**Graduate School of Westminster International University in Tashkent**
**Professorship of Statistics and Econometrics**

# Simulation study on double bootstrap confidence intervals in linear models: case of outliers

**Zarrukh Rakhimov**
PhD candidate in Econometrics and Statistics
Email: zrakhimov@wiut.uz
Westminster International University in Tashkent
Istiqbol str. 12, 100047 Tashkent, Uzbekistan

## Abstract

*The marginal effects in linear models have been of considerable interest in social science. Inferences about marginal effects have relied largely on asymptotic methods which have an assumption that the limiting distribution of the estimator is normal. We introduce bootstrap approach as an alternative way to construct confidence intervals and to estimate the sampling distributions of estimators of marginal effect in linear model. We illustrate the performance of traditional method and bootstrap procedure in case of bad outliers. We make use of double bootstrap procedure for confidence interval estimation. Results indicate that double bootstrap confidence intervals outperform traditional OLS intervals in presence of severe outliers in small samples.*
*Key words: Double bootstrap; Simulation study; Lineal Model; Confidence Internal*

## Introduction

Since the introduction of Regression models, social science has heavily relied on its usage. Regression model was first introduced by Francis Galton in his famous papers regarding children’s height. Despite widespread belief that tall parents tend to have tall children and short parents will have short children, he found that average height of children tend or regress towards the mean in the population as a whole (Francis.C, 1886). Current understand of regression changed since then. Regression analysis today studies the impact of one or more variables (referred to explanatory variable hereafter) on another variable (referred to dependent variable hereafter). Regression outcomes are used for two main reasons: for forecasting and for interpretations of impact of one or more variables on another.

The most widespread type of regression model today is linear model. Linear models assume that the relationship of the dependent variable and independent variable is linear. Although linearity assumption is almost always approximation of reality, linear models has proven to be quite good in approximating the relationships. The presence and scale of impact in explanatory variables is revealed via coefficient estimates of the linear model.

However, like almost any model, linear models have a set of assumptions in order to have the coefficients of explanatory variables to be best unbiased linear estimators. One the cases when it might causes estimation issues is presence of outliers. While mild outliers can cause slight inaccuracy of coefficient estimates, severe outliers can lets to bias in estimation. This is especially true when outliers either difficult to spot or cannot be removed as they bear some important information in the study.

In this study we will introduce double bootstrap for estimating confidence intervals of coefficients in linear models. For comparison we will show confidence interval from usual OLS estimates keeping those outliers as well as removing outliers. The paper is structured in the following way. In the next chapter we will review existing papers on this topic. Afterwards, we explain how simulation is carried out together with how outliers are created. Finally, we will look at outcomes of double bootstrap confidence intervals compared to traditional confidence intervals. Lastly, we will present implications of this study together with topic for further studies.

## Literature review

Under correct specification of the model and sample size large enough, OLS estimates are unbiased and have minimum variance among all unbiased estimators (Gujarati, 2012). Yet, the estimates are not robust to the violation of the OLS assumptions such as no severe outliers in the data. Imagine that a regression data contain certain *F* number of outliers. Most of the studies suggest identifying and excluding those outliers before building the regression model. Gentleman and Wilk (1975) suggest identifying *F* number of outliers if they result in largest reduction of Mean Squared errors when excluded. However, it is not always possible to identify outliers for the information they carry or simply difficulty of identifying.

With the improvement of computing powers bootstrap approach is getting larger popularity. Bootstrap is a resampling method very often used to create confidence intervals for a specific statistic. Bootstrap does not require any distribution assumptions about the data which is its biggest advantage over traditional approaches (e.g. CLT). There are two approaches in bootstrapping in the context of linear models, residual bootstrap and bootstrapping pairs (dependent and independent variables) (Chernick, 2011). Each of them fit well in different model specifications. Efron and Tibshirani (1986) claim that bootstrapping pairs perform better than residual bootstrap. Yet, some simulations by Horowitz (2000) show that bootstrapping pairs is not always accurate. Liu (1988), Mammen (1993), Wu (1986) implemented different variations of residual transformation bootstrap (called wild bootstrap) and shown that it performs better than bootstrapping pairs. Stephen (2009) implemented double block (paired) bootstrap in case different autocorrelations and showed that it substantially reduced coverage error. David (1998) showed that higher order iterations do not necessarily result in significant improvements in coverage error and sometimes single bootstrap iterations can already give confidence intervals that reach benchmark coverage of population parameter. Very limited number of studies are carried out on bootstrap approach to handling outliers in the dataset.

It should be noted that almost all studies used Monte Carlo simulation as coverage level can be estimated only by knowing the true parameter. Additionally, all unanimously agree that as sample size gets bigger, iterated bootstrap get computationally expensive and lower level of

iterations is more suitable. In this study, we imitate simulation of Chang (2015) in sample size selections and model evaluations.

## Methodology and Data: Simulation approach

### Simulation of Linear model

We have two solid reason for simulating the data in our analysis. Firstly, we need to know the true marginal effects of the population. In practice we quite rarely know the true parameters and we simply take a representative random sample from the population and make inferences about the population based on that sample. Inference can be accurate or inaccurate depending on the data type and analysis procedure. In this study, it is crucial to know the exact true parameters in order to analyze the properties of confidence intervals based on traditional t-distribution based confidence interval and bootstrap procedure. Secondly, we intend to see the performance of confidence intervals under presence of outliers. Although real data can sometimes have outliers, it usually difficult to exactly define to what extend those outliers are severe. For example, some leverage points can be rather bad that they can make the estimation quite misleading, whereas some leverage points can be good enough not to harm the accuracy of estimation. In practice, it is quite difficult to see what type of leverage points a given real data has unless we know the true parameters. In contrast, simulation allows not only knowing the population parameters, but also creating misspecifications of any choice quickly and easily. Therefore, we use simulation in our study that enables us to generate the whole population of any size and any functional form and to know the true parameters of the model. We start with the design of our population model. We choose the simplest form of true latent model with one explanatory variable.

$Y = {}_0 + \boldsymbol{\beta}_1 * X_1 + \mathcal{E}$

where

$X1 \sim N(5, 4)$

$\mathcal{E} \sim N(0,15)$

where intercept ( $\boldsymbol{\beta}_0$ ) and $\boldsymbol{\beta}_0$ are defined by us. Independent variables ($X_1$) and the error term ($\mathcal{E}$) is generated based on the specification of the model and sample size.

The disturbance term is generated from the standard normal distribution which is the assumption of linear model. Our model with normally distributed explanatory variable and disturbance are intended to mimic cross-sectional data. After the dependent variable $Y$ is calculated based on above equation. Finally, it should be noted that the population model is correctly specified and we introduce misspecifications only in random samples. We analyze the performance of various confidence intervals in different sample sizes. Therefore, we start with sample size of 30 increase by 10 observations up to 200 observations. The sample observations are taken only from the explanatory variable $X_1$ of the population and the values of the dependent variable $Y$ are generated using the same equation as for the population. In random samples, the only source of randomness is $\mathcal{E}$. It means that the values of sample explanatory variables are the same, but the values of disturbance term differ across all replications of random samples. As outliers are introduced, we estimates their confidence interval using bootstrap method and traditional method that uses standard errors of estimated coefficient. Finally we evaluate performance of both approaches of estimating confidence intervals. All of the simulations are carried out in R software.

**We take the following steps for simulation of linear model with outliers 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 $X1 \sim N(5, 4)$ starting with sample size n
Step 4: generate $\mathcal{E} \sim N(0,15)$ with size n
Step 5: generate $Y$ with $Y = 4 + 5 * X_1 + \mathcal{E}$
Step 6: replace few values of $Y$ with bad outliers. In our case, multiply 10 percent of by 5.
Step 7: estimate confidence intervals using traditional and bootstrap methods in repeated simulations (1000 times)
Step 8: evaluate how many times (out of 1000), true parameters were within estimated OLS and bootstrap confidence intervals
Step 9: repeat step 2 to step 8 by adding 10 observations to sample size (n=n+10). Finish when sample size reaches 200 observations

**Traditional confidence interval estimation**

Central Limit Theorem states that the coefficient of linear model if estimated in repeated sampling follows normal distribution. Therefore, we can make use of standard normal z-distribution to make probabilistic conclusions about $\boldsymbol{\beta}_1$ given that we know the true population variance $\sigma^2$ of coefficient.

However, we rarely know true population variance $\sigma^2$ and it is estimator by an unbiased sample variance $\hat{\sigma}^2$. And instead of using standard normal distribution, we can use of t distribution which closely imitates z-distribution (Gujarati, 2004).

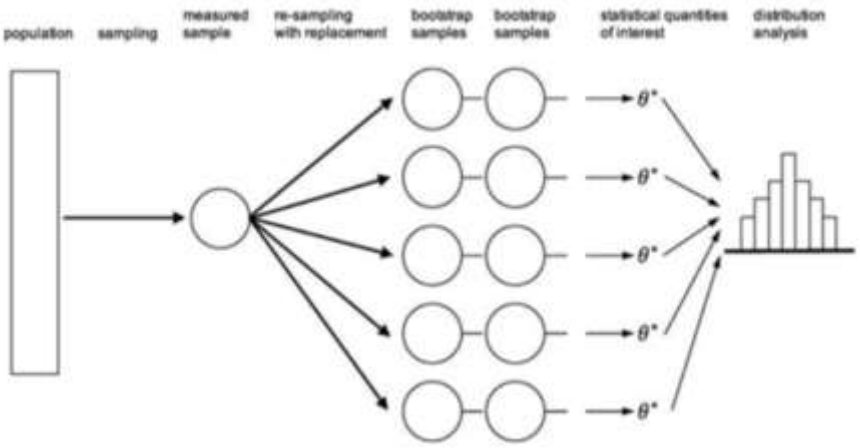
$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} * se(\hat{\beta}_1)$$

This confidence interval is currently provided by almost any statistical package that runs linear models.

**Bootstrap confidence interval estimation**

While traditional OLS confidence interval estimation is relatively easy to understand and given in almost any regression packages, it is necessary for us to explain bootstrap approach of getting confidence intervals.

Bootstrap is a simple but powerful resampling methods that generates distribution of parameter estimates out of a single sample. Taking 2.5th and 97.5th percentiles from the resulting distribution will provide us with 95% confidence interval. Below, a simple illustration is provided. There are many variations of bootstrap that might suitable for different cases (heteroscedasticity, outliers, multicollinearity etc). Among them are double bootstrap, bootstrap of residuals, block bootstrap, bootstrap pairs (Chernick, 2011).



In our study, we will introduce double bootstrap which we consider should decrease our exposure to outliers. This is coming from the logit that if 10 percent of data is outliers, resampling that dataset twice will reduce exposure to outliers.

## 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.

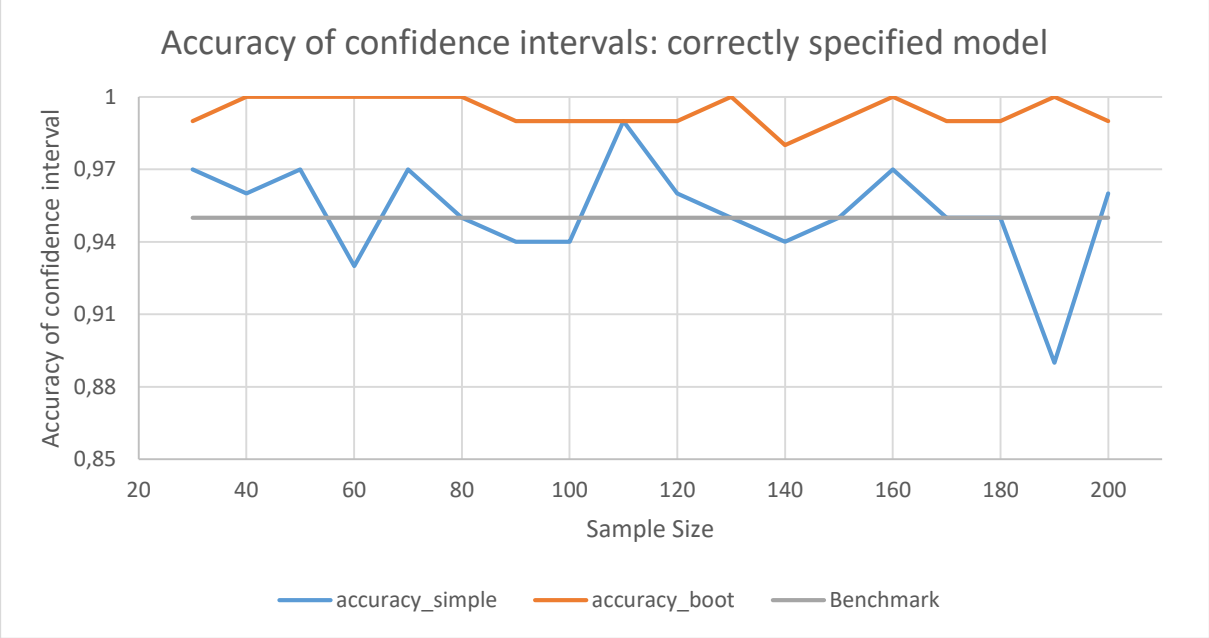
**Correctly specified model**

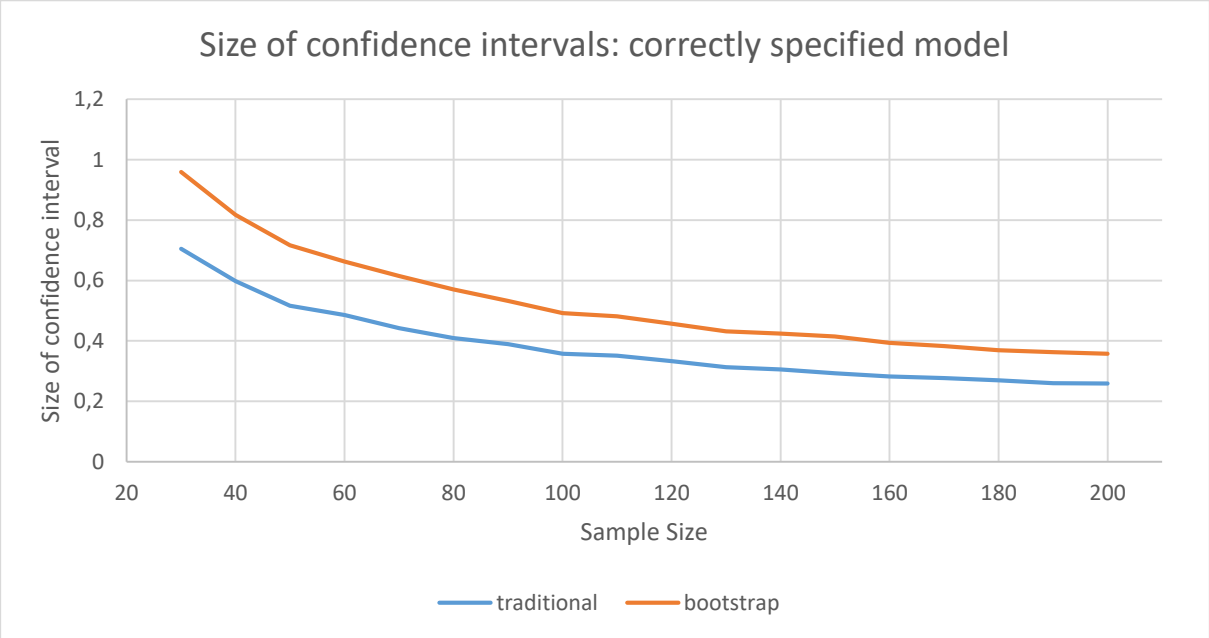
At first we want to see how bootstrap confidence intervals perform compared to traditional OLS

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 traditionals OLS intervals due to the fact that bootstrap intervals are simply larger in size across all simulated sample sizes.



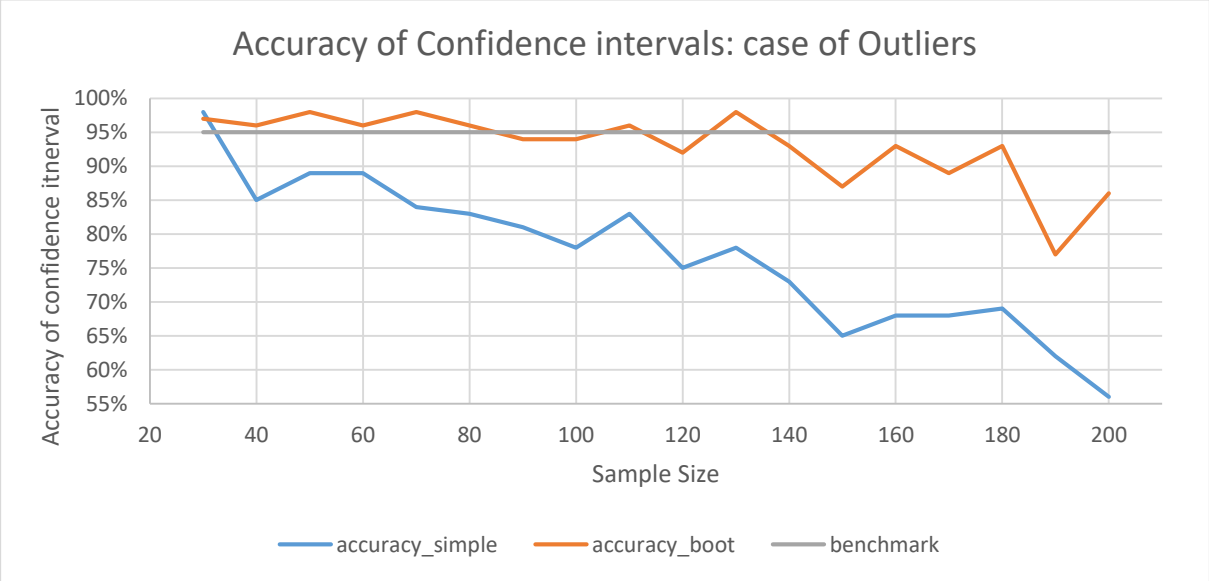

**Misspecified model: case of bad outliers**

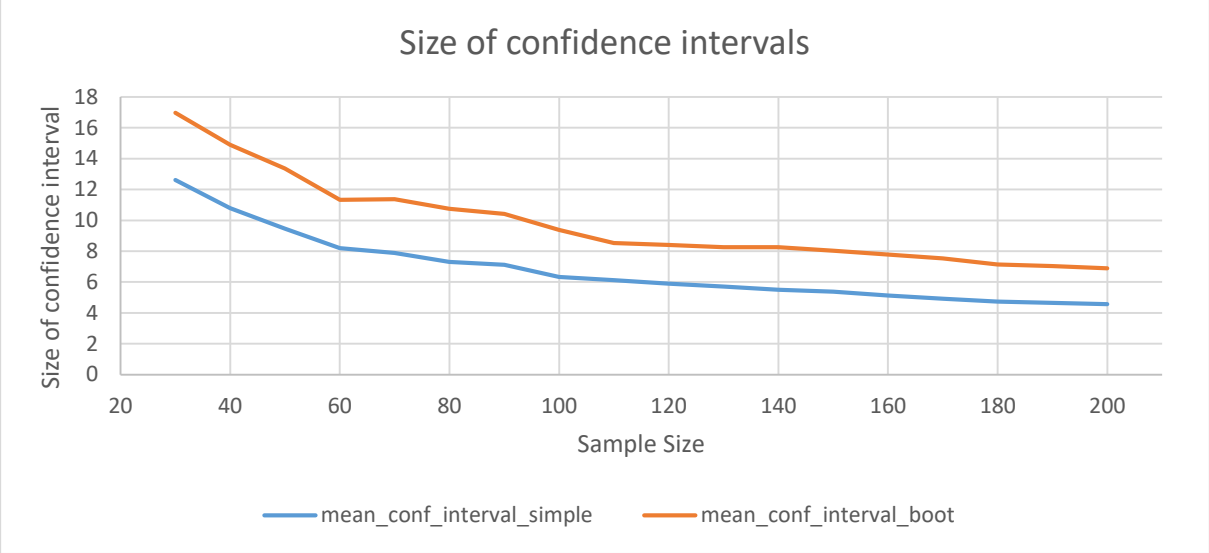
As mentioned in previous chapters, we introduce bad outliers by taking first 10 per cent of response variable and multiplying it by 5. At this point we expect traditional and bootstrap intervals still being affected by outliers, but at different degrees.

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 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.

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.

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 paper, we investigated usage of double bootstrap as an alternative way of handling outliers in cases when they are difficult to identify or carry important information. We fist reviewed performance of traditional OLS confidence intervals compared double bootstrap intervals in case of correctly specified model. As expected, both approaches performed well having accuracy (measured on how often true parameter is within the estimated intervals) above the benchmark 95 per cent under no outliers scenario. Afterwards, we introduced outliers in the response variable. Results indicate that traditional OLS intervals heavily suffer from outliers as accuracy/coverage rate are below benchmark 95 per cent. In contrast, double bootstrap intervals' coverage rate oscillate around benchmark level with different sample size. This is explained by increased size of bootstrap interval.

## Bibliography

Carroll , R. J. , and Ruppert , D. ( 1988 ). Transformations and Weighting in Regression. Chapman & Hall , New York

Chang, J., Hall, P. (2015). Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, vol. 102, pp. 203-214

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

David, L, McCullough, B. (1998). Better Confidence Intervals: The Double Bootstrap with No Pivot, *American Journal of Agricultural Economics,* Vol. 80, pp. 552-559

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

Francis, G (1886), “Family Likeness in Stature,” *Proceedings of Royal Society, London*, vol. 40, 1886 , pp. 42–72.

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer Science & Business Media.

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

Stephen, M, Lee, S., Lai, P. (2009). Double bootstrap confidence intervals for dependent data, *Biometrika*, vol. 96, pp. 427-443

Mammen , E. ( 1993 ). Bootstrap and wild bootstrap for high dimensional linear models . *Annals of Statistics* 21 , 255 – 285

Wu , C. F. J. ( 1986 ). Jackknife, bootstrap and other resampling plans in regression analysis (with discussion) . *Annals of Statistics* 14 , 1261 – 1350