Bootstrapping estimators from balance sheet data

Pham Viet Hung

MŰEGYETEM 1782

Content

- Background
  - Problems of traditional methods
  - Bootstrapping
- Application on balance sheet data
  - Data exploration
  - Results of the bootstrap simulations

Background

Problems of traditional methods (1)

The two traditional inference methods are **exact statistics** and **asymptotic theory**.

Problems of traditional methods (2)

Pro:

- Useful for understanding the underlying concepts and theories
- Working well having a lot of data points

Con:

- Exact statistics assume a lot of unrealistic and highly restrictive conditions such as homoscedasticity.
- Asymptotic theory may approximate estimators with uncertain accuracy.
- Sometimes not applicable if having a small sample size of data

Bootstrapping (1)

A technique developed by Bradley Efron in 1979.

- Sampling with replacement from given a sample size of n —> new samples are generated
- $P(\text{an observation in bootstrap sample}) = 1 - \left(1 - \frac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \approx 0.632$
- Usually 10000 resampling number is enough

The result is bootstrap sampling distribution of statistic of interest.

Bootstrapping (2)

Advanatges:

- Easy to derive statistics of our interest and confidence intervals.
- Not much assumption is needed
- Avoid the cost of repeating the experiment to obtain other groupsof sample data

Disadvantages:

- The result strongly depends on the quality of sample
- It used to be time-consuming and computationally expensive.

Algorithm 1: Classic Bootstrap Algorithm for estimating a parameter

Input : $S = \{x_1, x_2, \ldots, x_n\}$ – a size n original samples
Output: $\hat{\theta}^* = (\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*)$ – bootstrapped values
Init: B – num of bootstrap repetitions
$r := 1$
$L_{est} := []$ – an (empty) array of parameter estimate ;
T – A function of a statistic ;

while $r < B + 1$ **do**
  $S_k :=$ sampling a size n from S by *sampling with replacement*;
  $\hat{\theta}_r^* := T(S_k)$;
  $L_{est}[r] := \hat{\theta}_r^*$;
  r = r + 1
end

Application on balance sheet data

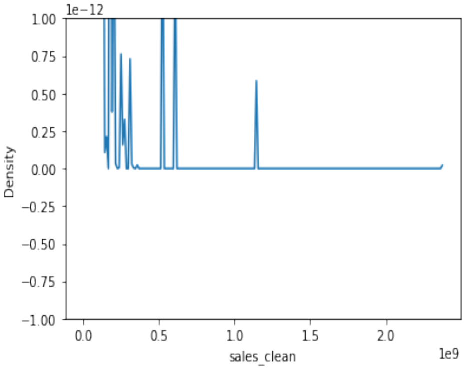
Data exploration – Before data transformation (1)

	sales_clean	tanass_clean	tax
Mean	314854.888	135862.025	1969.049
Std	8137357.078	11129327.382	118160.019
Min	0.537	0	0
25%	4642	70	18
Median	18034.500	1695	114
75%	70810.250	13675.250	576
max	2371623338	6094569000	63639000
RSD	*25.845*	*81.916*	*60.009*

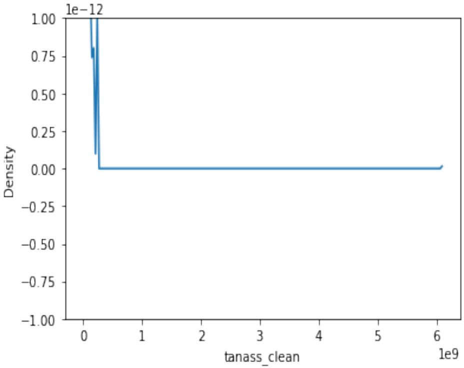
Data exploration – After data transformation (1)

	sales_clean	tanass_clean	tax
Mean	0	0	0
Std	1	1	1
Min	-3.799	-1.492	-1.846
25%	-0.604	-0.643	-0.694
Median	-0.003	0.124	0.037
75%	0.622	0.703	0.710
max	6.070	6.158	6.082

Data exploration – Before data transformation (2)

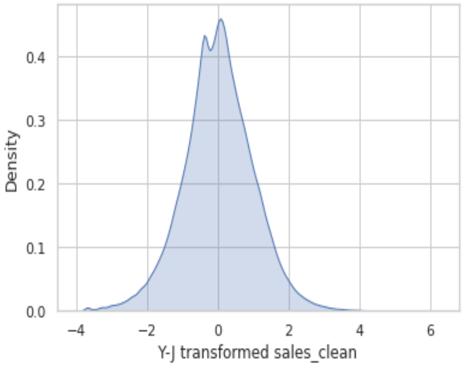


(a) sales_clean

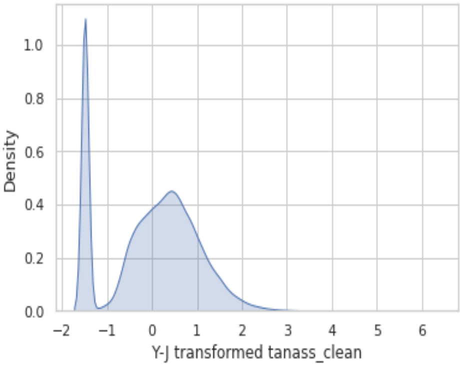


(b) tanass_clean

Data exploration – After data transformation (2)



(a) Y-J sales_clean



(b) Y-J tanass_clean

Data exploration – Yeo-Johnson data transformation

$$\psi^{\textbf{YJ}}(\lambda, x_i) = \begin{cases} ((x_i+1)^{\lambda}-1)/\lambda & \text{if } \lambda \neq 0,\ x_i \geq 0, \\ log(x_i+1) & \text{if } \lambda = 0,\ x_i \geq 0, \\ -[(-x_i+1)^{2-\lambda}-1]/(2-\lambda) & \text{if } \lambda \neq 2,\ x_i < 0, \\ -log(-x_i+1) & \text{if } \lambda = 2,\ x_i < 0, \end{cases} \tag{1}$$

where x_i is the data point in our data set for all i, and λ is the parameter approximated by the *Maximum Likelihood Estimation*.

Sampling distributions of bootstrapped statistics (1)

- *xbar_sales*: bootstrapped sampling distribution of sales mean
- *std_sales*: bootstrapped sampling distribution of sales standard deviation
- *median_sales*: bootstrapped sampling distribution of sales median
- *corr_sales_tanass*: bootstrapped sampling distribution of corr. bw. sales and tanass
- *corr_sales_tax*: bootstrapped sampling distribution of corr. bw. sales and tax
- *corr_tanass_tax*: bootstrapped sampling distribution of corr. bw tanass and tax
- *ols_tanass*: bootstrapped sampling distribution of tanass parameter in OLS
- *ols_tax*: bootstrapped sampling distribution of tax parameter in OLS
- *ols_r2*: bootstrapped sampling distribution of r-squared in OLS

Sampling distributions of bootstrapped statistics (2)

The plot of:

- *xbar_sales*
- *std_sales*
- *ols_r2*

Bootstrapped OLS parameters (1)

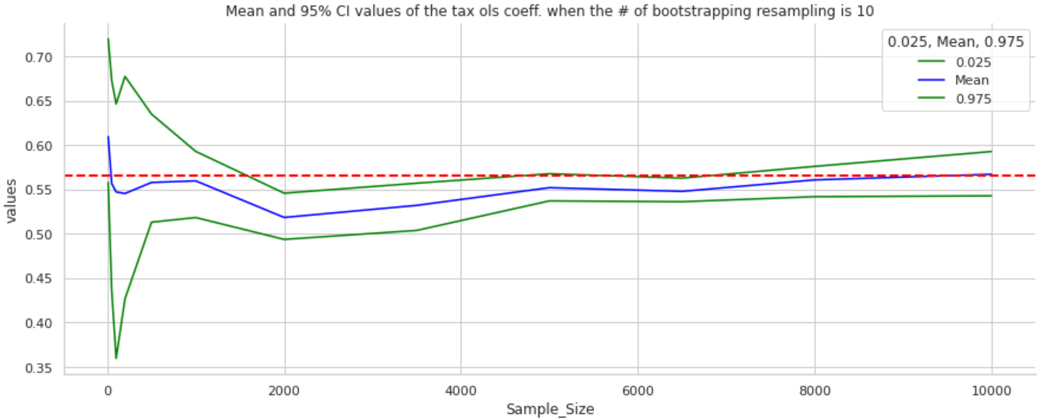


Figure: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 10

Bootstrapped OLS parameters (2)

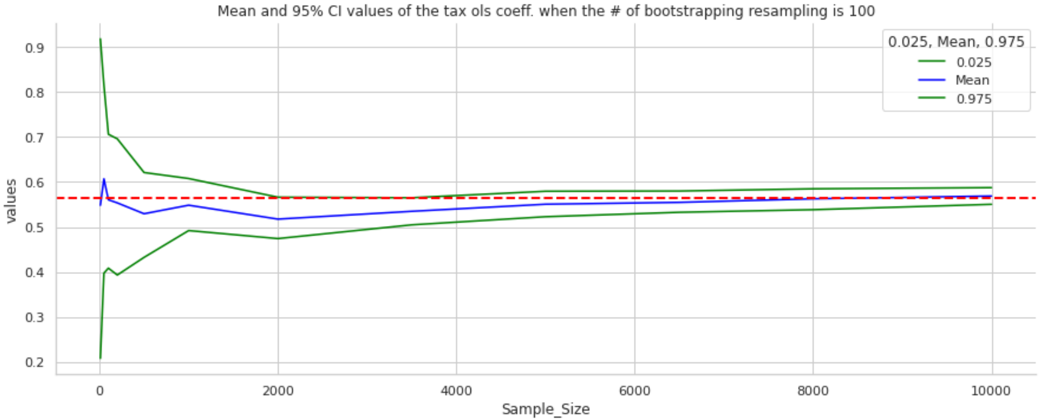


Figure: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 100

Bootstrapped OLS parameters (3)

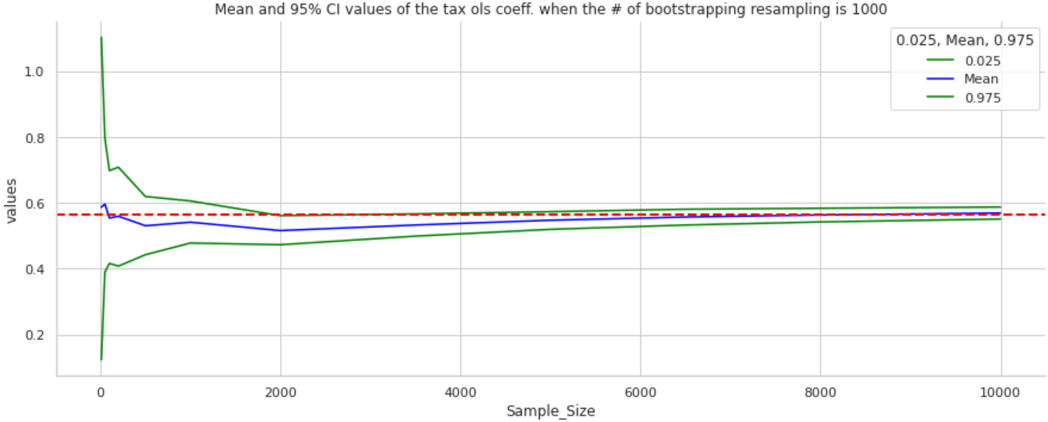


Figure: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 1000

Bootstrapped OLS parameters (4)

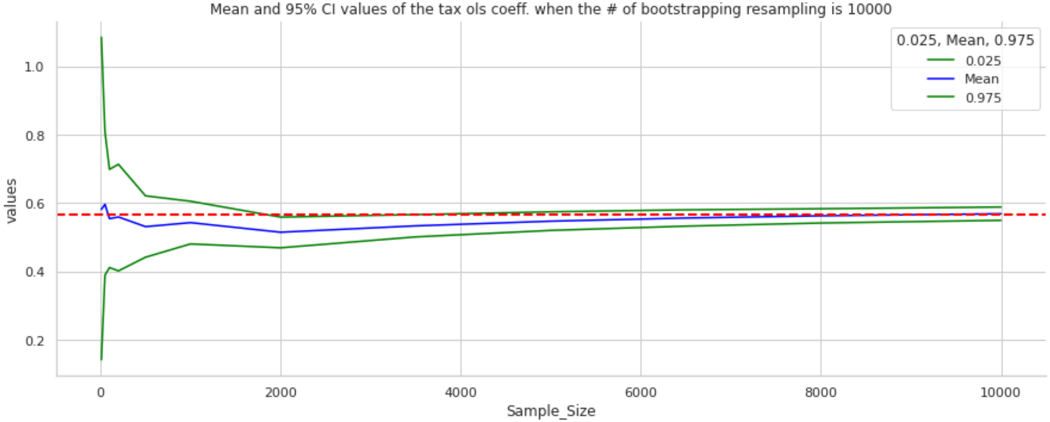
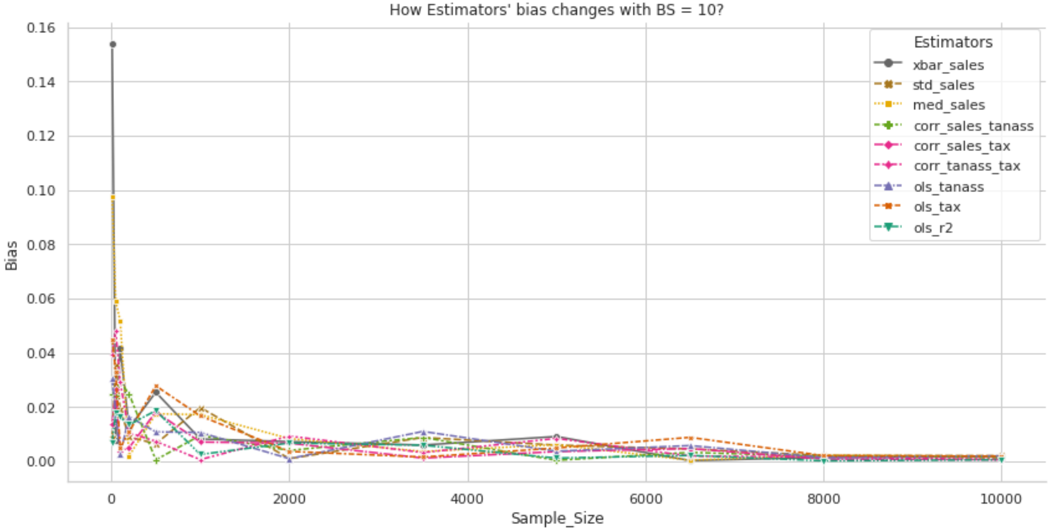
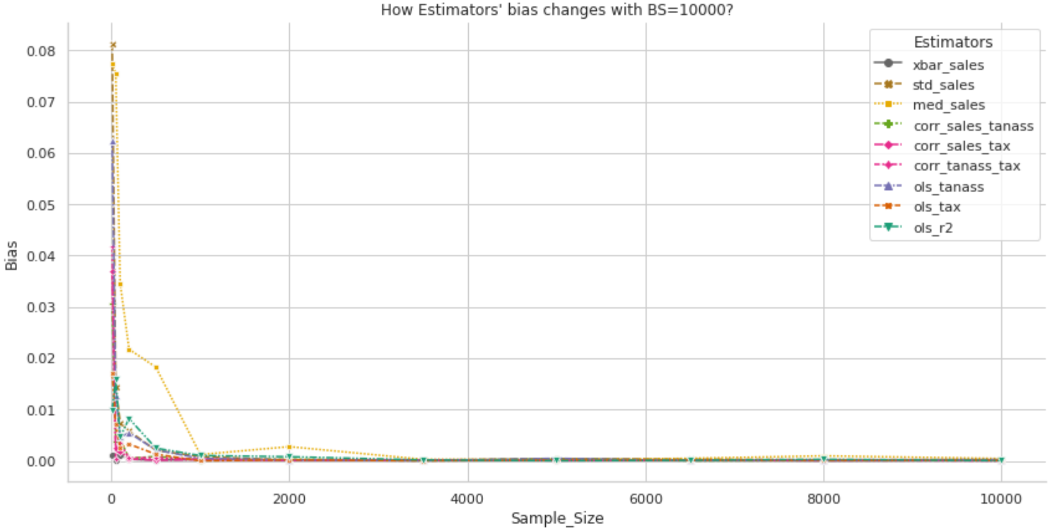


Figure: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 10000

Bias of the bootstrapped estimators (1)



Bias of the bootstrapped estimators (2)



Source code

The Python source code for this project can be found here.

Thank you very much for your attention!

Reference 1

- Wasserman, L. (2006). *All of nonparametric statistics*. New York, NY: Springer.
- Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- Shao, J. & Tu, D. (1995). *The jackknife and bootstrap*. New York, NY, USA: Springer Verlag.
- Rao, C. R. (1989). *Statistics and Truth. Putting Chance to Work*. International Co-operative Publishing House, Burtonsville, Md.
- Hansen, Bruce E. *Econometrics*. https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf

Reference 2

- Yeo, I., & Johnson, R. (2000). *A New Family of Power Transformations to Improve Normality or Symmetry*. Biometrika, 87(4), 954-959. http://www.jstor.org/stable/2673623

- Box, G., & Cox, D. (1964). *An Analysis of Transformations. Journal of the Royal Statistical Society*. Series B (Methodological), 26(2), 211-252. http://www.jstor.org/stable/2984418

- Weisberg, S. (2001). *Yeo-Johnson Power Transformations*. https://www.stat.umn.edu/arc/yjpower.pdf.

- Csörgő, S. (1990). *On the law of large numbers for the bootstrap mean*. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/30055/0000423.pdf?sequence=1