IQTISODIY TARAQQIYOT VA TAHLIL ILMIY ELEKTRON JURNAL

UO'K: 330.43

МАЪЛУМОТЛАР ТАСОДИФИЙ ЙЎҚОЛГАН ЧИЗИҚЛИ РЕГРЕССИЯ: БООТСТРАП ЁНДАШУВИ

PhD ***Рахимов Заррух Аминович***
Тошкент Халкаро Вестминстер Университети
ORCID: 0009-0001-0583-4819
zrakhimov@wiut.uz
Рахимова Нилуфар Аминовна
«Ипак йўли» туризм ва маданий мерос халқаро университети
ORCID: 0000-0002-8648-2543
nrahimova@wiut.uz

Аннотация. *ОЛС регрессиялари нуқта ва интервалларни холис ва самарали баҳолаш учун бир қатор фаразларга эга. Тасодифий йўқолган маълумотлар (MNAR) чизиқли регрессияни баҳолашда жиддий муаммоларни келтириб чиқариши мумкин. Ушбу тадққотда биз MNAR маълумотлари билан ОЛС ишонч оралиғи баҳоларининг ишлашини баҳолаймиз. Биз, шунингдек, бундай маълумотлар ҳолатлари учун восита сифатида юклашни таклиф қиламиз ва анъанавий ишонч оралиқларини бootстрап билан солиштирамиз. Ҳақиқий параметрларни билишимиз кераклиги сабабли, биз симуляция тадқиқотини ўтказамиз. Тадқиқот натижалари шуни кўрсатадики, иккала ёндашув ҳам ўхшаш оралиқ ўлчамига эга ўхшаш натижаларни кўрсатади. Бootстрап жуда кўп ҳисоб-китобларни талаб қилишини ҳисобга олиб, анъанавий усулларни MNAR ҳолатида ҳам қўллаш тавсия этилади.*

Калит сўзлар: *чизиқли модел, намуна ўлчами, ишонч интервал, юклаш чизиғи, аниқлик, интервал ўлчами, тасодифий эмас*

ЛИНЕЙНАЯ РЕГРЕССИЯ С ОТСУТСТВИЕМ ДАННЫХ НЕ СЛУЧАЙНО: МЕТОД БУТСТРАПА

PhD ***Рахимов Заррух Аминович***
Международный вестминстерский университет в Ташкенте
Рахимова Нилуфар Аминовна
Международный университет туризма "Шёлковый Путь"

Аннотация. *Регрессии OLS имеют набор допущений, чтобы точечные и интервальные оценки были несмещенными и эффективными. Отсутствие данных не случайно (MNAR) может создать серьезные проблемы с оценками в линейной регрессии. В этом исследовании мы оцениваем эффективность оценок доверительного интервала OLS с данными MNAR. Мы также предлагаем загрузку как средство решения таких случаев данных и сравниваем традиционные доверительные интервалы с загрузочными интервалами. Поскольку нам необходимо знать истинные параметры, мы проводим моделирование. Результаты исследования показывают, что оба подхода показывают схожие результаты при одинаковом размере интервалов. Учитывая, что бутстрап требует большого количества вычислений, традиционные методы по-прежнему рекомендуется использовать даже в случае MNAR.*

Ключевые слова: *линейная модель, размер выборки, доверительный интервал, бутстрап, точность, размер интервала, отсутствие не случайно.*

LINEAR REGRESSION WITH DATA MISSING NOT AT RANDOM: BOOTSTRAP APPROACH

PhD ***Rakhimov Zarrukh Aminovich***
Westminster International University in Tashkent
Rahimova Nilufar Aminovna
Silk Road International University of Tourism and Cultural Heritage

Abstract. *OLS regressions have a set of assumption in order to have its point and interval estimates to be unbiased and efficient. Data missing not at random (MNAR) can pose serious estimations issues in the linear regression. In this study we evaluate the performance of OLS confidence interval estimates with MNAR data. We also suggest bootstrapping as a remedy for such data cases and compare the traditional confidence intervals against bootstrap ones. As we need to know the true parameters, we carry out a simulations study. Research results indicate that both approaches show similar results having similar intervals size. Given that bootstrap required a lot of computations, traditional methods is still recommended to be used even in case of MNAR*

Key words: *linear model, sample size, confidence Interval, bootstrap, accuracy, interval size, missing not at random*

Introduction.

Since the introduction, OLS regression has become one of the widely used modelling techniques to show an impact of one or more variables to another dependent variable. This linear modelling approach used primarily for two goals. Firstly, OLS regressions can explain the relationship between two or more variables. Secondly, one can use OLS for simple forms of forecasting. Though they never perfectly imitate the real world, linear models is very widely used given its simplicity to build and ease of interpretability. Linear regressions provide almost always an approximation of real life relationships. In order for our OLS regression give reliable estimations, we must meet a set of OLS assumptions. These requirements are:

1. Equal variance of the error term
2. No strong multicollinearity between explanatory variables
3. No severe outliers
4. Sample size to be larger than 30 observation
5. Linearity in relationship
6. Normality of residuals
7. Stationarity or no autocorrelation of residuals (in case of time series data)
8. No important data missing in our dataset

In case any of these assumptions are violated, OLS confidence intervals might give misleading outcomes and inferences. Interested researchers can refer to Gujarati (2004) for more in-depth discussions of these assumptions and outcomes when they are violated. In this study however, we will concentrate on the case when an important data points are missing not at random. This case appears relatively often in cross sectional data when data collection in certain segments of the society is quite difficult or impossible. The results of this study will be of great benefits for cross sectional analysis which is applied not only in economic studies but also in many other social sciences. As we need to know the true coefficient in order to evaluate estimated intervals we will carrying out a simulation study and comparing both methods. In later chapters, we are going to look at how OLS confidence intervals may behave when data is missing not at random (hereafter referred as MNAR) and whether bootstrapping can serve as a remedy for such cases.

The paper is structured in the following way. First, we will discussing any existing studies on this topic and look at their findings. Afterwards, we will look at theoretical side of traditional confidence interval estimation and bootstrapping of the data and building bootstrap intervals.

Next, we will have a look at simulation approached carried out in R. Lastly, we will look into the results of the simulation and draw our conclusion.

Literature review.

Bootstrapping is a simple et a powerful resampling tool for estimating the properties of a certain statistic or parameter. The idea of bootstrapping lies in repeatedly resampling the sample data. This approach has been pioneered first by Efron (1979) and since then, bootstrap resampling has been widely used in many social sciences.

Bootstrap resampling can also be used in the context of linear models. In the literatures, two types of bootstrapping is used in linear models, bootstrapping residuals and bootstrapping pairs (Chernick and LaBudde, 2011).

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.

When comparing two approaches, a paper by Efron and Tibshirani (1986) come to conclusion that both approaches are equivalent when all assumption of the OLS are met, but each approach can perform differently when number of observations is small. Comparing 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. Another paper by Chernick and LaBudde (2011) finds 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.

Bootstrapping residuals: As noted earlier, this is a resampling technique first introduced by Efron in 1982. Let us consider 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.

Other methods are also considered in further literature such as the percentile-t bootstrap (Diciccio and Efron, 1992), stationary bootstrap (Politis and Roman, 1994) and each used under different scenarios of non-constant variance of the residuals.

This study wants to shed further light into the method of bootstrapping pair in the context of OLS models with data missing not at random.

Linear regression models.

Now, we will look into the method of building of linear models in more details. As mentioned on earlier chapters, linear regressions try to reveal relationship between one y (often referred as dependent variable) and one or more x variables (often referred as explained or dependent variable). The principle of linear model lies in mathematically calculating the beta coefficients of those x variables. For example, somebody wants to evaluate whether having a university degree influences ones income and if es, by how much. Linear regression as intended

to answer exactly these questions using so called ordinary least squared (referred as OLS) method where income is dependent "Y" variable, and ear 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 influence.

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

Where Y – dependent variable, β_0 – intercept (should not necessarily have meaning) β_1 – coefficient of first explanatory variable, X_1 – explanatory or independent variable, e – error or residual term

Given formula is a clear example of linear relationship between X and Y variables. Although, relationship between two variables is almost never linear in real life, linear approximation has proven to work well in many domains. In practice, researchers take more X variables that have been theoretically proved to affect selected depended variable Y. In order to evaluate the correctness and accuracy of the model, a set up statistics such as R squared, adjusted R squared, AIC or BIC are used in practice. This part is out of the score of our research, although interested readers can refer to Greene (2004).

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.

We are very often interested in not only coefficient estimates of, but also interval of possible values of the coefficient with certain level of confidence. In literature, the latter is knows as confidence intervals. Researchers are interested in interval estiamtes 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 software 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.

Confidence interval construction takes its origin from the core theory in statistics, Central Limit Theorem (referred to CLT). CLT indicates that 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) (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.

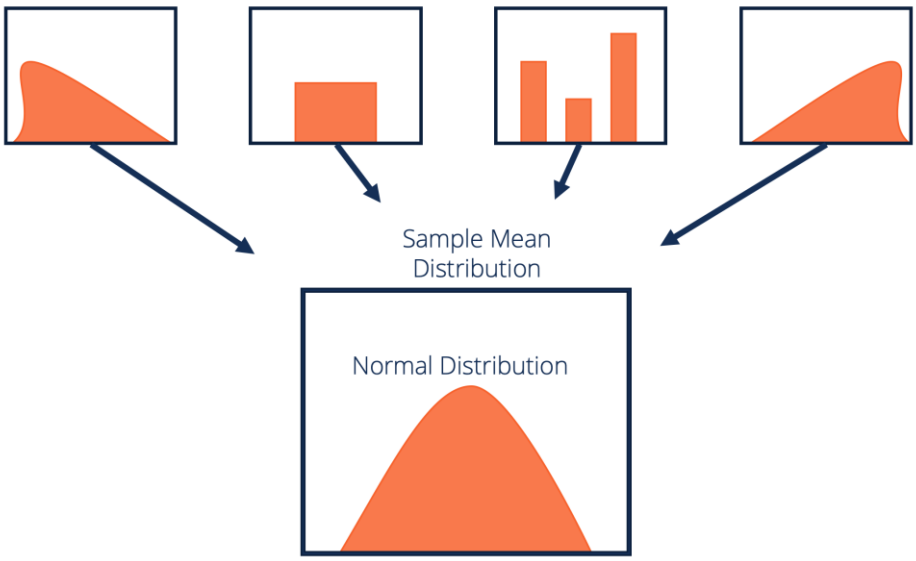

Figure 1

In reality, we almost never can take many samples from the same population due to size of the population (imagine taking 1000 samples of 10 000 size each) 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.

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

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 confidence interval can be understood 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)

Yet, the estimation of intervals and coefficients depends on the completeness of the data which is one of the assumptions of the linear model. Intervals estimates may give inaccurate or even biased calculations if certain portion of very important data is missing. In this study we look at this case also known as Data Missing Not at Random.

In the next section, we suggest another way, bootstrapping, of handling in residuals for construction of our confidence intervals for coefficients.

BOOTSTRAP CONFIDENCE INTERVAL ESTIMATION

Bootstrap confidence intervals offer alternative ways of building intervals which is rather simple approach. 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.

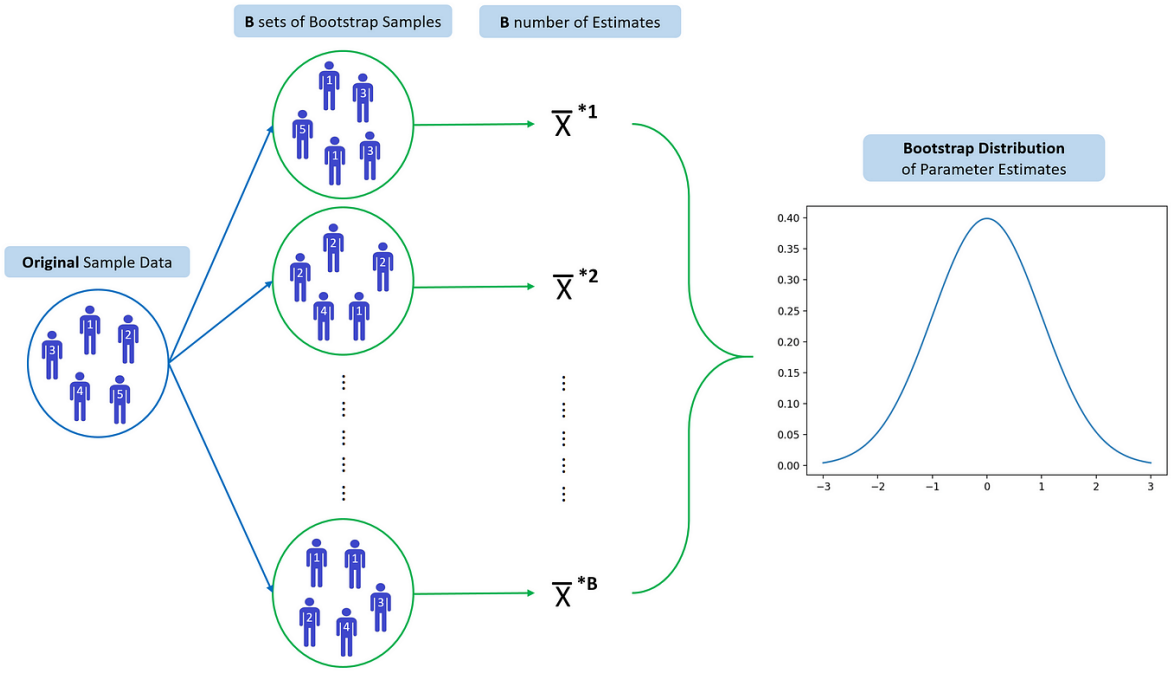

Figure 2

SIMULATION

In this section, we discuss simulation of linear regression and introduce case of data missing not at random. We do not use real life data, but we rather simulate for two reasons. In the first place, true population coefficient β_1 should be known to us and in real life we almost never know the true parameters. In the second place, we need to be aware of the form of data missing not at random, i.e. what share of data is missing and from which variable. We rely on existing papers to imitate a similar form of data missing not at random. Our simulation starts with the simplest form of linear model with one explanatory variable as given below

$$Y = \beta_0 + \beta_1 * X1 + \varepsilon$$

where

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

$$\varepsilon \sim N(0,50)$$

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 of 50.

In order to simulate data missing not a random, we follow the approach of Schafer et al (2002) where certain part of upper percentile of X variable is removed. In our case we take above 80th percentile data from X and remove 90 per cent of that data. Those values will be labelled as NA or Null (in R studio, both are treated equally). 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 β_0= 4 and coefficient β_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 $\varepsilon \sim N(0,50)$ starting with sample size n

Step 5: generate Y with $Y = \beta_0 + \beta_1 * X1 + \varepsilon$

Step 6: take X observations that are above 80th percentile and remove 90 per cent of that data.

Step 7: Estimate confidence intervals using traditional and bootstrap methods in repeated simulations (1000 times). Here we construction 95 percent confidence intervals

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 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

This part will introduce us with the outcomes of different simulations carried out in R studio software. One simulation is with correctly specified model with no missing data and second is with MNAR data. We also take a look at how estimated intervals change as we change our sample size.

Correctly specified model

In the first place, it is necessary to evaluate how traditional confidence interval and bootstrap confidence intervals perform when all data is present and we don't have any violation of regression assumptions. According to theory and many revised studies, it is expected that both methods will perform relatively similar to each other. In other words, for 95 percent confidence intervals, we expect true parameters to fall within estimated intervals at least 95 per cent of cases.

Figure 3 below illustrates how often true coefficients fall within estimated confidence intervals built using traditional and bootstrap methods. We can observe that both approaches are doing pretty good, that is constructed intervals are containing true coefficient at least 95 per cent of the cases with different sample size. In other words, 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.

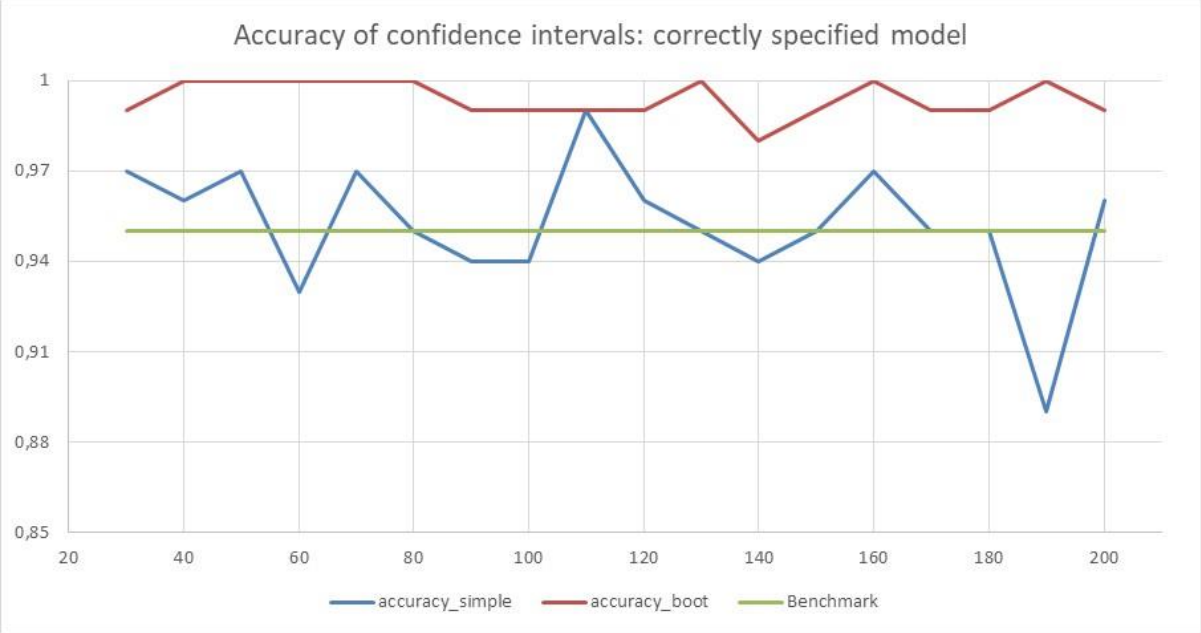

Figure 3

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)

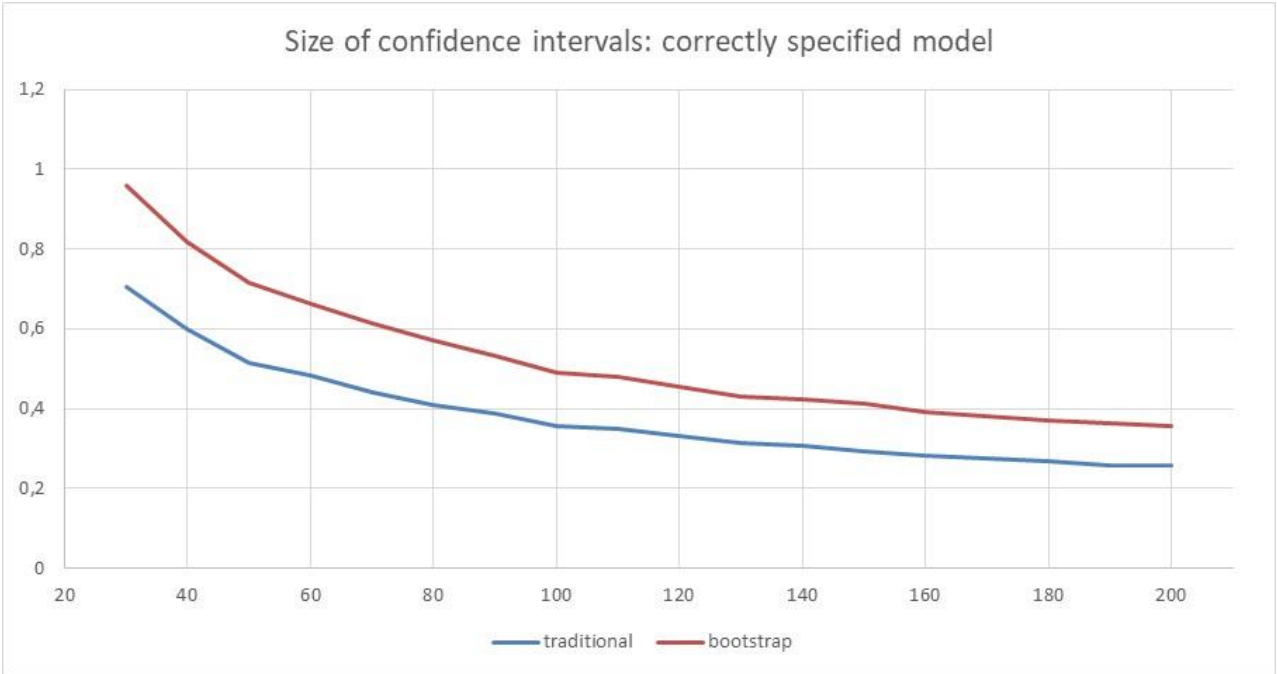

Figure 4

Data missing not at random

Here we will be looking at performance of traditional and bootstrap interval estimations when large portion of upper percentile of explanatory variables is missing. To remind the reason, we tool upper 80th percentile of X variables and removed 90 per cent of that data. Afterwards, we estimated confidence intervals using traditional and bootstrap approaches.

Lastly we evaluated how often the true coefficient from our simulation was falling within the given interval. Ideally, the true coefficient must fall in 95 per cent of simulated cases.

The results in Figure 5 indicate that accuracy of traditional and bootstrap intervals estimates are oscillating around 95 per cent which is out benchmark. This indicates that both approaches are doing pretty well in term of interval estimates even when quite important portion of data is missing. This is a very strong and good finding in favor of traditional approaches.

This tells us that even when large share of important data is missing, traditional central limit theorem based interval estimation is doing a pretty good work.

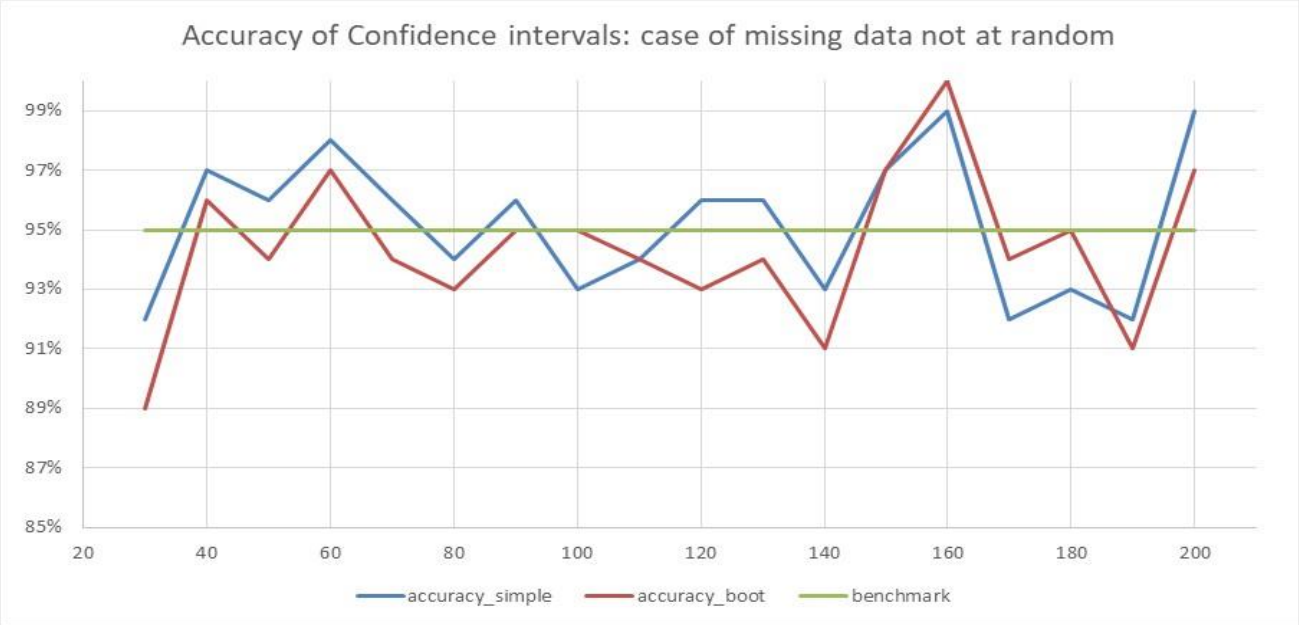

Figure 5

If we compare sizes of confidence intervals from Figure 6 estimated using traditional and bootstrap methods, one can see that both approaches have a very similar size.

Given that bootstrap requires a lot of computing power and both approaches are showing similar results, we can conclude that traditional approach is still reliable even when good share of important data is missing not at random.

Figure 6

Conclusion

This study looked into cases when important data is missing not at random and looked at two ways of interval estimations of coefficients of linear regression. In the first place, we revised related literature on topic of data MNAR. Based on our investigation, there is limited literature on application of bootstrap approach in case of data missing not at random. Afterwards, we investigated the theoretical side of linear models and traditional way of building confidence intervals that is based on central limit theorem. Along with that, we also looked into bootstrap approach of constructing confidence intervals. 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. For this reason, 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 MNAR data with different sample size, spanning from 30 to 200 observations. Simulation results indicate that even when important data is missing not at random, both, traditional and bootstrap methods are building rather good intervals. In other words, both interval estimates have been including the true coefficient in around 95 per cent of the cases. In additional, interval sizes of both, traditional and bootstrap confidence intervals are quite similar. This is rather strong finding in favor of both approaches. Yet, as bootstrap requires intense computational power while traditional methods is estimated in a fast way, we conclude that researchers are recommended to still use traditional method even when good share of important data is not missing at random.

Reference:

Carpenter, J. R., & Kenward, M. G. (2012). Missing data in clinical trials: a practical guide. Practical Guides to Biostatistics and Epidemiology. Cambridge University Press.

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.

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

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

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

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

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

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

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. Structural Equation Modeling, 10(1), 80-100.

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

He, Y., & Zaslavsky, A. M. (2012). Diagnostics for multiple imputation in surveys with missing data. Biometrika, 99(4), 731-745.

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

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

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

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. Wiley.

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

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

Schafer, J. L., & Graham, J. W. (2002). Multiple imputation for missing data: A cautionary tale. Sociological Methods & Research, 31(4), 445-454.