GRADUATE SCHOOL OF WESTMINSTER INTERNATIONAL UNIVERSITY IN TASHKENT PROFESSORSHIP OF STATISTICS AND ECONOMETRICS

Bootstrap standard errors and intervals under different sample size in OLS

Zarrukh Rakhimov
PhD candidate in Econometrics and Statistics Module leader in Data Analytics
Email: zrakhimov@wiut.uz
Westminster International University in Tashkent Istiqbol str. 12, 100047 Tashkent, Uzbekistan
Nilufar Rahimova
Silk Road International University of Tourism and Heritage Module leader in Economics of Tourism
Email: zrakhimov@wiut.uz
University Boulevard 17, Samarkand, Uzbekistan

Abstract

Linear regression is one of the widely used statistical methods in social sciences. The core part of the regressions are coefficients which bring some inference. Yet, we rely on hypothesis testing or confidence intervals and certain assumptions underlying linear models such as sample size being large enough. In this study, we suggest alternative way of constructing confidence intervals using bootstrap which is expected to work well even when the sample size is smaller than required per OLS assumptions. We find that even in small samples, bootstrap confidence intervals can perform better than traditional interval estimations

***Key words**: sample size, linear model, confidence Interval, bootstrap, accuracy, interval size*

Contents

INTRODUCTION

Linear regression is quite broadly used methodology to explain relationships between different variables in many domains. Linear model (often referred as OLS model) is used primarily for two purposes. First, this model can explain the relationship between two or more factors. Second, linear models are often used to make simple and still efficient forecasting. Linear models are very popular due to the fact that they are relatively easy to learn, build and interpret. Yet, we almost never meet a perfect linear relationship between two or more factors in real life, thus linear regression is almost always an approximation of real life relationships. Linear regression, sometimes referred to as OLS, has a set of assumptions that should be met in order to make the outcomes of the OLS model reliable. These assumptions are:

1. Homoscedasticity (or no heteroscedasticity)
2. Stationarity or no autocorrelation of residuals (in case of time series data)
3. No strong multicollinearity between explanatory variables
4. No severe outliers
5. Sample size to be larger than 30 observation
6. Linearity in relationship
7. Normality of residuals

Violations of one or more of the above assumptions can lead to inaccuracy or even bias in the estimation. Interested readers are encouraged to explore more details of each assumption, but in this study we will discuss in more detail the presence of heteroscedasticity, how OLS estimates can suffer and how bootstrap can be a remedy in absence of homoscedasticity.

Literature Review

Bootstrap method is a resampling method of a given dataset to build a sampling distribution of a specific statistic. Bootstrapping has become popular because it has proven to provide reliable inferences in many cases even when underlying assumptions are not satisfied. This also applied to cases of heteroscedastic residuals which is first

discussed in papers of Efron (1979). Since then, theoretical foundations have been concentrated on justifying validity and efficiency of bootstrap confidence intervals with non-constant variance of errors (Davison and Hinkley, 1997).

In the context of linear models, there have been primarily two types of bootstrapping used for estimating point and interval estimates, bootstrapping residuals and bootstrapping pairs (Chernick and LaBudde, 2011).

Bootstrapping residuals: This method of bootstrapping was first introduced by Efron (1982). Imagine we have the following model

$Y_i = g_i(\beta) + e_i$, for $i=1,2,....,n$

where $g_i(\beta)$ is a function with a known form. To estimate β, we minimize distance between our true dependent variable Y_i and estimated function $g_i(\beta)$. These distances are expressed in terms of residuals $\hat{e_i} = Y_i - g_i(\hat{\beta})$. The idea behind Wild bootstrap is to take the distribution of residuals each having probability of *1/n* for $i=1,2,....,n$ and sample *n* times from this distribution to get bootstrap sample of residuals which can be denoted as $(e_1, e_2, e_3, \dots, e_n)$. Afterwards, bootstrap dependent variable can be generated using $Y_i^* = g_i(\hat{\beta}) + e_i^*$. Now, as we have our bootstrap dataset, we use simple OLS method to estimate β^*. We repeat the above procedure *B* times to get a distribution of β_j^* estimates for $j=1,2,....,B$. One can get standard deviation of β^* to build bootstrap confidence intervals.

Bootstrapping pairs: bootstrapping pairs is a rather simple but powerful approach proposed first by Freedman (1981). Under this approach, we resample independent and dependent variables from the original sample which results in a bootstrap sample. We then use usual OLS method to estimate β^* from the bootstrap sample. This procedure is repeated B times in order to get distribution of coefficients β_j^* estimates for $j=1,2,....,B$. This distribution in turn can give bootstrap standard deviation.

Efron and Tibshirani (1986) conclude that two approaches are equivalent when the model is correctly specified, but they can perform differently when the sample is small. Flachaire (2003) compared bootstrapping residuals and bootstrapping pairs when the model is correctly specified and when heteroscedasticity is present in the linear models. Flachaire (2003) concludes that when a proper transformation to the residual term is applied (wild bootstrap), residuals bootstrap performs better than bootstrapping pairs. Chernick and LaBudde (2011) conclude however that bootstrapping vectors are less sensitive to violations of model assumptions and can still perform well if those assumptions are not met. This can be explained by the fact that the vector method does not depend on model structure while bootstrapping residuals do.

Other approaches are stationary bootstrap (Politis and Roman, 1994), and the percentile-t bootstrap (Diciccio and Efron, 1992) each used under different scenarios of non-constant variance of the residuals.

This study tries to shed further light into implementing bootstrapping pair in the context of linear models with heteroscedastic residuals and test bootstrap interval performance under different sample sizes.

Linear regression models

First of all, let's look into how linear models are built and how coefficients as well as their intervals are estimated. As mentioned earlier, the linear model evaluates the impact of one or more variables (explanatory variables) to another variable (explained or dependent variable). This is done by estimating coefficients of estimates of each explanatory variable. For instance, imagine that we want to evaluate whether your year of education affects your income and by how much. If we build our simple OLS model where income is dependent "Y" variable, and year of education is "X_1" explanatory variable, then coefficient of "years of educations" (β_1) shows the size and direction (positive or negative) of the impact.

$$Y = \beta_0 + \beta_1 * X_1 + e$$

Where

$Y -$ dependent variable,

β_0 – intercept,

$\beta_1 -$ coefficient of first explanatory variable

$X_1 -$ explanatory or independent variable

$e -$ error or residual term

The above model is the simplest one variable example of linear regression and usually most studies take into account more explanatory variables that will improve the model (there are metrics to evaluate whether a model is improving or not, e.g. adj. R squared, AIC, MSE).

Estimation of coefficients in the above model is done with the method of least squares commonly known as OLS (ordinary least squares). Least squares estimate of β_1 is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \underline{X})(Y_i - \underline{Y})}{\sum_{i=1}^{n} (X_i - \underline{X})^2}$$

where

$n -$ number of observations

X_i – value of the independent variable for the i-th observation
Y_i – value of the dependent variable for the i-th observation
$\underline{X}$ – mean of the independent variable X
$\underline{Y}$ – mean of the independent variable Y

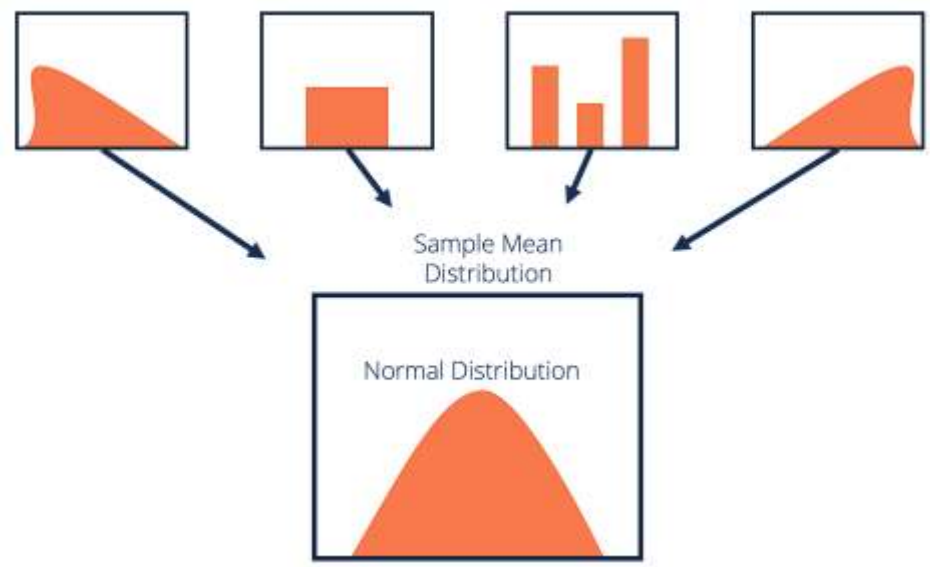
Traditional Confidence intervals

Researchers are often interested not only in point estimates of coefficient, but also interval estimations. This is because point estimates of coefficients are always an approximation to true population value. In contrast, interval estimations, commonly known as confidence intervals, have a set of advantages. Firstly, it gives a range of values where true population value can be located. Secondly, confidence intervals will indicate whether the true population parameter might be equal to 0. In other words, whether the effect of that specific explanatory/independent variable to dependent variable is insignificant. Currently, all statistical softwares provide both point and interval estimates by default. Below, we will look at the theoretical side of building confidence intervals of coefficients of linear models.

Central Limit Theorem

Central Limit theorem (CLM) is the core concept of statistics that is employed also in building confidence intervals. The theory says that irrespectful of the true population dataset, if one derives many sample averages from many samples generated from the same population, then the distribution of sample averages is approximately normal (also referred as Gaussian, see graph below) (Lind et al, 1967). The midpoint of resulting distribution of sample averages will be equal to the true population mean (see Figure 1). This is a very strong finding that can also be applied in confidence interval construction.

Figure 1



In practice, we often cannot take many samples from the same population and very often left to work with only one sample. Nevertheless, one can still make some estimation regarding the population value (e.g. mean, coefficient) using the central limit theorem even when the distribution of the population dataset is not known.

Confidence interval based on CLT

Consider we have only one sample from the population data. Firstly, we can estimate the sample coefficient using the method of ordinary least squares (discussed in previous chapter). Afterwards, we can estimate standard error of the estimated coefficient using the following formula also arising from the method of least squares.

$$se(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} \left(X_i - \underline{X}\right)^2}}$$

where

s – standard deviation of the residuals (residual standard error)

n – number of observations

X_i – value of the independent variable for the i-th observation

$\underline{X}$ – mean of the independent variable X

As distribution of $\hat{\beta}_1$ coefficient is approximately normal distribution based on central limit theorem, we employ properties of standard normal distribution (z-distribution) and build 90%, 95% or 99% confidence intervals.

$$\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} * se(\hat{\beta}_1)$$

where

$\hat{\beta}_1$- is sample coefficient estimate

$z_{\frac{\alpha}{2}}$ – is a value from the standard normal distribution the give an area of $\frac{\alpha}{2}$

$se(\hat{\beta}_1)$ - sample variance of the coefficient

The above interval estimation is interpreted in the following way. 97% interval indicates that if we construct 100 confidence intervals from 100 random samples generated from the true population, then 97 of those confidence intervals will contain true population coefficient β_1. Also, employing this confidence interval you can verify whether population coefficient is insignificant. If estimated confidence interval contains zero, then one can suspect that the true population parameter can be equal to zero (Gujarati, 2004)

However, one can see that estimation of the standard error of the same coefficient depends on the normality of the residual term. In the presence of heteroscedasticity, standard deviation of the error term can be inflated which will result in inaccuracies in confidence interval constructions using the CLT approach (Gujarati, 2004).

Heteroscedasticity can arise from various sources, such as:

1. Omitted variables
2. Measurement error
3. Non-linearity of the relationship of dependent and independent variable
4. Outliers
5. Residual variance that deviates with time
6. Endogeneity
7. Model misspecification

If no remedy is applied to heteroscedasticity in residuals, it will make the standard error of the residuals biased and can lead to wrong conclusions in hypothesis testing. Academia suggested a set of way on how heteroscedasticity, such transforming variables, weighted least squares, including important variables and many others (Greene, 2021)

Below, we suggest another way, bootstrap, of handling heteroscedasticity in residuals for construction of our confidence intervals for coefficients.

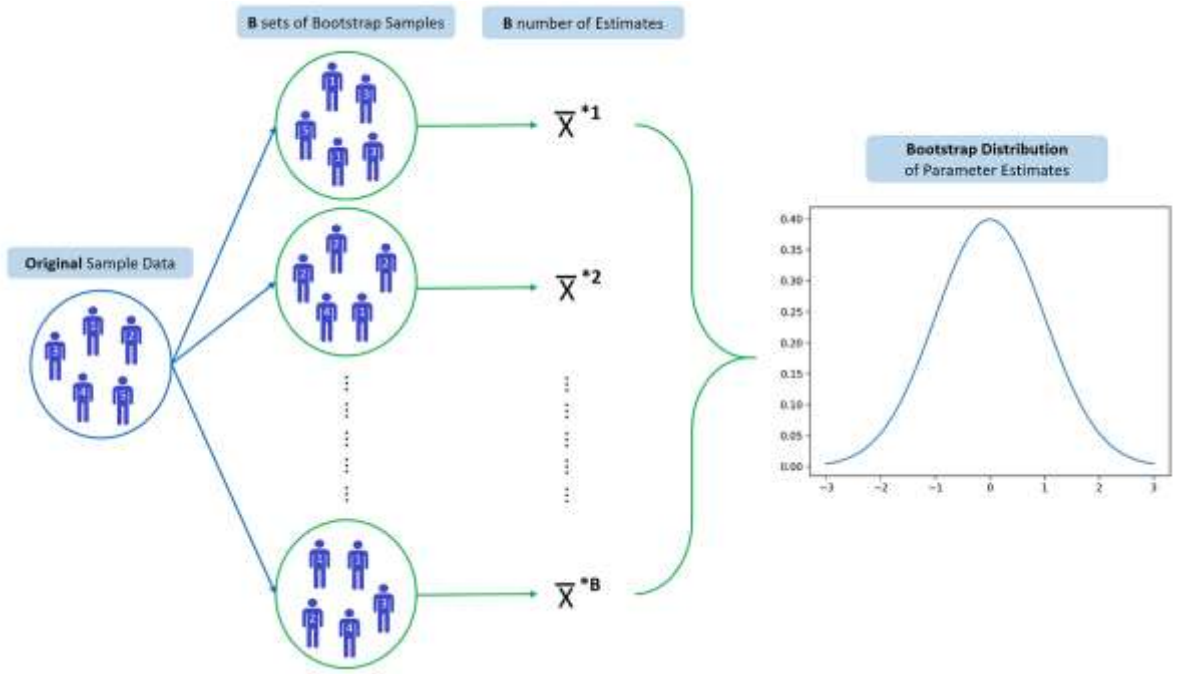
Bootstrap confidence interval estimation

In the first place, it is necessary to explain the concept of bootstrapping. Bootstrap is a relatively easy resampling technique that can offer alternative ways of building confidence intervals. Bootstrap implies selecting one sample and generating many other different samples from this single original sample and estimating your parameter of interest in each newly created sample. Under the bootstrap approach, the original

sample is considered as a population and we generate many other samples (known as bootstrap samples) out of it. When a large number of bootstrap samples are created, we estimate sample parameters (e.g. coefficient) from every bootstrap sample. Consequently, we will have a distribution of bootstrap sample estimates.

This distribution of bootstrap sample estimates can be used to construct our confidence intervals. For example, if we want to construct a 95 percent interval, we take 2.5th and 97.5th percentiles from bootstrap distribution. Figure 2 explains visually the method of bootstrapping.

Figure 2



Simulation

In order to evaluate performance of bootstrap confidence intervals when heteroscedasticity is present, it is necessary to carry out a simulation of a linear model. Simulation is necessary for two reasons. First, we need to know the true population coefficient β_1 and in practice we rarely know the true population parameter. Secondly, we need to evaluate performance of estimated confidence intervals in presence of heteroscedasticity. Although real data can have heteroscedasticity of residuals, we do not know the true form of residuals distribution. For these two reasons we need to model our linear model with heteroscedastic residuals. We select the simplest form of linear model with one explanatory variable that is correlated with the error term.

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 * X1 + X1^2 * \varepsilon$$

where

$X1 \sim N(5, 4)$

$\mathcal{E} \sim N(0, X1)$

where intercept (β_0) and β_1 are defined by us. Independent variables (X_1) come from normal distribution with mean of 5 and standard deviation of 4. Error term ($X1^2 * \mathcal{E}$) is simulated following the approach suggested by Flachaire (2003). Under this scenario, error term is correlated with explanatory variable and its variance grow as the value of $X1^2$ grows.

We check the performance of bootstrap confidence intervals in different sample sizes. Thus, we have a first sample size of 30 and then we increase it by 10 observations up to 200 observations. All of the simulations are carried out in R software.

We take the following steps for simulation of linear model with heteroscedasticity with different sample sizes

Step 1: set intercept $\boldsymbol{\beta_0}$= 4 and coefficient $\boldsymbol{\beta_1}$=5

Step 2: Set sample size to n=30

Step 3: generate *X1 ~ N(5, 4)* starting with sample size n

Step 4: generate *Y* with $Y = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * X1 + X1^2 * \mathcal{E}$

Step 5: estimate confidence intervals using traditional and bootstrap methods in repeated simulations (1000 times). Here we construction 95 percent confidence intervals

Step 6: evaluate how many times (out of 1000), true parameters were within estimated OLS and bootstrap confidence intervals

Step 7: repeat step 2 to step 8 by adding 10 observations to sample size (n=n+10). Finish when sample size reaches 200 observations

Traditional and bootstrap confidence intervals estimations are discussed in above sections. For traditional intervals, we use the following formula which is estimated in any statistical package when we construct our linear model.

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} * se(\hat{\beta}_1)$$

Bootstrap confidence intervals are built taking values in certain percentiles of parameter distributions that were generated as a result of bootstrapping.

Results

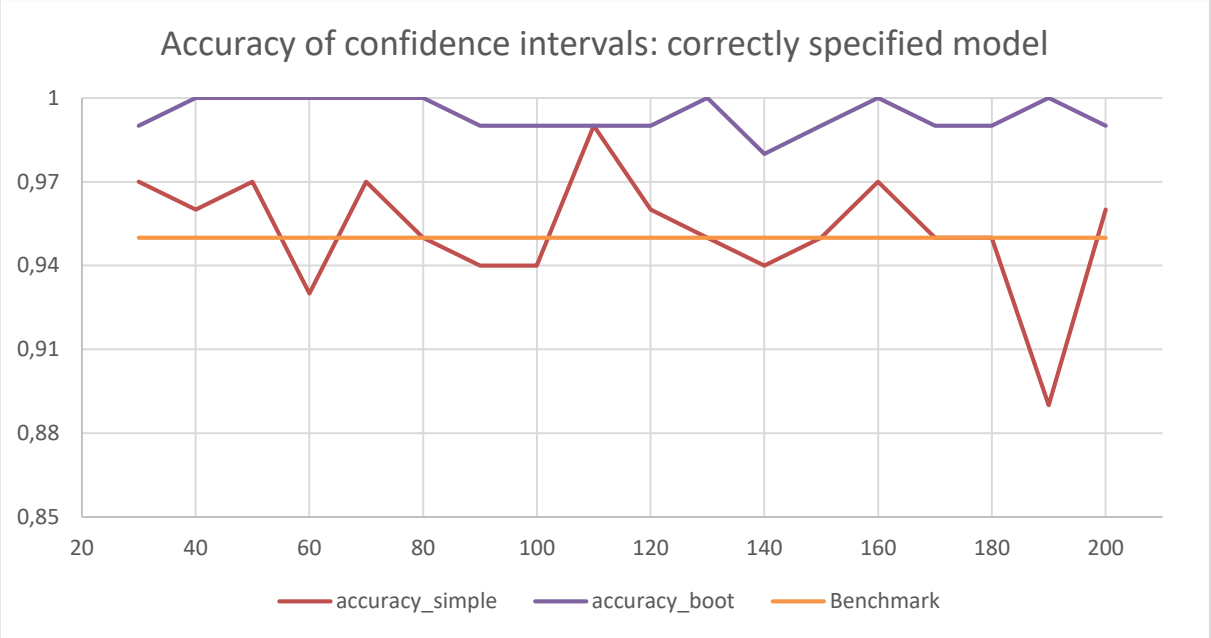
In this part, we will look into two results of the simulation. One is with homoscedastic residuals and second is with different size. We also take a look at how estimated intervals change as we change our sample size.

Correctly specified model

First of all, we want to see how traditional CLT based and bootstrap confidence intervals perform when no violations of OLS assumptions are present. We expect that both approaches will do relatively good work in building interval estimates. In other words, for 95 percent confidence intervals, we expect true parameters to fall within estimated intervals at least 95 per cent of cases.

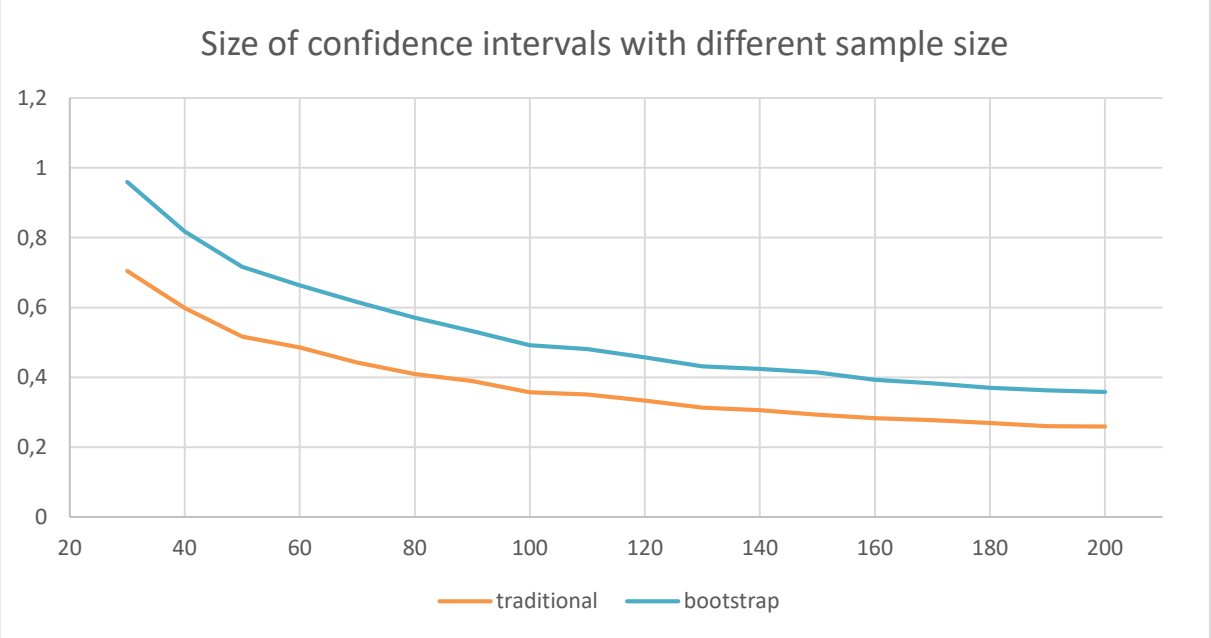
The first graph below shows often true coefficients fall within estimated confidence intervals built using traditional and bootstrap methods. One can see that both methods are doing relatively well, that is constructed intervals are containing true coefficient at least The chart clearly shows that both traditional and bootstrap confidence intervals contain true parameter in 90-100 percent of the cases which is expected outcomes (see Figure 3)

Figure 3



Bootstrap confidence intervals contain true coefficients more often compared to traditional OLS intervals. This is explained in the second graph which shows that bootstrap intervals are larger in width compared to OLS intervals across all sample sizes (see Figure 4)

Figure 4



CONCLUSION

In this paper, we carried out a simulation study of building bootstrap confidence intervals in linear models when variance of residuals is constant. We first looked at existing literature on this topic and then looked at the theoretical side of linear models with heteroscedasticity. We explained that traditional confidence intervals might be biased when heteroscedasticity is present in data and therefore suggested using bootstrapping pairs for building confidence intervals which do not have any assumptions of residual distribution. Our simulation study shows that bootstrap confidence intervals outperform traditional ones though they are still not reaching targeted 95 percent coverage rate. In contrast, traditional intervals are highly inaccurate as they contain true coefficients in less than 80 per cent of the cases compared to targeted 95 per cent.

BIBLIOGRAPHY

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26.

Efron , B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia

Efron , B., and Tibshirani , R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science.* Vol. 1 , 54 – 77

Davison , A. C. , and Hinkley , D. V. (1997). Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge .

Chernick, M. R., and LaBudde, R. A. (2014). An introduction to bootstrap methods with applications to R. John Wiley & Sons.

Chernozhukov, V., and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2), 293-346.

DiCiccio , T., and Efron , B. (1992). More accurate confidence intervals in exponential families. *Biometrika* 79, 231 – 245 .

Fan, Y., and Li, Q. (2004). A consistent model specification test based on the kernel density estimation. *Econometrica*, 72(6), 1845-1858.

Freedman , D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, 9, 1218 – 1228

Flachaire, E. (2007). Bootstrapping heteroscedastic regression models: wild bootstrap vs pairs bootstrap. *Computational Statistics and Data Analysis*, 49 (2), 361-376

Horowitz, J. L., and Markatou, M. (1996). Semiparametric estimation of regression models for panel data. *Review of Economic Studies*, 63(1), 145-168.

Greene, W. H. (2021) Econometric Analysis, 8th edn, Pearson

Gujarati, D. N., Porter, D. C., and Gunasekar, S. (2012). Basic econometrics. McGraw-Hill Higher Education

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). An Introduction to Statistical Learning. Publisher.

Politis, D. and Romano, J, (1994). The Stationary bootstap. *The journal of American Statistical Association.* 89 (428), 1303-1312

Lind, D. A., Marchal, W. G., and Wathen, S. A. (1967). Statistical Techniques in Business and Economics (2nd ed). Publisher

Liu , R. Y. (1988). Bootstrap procedures under some non i.i.d. models . *Annals of Statistics* 16, 1696 – 1708