IQTISODIY TARAQQIYOT VA TAHLIL
ILMIY ELEKTRON JURNAL

UO‘K: 330.43

BOOTSTRAP CONFIDENCE INTERVALS IN LINEAR MODELS: CASE OF OUTLIERS

PhD ***Rakhimov Zarrukh Aminovich***
Westminster International University in Tashkent
ORCID: 0009-0001-0583-4819
PhD ***Rahimova Nilufar Aminovna***
Westminster International University in Tashkent
ORCID: 0000-0002-8648-2543

Annotation. *Confidence interval estimations in linear models have been of large interest in social science. However, traditional approach of building confidence intervals has a set of assumption including dataset having no extreme outliers. In this study, we discuss presence of severe outliers in linear models and suggest bootstrap approach as an alternative way to construct confidence intervals. We conclude that bootstrap confidence intervals can outperform traditional confidence intervals in presence of outliers when sample size is small or population distribution is not normal. Lastly, we encourage researchers to run a computer simulation to evaluate conclusions of this study.*

Key words: *bootstrap, lineal model, confidence Interval, extreme outliers, resampling.*

ЧИЗИҚЛИ МОДЕЛЛАРДА БООТСТРАП ИШОНЧЛИК ИНТЕРВАЛЛАРИ: ЧЕТ ҚИЙМАТЛАР ҲОЛАТИ

Рахимов Заррух Аминович
Тошкент халқаро вестминстер университети
Рахимова Нилуфар Аминовна
Тошкент халқаро вестминстер университети

Аннотация. *Чизиқли моделлардаги ишонч оралиғини баҳолаш ижтимоий фанларда катта қизиқиш уйғотди. Бироқ, ишонч оралиқларини қуришнинг анъанавий ёндашуви бир қатор тахминларга эга, шу жумладан маълумотлар тўплами ҳеч қандай ҳаддан ташқари чегараларга эга эмас. Ушбу тадқиқотда биз чизиқли моделларда жиддий чегаралар мавжудлигини муҳокама қиламиз ва ишонч оралиқларини қуришнинг муқобил усули сифатида юклаш усулини таклиф қиламиз. Намуна ҳажми кичик бўлса ёки популяция тақсимоти нормал бўлмаса, юклашнинг ишонч оралиғи анъанавий ишонч оралиқларидан устун бўлиши мумкин деган хулосага келдик. Ниҳоят, тадқиқотчиларни ушбу тадқиқот натижаларини баҳолаш учун компютер симуляциясини ишга туширишни тавсия қиламиз.*

Калит сўзлар: *бootстрап, чизиқли модел, ишонч оралиғи, экстремал чегаралар, қайта намуна олиш.*

БУТСТРАП-ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ В ЛИНЕЙНЫХ МОДЕЛЯХ: СЛУЧАЙ ВЫБРОСОВ

Заррух Рахимов Аминович
Международный вестминстерский университет в Ташкенте
Нилуфар Рахимова Аминовна
Международный вестминстерский университет в Ташкенте

Аннотация. *Оценки доверительных интервалов в линейных моделях представляют большой интерес в социальных науках. Однако традиционный подход к построению доверительных интервалов предполагает ряд допущений, включая набор данных, не имеющий экстремальных выбросов. В этом исследовании мы обсуждаем наличие серьезных выбросов в линейных моделях и предлагаем метод начальной загрузки в качестве альтернативного способа построения доверительных интервалов. Мы пришли к выводу, что доверительные интервалы начальной загрузки могут превосходить традиционные доверительные интервалы при наличии выбросов, когда размер выборки невелик или распределение популяции не является нормальным. Наконец, мы призываем исследователей провести компьютерное моделирование, чтобы оценить выводы этого исследования.*

Ключевые слова: *бутстрап, линейная модель, доверительный интервал, экстремальные выбросы, повторная выборка.*

Introduction.

Regression model has become one of widely used econometrics models across various disciplines. One of the simplest and widespread version of regression is linear regression. Linear regression builds linear relationship between dependent and explanatory variables. Although linearity is almost always an approximation to real life scenario, it has proven to be good enough to evaluate relationship of different variables. Linear regressions have been primary used for two purposes. First of all, linear models are used to evaluate whether a certain factors really has an impact on a dependent variable and what is the impact. Secondly, linear model is used to make predictions of dependent variable. Compared to other econometric models, linear model is easy to build and to interpret.

In this study we concentrate on the first usage of the linear models, i.e. impact of one variable to another. This is done by estimating coefficients of estimates of each explanatory variables. For example, if we want to evaluate what factors determine salary and we check years of educations as one of the factors, then coefficient of "years of educations" (β_1) show the direction and size of the impact of this variable on salary.

$$Salary = \beta_0 + \beta_1 * Years\ of\ Education + \beta_2 * Age + \beta_3 * Gender + \beta_4 * Region + \cdots$$

However, before evaluating an impact of each variable to dependent variable, we always check for the significance of impact. In other words, we carry hypothesis testing of checking whether each coefficient is significantly different from zero

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

In other words, in the above hypothesis testing, we evaluate whether "Years of education" have any impact on Salary. In order to decide on this hypothesis, most statistical packages make use of traditional confidence intervals that are based on central limit theorem. However, making a decision on such hypothesis using traditional confidence intervals relies on a set of assumptions such as no outliers, no strong multicollinearity, stationarity of data sets, heteroscedasticity to name just a few.

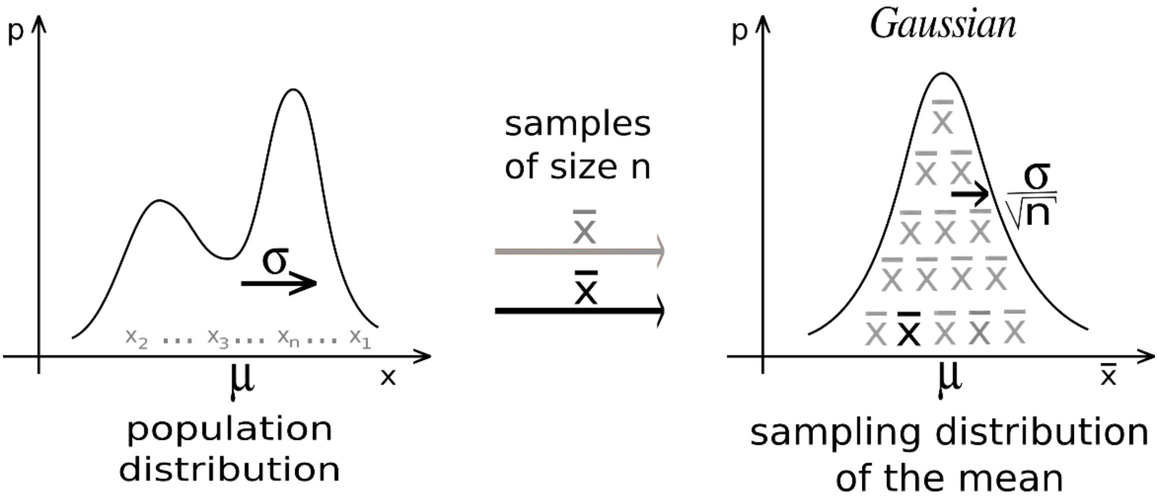
In this study, we consider linear model estimation in presence of severe outliers and suggest bootstap as alternative way of building confidence intervals which does not have theoretical assumptions. We give theoretical background of bootstrap and theoretically explain why it can lead to more accurate than the traditional boostrap interval.

The study in structured with the following sections. Firstly, we will have a look at theoretical background of traditional confidence interval that are based on Central Limit Theorem and explain why it can suffer in presence of large outliers. Secondly, we discuss method of bootstrap and bootstrap as a resampling method. Thirdly, we review how confidence intervals can be derive from bootstrap and how it can improve our estimation in the presence of severe outliers.

Literature review.

Confidence intervals of coefficients provide interval estimates for the regression coefficients. Modern statistical packages mostly provide traditional confidence intervals that rely on Central Limit theorem.

Central Limit theorem (CLM) is a key concept in statistics and econometrics that is widely used on modellings. It states that no matter what distribution your population has, if you get sample averages from relatively large number of identically and independently samples, then the distribution of sample means will be approximately normal or Gaussian (see graph below) (Lind et al, 1967). The center of this normal distribution of sample means will be population mean. This is a very valuable theorem that can applied in point and interval estimations of regression models.


p
σ
x_2 ... x_3 ... x_n ... x_1
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
p
Gaussian
$\frac{\sigma}{\sqrt{n}}$
μ
$\bar{x}$
sampling distribution of the mean

Having only one sample, you can already make some inference about the population parameter using the central limit theorem even when the distribution of population dataset is not known.

Confidence interval based on CLM. If distribution of sample means is normally distributed based on central limit theorem, we can make use of properties of standard normal distribution, namely standard normal distribution (z distribution), and build 90%, 95% or 99% confidence intervals (Tibshirani et al, 2023)

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

where

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

$z_{\frac{\alpha}{2}}$ – is a precalculated statistic from standard normal distribution for any probability area from standard normal distribution

$se(\hat{\beta}_1)$ - sample variance which serves as an unbiased proxy for population variance

We interpret confidence interval in the following way. 95% confidence interval mean that if we build 100 confidence intervals from 100 samples taken from the population, then 95 of those intervals will contain true population parameter β_1. As a results, one can also check whether population parameter is equal to zero by checking whether your estimated confidence interval contains zero (Gujarati, 2012)

Yet, linear models have a set of assumptions that need to be satisfied so that its point and interval estimates will be best unbiased efficient estimates. These assumptions are:

1. No severe outliers
2. No strong multicollinearity between explanatory variables
3. Mean value of error term and its constant variance (no heteroscedasticity)
4. Number of observations must be larger than 30
5. No autocorrelation of the error term (or stationarity)

Violation of any of these assumptions can make our coefficient or interval estimations highly inaccurate or biased (Gujarati, 2012).

In this paper we consider presence of severe outliers, its impact on estimation if no remedy is applied and consideration of bootstrap approach as a way of reducing impact of large outliers.

Severe outliers can be the result of multiple sources, such as measurement error (e.g. few observations are measured in thousands while all should be measured in million, one variable), data entry errors, sampling errors or natural variations.

If no remedy is applied, outliers can lead to heteroscedasticity in residuals, make coefficients biased and distort model accuracy. One of the common approaches of handling outliers are removing them, capping extreme values at certain range/percentiles, removing observations with high z-scores, log transforming or simply leaving them in the estimation as they bear useful information (Greene, 2021)

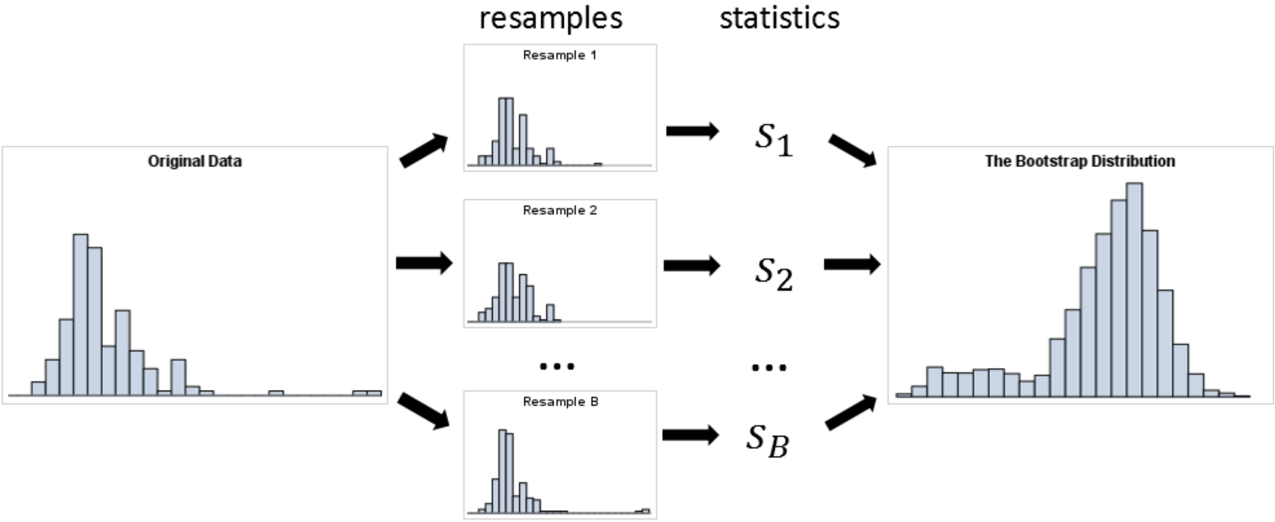
However, sometimes it is not so easy to spot outliers or apply the correct approach of removing or reducing impact of outliers. For this reason, we suggest alternative way of estimating confidence intervals that will reduce impact of severe outliers which we will discuss in next sections.

Methodology.

Bootstrap confidence interval estimation

OLS confidence intervals are relatively easy to estimate and is provided by any statistical package, yet it is necessary to explain the concept of bootstrapping and how it can be used to estimate confidence interval.

Bootstrap is rather a simple resampling methods that can be powerful when applied intelligently. Bootstrap takes one sample and creates distribution of sample estimates by taking creating other samples out of original sample. In other words, bootstrap treat original sample as population and generates many samples out of it. Once a poll of boostrap samples are generated, one can get parameter estimates from each bootstrap sample. As a result, we can have a distribution of bootstrap sample estimates. Taking 2.5th and 97.5th percentiles from this distribution will provide us with 95% confidence interval. Below, you can find visual explanation of bootstrap. There are many types of bootstrap that might suitable for different situations/violations of linear model (heteroscedasticity, outliers, multicollinearity etc). Among them are regular bootstrap, iterated bootstrap, block bootstrap, bootstrap pairs or bootstrap of residuals (Chernick, 2014).



Bootstrap in case of severe outliers

Imagine that we have a random sample where ten percent of data to be severe outliers. That is some observations in response variable is highly dispersed from the remaining observations. Applying bootstrap resampling 1000 times, we will have 1000 bootstrap samples. Afterwards, we can apply z-score filtering and removing samples that have high z-score values in the dataset (instead of removing only observations). Afterwards, we can estimates coefficients of linear models constructed on remaining bootstrap samples that do not contain high z-scores of observations. Lastly, once we have a distribution of coefficient estimates derived from bootstrap samples, we can take 2.5th and 97.5th percentiles to construct our 95 per cent confidence intervals.

There are set of advantages of this approach over traditional method. Firstly, if sample size is smaller than 30 if we remove outliers from the original dataset, bootstrap interval estimation can still be derived. In contrast, traditional method required sample size to be larger than 30 for estimates to be reliable enough. Secondly, by removing samples that contain potential outliers, our distribution of estimates should not be influences by extreme outliers. Lastly, bootstrap distributions of estimates does not have any assumptions of true distribution of population dataset.

Results.

In this section we will present two outcomes of the simulation. One with case of no outliers and the other with 10 per cent of data being as outliers. We will show also how size of confidence intervals change as we grow our sample size.

Correctly specified model

At first we want to see how bootstrap confidence intervals perform compared to traditional OLS

Confidence intervals. We expect that both perform relatively as good since this models satisfies all assumptions of OLS models.

In the first chart below you can see how often true coefficient is falling within the estimated confidence intervals. In case of all OLS assumptions satisfied, we expect true coefficient to fall within estimated confidence intervals in 95 per cent of the cases. 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.

Bootstrap intervals are slightly outperforming traditional OLS intervals due to the fact that bootstrap intervals are simply larger in size across all simulated sample sizes.

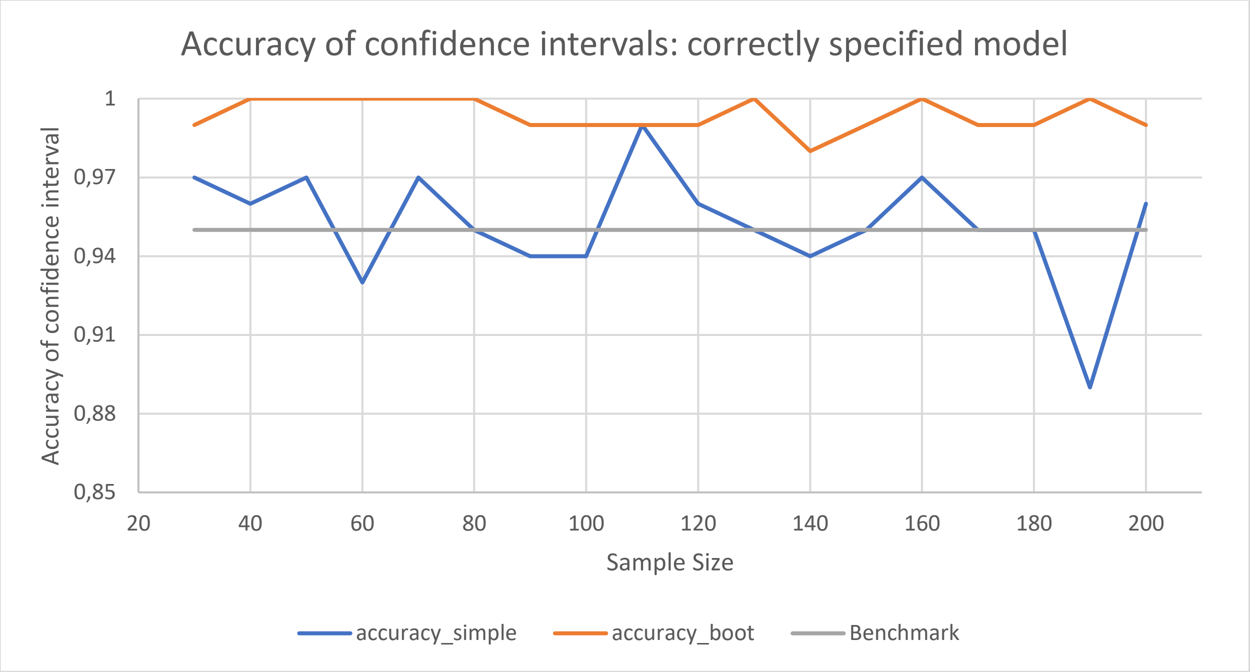


Figure 1. Accuracy of confidence intervals: correctly specified model[75]

Misspecified model: case of bad outliers

As mentioned in previous chapters, we introduce bad outliers by taking first 10 per cent of response variable and multiplying it by 5. At this point we expect traditional and bootstrap intervals still being affected by outliers, but at different degrees.

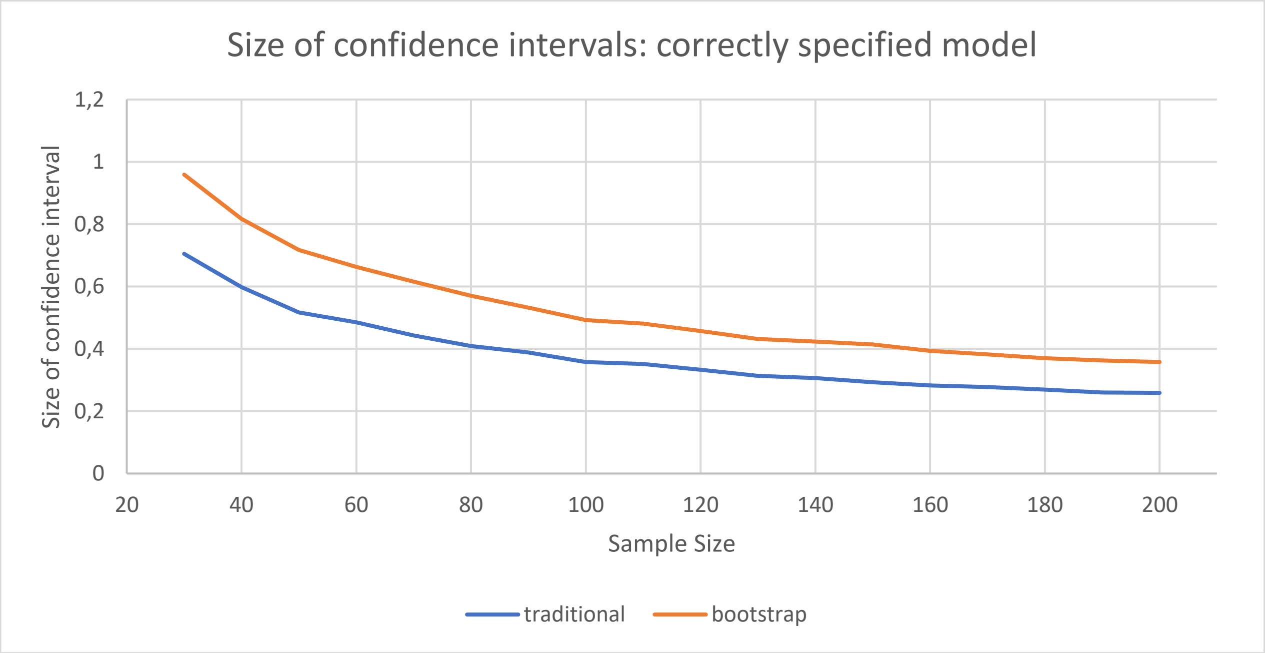


Figure 2. Size of confidence intervals: correctly specified model [76]

In the graph below, you can see that accuracy of traditional OLS confidence interval is far below 95 per cent benchmark especially with large sample size. This means that in presence of bad outliers, OLS confidence intervals will reject the null hypothesis when the null is true more than 5 per cent of the cases. In a similar way, probably of accepting null hypothesis when it is

[75] created by the author
[76] created by the author

false will also be larger than 5 per cent. In line with OLS assumptions, presence of these bad outliers make inferences based on derived confidence intervals inaccurate.

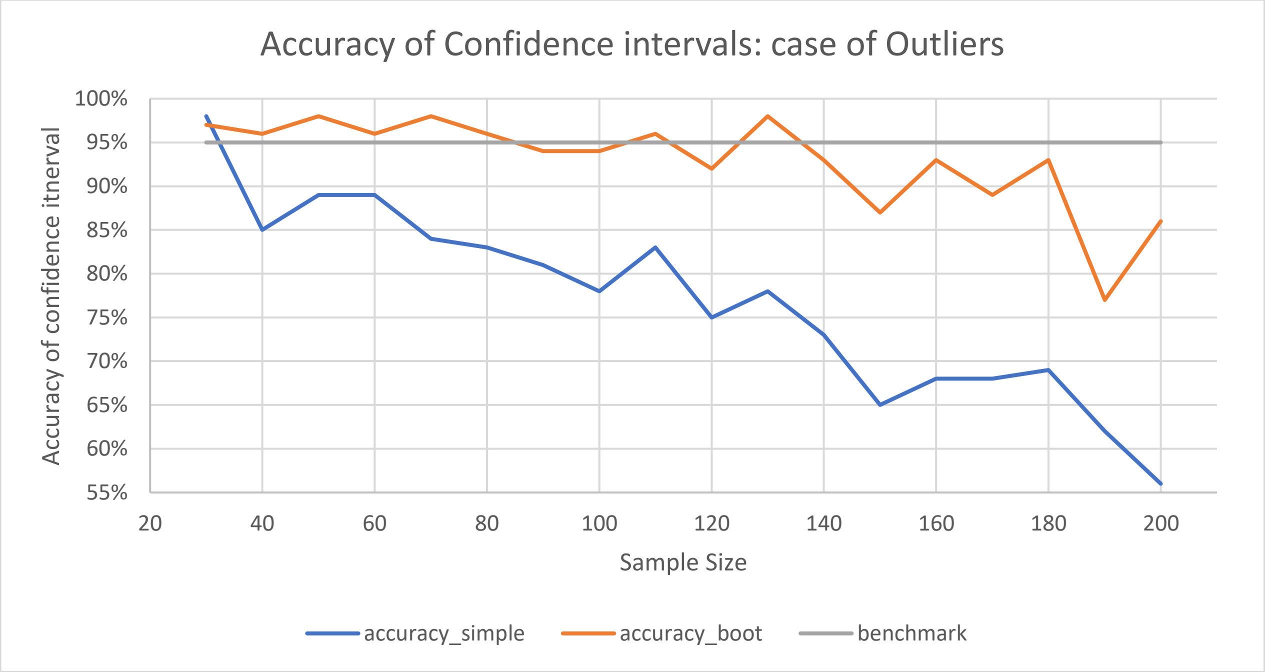


Figure 3. Accuracy of Confidence intervals: case of Outliers[77]

In contrast, accuracy of bootstrap confidence interval is oscillating around 95 per cent benchmark up to sample size of 150. This explained by the fact that number of outliers decrease with double bootstrapping as well as size of intervals increase. As sample size increases over 150, absolute number of outliers are also larger, making chances of getting outliers in iterated bootstrap higher. That explains why accuracy of double bootstrap intervals in sample sizes above 150 are slightly below the benchmark of 95 per cent.

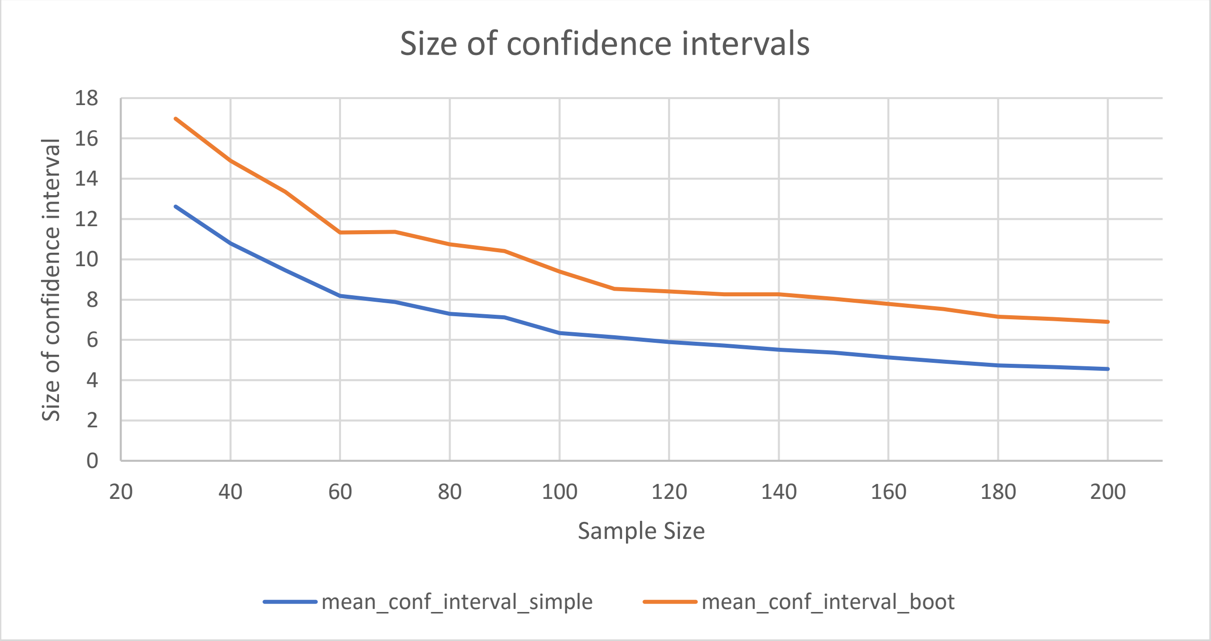


Figure 4. Size of confidence intervals [78]

[77] created by the author
[78] created by the author

As a results, if outliers are difficult to detect or cannot be removed as they carry important information, higher levels of iterations are suggested in future studies when absolute number of outliers are a lot.

Conclusion.

In this study, we considered theoretical side of applying bootstrap confidence intervals in the context of linear models when severe outliers are present. As linear model estimates are prone to inaccuracy when influence by extreme outliers, we suggest applying bootstrap intervals by removing samples with large z-score of some observations. Bootstrap confidence interval can do a better estimation in this context if sample size is small or distribution of population dataset is unknown or non-normal.

As a result, this study revealed a new alternative way of handling with extreme outliers in our dataset when building linear regression. As mentioned in the earlier chapters, one if the way to handle outliers is to remove them. If those outliers carry some useful information, then this study have shown that with the method of bootstrap, we can reduce impact of severe outliers and still have relatively good coverage of confidence intervals. Researchers can use the method shown in this study especially for the cases of small sample which can be often the case when carrying a small survey. Up the sample size of 140 observations, bootstrap confidence intervals have proven to have quite good coverage rate. In other words, bootstrap confidence intervals are include the true population parameter in at least 95 percent of the cases when sample size is up to 140 observations. In case of higher sample size, researchers are advised to consider alternative ways of handling outliers.

Lastly, there are some more topics arising from this study. One should evaluate performance of traditional confidence intervals when outliers are still present and when a remedy is applied. Then bootstrap intervals should be estimated based on the above explained approach. Lastly, all three outcomes should be compared to conclude what approach is performing best in presence of severe outliers. Further areas of research could be to application of bootstrap approach in case of small samples where traditional approach can be sensitive to samples smaller than 30 observations.

Reference:

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

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

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

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

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

Tibshirani, R., Hastie, T., Witten, D., James, G. (2023). An introduction to statistical learning, 2nd Ed. Springer