mu$ *and* $0 < Var(X_1) < \infty$*. Then,*

$$\sup_x |P_*(T_n^* \leq x) - \phi(x/\sigma)| = o(1) \text{ as } n \to \infty, \text{ a.s.}$$

Proof. Proof is similar to the convergence of bootstrap distribution of T_n in i.i.d case. Berry-Esseen Theorem still applies here and $s_n^2 \to \sigma^2$. But Marcinkiewicz-Zygmund SLLN is for i.i.d. case. Here we use the modified version of this theorem for m-dependent variables to show that

$$n^{-3/2}\sum_{i=1}^{n} |X_i|^3 \to 0 \text{ as } n \to \infty, \text{ a.s.}$$

This can be proved using following lemma:

Lemma 3.2.2. *Let $\{X_1, X_2, \dots, \}$ be the stationary m-dependent sequence of random variables. Let $f : \mathbb{R} \to \mathbb{R}$ be a borel measurable function such that $E(|f(X_1)^p|) < \infty$ for $o < p < \infty$ and $E(f(X_1)) = 0$ if $p \geq 1$. Then,*

$$n^{-1/p} \sum_{i=1}^{n} |f(X_i)| \to 0 \text{ as } n \to \infty, \text{ a.s.}$$

Proof. Idea of the proof is as follows. Break $\{X_1, X_2, \dots, \}$ into different subsequence. Let $Y_{ji} = X_{j+(i-1)(m+1)}$ for $j = 1, \dots, m+1$. Then Y_{ji} for $i \geq 1$ is an i.i.d sequence. Thus, $n^{-1/p} \sum_{i=1}^{n} |f(X_i)|$ can be divided into two summations over i and over j. As, Y_{ji} for $i \geq 1$, is an i.i.d sequence, we can apply Marcinkiewicz-Zygmund SLLN there with little modifications in constants. This proves the lemma. □

□

Thus we see that bootstrap distribution of T_n^* converges to a normal distribution with a different variance. Thus i.i.d. bootstrap method proposed by Efron fails to work in the case of m-dependent sequence. Intuitive idea for this failure is that bootstrap samples X_i^* fails to capture the dependency in the original sequence X_i. Next couple of sections will focus on different approaches to modify the bootstrap method for certain kind of dependent models.

3.3 Bootstrap Based on lID Innovations

I.i.d. bootstrap can be modified based on the specific kind of model that is under consideration. Let $\{X_n : n \geq 1\}$ be the sequence such that,

$$X_n = h(X_{n-1}, \dots, X_{n-p}; \beta) + \epsilon_n$$

$n > p$ and where β is a $q \times 1$ vector and $h : \mathbb{R}^{p+q} \to \mathbb{R}$ is a borel measurable functions. Let ϵ_i for $i \geq p$ be i.i.d sequence of random variables independent of $\{X_1, \dots, X_p\}$ and having distribution F. For simplicity let us assume that $E(\epsilon_1) = 0$. This model implies that X_n is dependent upon last p X_is below X_n and is basically driven by independent random errors ϵ_i. In such kind of model, i.i.d. bootstrap approach proposed by Efron can easily be extended.

Suppose that we have sample $X = \{X_1, \ldots, X_n\}$ and we have to estimate the sample distribution of $T_n = t_n(X; \beta, F)$ which is based on some function t_n. Now let $\hat{\beta}_n$ be some estimator of β based on sample X, it may be LSE or some other kind of estimator. Define residuals, $\hat{\epsilon}_i$ as follows,

$$\hat{\epsilon}_i = X_i - h(X_{i-1}, \ldots, X_{i-p}; \hat{\beta}_n);\ p < i \leq n$$

If, $\bar{\epsilon} = 1/(n-p) \sum_{i=p+1}^{n} \hat{\epsilon}_i$, them centered residuals can be given as, $\tilde{\epsilon}_i = \hat{\epsilon}_i - \bar{\epsilon}$. Centering is done just to match our assumption of errors having mean 0, otherwise it produces nonzero bias later.
Let $X_i^* = X_i$ for $i = 1, \ldots, p$ and

$$X_i^* = h(X_{i-1}^*, \ldots, X_{i-p}^*; \hat{\beta}_n) + \tilde{\epsilon_i} \text{ for } n \geq i > p$$

Then bootstrap version for T_n can be defined as $T_n^* = t_n(X^*; F_n, \hat{\beta}_n)$ where $X^* = \{X_1^*, \ldots, X_n^*\}$ and F_n is an emperical distribution based on centered residuals $\tilde{\epsilon}_i,\ p < i \leq n$. This sampling distribution of T_n is estimated by sampling distribution of T_n^* conditioned on X under this innovation of i.i.d. bootstrap approach. Autoregressive time series model is special case of the model we considered in this section and we will dive more in depth for this special case in Time series section.

3.4 Moving Block Bootstrap

In the last section, the idea was to use independence of residuals to apply bootstrap method. But most of the time, we don't know the underlying structure or model assumptions for the model. In that case we can't use any underlying independence structure to generate bootstrap method. In the view of this lack of knowledge of underlying structure, Kiinsch (1989) and Liu and Singh (1992) seperately formulated new resampling scheme known as Moving Block Bootstrap(MBB). In this section we will describe this method and give intuitive idea about why this approach works for dependent data.

Let $\{X_1, X_2, \ldots, \}$ denote the stationary sequence of random variables and let us say we have obtained observations $X = \{X_1, \ldots, X_n\}$. We will provide the moving block bootstrap version of statistics of the form, $\hat{\theta}_n = T(F_n)$, where F_n is the emperical distribution based

on data X and $T(.)$ is some real valued function.
We will not need following assumptions while introducing the theory of MBB, but in the literature standard assumptions are as follows: Let $l \in [1, n)$ be an integer, then we require

$$l \to \infty \text{ and } l/n \to 0 \text{ as } n \to \infty$$

Introduction of MBB however does not depend on these assumptions. Let $B_i = \{X_i, \ldots, X_{i+l-1}$ denote the l-sized block starting with X_i and $1 \leq i \leq N$ where $N = n-l+1$(Basically this is to assure that we don't go beyond X_n.). Also assume that k is the smallest integer such that $n \leq kl$. We now have N blocks $B_1, \ldots, B_N$. To generate MBB samples , we randomly select k blocks from these N blocks. Let us denote MBB sample as $B_1^*, \ldots, B_k^*$. Let us denote elements in B_i^* as $X_i^*, \ldots, X_{i+l-1}^*$. Then if we assume sequence $B_1^*, \ldots, B_k^*$ in a sequence then first n elements will be $X_1^*, \ldots, X_N^*$. This is known as MBB sample. Then MBB version of $\hat{\theta}_n$ is given as

$$\hat{\theta}_n^* = T(F_n^*)$$

here, F_n^* denotes the emperical distribution based on MBB sample $X_1^*, \ldots, X_n^*$.

Note that, if $l = 1$ then this is basically i.i.d. bootstrap approach introduced by Efron(1979). MBB basically draws blocks from given sample. So dependency structure of underlying dependent data is still captures within each block. If $\{X_i\}_{i \geq 1}$ have dependency stucture of finite lags, then asymptotically, as $l \to \infty$, MBB captures this dependency structure successfully. Typical choices for l includes $l = Cn^\delta$ where $\delta \in (0, 1/2)$.

3.5 Non-overlapping Block Bootstrap

Note that, in MBB method, resampled blocks can overlap with each other. Carlstein (1986) changed the blocking rule we used for MBB method. Initial Set-up is mostly similar to MBB. Let $l \in [1, n]$ and b be the largest integer satisfying $lb \leq n$. We define blocks as follows:

$$B_i^{(2)} = (X_{(i-1)l+1}, \ldots, X_{il}) \; i = 1, \ldots, b$$

Where superscript (2) denotes NBB (Non-overlapping Block Bootstrap) blocking. Note that these blocks don't overlap with each other unlike MBB approach. Then we we randomly

select k blocks from these non-overlapping blocks, where k is same as that of in last section. Let us denote resampled blocks as $B_1^{*(2)}, \ldots, B_k^{*(2)}$. We get sequence of random variables $X_{2,1}^*, \ldots, X_{2,n}^*$ if we arrange resampled blocks in order. Let $F_n^{(2)*}$ be the emperical distribution based on $X_{2,1}^*, \ldots, X_{2,n}^*$. Then bootstrap version of $\hat{\theta}_n$ is

$$\hat{\theta}_n^{(2)*} = T(F_n^{(2)*})$$

It has been shown that, if the stationary sequence $\{X_1, X_2, \ldots, \}$ satisfies some moment conditions then,

$$E(E_*(\hat{\theta}_n^*) - E_*(\hat{\theta}_n^{(2)*}))^2 = O(l/n^2)$$

This means that bootstrap estimators produced by MBB and NBB will have small difference if the sample size is large.

This methods are useful if we don't know the underlying model structure possessed by the stationary sequence under considerations. However if the model structure is known then it will be to use that particular structure while resampling in bootstrap method. Next section will deal with bootstrap modifications for the time series model.

3.6 Time series

In this section we will consider the time series model widely known as Autoregressive Processes. In this type of model, ideas behind application of bootstrap on linear models are modified so that it provides a consistent bootstrap estimates. We will consider Stationary case. We will describe the modification of Efron's bootstrap to this particular case and provide some theoretical results found in literatures. Autoregressive process $\{X_i\}_{i \in \mathbb{Z}}$ of order p ($p \in \mathbb{Z}$) has following structure:

$$X_i = \mu + \beta_1 X_{i-1} + \ldots + \beta_p X_{i-p} + \epsilon_i$$

where $\mu, \beta_1, \ldots, \beta_p$ are unknown autoregressive parameters and ϵ_i are i.i.d. random variables with mean 0, variance σ^2 and same distribution, say F. For simplicity we consider, $\mu = 0$.

3.6.1 Stationary Autoregressive Model

Autoregressive process is called stationary if

$$\beta(z) = 1 - \sum_{i=1}^{p} \beta_i z^p \neq 0, \text{ for } z \in \mathbb{C} \text{ and } |z| \leq 1$$

Note that here X_i are dependent, so we can use block bootstrap methods. But our approach in this section will be to use model structure of autoregressive time series to generate bootstrap data.
Suppose from the process $\{X_i\}_{i\in\mathbb{Z}}$, data $X_1, \ldots, X_n$ is observed. Then least square estimates $\hat{\beta}_{p1}, \ldots, \hat{\beta}_{pn}$ of $\beta_1, \ldots, \beta_p$ based on n observation satisfy

$$(\hat{\beta}_{p1}, \ldots, \hat{\beta}_{pn})' = (V_n' V_n)^{-1} V_n' (X_{p+1}, \ldots, X_n)' \tag{3.1}$$

Here V_n is a matrix whose i-th row is $X_{i+p-1}, \ldots, X_i$ and i varies from 1 to $n-p$. Residuals are given by $\hat{\epsilon}_i = X_i - \hat{\beta}_{1n} X_{i-1} + \ldots + \hat{\beta}_{pn} X_{i-p}$ for $i = p+1 \ldots, n$. Note that $\hat{\epsilon}_i$, using model structure of autoregressive process, can also be written as $\hat{\epsilon_j} = \epsilon_j - \sum_{i=1}^{p} (\hat{\beta}_{in} - \beta_i) X_{j-i}$ for $p+1 \leq i \leq n$. If suppose $\hat{\beta}_{in} \to_p \beta_i$ for $i = 1, \ldots, p$ as $n \to \infty$, then $\hat{\epsilon_j}$ can be approximated by ϵ_i, and thus are i.i.d under our assumptions. So we can use Efrons i.i.d bootstrap on these residuals. But before that, to adapt our assumption of mean of ϵ_i being 0 we need to center the residuals. New residuals are given by,

$$\tilde{\epsilon}_i = \hat{\epsilon}_i - \bar{\epsilon} \text{ for } i = p+1, \ldots, n$$

where $\bar{\epsilon} = (n-p)^{-1} \sum_{i=p+1}^{n} \hat{\epsilon}_i$. From the emperical distribution based on $\tilde{\epsilon}_i$, $i = p+1, \ldots, n$, draw random sample ϵ_i^*, $i \in \mathbb{Z}$. Each of this ϵ_i^* will have probability of being equal to $\tilde{\epsilon}_i$, conditioned upon $X_1, \ldots, X_n$, $1/(n-p)$
Bootstrap version of autoregressive process is given by

$$X_i^* = \hat{\beta}_{1n} X_{i-1}^* + \ldots + \hat{\beta}_{pn} X_{i-p}^* + \epsilon_i^* \ , \ i \in \mathbb{Z} \tag{3.2}$$

Let X_i^* be the stationary solution of this process. When the condition $\hat{\beta}_{in} \to_p \beta_i$ is satisfied, stationary solution for above equation exists with high probability for large n. But in practice we derive such stationary solution by following method. Let $X_i^* = X_i$ for $i = 1, \ldots, p$. Now if the $\hat{\beta}(z) = 1 - \sum_{i=1}^{p} \hat{\beta}_{in} z^p \neq 0,$ for $z \in \mathbb{C}$ and $|z| \leq 1$, then in the long run effects of

first p coefficients die out fast enough. So we use bootstrap version of autoregressive process recursively many times so that the desired stationary stage is reached.

The bootstrap version for autoregressive process for statistic $T_n = t_n(X_1, \ldots, X_n, \beta_1, \ldots, \beta_p, F)$ is given by

$$T_n^* = t_n(X_1^*, \ldots, X_n^*, \hat{\beta}_{1n}, \ldots, \hat{\beta}_{pn}, \hat{F}_n)$$

where $\hat{F}_n$ is emperical distribution described earlier based on centered residuals. Bose(1988) showed that ARB (Auto-regressive Bootstrap) approximation is better than normal approximation which has error of $O(n^{-1/2})$, while ARB has error rate of order $o(n^{-1/2})$.

Let $Cov(X_i, X_j)/\sigma^2$ be (i,j)th element of $p \times p$ matrix Σ. Let $\{X_i^*\}_{i\in\mathbb{Z}}$ be the stationary solution to (3.2) and let $Cov_*(X_i^*, X_j^*)/Var_*(X_i^*)$ be the (i,j)th element of $\hat{\Sigma}$. Assume that $E|\epsilon_1^{2(s+1)}| < \infty$ for some $s \geq 3$ and $(\epsilon_1, \epsilon_1^2)$ satisfies following condition : For every $c > 0$ there is $\delta_c > 0$ which satisfies $\underset{|u|>c}{sup}|\psi(u)| \leq e^{-\delta_c}$, where $(\epsilon_1, \epsilon_1^2)$ has charecteristic function $\psi(u)$. Let $\beta = (\beta_1, \ldots, \beta_p)$, $\hat{\beta}_n = (\hat{\beta}_{p1}, \ldots, \hat{\beta}_{pn})$ and $\hat{\beta}_n^*$ be the p-vector obtained by putting X_i^* in place of X_i in equation (3.1). If, H_n is the distribution of $\sqrt{n}\Sigma(\hat{\beta}_n - \beta)$ and H_{Boot} is the distribution of $\sqrt{n}\hat{\Sigma}(\hat{\beta}_n^* - \hat{\beta}_n)$ then Bose(1988) proved that

$$\underset{x}{sup}\,|H_{Boot}(x) - H_n(x)| = o(n^{-1/2}) \text{ a.s.}$$

ARB tends to have more accuracy when data assumes the above given model structure. When we know that data does not assume the stationary autoregressive structure, it is better to apply moving block methods for greater accuracy.

3.7 Simulations

In this section we will provide some basic simulations to give insight on how the bootstrap works.

We randomly generate 2-dimensional data having 50 data points. We can think of this as 2 test scores of the 50 students in the class. Suppose i-th data point is denoted as $x_i = (x_{i1}, x_{i2})$. We are interested in the correlation of vector x_{i1} and vector x_{i2}, say $\hat{\theta}$. When we apply bootstrap to this data. It resamples 50 data point x_i^* from original data and calculate

correlation coefficient for this new data. We carried out 1000 bootstrap replications for the statistic $\hat{\theta}$. Following image shows their distribution. Original value of $\hat{\theta}$ is 0.138. Standard error and bias estimates given by bootstrap method are 0.111 and -0.009 respectively. So if in practice we need to estimate distribution of $\hat{\theta}$, we can use the simulated distribution to get a nice approximation.

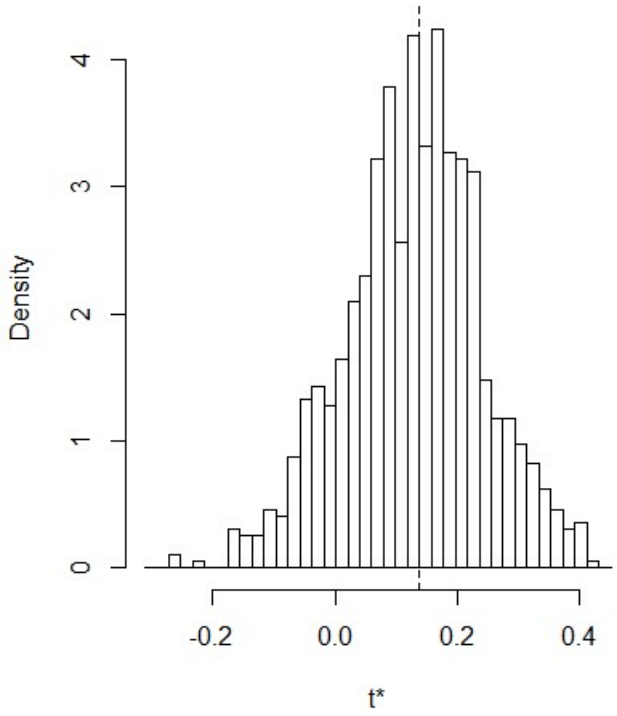

Figure 3.1: Correlation coefficient

Next, we will consider autoregressive time series model of order 1. So our model is $X_i = \beta_1 X_{i-1} + \epsilon_i$. Here $\epsilon_i \sim N(0, 1)$. We generate data set of 50 data points using this model and $\beta_1 = 0.5$. We will apply ARB method to generate a bootstrap version of above time series. Original time series is shown in *Figure 3.2*.

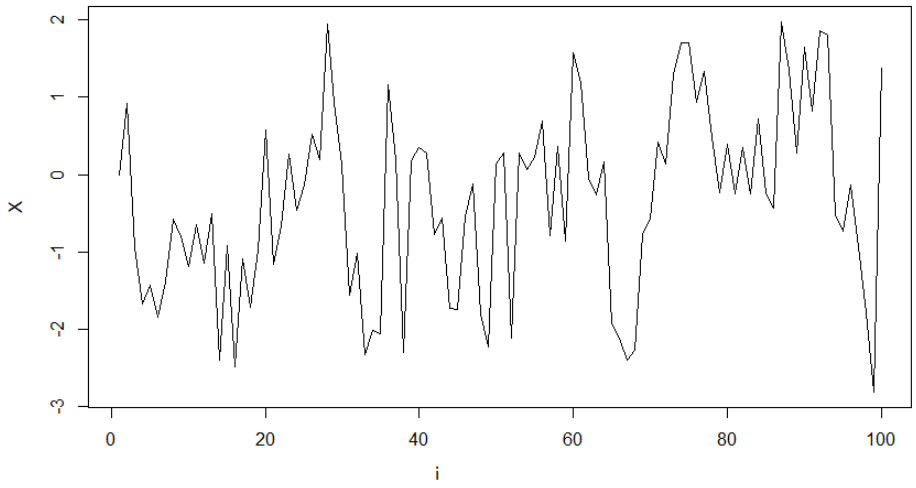

Figure 3.2: Original Time series

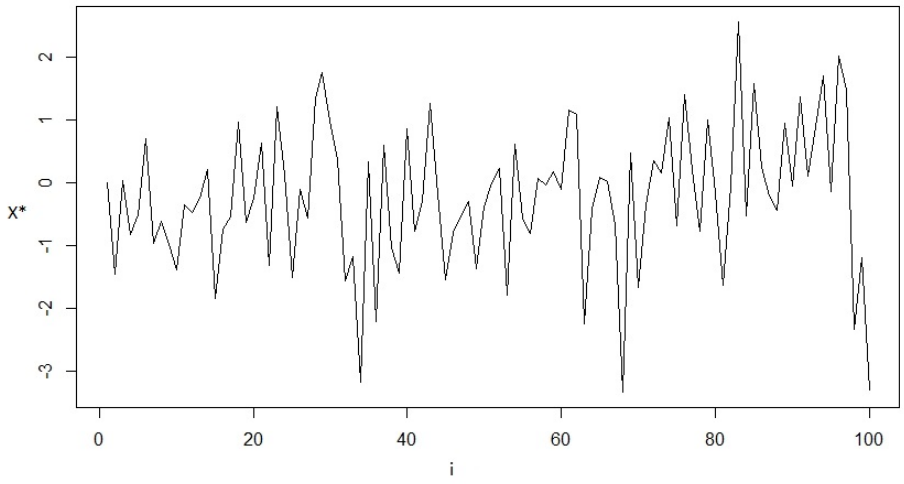

Figure 3.3: Bootstrap replication

Here least square estimate for β_1 is $\hat{\beta}_1 = 0.519$. Corresponding to this estimate, we obtain residuals and resample them after centering. The bootstrap replication of above time series obtained using ARB method described in previous section is shown in *Figure 3.3.*
Let us consider statistic $T = (\sum_{i=1}^{99} X_i^2)^{1/2}(\hat{\beta}_1 - \beta_1)$. Bootstrap replica using ARB method

would be $T^* = (\sum_{i=1}^{99} X_i{*}^2)^{1/2}(\hat{\beta}_1^* - \hat{\beta}_1)$. Sampling Distribution of T^* using 500 bootstraps is shown in the following figure. Thus, in practice, distribution of T can be approximated by following simulated distribution.

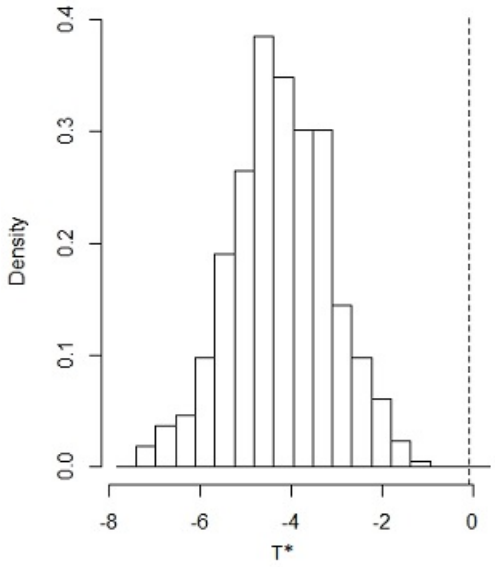

Figure 3.4: Bootstrap Distribution T^*

Chapter 4

Conclusion

We learned the variance and the bias estimates produced by the jackkife and the bootstrap method in case of independent data set. We approximated the distribution of any given statistic with its bootstrap distribution and showed one way of proving the consistency of this bootstrap distribution. Based on this bootstrap distribution, there are two ways to obtain confidence intervals. Based on certain assumptions, we showed why this confidence interval will work. But in case, this assumption is violated, we improved its accuracy using better approximation. Through this we obtained BC and BCa confidence intervals. Later we showed the consistency of all of these intervals and showed that as sample size gets larger, we will get closer to desired level of confidence.

Further, applications of these methods to the linear models and dependent data set have also been studied. These resampling procedures are not only used to estimate variance and bias but can also be used in prediction problem. We used bootstrap to estimate the prediction error and by this procedure we can make sure if our linear model is accurate or not. One important problem was model selection. And we looked at the resampling technique known as cross validation to estimate the give model.

In dependent data structure, we first looked at why the Efron's bootstrap does not work. Later we looked at MBB and NBB methods and their relation with each other. Time series model was also studeied to give some idea of how bootstrap method can be modified to the specific structure of dependent data.

Bibliography

[1] A. Bose: Edgeworth correction by bootstrap in autoregressions, The Annals of Statistics, **16**, 1709-1722, 1988.

[2] B. Efron: Bootstrap methods: Another look at the jackknife, Annals of Statistics, **7**, 1-26, 1979.

[3] B. Efron: The Jackknife, the Bootstrap, and Other Resampling Plans, SIAM, Philadelphia, 1982.

[4] B. Efron: Estimating the error rate of a prediction rule: Improvement on cross-validation, J. Amer. Statist. Assoc., **78**, 316-33l, 1983.

[5] B. Efron: The Jackknife, the Bootstrap and Other Resampling Plans, CBMS-NSF Regional Conference Series in Applied Mathematics, 1987.

[6] B. Efron: Better bootstrap confidence intervals , J. Amer. Statist. Assoc., **82**, 171-200, 1987.

[7] B. Efron, R.J. Tibshirani: An Introduction to the Bootstrap, Chapman and Hall/CRC, 1994.

[8] C. F. J. Wu: Jackknife, bootstrap and other resampling methods in regression analysis, Annals of Statistics, **14**, 1261-1350, 1986.

[9] D. M. Allen: The relationship between variable selection and data augmentation and a method for prediction, Technometrics, **16**, 125-127, 1974.

[10] D. V. Hinkley: Jackknifing in unbalanced situations, Technometrics, **19**, 285-292, 1977.

[11] E. Carlstein: The use of subseries methods for estimating the variance of a general statistic from a stationary time series, The Annals, of Statistics, **14**, 1171-1179, 1986.

[12] H. R. Kiinsch: The jackknife and the bootstrap for general stationary observations, The Annals of Statistics, **17**,1217-1261, 1989.

[13] J. Shao: Bootstrap variance and bias estimation in linear models, Canadian J. Statist., **16**, 371-382, 1988(c).

[14] J. Shao, D. Tu: The Jackknife and Bootstrap, Springer-Verlag, 1995.

[15] K. Singh: On the asymptotic accuracy of the Efron's bootstrap, The Annals of Statistics, **9**, 1187-1195, 1981.

[16] R. G. Miller: An unbalanced jackknife, Annals of Statistic, **2**, 880-89l, 1974.

[17] R. Y. Liu and K. Singh: Moving blocks jackknife and bootstrap capture weak dependence, in R. Lepage and L. Billard, eds, 'Exploring the Limits of the Bootstrap', Wiley, New York, pp. 225-248, 1992.

[18] S. N. Lahiri: Resampling Methods for Dependent Data, Springer-Verlag, 2003.