# BOOTSTRAPPING IN REGRESSION ANALYSIS OF TERTIARY EDUCATION ATTAINMENT IN EUROPEAN COUNTRIES

**Dagmar Blatna**

**Abstract**

The present paper demonstrates the applicability of a bootstrap approach in a regression analysis. Bootstrapping is an approach to statistical inference adherent to computationally intensive statistical techniques. Since it does not require the classical distributional assumption, the bootstrap can provide more accurate inferences when the data are not well behaved. The Ordinary Least Squares (*OLS*) method, often used to estimate the parameters of regression models in the bootstrap procedure, is extremely sensitive to outliers and non-normality of errors. The robust bootstrapping method replaces the classical bootstrap mean and standard deviation with robust estimates by using robust regression estimates with a high breakdown-point. Values of the indicator of tertiary education attainment in the European countries depend on many indicators of the general economic background, employment, etc. The values of these indicators vary between the European countries and, consequently, the occurrence of outliers can be dealt with in an analysis. The research results obtained by using bootstrapping, *OLS* and robust regression analysis are compared.

**Key words:** bootstrapping, robust regression, outliers, tertiary education attainment.

**JEL Code:** C190, C490, O570

## Introduction

The regression analysis is the most commonly used statistical tool for analyzing dependences. The classical statistical approach – the least squares method (*OLS*) – can be highly unsatisfactory due to the presence of outliers that are likely to occur in an analysis of data from the European countries. In such a case, the robust regression becomes an acceptable and useful tool, since it provides a good fit to the bulk of the data, the outliers being exposed clearly enough. The aim of this paper is to verify the applicability of the bootstrapping (resampling) technique in both *OLS* and robust regression.

# 1 The principle of bootstrapping in regression

The bootstrap was introduced by (Efron,1979). The essence of the bootstrapping method is to create a large number of sub-samples by randomly drawing observations with replacement from the original dataset. Each element of the original dataset is selected for the bootstrap sample with probability 1/*n*, mimicking the original selection of the bootstrap sample from the original one. We repeat this procedure a large number of times (*R*), receiving *R* artificial samples of *n* observations from the data in the original sample. These artificial sub-samples are termed as bootstrap samples, being used to recalculate the estimates of the statistic (e.g. regression coefficients). The resampling distribution of a statistic is then constructed empirically by resampling from the sample. The bootstrap gives slightly different results when repeated on the same data. (More see e.g. in (Cole,1999), (Efron,1993), (Stine, 1990).

We consider a linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y} = (Y_1,\ldots,Y_n)'$ is a response variable, $\boldsymbol{X} = (x_{ij})'$; $i = 1,\ldots, n$; $j = 1,\ldots p$; is a design matrix, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ is a vector of unknown regression coefficients (matrix of regressors) and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)'$ is a vector formed by an initial part of the sequence of independent identically distributed residuals distributed according to the distribution *F, n* being the sample size. There are two general ways to bootstrap a regression (see in detail e.g. in (Stine, 1990),(Fox, 2002)). In our analysis, random *x*-resampling (case resampling) was used. This approach treats the regressors as random, potentially changing from sample to sample, and selects bootstrap samples directly from the observations. Thus, we get a sample of *n* observations $z_i' = (y_i, x_i)$; $i = 1,...,n$, the data being ready for a resampling-with-replacement procedure, the resample size having to be equal to that of the original data set. Then the regression coefficient is computed from the resample in the first step. We repeat this routine *R* times to get a more precise estimate of the bootstrap distribution which represents an "empirical bootstrap distribution" of sample regression coefficients $\beta_b^*, b = 1,2,..., R.$. The average of the bootstrapped regression coefficients $\beta_b^*$ is an estimate of $\boldsymbol{\beta}$ ($\hat{\beta}$). A measure of accuracy for $\hat{\beta}$ is the standard error *SE* (a boot of estimated bootstrap variance of ($\hat{\beta}$) and the bias of the estimator $\hat{\beta}$ which can be estimated as the difference between an average bootstrapped value of the regression coefficient and its original-sample value. There are two basic approaches to constructing bootstrap confidence intervals: the bootstrap percentile interval (*EP*) based on the empirical quintiles of bootstrap regression coefficients $b_b^*$ and the bias-corrected, accelerated percentile interval ($BC_\alpha$) with correction factors for lower and upper percentiles of the statistic $\beta$ based on jackknife values

of the statistic $\beta$ (see e.g. in (Cole, 1999), (DiCiccio & Efron, 1996), (Freedman, 1981), (Efron, 1993, 2000).

The classical approach to bootstrapping regression is based on the *OLS* method which assumes that the error terms are normally distributed. The classical bootstrap procedure is sensitive to outliers, i.e. the points deviating relative to the response variable as well as those distant from the bulk of the data relative to the factor space (so-called leverage points). However, the classical bootstrap is extremely sensitive to influential points which are deviated simultaneously both in the explanatory variables and the response variable respectively.

The robust bootstrapping method modifies the classical *OLS* bootstrap algorithm by using robust regression estimates with a high breakdown-point, e.g. the *MM* procedure with an initial *S*-estimate or *LTS*-estimates (see more details in (Hamadu, 2012), (Silibian-Barrera & Zamar, 2002).

## 2 Robust regression methods and diagnostic tools

*MM*-estimates (proposed by (Yohai, 1987)) are defined by a three-stage procedure. At the first stage, an initial regression estimate is computed; it is consistent, robust, with a high breakdown-point, but not necessarily efficient. At the second stage, an *M*-estimate of the error scale is computed, using residuals based on the initial estimate. Finally, at the third stage, a (final) *M*-estimate of *MM* estimates represents a combination of high breakdown value estimation and an efficient estimate of the regression parameters based on a proper redescending $\psi$-function which is the derivative of a proper loss function $\rho$ ($\psi=\rho'$). (See more details in (Yohai, 1987),(Rousseeuw & Leroy,2003)).

The least trimmed squares (*LTS*) estimator proposed by Rousseeuw in 1984 is obtained by minimizing $\sum_{i=1}^{h} r_i^2$ , where $r_{(i)}$ is the *i*th order statistic among the squared residuals written in the ascending order, *h* being the largest integer between ($n$/2) and ($3n$+$p$+1)/4 and $p$ being the number of predictors. The *LTS* regression is a reliable data analytic tool that may be used to discover regression outliers both in simple and multivariate situations. A more detailed description can be found e.g. in (Ruppert,D. & Carroll, R.J.,1990). (Rousseeuw, 2003), (Fox, 2002).

Various numerical and graphic diagnostic methods for detecting outliers, leverage points and influential observations can be employed. In this paper, the following ones have been used: *Residuals associated with LTS regression*, *Standardized*, *Studentized residuals* (a

type of standardized *t* distribution residuals with *n-p*-2 Df), *Robust distance, Diagnostic plots, Normal Q-Q plot of the standardized residuals*, *Plot of Kernel density of residuals.*

For the selection of a proper regression model, the following diagnostic tools were used: *The significance robust tests: robust t-test***,** *robust F-tests***,** *robust Wald test* and the Robust selection information criteria: *Robust Akaike's Information Criterion* (*AICR),Robust Bayesian Information Criterion* (*BICR*) and *Robust Final Prediction Error* (*RFPE*).

## 3 Results of Analyses and Discussion

The tertiary educational attainment (**TEA)** indicator is constructed as a share of the population aged 30-34 years who have successfully completed tertiary level education (Eurostat). This indicator is one from the set of Europe 2020 indicators used by the European Commission to monitor headline targets of a strategy for the next decade – Strategy for smart, sustainable and inclusive growth. Education plays a key role in Europe 2020 and particularly in the Inclusive Growth agenda. The value of TEA in the European countries depends on numerous indicators of the general economic and social background, employment, etc.

There are two robust models which fulfil the aforementioned criterion. Both the models include a general government expenditure on education (GEE) variable (based on COFOG classification) as a percentage of GDP. Two fitting models were found, the first one including GDP per capita as a second explanatory variable, the other one containing the net national investment (NNI) instead. (See the goodness-of-fit test of this model in Table 1.)

In the first regression **TEA~ GEE + GDPc** model, the robust diagnostics identified four leverage points (2 Belgium, 4 Denmark, 15 Luxembourg, 22 Romania), but none of them was also an outlier. Due to the absence of influential points, classical regression modelling can be considered fully appropriate. Table 1 shows that the regression parameters in both LS and robust model are the same. (As a matter of course*, t*-test statistics are not the same but very similar.) Robust diagnostics (see Table 2) revealed two influential points (7 Ireland and 15 Luxembourg) in the second suitable regression **TEA~ GEE+NNI** model. The same can be seen from graphical diagnostics in Figures 1 and 2.

**Tab. 1: Final regression models for TEA - Goodness-of-fit tests**

| Outliers/*Lev.points* | Robust MM fit / LS fit | R-sq. | AICR | BICR | Deviation | RFPE |
|---|---|---|---|---|---|---|
| -;<br>*2,4,8,15.22* | -2.921+5.063 GEE +0.089 GDPc<br>-2.921+5.063 GEE +0.089 GDPc | 0.540<br>0.541 | 21.356 | 27.655 | 1078.06 | 14.182 |
| 7,15;<br>*3,5,6,7,15,17.21* | -104.94+4.116 GEE + 1.385 NNI<br>12.574 +5.786 GEE -0.133 NNI | 0.532<br>0.408 | 22.289 | 29.579 | 1050.56 | 14.263 |

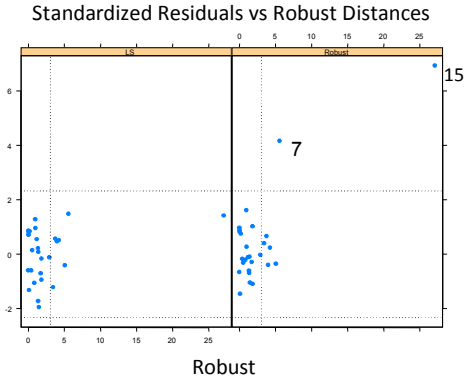
Source: data EUROSTAT, author's own calculations

**Tab. 2: Robust diagnostics for TEA~ GEE+NNI model**

| Observation | Mahalanobis distance | Robust MCD Distance | Leverage | Stand.Robust Residual | Outlier |
|---|---|---|---|---|---|
| 3 Czech Republic | 1.2974 | 3.9202 | * | 0.3989 | |
| 5 Germany | 1.6744 | 2.1311 | * | -0.3680 | |
| 6 Estonia | 1.1599 | 2.9332 | * | 0.9970 | |
| 7 Ireland | 1.6386 | 5.2813 | * | 4.0653 | * |
| 15 Luxembourg | 3.8325 | 11.0645 | * | 6.7694 | * |
| 17 Malta | 0.7066 | 2.8946 | * | -0.6784 | |
| 22 Romania | 0.9511 | 2.8355 | * | -1.0292 | |

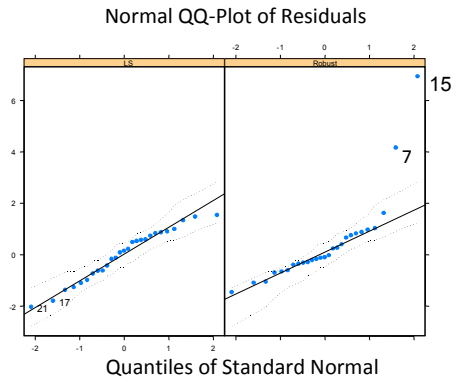
Source: data EUROSTAT, author's own calculations

**Fig. 1: Diagnostic Graph (TEA~ GEE+NNI)**




Source: data EUROSTAT, author's own elaborations

**Fig. 2: Normal Q-Q Plot**




Source: data EUROSTAT, author's own elaborations

For these two selected regression models, both classical and robust bootstrapping was performed with a various number of resamplings (R = 500, 1000 and 1500). For each of the sets, the boostrap mean as well as the standard deviation (*SE*), bias and both empirical 95% *EP* and 95% *BCa* confidence intervals were computed.

In order to compare the results of bootstrap regressions, *OLS* and robust *MM* methods were employed as well. Values of regression coefficients, *SE*s, *t*-tests, *p*-values for *t*-tests, *R*-squares, 95% confidence O*LS* intervals and 95% confidence *MM*-intervals for comparison

with robust bootstrapped ones were calculated. Robust regression was taken as a tool to identify outliers.

All results of analyses were obtained using S-Plus 6.2 and SAS 9.1 systems. The results of all analyses are presented in Tables 3 (for **TEA~ GEE + GDPc** model) and 4 (for **TEA~ GEE+NNI** model). Since there are numerous graphs, only histograms for the 1000 bootstrap replications of data sets parameters are presented (see Figs. 3-6).

**Tab. 3: Classical bootstrap regression, robust bootstrap regression, OLS and MM regression for TEA~ GEE+GDPC model**

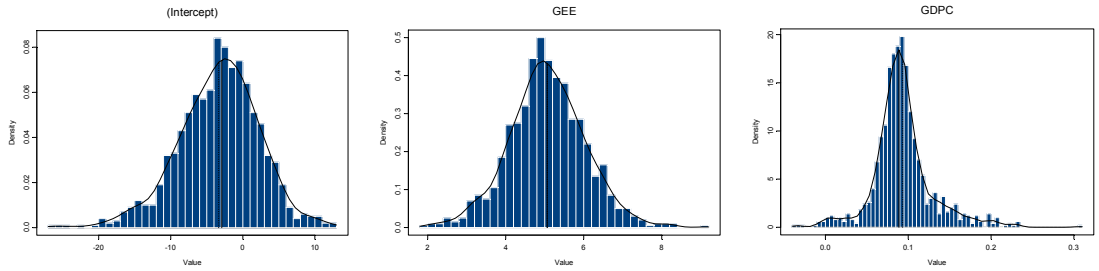
| | | Observed | Bias | Mean | SE | 95% EP | 95% BCa |
|---|---|---|---|---|---|---|---|
| Bootstrap R=500 | intercept | -2.9209 | -0.6004 | -3.5212 | 6.0657 | -17.270;6.585 | -16.011;7.073 |
| | GEE | 5.0630 | 0.0058 | 5.0688 | 0.9878 | 3.210;7.034 | 3.166;7.025 |
| | GDPC | 0.0888 | 0.0057 | 0.0945 | 0.0365 | 0.032;0.187 | 0.0187;0.161 |
| Robust Bootstrap R=500 | intercept | -2.9209 | 0.2434 | -2.6774 | 11.3230 | -23.944;25.461 | -21.618;30.9362 |
| | GEE | 5.0630 | -0.3311 | 0.1134 | 0.0896 | -0.048;8.0619 | 0.1912;8.4185 |
| | GDPC | 0.0888 | 0.0246 | 0.1134 | 0.0896 | -0.0244;0.3475 | -0.0892;0.2457 |
| Bootstrap R=1000 | intercept | -2.9209 | -0.5845 | -3.5054 | 5.7524 | -16.646;6.8798 | -15.4893;-7.1924 |
| | GEE | 5.0630 | -0.0043 | 5.0587 | 0.9705 | 3.0431;6.9966 | 2.9120;6.8672 |
| | GDPC | 0.0888 | 0.0065 | 0.0953 | 0.0409 | 0.0216;0.1990 | 0.0128;0.1881 |
| Robust Bootstrap R=1000 | intercept | -2.9209 | 0.3462 | -2.5747 | 11.022 | -24.719;25.032 | -22.1503;29.6595 |
| | GEE | 5.0630 | -0.3894 | 4,6736 | 1,8506 | -0.2242;7.7333 | 0.1641;7.7739 |
| | GDPC | 0.0888 | 0.0277 | 0,1165 | 0,0954 | 0.0051;0.4055 | -0.0377;0,2926 |
| Bootstrap R=1500 | intercept | -2.9209 | -0.6927 | -3.6136 | 5.8770 | -17.289;7.196 | -14.866;8.4201 |
| | GEE | 5.0630 | 0,0075 | 5.0705 | 1.0031 | 2.9495;7.0549 | 2.8866;6.9617 |
| | GDPC | 0.0888 | 0,0061 | 0.0949 | 0.0390 | 0.0222;0.1918 | 0.0054;0.1723 |
| Robust Bootstrap R=1500 | intercept | -2.9209 | 0.4523 | -2.4685 | 10.900 | -26.425;19.627 | -26.453;19.6215 |
| | GEE | 5.0630 | -0.4202 | 4.6428 | 1.8700 | -0.401;8.0981 | -0.4926;8.5323 |
| | GDPC | 0.0888 | 0.0291 | 0.1178 | 0.1020 | -0.0060;0.4261 | -0.032;0.3511 |
| | | Parameter | SE | T | p-value | 95 % conf.interval | |
| *OLS* R-sq 0.5413 | intercept | -2.9209 | 7.4184 | -0.3937 | 0.6973 | -18.2317; 12.3900 | |
| | GEE | 5.0630 | 1.2331 | 4.1059 | 0.0004 | 2.518; 7.6080 | |
| | GDPC | 0.0888 | 0.0329 | 2.6998 | 0.0125 | 0.0209; 0.1567 | |
| *MM* R-sq 0.5400 | intercept | -2.9209 | 8.4535 | -0.3555 | 0.7327 | -17.4607; 11.6190 | |
| | GEE | 5.0630 | 1.4057 | 3.6018 | 0.0014 | 2.6462; 7.4799 | |
| | GDPC | 0.0888 | 0.0375 | 2.3679 | 0.0263 | 0.0243;0.1532 | |

Source: author's own calculations

In **TEA~ GEE+GDPC** model, the assumption of normality of errors is met. It is known that if errors are normally distributed, the *OLS* estimator will have a minimum variance among all unbiased estimators. The best results are provided by the classical bootstrap based on the *OLS* method. Standard deviations (*SE*) are even lesser than those acquired by the *OLS* method, confidence intervals being narrower. Since the classical

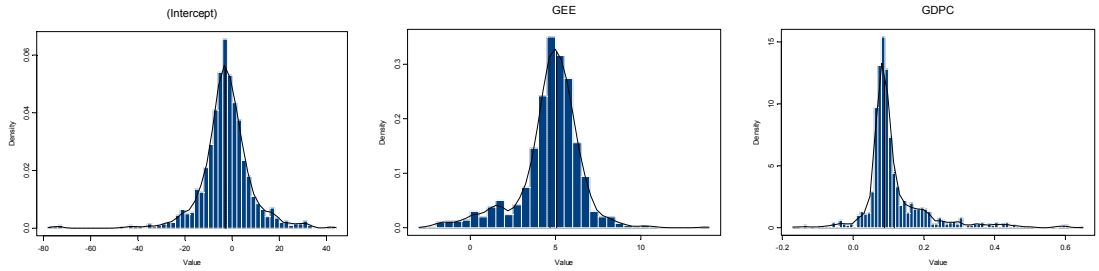
bootstrap produces better results than the robust one, it ought to be given due preference. Both classical and robust bootstrap distributions of both the intercept and regression coefficient are adequately symmetric (see Figs. 3 and 4).

**Fig. 3 Histograms for classical replications of regression coefficients for TEA~ GEE+GDPc model (R=1000)**




Source: author's own elaborations

**Fig. 4 Histograms for robust replications of regression coefficients for TEA~ GEE+GDPc model (R=1000)**




Source: author's own elaborations

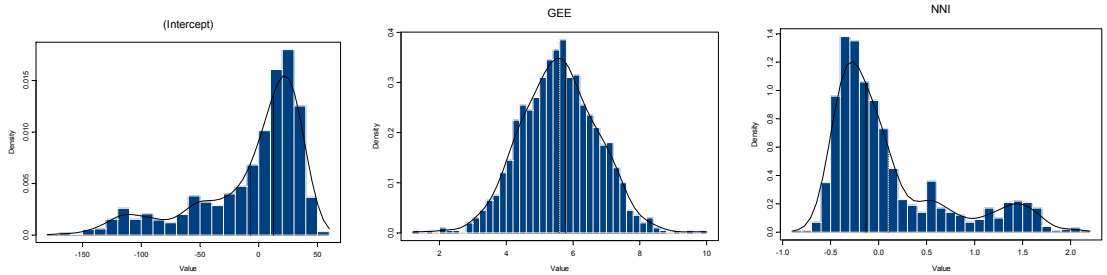
In **TEA~ GEE+NNI** model, two influential points were identified (seven leverage points and two outliers). *OLS* methods fail totally, NNI indicator having an opposite sign. The results are distorted, which means that in this case the linear regression model is absolutely inapplicable. The robust *MM* regression provides the best results, the smallest *SE*s and the narrowest confidence intervals. The bootstrap distribution in this case might be a very poor estimator of the distribution of regression estimates, since the proportion of outlier points in a bootstrap sample can be larger than the fraction of contamination in the original sample. In bootstrap, both outlying and non-outlying observations have the same chance of belonging to any bootstrap since such samples are drawn from the original sample with replacement.

**Tab. 4: Classical bootstrap regression, robust bootstrap regression, OLS and MM regression for TEA~ GEE+NNI model**

| | | Observed | Bias | Mean | SE | 95% EP | 95% Bca |
|---|---|---|---|---|---|---|---|
| Bootstrap R=500 | Intercept | 12.5737 | -19.756 | -7.1828 | 46.2194 | -121.727;44.01 | -100.428;50.802 |
| | GEE | 5.7860 | -0.7565 | 5.6115 | 1.1558 | 3.354;7.8077 | 3.5977;8.2898 |
| | NNI | -0.1330 | 0,2479 | 0.1149 | 0.5999 | -0.5239;1.6143 | -0.5987;1.3528 |
| Robust Bootstrap R=500 | Intercept | -104.94 | 30.0766 | -74.864 | 75.456 | -217.66;59.309 | -225.12;49.8567 |
| | GEE | 4.1163 | 0.1878 | 4.3040 | 2.0360 | -0.964;7.4567 | -3.9912;6.7726 |
| | NNI | 1.3846 | -0.3602 | 1.0240 | 0.9700 | -0.624;3.2632 | -0.537;3.4355 |
| Bootstrap R=1000 | Intercept | 12.5737 | -19.740 | -7.166 | 48.3005 | -125.62;42.439 | -108.061;49.544 |
| | GEE | 5.7860 | -0.225 | 5.562 | 1.2473 | 2.9358;8.0817 | 3.5530;8.6437 |
| | NNI | -0.1330 | 0.251 | 0.118 | 0.6283 | -0.5288;1.6985 | -0.6308;1.4010 |
| Robust Bootstrap R=1000 | Intercept | -104.94 | 32.6266 | -72.313 | 73.0623 | -187.607;61.01 | -227.623;46.913 |
| | GEE | 4.116 | 0.2986 | 4.4149 | 2.1059 | -0.757;7.5976 | -3.929;6.6649 |
| | NNI | 1.385 | -0.3997 | 0.9879 | 0.9389 | -0.662;2.4691 | -0.488;3.4677 |
| Bootstrap R=1500 | Intercept | 12.5737 | -16.442 | -3.8687 | 44.4225 | -118.67;43.549 | -96.766;51.4324 |
| | GEE | 5.7860 | -0.1953 | 5.59071 | 0.5721 | 3.3782;7.9016 | 3.8087;8.6007 |
| | NNI | -0.1330 | 0.2098 | 0.0768 | 0.5721 | -0.5503;1.5604 | -0.6464;1.3017 |
| Robust Bootstrap R=1500 | Intercept | -104.94 | 33.2008 | -71.739 | 73.1015 | -197.06;60.658 | -222.77;47.0736 |
| | GEE | 4.116 | 4.116 | 0.3873 | 2.0767 | 0.712;8.0660 | -3.677;6.8228 |
| | NNI | 1.385 | -0.4127 | 0.9719 | 0.9401 | -0.675;2.3545 | -0.507;3.4562 |
| | | parameter | SE | t | p-value | 95 % conf..interval | |
| *OLS* R-sq 0.3691 | Intercept | 12.5737 | 21.4027 | 0.5875 | 0.5624 | -31.599;56.748 | |
| | GEE | 5.7860 | 1.4429 | 4.0101 | 0.0005 | 2.8081;8.7639 | |
| | NNI | -0.1330 | 0.2710 | -0.4907 | 0.6281 | -0.6922;0.4263 | |
| *MM* R-sq 0.5320 | Intercept | -104.940 | 26.8624 | -3.9066 | 0.0007 | -156.894;-52.986 | |
| | GEE | 4.1163 | 1.0427 | 3.9477 | 0.0006 | 2.0996;6.1329 | |
| | NNI | 1.3846 | 0.3440 | 4.0253 | 0.0005 | 0.7193;2.0498 | |

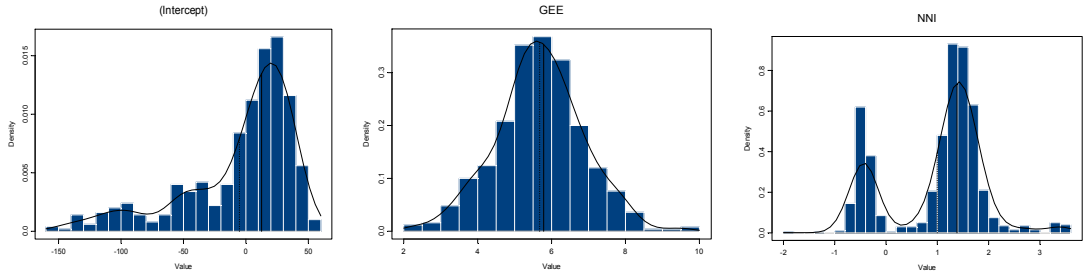
Source: author's own calculations

**Fig. 5 Histograms for classical replications of regression coefficients for TEA~ GEE+NNI model (R=1000)**




Source: author's own elaborations

**Fig. 6 Histograms for robust replications of regression coefficients for TEA~ GEE+NNI model (R=1000)**




Source: author's own elaborations

The bimodality in the histograms of regression coefficients confirms the presence of outliers. Robust bootstrap distributions of the regression coefficient are more tightly concentrated than those of *OLS* bootstrap estimators, being considerably heavy-tailed due to the existence of outliers.

## Conclusion

In **TEA~ GEE+GDPC** model, the assumption of normality of errors was met and only leverage points were identified. The classical bootstrap produces better results than the robust one, thus the classical bootstrap could be preferred. *SE*s are even lesser than in *OLS* fits and confidence intervals are narrower. Both the classical and robust bootstrap distributions of the regression coefficient are adequately symmetric.

In **TEA~ GEE+NNI** model, seven leverage points and two outliers were identified. The *OLS* method totally fails, the NNI indicator has the opposite sign. The classical bootstrap provides better results than the *OLS* method, the sign of the regression coefficient for NNI parameter being right. The bootstrap – both classical and robust – provides regression coefficient estimators with broader confidence intervals than the *MM* regression owing to the fact that the outlier proportion in bootstrap samples can be higher than in the original dataset. The robust *MM* regression provides the best results, with the smallest *SE*s and the narrowest confidence intervals.

## References

Cole, S. R. (1999). Simple bootstrap statistical inference using the sas system. *Computer Methods and Programs in Biomedicine.*, *60*, 79-82.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science.*, *11*(3), 189-212.

Efron, B. (1979a). Bootstrap methods: another look at the jackknife. *Ann. of Statistics*,7, 1-26.

Efron, B., & Tibshirani, R. J. (1993). CHAPMAN&HALL/CRC (Eds.), *An Introduction to the Bootstrap.* New York, U.S.A.

Efron, B. (2000). The bootstrap and Modern Statistics. *JASA*, *95*(452), 1293/1296.

Fox,J.(2002). *Bootstrapping regression models*. Retrieved from http://cran.r/project.org/doc/contrib/FOX.copenion/appendix/bootstrapping.pdf

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*(6), 1218-1228.

Hamadu, D. (2012). A bootstrap approach to robust regression. *International Journal of Applied Sciences and Technology*, *2*(9), 114-119.

Hampel, F. R., &Ronchetti, E. M., & Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* New York: J. Willey.

Hubert, M. & Rousseeuw, P.J.& Van Aelst. (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science* 2008, *23* (1), 92-119.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust Regression and Outlier Detection.* New Jersey: J Willey.

Ruppert,D. & Carroll, R.J. (1990). Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association, 75*, 828-838.

Salibian-Barrera,, M., & Zamar,, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statitics*, *30*(2), 556-582.

Stine, R. (1990). An introduction to bootstrap methods. examples and ideas. *Sociological Methods and Research*, *18.*(2-3), 243-291.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates in regression. *The Annals of Statistics.*, *15.*(2.), 642-656.

**Contact:**

Dagmar Blatná

University of Economics, Prague

Prague 3, W. Churchill sq. 4, Czech Republic

blatna@vse.cz