IQTISODIY TARAQQIYOT VA TAHLIL
ILMIY ELEKTRON JURNAL

UO‘K: 330.43

OLS CONFIDENCE INTERVALS IN NON-LINEAR MODELS: BOOTSTRAP APPROACH

PhD ***Rakhimov Zarrukh Aminovich***
Westminster International University in Tashkent
ORCID: 0009-0001-0583-4819
zrakhimov@wiut.uz

Abstract. *Linear models has been a powerful econometric tool used to show the relationship between two or more variables. Many studies also use linear approximation for nonlinear cases as it still might show valid results. OLS method requires the relationship of dependent and independent variables to be linear, although many studies employ OLS approximation even for nonlinear cases. In this study, we are introducing alternative method of intervals estimation, bootstrap, in linear regressions when the relationship is nonlinear. We compare the traditional and bootstrap confidence intervals when data has nonlinear relationship. As we need to know the true parameters, we carry out a simulation study. Our research findings indicate that when error term has non-normal shape, bootstrap interval will outperform the traditional method due to no distributional assumption and wider interval width.*

Key words: *OLS, nonlinear model, sample size, confidence Interval, bootstrap, accuracy, interval size, variance.*

CHIZIQLI BO'LMAGAN MODELLARDA OLS ISHONCH INTERVALLARI: BOOTSTRAP YONDASHUV

PhD ***Raximov Zarrux Aminovich***
Toshkent Xalqaro Vestminster Universiteti

Аннотация. *Чизиқли моделлар икки ёки ундан ортиқ ўзгарувчилар ўртасидаги муносабатни кўрсатиш учун ишлатиладиган кучли економетрик восита бўлган. Кўпгина тадқиқотлар, шунингдек, чизиқли бўлмаган ҳолатлар учун чизиқли яқинлашувдан фойдаланади, чунки у ҳали ҳам ҳақиқий натижаларни кўрсатиши мумкин. ОЛС усули боғлиқ ва мустақил ўзгарувчилар муносабатларининг чизиқли бўлишини талаб қилади, гарчи кўплаб тадқиқотлар ҳатто чизиқли бўлмаган ҳолатлар учун ҳам ОЛС яқинлашувидан фойдаланади. Ушбу тадқиқотда биз чизиқли регрессияларда, агар муносабатлар чизиқли бўлмаса, интервалларни баҳолашнинг муқобил усули, юклаш усулини киритамиз. Маълумотлар чизиқли бўлмаган муносабатларга ега бўлса, биз анъанавий ва юклаш ишонч оралиқларини солиштирамиз. Ҳақиқий параметрларни билишимиз кераклиги сабабли, биз симуляция тадқиқотини ўтказамиз. Тадқиқот натижаларимиз шуни кўрсатадики, агар хато атамаси ноодатий шаклга ега бўлса, юклаш оралиғи тақсимот тахмини ва кенгроқ интервалли кенглиги туфайли анъанавий усулдан устун бўлади.*

Калит сўзлар: *ОЛС, чизиқли бўлмаган модел, намуна ҳажми, ишонч оралиғи, юклаш чизиғи, аниқлик, интервал ўлчами, дисперсия.*

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ OLS В НЕЛИНЕЙНЫХ МОДЕЛЯХ: МЕТОД БУТСТРАПА

PhD ***Рахимов Заррух Аминович***
Международный Вестминстерский Университет в Ташкенте

Аннотация. *Линейные модели стали мощным эконометрическим инструментом, используемым для демонстрации взаимосвязи между двумя или более переменными. Многие исследования также используют линейную аппроксимацию для нелинейных случаев, поскольку она все еще может дать достоверные результаты. Метод МНК требует, чтобы отношения зависимых и независимых переменных были линейными, хотя во многих исследованиях используется приближение МНК даже для нелинейных случаев. В этом исследовании мы представляем альтернативный метод оценки интервалов, бутстрап, в линейных регрессиях, когда взаимосвязь нелинейна. Мы сравниваем традиционные доверительные интервалы и доверительные интервалы начальной загрузки, когда данные имеют нелинейную зависимость. Поскольку нам необходимо знать истинные параметры, мы проводим моделирование. Результаты нашего исследования показывают, что, когда член ошибки имеет ненормальную форму, бутстрап-интервал превосходит традиционный метод из-за отсутствия предположения о распределении и более широкой ширины интервала.*

Ключевые слова: *МНК, нелинейная модель, размер выборки, доверительный интервал, бутстрап, точность, размер интервала, дисперсия.*

Introduction.

Ordinary Least Squares (OLS) regression is a cornerstone of statistical analysis. It establishes a linear relationship between a dependent variable and one or more independent variables, allowing us to estimate the effect of changes in the independent variables on the dependent variable. However, OLS relies on several key assumptions to ensure the validity of its results. One of the most critical assumptions is linearity in the relationship between the independent and dependent variables. This implies that the change in the dependent variable is constant for each unit change in the independent variable.

In many real-world scenarios, however, the relationship between variables is not perfectly linear. This limitation can lead to biased estimates and unreliable conclusions when using OLS for non-linear models. For instance, imagine analyzing the effect of fertilizer on crop yield. Initially, as fertilizer increases, crop yield might rise proportionally. But beyond a certain point, adding more fertilizer might have diminishing returns or even negative effects. A linear model would not capture this complexity.

This paper explores the use of the bootstrap approach to construct confidence intervals for parameters estimated in non-linear models. Confidence intervals provide a range of plausible values for a population parameter, with a certain level of confidence (e.g., 95%). While traditional methods for constructing confidence intervals in non-linear models can be complex and rely on specific assumptions about the error distribution, the bootstrap offers a more robust and flexible alternative.

The bootstrap method is a resampling technique that utilizes the data you already have to create new datasets (called bootstrap samples). By analyzing these bootstrap samples, we can estimate the sampling distribution of the parameter estimates, which allows us to construct reliable confidence intervals even for non-linear models.

This chapter outlines the limitations of OLS for non-linear models, introduces the concept of confidence intervals, and emphasizes the importance of the linearity assumption in OLS. It then highlights the advantages of the bootstrap approach for constructing confidence intervals in scenarios where the linearity assumption is violated. The subsequent chapters will delve deeper into the details of the bootstrap method, its application to non-linear models, and the interpretation of the resulting confidence intervals.

Literature review.

Limitations of OLS for Non-Linear Models: Ordinary Least Squares (OLS) regression remains a fundamental tool for statistical analysis. However, its core assumption of linearity between the independent and dependent variables can lead to significant drawbacks when applied to non-linear models (Weisberg, 2014). When the true relationship is non-linear, OLS estimates become biased and unreliable (Fox, 2016). This bias arises because the linear model fails to capture the true curvature or non-linear trend in the data (Montgomery and Chatterjee, 2015).

Consequently, relying solely on OLS for non-linear models can lead to misleading interpretations of the relationships between variables and inaccurate predictions (Pindyck and Rubinfeld, 2013). For example, imagine analyzing the effect of temperature on ice cream sales. An OLS model might suggest a constant increase in sales with rising temperature. However, in reality, sales might peak at a certain temperature as consumers turn to other options in extreme heat.

Confidence Intervals and Their Importance: Confidence intervals are a crucial aspect of statistical inference. They provide a range of plausible values for a population parameter, with a specific level of confidence (e.g., 95%). In regression analysis, confidence intervals are typically constructed around the estimated coefficients of the model (Faraway, 2014). These intervals help us assess the precision of our estimates and the degree of uncertainty associated with them.

Narrower confidence intervals indicate more precise estimates, while wider intervals suggest greater uncertainty. By interpreting confidence intervals alongside the estimated coefficients, we can gain valuable insights into the statistical significance of the relationships between variables (Gamerman and Lopes, 2006).

Challenges of Constructing Confidence Intervals in Non-Linear Model: Traditional methods for constructing confidence intervals in non-linear models can be complex and rely on specific assumptions about the error distribution (Wu, 2004). For instance, delta methods or likelihood-ratio tests often require normality of the error terms, which might not always hold true in non-linear models (Harvey, 2013). Additionally, these methods can be computationally intensive, especially for intricate non-linear relationships.

The Bootstrap Approach: A Viable Alternative: The bootstrap method offers a robust and flexible alternative for constructing confidence intervals in non-linear models (Efron and Tibshirani, 1993). It is a resampling technique that leverages the data you already have to create new datasets, called bootstrap samples. These samples are generated by drawing observations from the original data with replacement, meaning an observation can be selected multiple times in a single bootstrap sample (Efron and Tibshirani, 1994).

By repeatedly fitting the non-linear model to these bootstrap samples and obtaining the corresponding parameter estimates, the bootstrap approach allows us to estimate the sampling distribution of the parameter estimates. This sampling distribution reflects the variability of the estimates had we collected different samples from the same population (Davison and Hinkley, 1997).

Once the sampling distribution is obtained, we can calculate percentiles to construct confidence intervals. For instance, the 2.5th and 97.5th percentiles of the bootstrap distribution would provide a 95% confidence interval for the parameter estimate (Efron and Tibshirani, 1998).

The beauty of the bootstrap lies in its ability to bypass the need for specific assumptions about the error distribution. It relies solely on the data itself, making it a more generalizable approach for constructing confidence intervals in non-linear models (Leek, 2010).

OLS Confidence Intervals and Limitations in Non-Linear Models

The Power of OLS Regression: Ordinary Least Squares (OLS) regression remains a workhorse in statistical analysis. It establishes a linear relationship between a dependent variable (Y) and one or more independent variables (X), allowing us to estimate how changes in X impact Y. The core of OLS lies in minimizing the squared deviations between the actual Y values and the Y values predicted by the fitted line. This minimization process yields estimates for the slope (β_1) and intercept (β_0) coefficients of the linear model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Here, ε represents the random error term, capturing the unexplained variation in Y. OLS assumes ε follows a normal distribution with a mean of zero and constant variance (homoscedasticity).

Confidence Intervals: Quantifying Uncertainty: A crucial aspect of regression analysis is constructing confidence intervals (CIs) for the estimated coefficients. These CIs provide a range of plausible values for the population coefficient, with a specific level of confidence (e.g., 95%). In OLS, CIs are typically constructed around the estimated slope (β_1) and intercept (β_0).

The width of the CI reflects the precision of the estimates. Narrower intervals indicate more precise estimates, with a higher degree of confidence that the true population coefficient falls within that range. Conversely, wider CIs suggest greater uncertainty about the true coefficient value. Interpreting CIs alongside the estimated coefficients provides valuable insights into the statistical significance of the relationships between variables.

Non-Linear Relationships: While OLS delivers powerful insights in linear scenarios, its core assumption of linearity can become a significant limitation when the true relationship between X and Y is non-linear. In non-linear models, the linear model fails to capture the true curvature or non-linear trend in the data. This can lead to several issues:

Biased Estimates: When the relationship is non-linear, the OLS estimates (β_0 and β_1) become biased. This means the estimates are systematically skewed away from the true population values. Imagine analyzing the effect of advertising spending on sales. A non-linear relationship might exist where initial advertising has a high impact, but the effect plateaus or even diminishes with further spending. An OLS model would likely underestimate the initial impact and overestimate the effect of higher spending levels.

Unreliable Inferences: Biased estimates lead to unreliable inferences about the relationships between variables. Significance tests based on OLS might incorrectly suggest a statistically significant relationship when none exists, or vice versa. This can lead to misleading conclusions about the true impact of X on Y.

Limited Generalizability: OLS predictions based on a non-linear relationship are only reliable for the range of X values observed in the data. Extrapolating beyond this observed range can lead to inaccurate predictions, as the linear model doesn't capture the true underlying trend.

Visualizing the Challenges: A Worked Example: Consider the scenario where we want to analyze the effect of fertilizer application (X) on crop yield (Y). Imagine the true relationship is a quadratic curve, where initially yield increases with fertilizer application, but eventually reaches a peak and starts to decline due to over-fertilization.

An OLS model would fit a straight line to this data. This line might intersect the true curve at two points, potentially leading to misleading interpretations. The estimated slope might suggest a positive relationship throughout the observed range, even though the true effect plateaus and then becomes negative.

Furthermore, OLS CIs constructed around the slope estimate would not accurately reflect the true uncertainty in the non-linear relationship.

The Need for Alternative Approaches: The limitations of OLS in non-linear models highlight the need for alternative approaches that can provide more reliable estimates and CIs. The next chapter will explore the bootstrap method, a robust and flexible technique for constructing CIs in non-linear models, even when the specific form of the non-linearity is unknown. The bootstrap approach offers a valuable tool for researchers and analysts to gain deeper insights from data exhibiting non-linear relationships.

The Bootstrap: A Resampling Rescue: The limitations of OLS confidence intervals (CIs) in non-linear models necessitate alternative approaches. The bootstrap method emerges as a powerful and flexible technique for constructing reliable CIs in these scenarios. Unlike traditional methods, the bootstrap does not rely on specific assumptions about the error distribution or the form of the non-linearity (Efron and Tibshirani, 1993).

The core idea of the bootstrap revolves around resampling the data you already have. Here's how it works:

Sample with Replacement: We create new datasets, called bootstrap samples, by drawing observations from the original data with replacement. This means an observation can be selected multiple times in a single bootstrap sample. The size of the bootstrap sample is usually equal to the size of the original data.

Repeat and Estimate: This resampling process is repeated a large number of times (e.g., 1000 times). For each bootstrap sample, we fit the non-linear model and obtain the corresponding estimates for the model coefficients.

Distribution of Estimates: By repeating step 2 numerous times, we generate a distribution of the estimates for each coefficient (e.g., the slope and intercept). This distribution reflects the variability of the estimates had we collected different samples from the same population.

Constructing Bootstrap CI: Once we have the distribution of the estimates from the bootstrap samples, we can construct CIs for the non-linear model coefficients. Here's a common approach:

Percentile Method: We identify the percentiles of the bootstrap distribution that correspond to the desired confidence level (e.g., 2.5th and 97.5th percentiles for a 95% CI). These percentiles represent the lower and upper bounds of the CI for the coefficient.

For instance, the 2.5th percentile of the bootstrap distribution for the slope estimate (β_1) would be the value below which 2.5% of the bootstrap estimates lie. Similarly, the 97.5th percentile would be the value above which only 2.5% of the estimates fall. The resulting interval represents the range of values within which we are confident (e.g., 95%) the true population value for β_1 lies.

Advantages of Bootstrap CIs in Non-Linear Models: The bootstrap approach offers several advantages for constructing CIs in non-linear models:

Robustness: Unlike traditional methods, the bootstrap doesn't require specific assumptions about the error distribution or the form of the non-linearity. It relies solely on the data itself, making it a more generalizable approach. (Efron and Tibshirani, 1994)

Flexibility: The bootstrap can be applied to various non-linear models, regardless of their complexity. This makes it a versatile tool for researchers working with diverse datasets. (Davison & Hinkley, 1997)

Interpretability: Bootstrap CIs can be easily interpreted alongside the estimated coefficients, providing a clear picture of the uncertainty associated with the estimates in a non-linear context. (Wasserman, 2004)

Applying Bootstrap CIs to Our Fertilizer Example: Recall our example where the true relationship between fertilizer application (X) and crop yield (Y) is non-linear (quadratic). OLS CIs would be unreliable in this scenario. However, the bootstrap method can be applied as follows: We would draw bootstrap samples from the original data on fertilizer application and crop yield. For each bootstrap sample, we would fit a non-linear model (e.g., a quadratic function) to estimate the coefficients. By repeating this process a large number of times, we would obtain a distribution of the estimated coefficients for the non-linear model (e.g., the intercept and the coefficient for the quadratic term). Using percentiles from this bootstrap distribution, we could construct CIs for each coefficient. These CIs would reflect the uncertainty in the estimates due to the non-linear relationship.

By interpreting the bootstrap CIs alongside the estimated coefficients, we can gain a more accurate understanding of the effect of fertilizer application on crop yield. For instance, a wide CI for the coefficient of the quadratic term might suggest that the exact peak yield and the point of diminishing returns are uncertain due to the non-linearity in the data.

Methodology: simulation.

In this part, we will look into simulation of a regression model with non-normal error term. Simulation is necessary in our study, as we need to know the true parameter in the first place. Using real live data will almost never allow us to know the true population coefficients. Secondly, we need to control the form of non-normality we are introducing in our case. Our simulation starts with the simplest form of linear model with one explanatory variable as given below

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 * X1 + \mathcal{E}$$

where

$$X1 \sim N(5, 4)$$

$$\mathcal{E} \sim N(0, X1/2)$$

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 has mean of 0 and variance dependent of X_1 variable. This way we simulate non normal distribution of error term known as heteroscedasticity.

Afterwards, we construction confidence intervals using both approaches, traditional and bootstrap ones. In order to evaluate the performance at difference sample size, first we start with 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 $X1 \sim N(5, 4)$ starting with sample size n

Step 4: generate $\mathcal{E} \sim N(0, X1/2)$ starting with sample size n

Step 5: generate Y with $Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 * X1 + \mathcal{E}$

Step 6: Estimate confidence intervals using traditional and bootstrap methods in repeated simulations (1000 times). Here we construction 95 percent confidence intervals

Step 7: evaluate how many times (out of 1000), true parameters were within estimated OLS and bootstrap confidence intervals

Step 8: 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.

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

In case of bootstrap confidence intervals, we get our bootstrap 95 per cent intervals by taking 2.5th and 97.5th percentiles from 1000 estimated bootstrap coefficients. This kind of bootstrap approach is known in literature as bootstrapping pairs.

Results.

In this section, we will be looking at the results of our simulation carried out in R. In Figure 1, one can see non-normal behavior of the error term with respect to *X1* explanatory variable. This gives a visual image of the violation of the one of the assumptions of the linear model, i.e. constant variance of the error term.

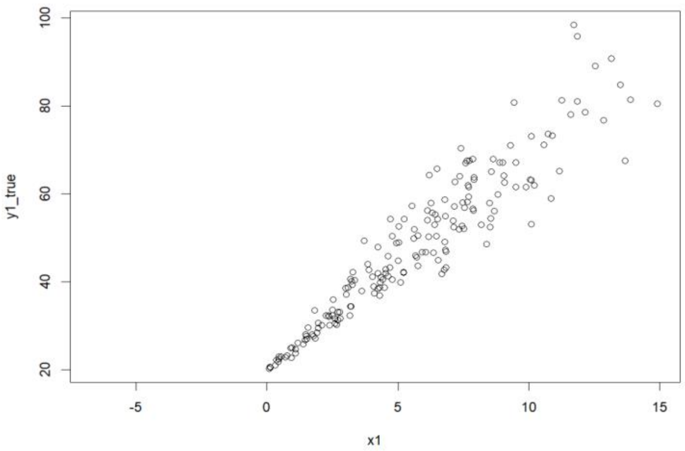

Figure1: non-normal variance of the error term in the linear model

We now look at the performance of the confidence intervals constructed using traditional and bootstrap approaches. In order to evaluate the performance, we calculate how often the true coefficient from our simulation was falling within the given interval. Ideally, the true coefficient must fall in at least 95 per cent of simulated cases. We will be terming frequency of true coefficient falling within our estimated intervals as accuracy.

The results in Figure 2 indicate that accuracy of traditional confidence intervals are below our benchmark of 95 per cent across all sample sizes (from 30 to 200). In other works, traditional approach is suffering a lot from the violations of the OLS assumption. In contrast, bootstrap confidence intervals at above 95 per cent benchmark, meaning that bootstrap intervals are including true coefficients in at least 95 per cent of cases.

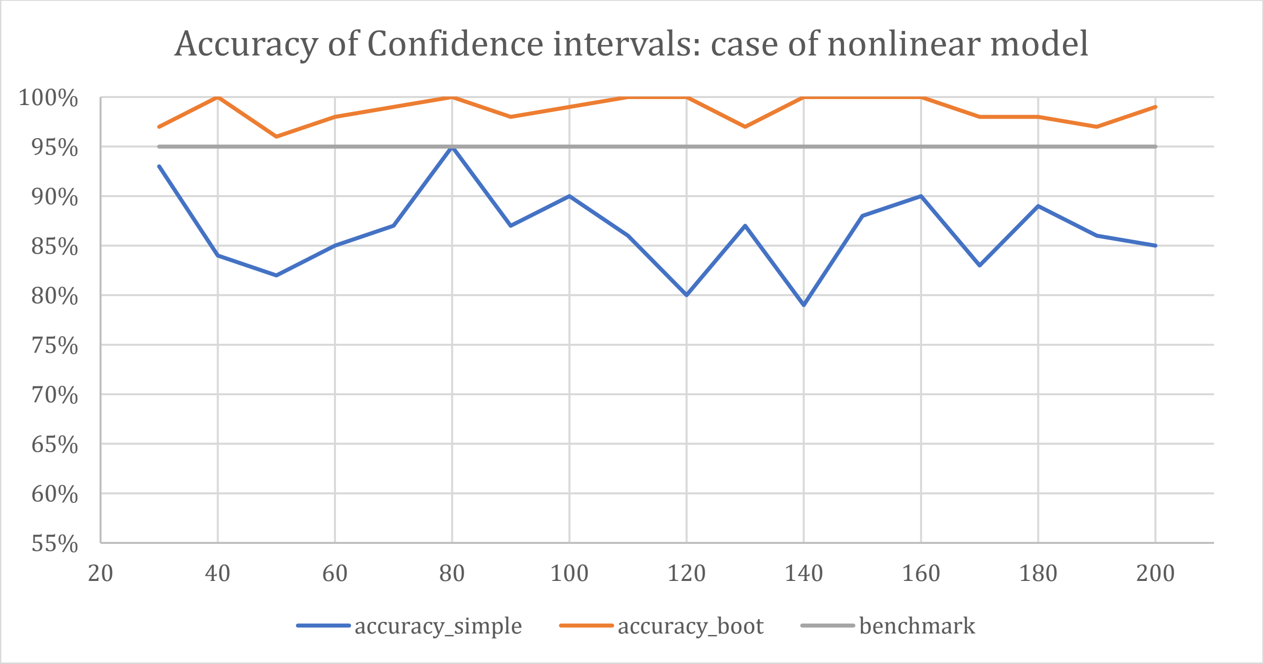

Figure 2. Accuracy of Confidence intervals: case of nonlinear model

One of the explanations of better performance of bootstrap method lies in the fact that this approach does not have any distributional or OLS assumptions in contrast to traditional method. Second reason of stronger performance is explained with Figure 3. One can see that the size of bootstrap intervals are simply larger which increases changes of including true coefficients.

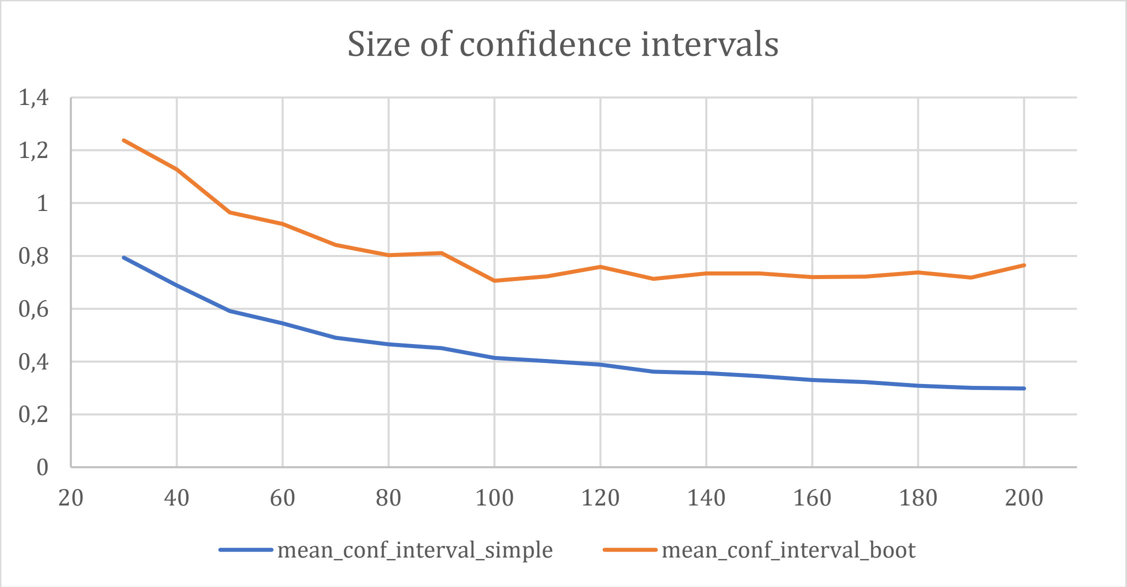

Figure 3. Size of confidence intervals

We should note here that bootstrap method might not always outperform traditional methods even in case of nonlinear models or violations of OLS assumptions. Our study show that specifically when error term has non-constant variance, bootstrap method is showing a more promising outcome. Thus, researchers are highly encouraged to consult with related papers that is suitable in their specific case before applying traditional or bootstrap methods.

Conclusion.

This paper looked into cases when data is nonlinear distributed in error term and investigated at two ways of confidence interval estimations of coefficients of OLS regression. In the first chapter, we gave an introductory guide to our topic explaining the relevance and applicability of our study. Afterwards, we revised related literature on topic of using OLS approach with nonlinear data and looked into possible pitfalls. Revision of existing papers indicate that there is limited literature on application of bootstrap approach in case of nonlinear data. Following this chapter, we looked into theory of ordinary least squares and traditional approach of constructing confidence intervals. In the same chapter, we also presented how bootstrap confidence intervals are applicable in this case and how they are estimated. We have employed bootstrapping pair approaches that does not have any distributional assumptions. In order to evaluate the performance, we need to know the true parameters. Therefore, we carried out a simulation of a simple linear model with one explanatory variable. In order to evaluate performance of both approaches we simulated our regression with error term that has different variance across independent variable. We simulated our model with different sample size, spanning from 30 to 200 observations. Our simulation indicates that traditional approach suffers from introduced nonlinearity as its 95 percent intervals include true coefficient in less than 95 per cent of the cases. In contrast, bootstrap method has been steadily performing at above 95 per cent accuracy across all sample sizes (from 30 to 200 observations). This indicates that bootstrap approach is performing better than traditional approach due to two core reasons. Firstly, bootstrap method has no distributional assumption in contrast to OLS method.

Secondly, our simulation showed that bootstrap intervals are usually wider in size increasing the chances of including the true coefficient.

We need to mention that future researcher should use this method with care, as it might not be suitable for all nonlinear cases. Thus, they are recommended to look into studies that are related to their case.

Reference:

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

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

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. J. (1993). An introduction to the bootstrap. Chapman and Hall/CRC.

Efron, B., and Tibshirani, R. J. (1994). An introduction to the bootstrap (Vol. 57). Chapman and Hall/CRC.

Efron, B., and Tibshirani, R. J. (1998). The art of statistical learning (Vol. 1). Springer.

Faraway, J. J. (2006). Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models. Chapman and Hall/CRC.

Faraway, J. J. (2014). Linear models and extensions (Vol. 14). Springer.

Fox, J. (2016). An R companion to applied regression (3rd ed.). Sage Publications.

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

Gamerman, J. A., and Lopes, H. F. (2006). Markov chain Monte Carlo: Stochastic simulation for Bayesian inference (2nd ed.). Chapman and Hall/CRC.

Harvey, D. I. (2013). Linear regression analysis for count data (2nd ed.). John Wiley and Sons.

Montgomery, D. C.,and Chatterjee, S. (2015). Design and analysis of experiments (8th ed.). John Wiley and Sons.

Pindyck, R. S., and Rubinfeld, D. L. (2013). Econometric models and economic forecasts (5th ed.). Pearson Education.

Wasserman, L. (2004). All of statistics: A concise course with applications. Springer.

Weisberg, S. (2014). Applied linear regression (4th ed.). John Wiley and Sons.

Wu, C. F. J. (2004). Asymptotic theory for nonlinear regression. Chapman and Hall/CRC.