Bootstrapping statistics from companies 2018's balance sheet data in Hungary

Author: Viet Hung Pham
Supervisors: Olivér Kiss (CEU)
Dr. Miklós Koren (CEU)
Dr. Edith Alice Kovács (BME)

Abstract

This article aims to present the effectiveness of the traditional bootstrapping technique on various statistics with application on hungarian balance sheet data. Firstly, the con of traditional approach (that, is asymptotic theory and exact statistics) is discussed. Then the brief introduction of bootstrap method is detailed with the algorithm how it can be implemented. Finally, the rest of this article is going to be about the simulation and result of running the bootstrap on 2018's balance sheet data. The focus is on cleaned sales data, cleaned tangible assets, and tax. Before running bootstrapping, these skewed data are transformed using Yeo-Johnson transformation. Simulations are based on various sample sizes with a smoother range from 10 to 10.000 and on 4 bootstrap resampling numbers: 10, 100, 1000, 10000. Some sampling distribution of sample statistics (such as sample mean) will be compared through the varying number of sample sizes and bootstrap resampling number with 95% conficdence interval. OLS coefficents of tax and tangible assets are also plotted with 95% confidence. The bias of these bootstrapped estimators are also measured and compared in this article. From the simulations, it can be concluded that bootstrapped estimators are quite close to the true population values when using the suitable amount of sample size and even with lower number of bootstrap resampling.

Keywords: sample size, resampling with replacement, bootstrap, sample mean, sample standard deviation, sample median, correlation, ols, r-sqaured, bias

Acknowledgement

I would like to pronounce gratitude to Olivér Kiss for his guidance and continuous help throughout the process. I also like to thank Dr. Miklós Koren for the bootstrapping topic advise with this specific application and Dr. Edith Alice Kovács for her contribution to this individual project.

1 Introduction

The main aim of the statistics is to "extracting all the information from the data" [4] to make inferences about the population from the sample data we have. This can be done by *estimating a statistic(s)*, which is a function of data (usually denoted by $T(X)$) and selected according to the question of interest. A closely-related concept is the sampling distribution which is the probability distribution of a given random-sample based statistic. Most statistical problems requires some task-dependent knowledge about the sampling distribution of a statistic. In estimation problem, the knowledge of accuracy measures such as the mean, variance, bias, and mean squared error of the estimator is required as they are the characteristics of the estimator's sampling distribution [3]. The sampling distribution of a statistic and its properties are usually not known as they are based on the population. They have to be estimated from the collected and/or observed data. There are many techniques for estimating the sampling distribution such as jackknife and bootstrap.

2 Background

2.1 The problem of traditional methods

There are two traditional inference methods: asymptotic theory/large sample theory and exact statistics. Exact statistics such as linear regression is really useful to understand a lot of problems. However, it assumes a lot of unrealistic and highly restrictive conditions such as homogeneity of variance of a regression. The asymptotic theory, which assumes that given that the sample size can be grown indefinitely, the properties of estimators and test are then evaluated under $n \to \infty$. Even though the large sample theory provides more flexibility in distribution theory, it may approximate with uncertain accuracy and may have some statistics that are highly complicated to approximate. [5]

To overcome these difficulties, a new set of statistical inference technique is introduced which are based on the concept of resampling. In short, it is a method of extracting sampling information from the empirical distribution. One of the resampling method is bootstrapping which is going to be detailed in the rest of this article with application on real-world data.

2.2 Bootstrapping

The name comes from the common English idiom "pull yourself up by your (own) bootstraps" which means according to the Oxford Dictionary improve your situation yourself, without help from other people. Here it means sampling from an existing sample itself without using external samples. In our context, it means sampling from an existing sample itself without using external samples Bootstrapping overcomes most of the obstacles in the previous. This technique was published by Bradley Efron in "Bootstrap methods: another look at the jackknife" (1979) [2]. The main idea of the bootstrapping is that one has a sample of size n from the population. New resamples of size n are generated from the original, given sample *with replacement*. The probability that an individual observation will occur at least once in the bootstrap sample after sampling *with replacement* from the original sample is

$$P(\text{an observation in bootstrap sample}) = 1 - \left(1 - \frac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \approx 0.632, \qquad (1)$$

as n tends to infinity. [5]
The number of resamples are subject to the preference, but usually a 10.000 resampling number is considered enough most of the time. Then, calculate a statistic of interest for each new sample thus creating a so-called bootstrap sampling distributions for that statistic. In the end, it is possible to either construct a bootstrap confidence interval for our statistic or aggregate of each resample statistic to provide the estimate statistic for the population.

There are many advantages to the bootstrap techniques including:

- their simplicity easy to derive estimate of mean, standard errors and confidence intervals.
- their goodness in controlling and checking the stability of the outcomes. Bootstrap methods are asymptotically more accurate than the standard intervals obtained using sample variance and assumption of normality.
- their practicality to avoid the cost of repeating the experiment to obtain other groups of sample data

Nevertheless, they also carry some disadvantages:

- They do not provide general finite-sample guarantees so the result may depend on the (representative) sample.
- Bootstrapping can be time-consuming.

The following figure shows the pseudocode for implementing the classic bootstrapping.

Algorithm 1: Classic Bootstrap Algorithm for estimating a parameter

Input : $S = \{x_1, x_2, \ldots, x_n\}$ – a size n original samples
Output: $\hat{\theta^*} = (\theta_1^*, \hat{\ldots}, \theta_n^*)$ – 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

3 Application on balance sheet data

2018's company balance sheet data was used for bootstrapping. Specifically, 3 features is going to be used: *sales_clean*, *tanass_clean*, *tax*, which are cleaned sales, cleaned tangible asset, tax data from 2018. Figures for these variables should be interpreted as *value* × *1000 Forint*.

3.1 Data Exploration

First, only companies with sales figure greater than 0 and companies with valid tax and tangible asset values are examined. Originally, there are 420018 instances. After conducting the above-mentioned restriction, this figure becomes 332604. This data is going to be treated as the population data.

	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*

Table 1: Summary statistics of the variables

Table 1 describe the general descriptive statistics about the features. What is striking is that the mean and median values especially for *tanass_clean* and *tax* differ significantly. Furthermore, relative standard deviation (which is calculated as std/mean) of these features is high. These two observations imply that the distribution of these variables are highly skewed. In this case, they are skewed towards lower figures as it can be seen in the quartile values. This fact can be demonstrated with Figure 1. The problem with very skewed data is when it comes to sampling from the population (the sampling method will be *without replacement* with uniform distribution), the resulting sample which is going to be used for bootstrapping is not representative. Therefore, data transformation is needed.

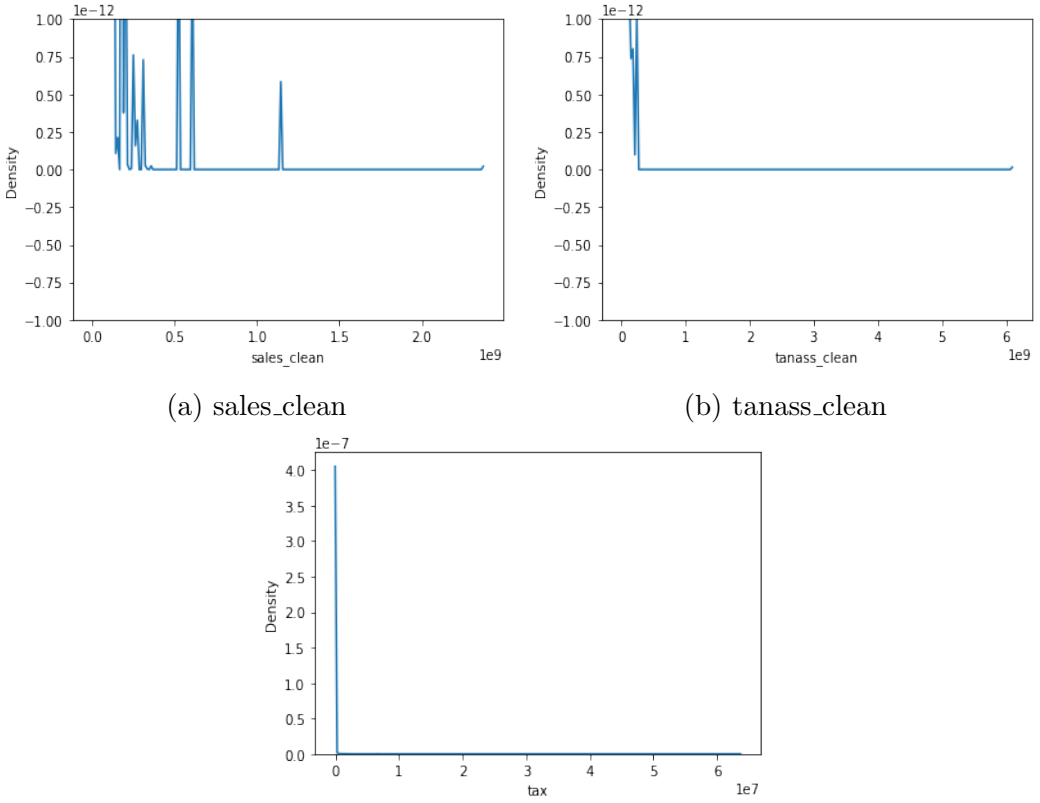
3.2 Data Transformation

Data transformation, which is defined on $\mathbb{R}$ for reducing skewness and making a distribution normal-like, has a significant role in regression tasks [6], [8]. Brief detail about the **Box-Cox** transformation should be outlined, because the Yeo-Johnson transformation is just the extension of the Box-Cox transformation for non-positive variables.

The transformation made by Box & Cox [7] is the $\psi^{\mathbf{BC}}$ given by:

$$\psi^{\mathbf{BC}}(\lambda, x_i) = \begin{cases} (x_i^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, \\ log(x_i) & \text{if } \lambda = 0, \end{cases} \tag{2}$$

for positive x_i. Those two are the data point in our data set for all i.



(a) sales_clean

(b) tanass_clean

(c) tax

Figure 1: Distribution of the original variables

The transformation proposed by Yeo & Johnson [6]:

$$\psi^{\mathbf{YJ}}(\lambda, x_i) = \begin{cases} ((x_i+1)^{\lambda}-1)/\lambda & \text{if } \lambda \neq 0,\ x_i \geq 0, \\ log(x_i+1) & \text{if } \lambda = 0,\ x_i \geq 0, \\ -[(-x_i+1)^{2-\lambda}-1]/(2-\lambda) & \text{if } \lambda \neq 2,\ x_i < 0, \\ -log(-x_i+1) & \text{if } \lambda = 2,\ x_i < 0, \end{cases} \tag{3}$$

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*. The Yeo-Johnson transformation is the same as the Box-Cox transformation of $(x_i + 1)$ when x_i is positive. However, when x_i is negative, the Yeo-Johnson transformation is the same as the Box-Cox transformation of $(-x_i + 1)$ with the power of $(2 - \lambda)$. When λ equals either 0 or 2, a similar method applies to them but now with the *log* transformation. It is also important to notice that +1 at both x_i and $-x_i$. Therefore, the case when x_i is equal to 0 can be handled. The power of the Yeo-Johnson transformation is that it enables us to extend the notion of normality approximation to variables containing 0 and negative values.

3.3 Analysis of the transformed data

We apply Yeo-Johnson transformation and standardization for the variables. It can be seen from figures and distribution plots that with these alterations, the data is much less skewed

(in the case of *sales_clean*, its distribution resembles somewhat normal) that is variance and extreme values are under control and are not going to influence severely when it comes to sampling from the population (without replacement) for bootstrapping. The sample is going to be more representative of the population.

	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

Table 2: Yeo-Johnson transformed summary statistics of the variables

Table 2 reinforces the above-mentioned statements. It can be recognised that, because of the standardization, the mean and standard deviation is 0 and 1, respectively. Mean and median values does not differ significantly and max values are not strikingly extreme at all. Figure 2 of distribution plots also confirm these findings. Spikes in the lower region values of the transformed *tanass_clean* and *tax* can be observed because as opposed to the original *sales_clean* value, 0s in the original*tanass_clean* and *tax* are not excluded.

3.4 OLS Regression and Correlation

OLS Regression is applied to the transformed variables in the following way:

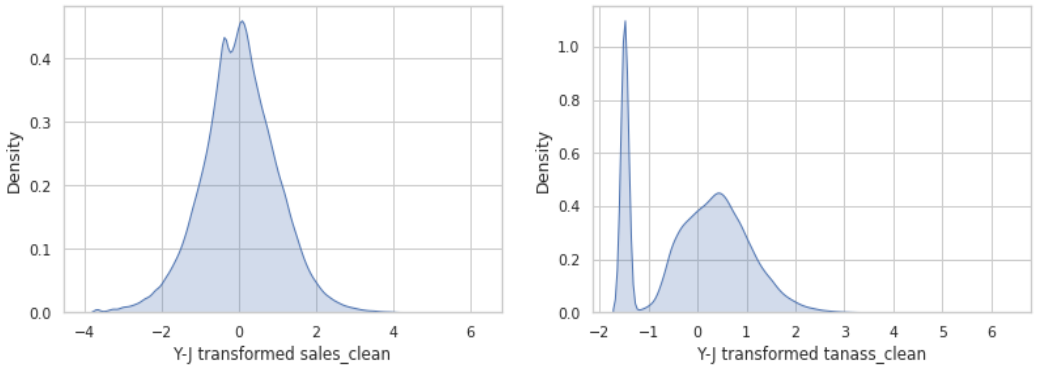
$$\text{sales\_clean} = \beta_0 + \beta_1 \times \text{tanass\_clean} + \beta_2 \times \text{tax}, \tag{4}$$

where β_0 is the intercept, β_1 is the parameter for *tanass_clean*, and β_2 is the parameter for *tax*. So the dependent variable is going to be the *sales_clean*. Table 3 shows the parameters of (our) interest of the OLS regression for the population data.

β_1	0.3324
β_2	0.5654
R-squared	0.569

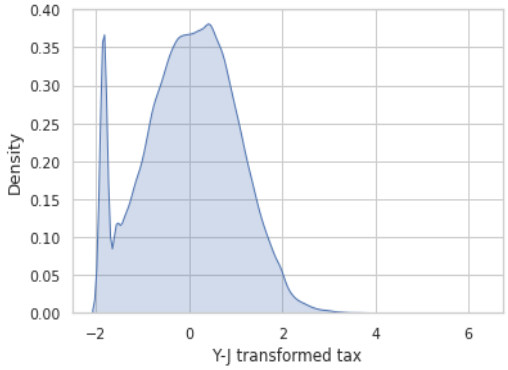
Table 3: OLS population parameters

Pearson correlations among the transformed variables are also discovered in Table 4. It can be observed that there are strong correlations between *tax* and *sales_clean* with coefficient of 0.688, and between *tanass_clean* and *sales_clean* with coefficient of 0.541.



(a) Y-J sales_clean

(b) Y-J tanass_clean



(c) Y-J tax

Figure 2: Distribution of the Yeo-Johnson transformed variables

3.5 Sample Sizes

As mentioned above, data transformation plays pivotal role when it comes sampling from the population. In the following link containing a series of plots demonstrates how the distribution of transformed variables changes as different sample sizes are selected ranging from 10 to 10000:
Figure 6 and 5 together in Appendix B contains a series of plots demonstrates how the distribution of transformed variables changes as different sample sizes are selected ranging from 10 to 10000. As expected, the larger the sample size, the more it is resemble to the population distribution of variables shown in Figure 2. The surprising observation is

Correlation	**sales_clean**	**tanass_clean**	**tax**
sales_clean	1	0.541	0.688
tanass_clean	0.541	1	0.370
tax	0.688	0.370	1

Table 4: Correlation between the variables

that even with the sample size of 500, 1000 and 2000, the original distributions are well approximated.

3.6 Sampling distributions of bootstrapped statistics

The following statistics is used for bootstrapping:

- *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 correlation between sales and tanass
- *corr_sales_tax*: bootstrapped sampling distribution of correlation between sales and tax
- *corr_tanass_tax*: bootstrapped sampling distribution of correlation between 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

Bootstrapped sampling distribution of sales mean, sales standard deviation and R-squared are plotted in Appendix C (Figure 8, 7), Appendix D (Figure 10, 9), Appendix E (Figure 12, 11), respectively. Mean of the bootstrapped estimator and the 95% intervals are marked in these plots. It can be observed that with greater number of sample sizes **and** bootstrap resampling the mean of those bootstrapped statistics are very close to the true population value. The reason for that is the underlying Law of Large Numbers.[9]

3.7 Bootstrapped OLS parameters

Figure 13, 14, 15, 16 from Appendix F present the behaviour of mean and 95% CI of bootstrapped *tax* OLS parameter with the red dash line showing the population OLS tax parameter value which is 0.5654. It can be pointed out that regard that with greater number of bootstrap resampling the 95% confidence interval is narrower and the convergence of mean bootstrapped tax OLS parameter to the true one is faster. The same applies to the bootstrapped tanass_clean OLS parameter.

3.8 Bias of the bootstrapped estimators

The following figures show how the (absolute) bias of each estimator evolves with growing sample sizes when the number of the bootstrap resampling is 10 and 10000.

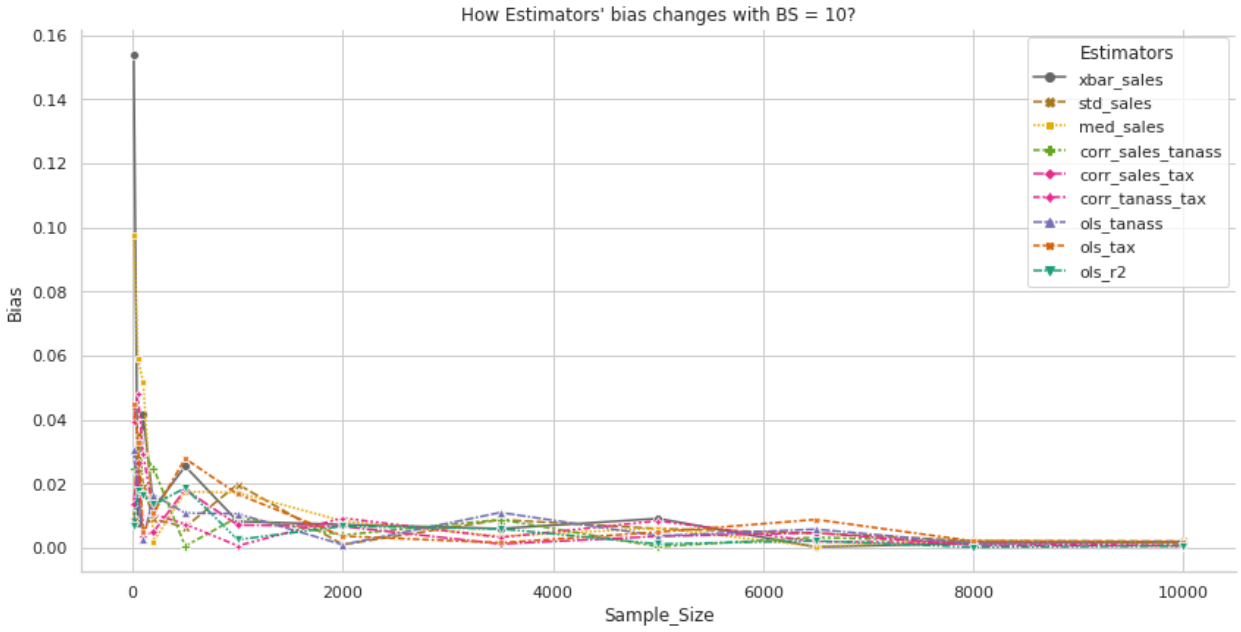


Figure 3: Bias when the # of bootstrap resampling is 10

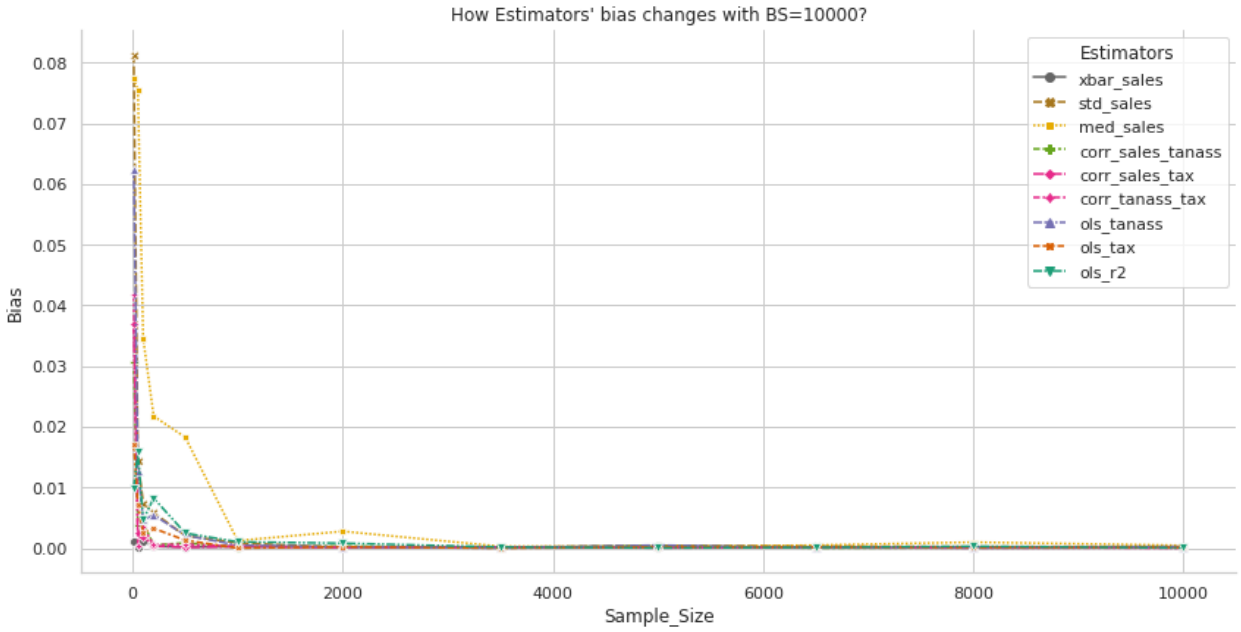


Figure 4: Bias when the # of bootstrap resampling is 10000

It can be easily noted that bias for all estimators converges to 0 much faster and more smoothly when the bootstrapping resampling number is higher.

3.9 Conclusion

With the right treatment of the data, the classic bootstrapping method is proved to be a very effective and nowadays very efficient tool (because of the increase in computational power) to estimate the true population value. We have shown the power of this resampling technique on several examples above. Since the introduction of the original one, many improved version of bootstrapping techniques have been developed. The possible continuation of this individual project is the examination of several (wild) cluster bootstrap methods on this data.

Appendices

Appendix A Source Code

The Python source code for this project can be found on the following link:
`https://github.com/pvh95/Individual_project2_BME/blob/main/bootstrap_on_bs.ipynb`

Appendix B Sample Size plot

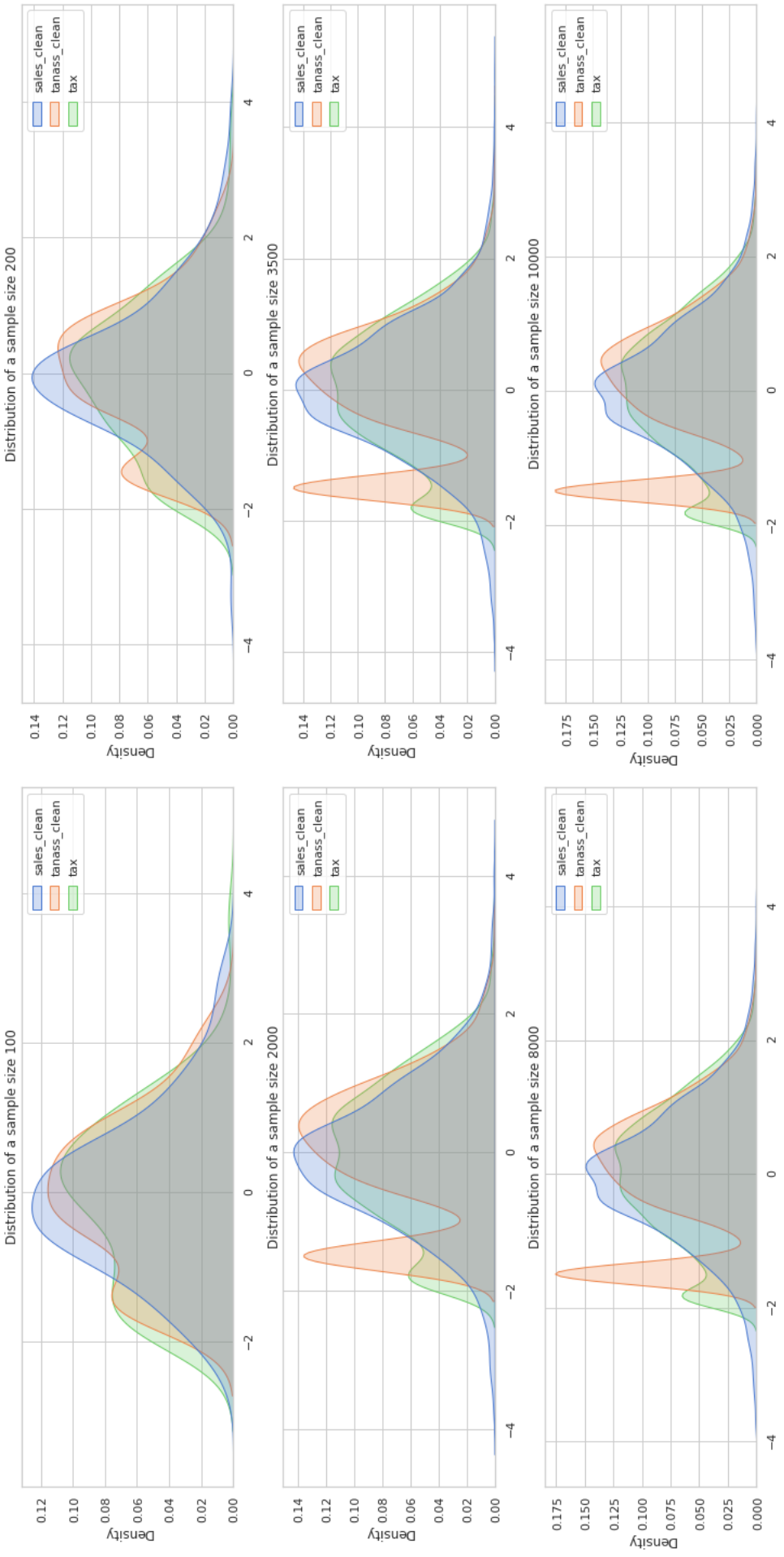


Figure 5: Distributions of variables under varying sample sizes Part 2

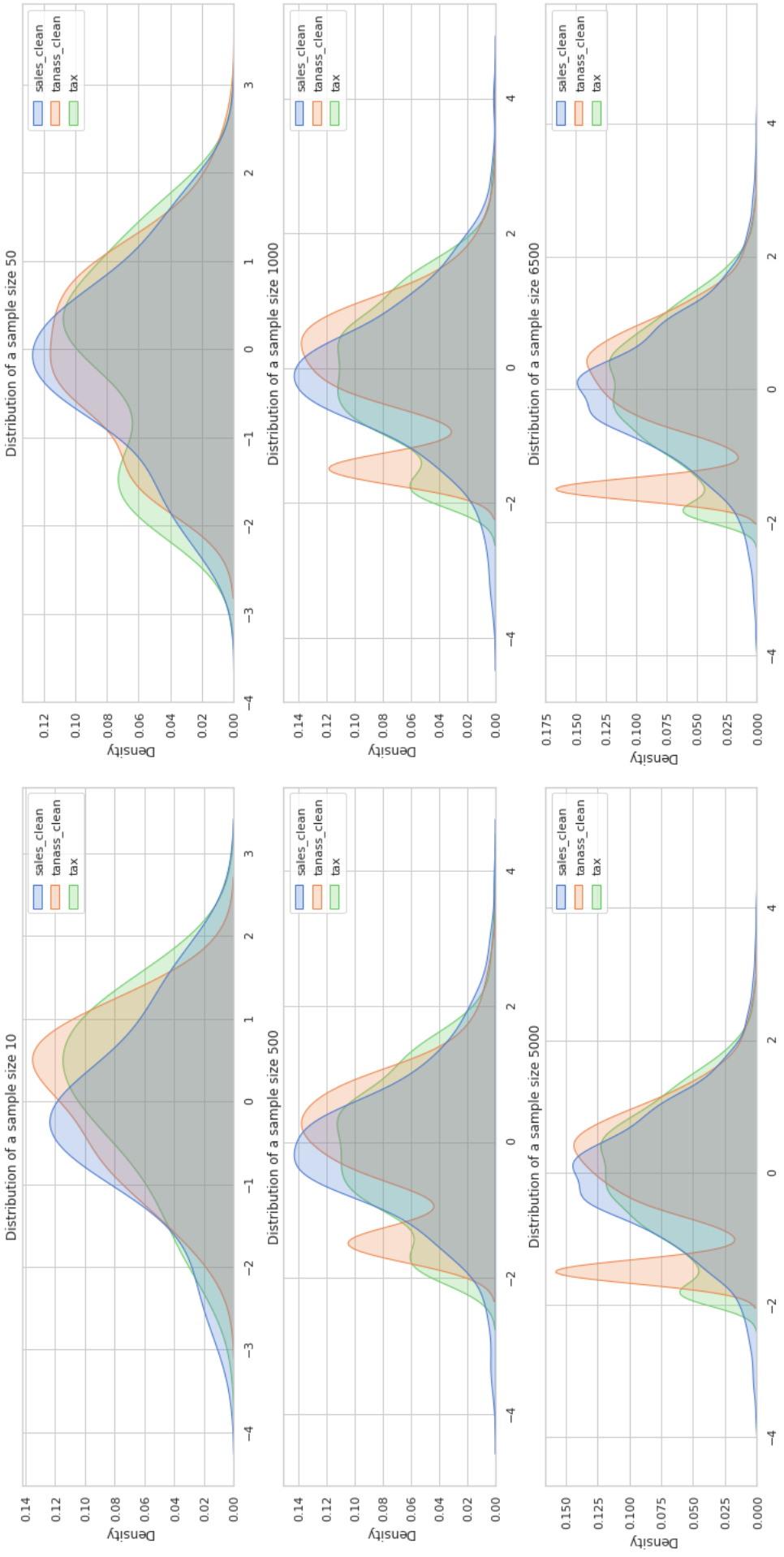


Figure 6: Distributions of variables under varying sizes Part 1

Appendix C Bootstrapped Sales Mean Plot

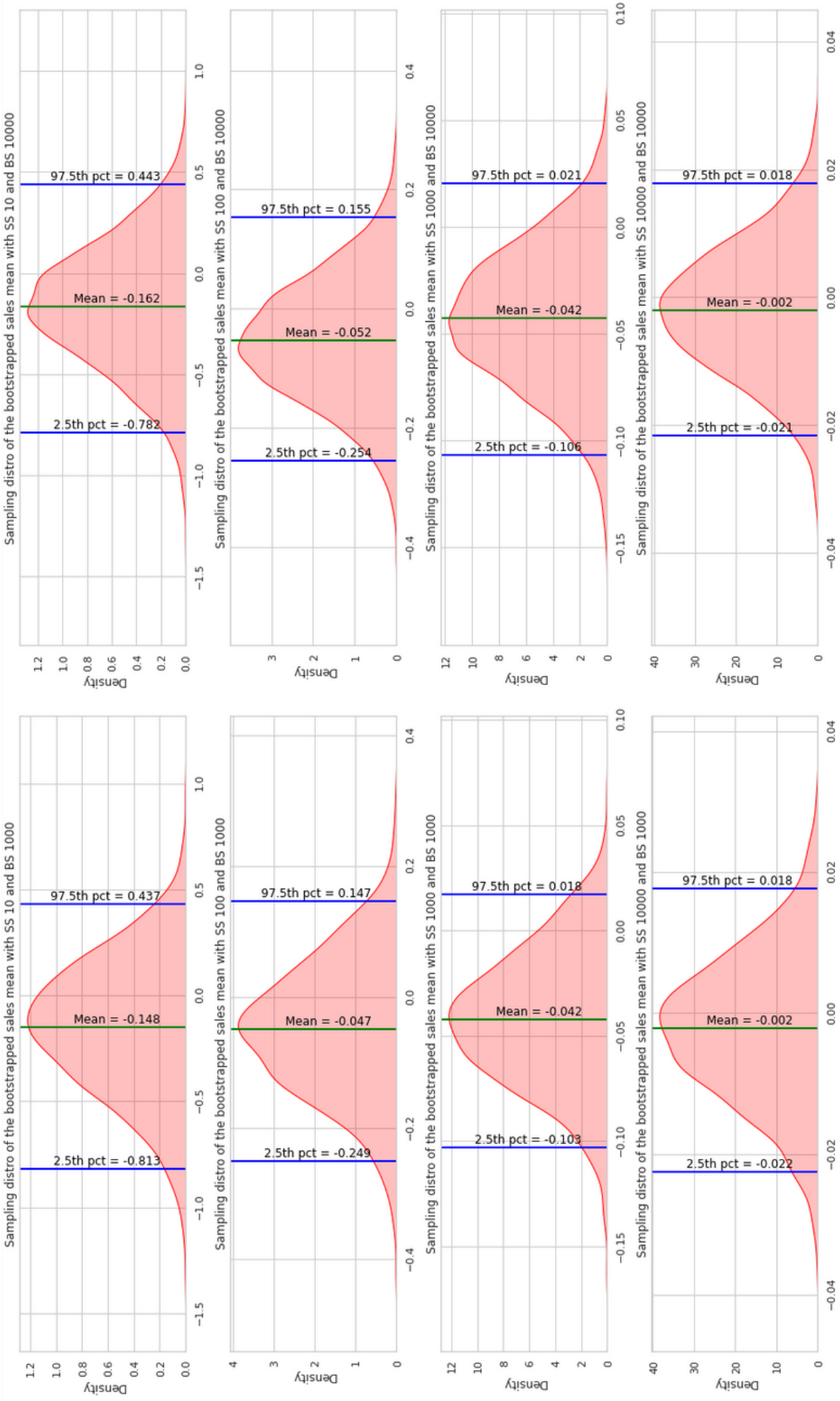


Figure 7: Sampling distributions of sales mean Part 2

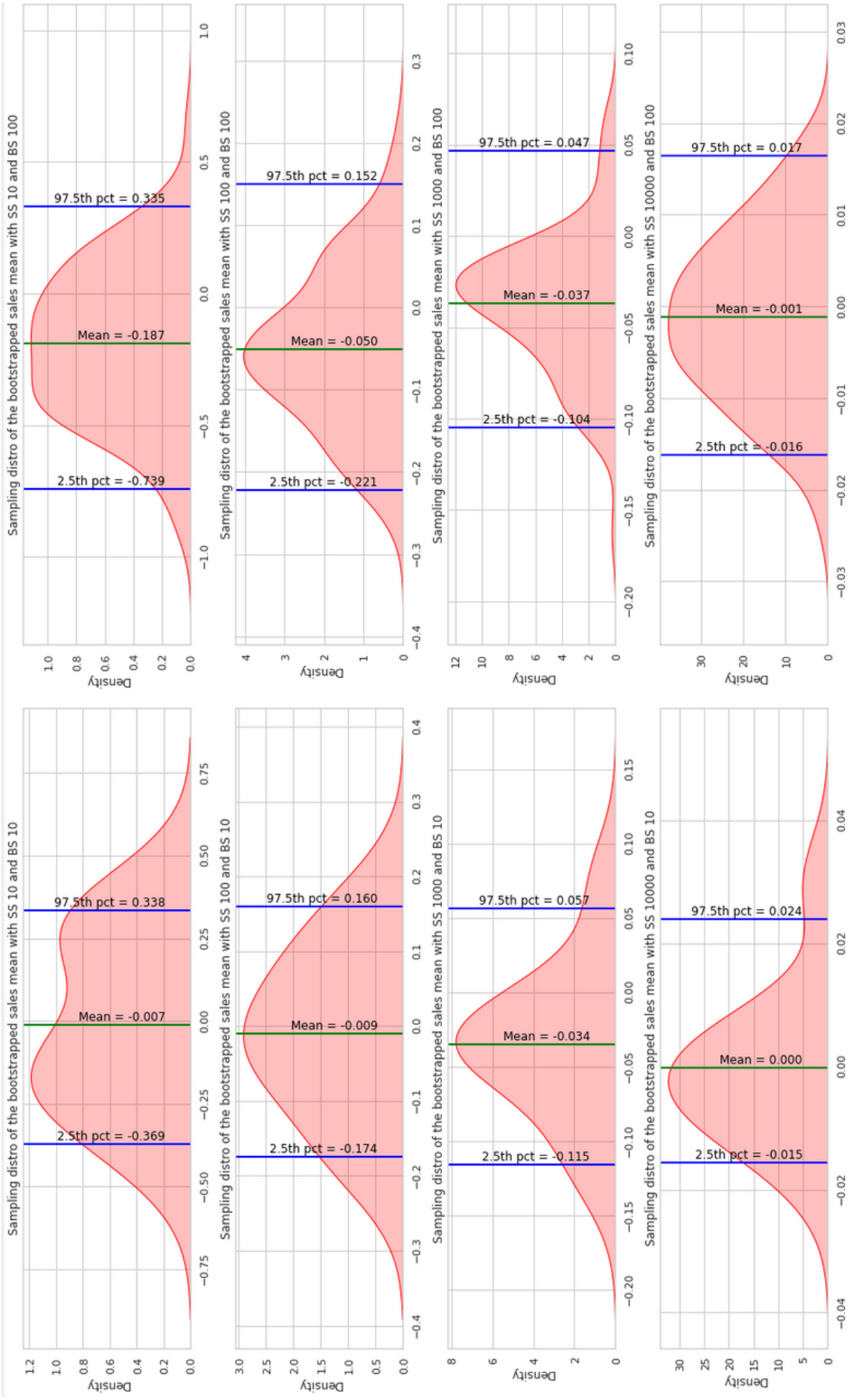


Figure 8: Sampling distributions of sales mean Part 1

Appendix D Bootstrapped Sales Std Plot

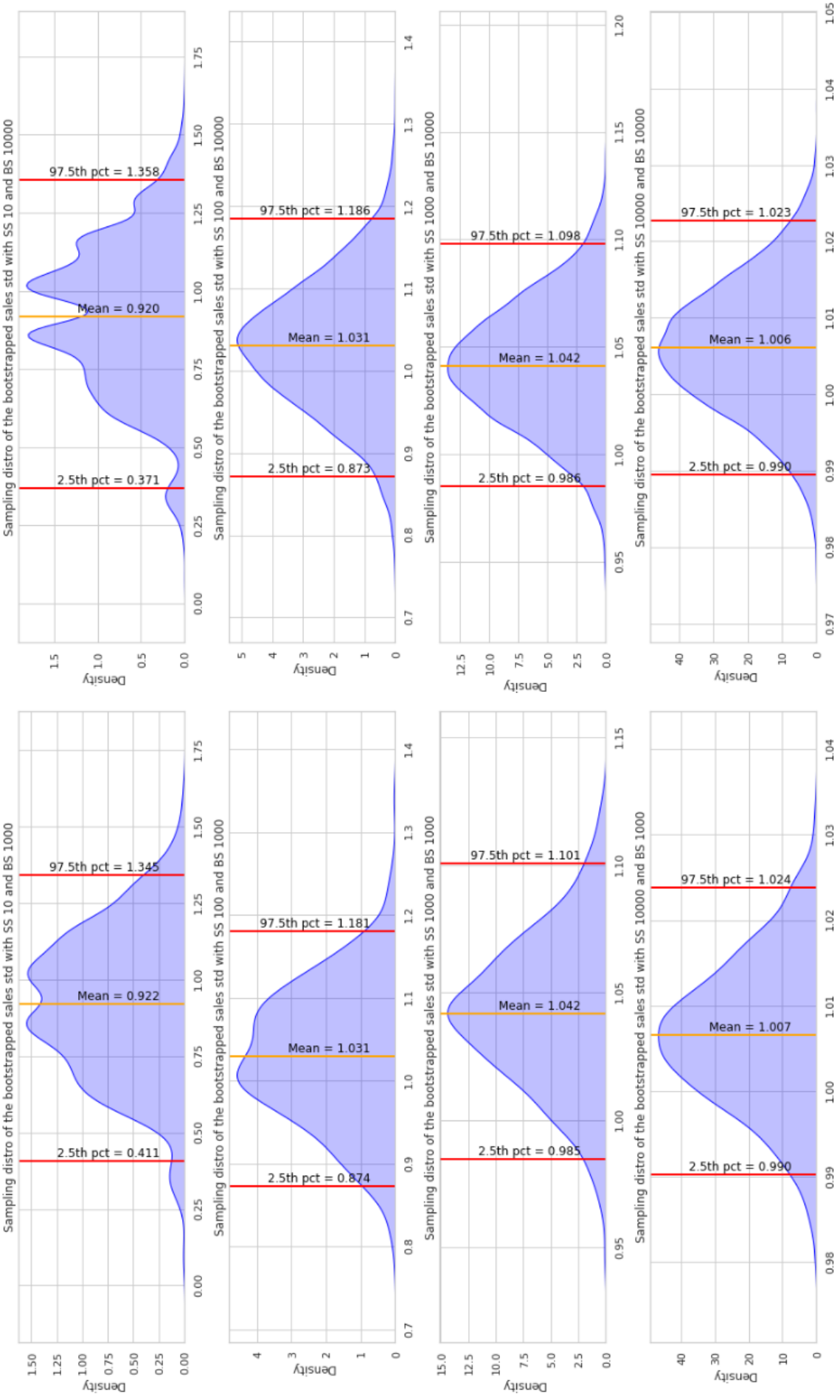


Figure 9: Sampling distributions of sales std Part 2

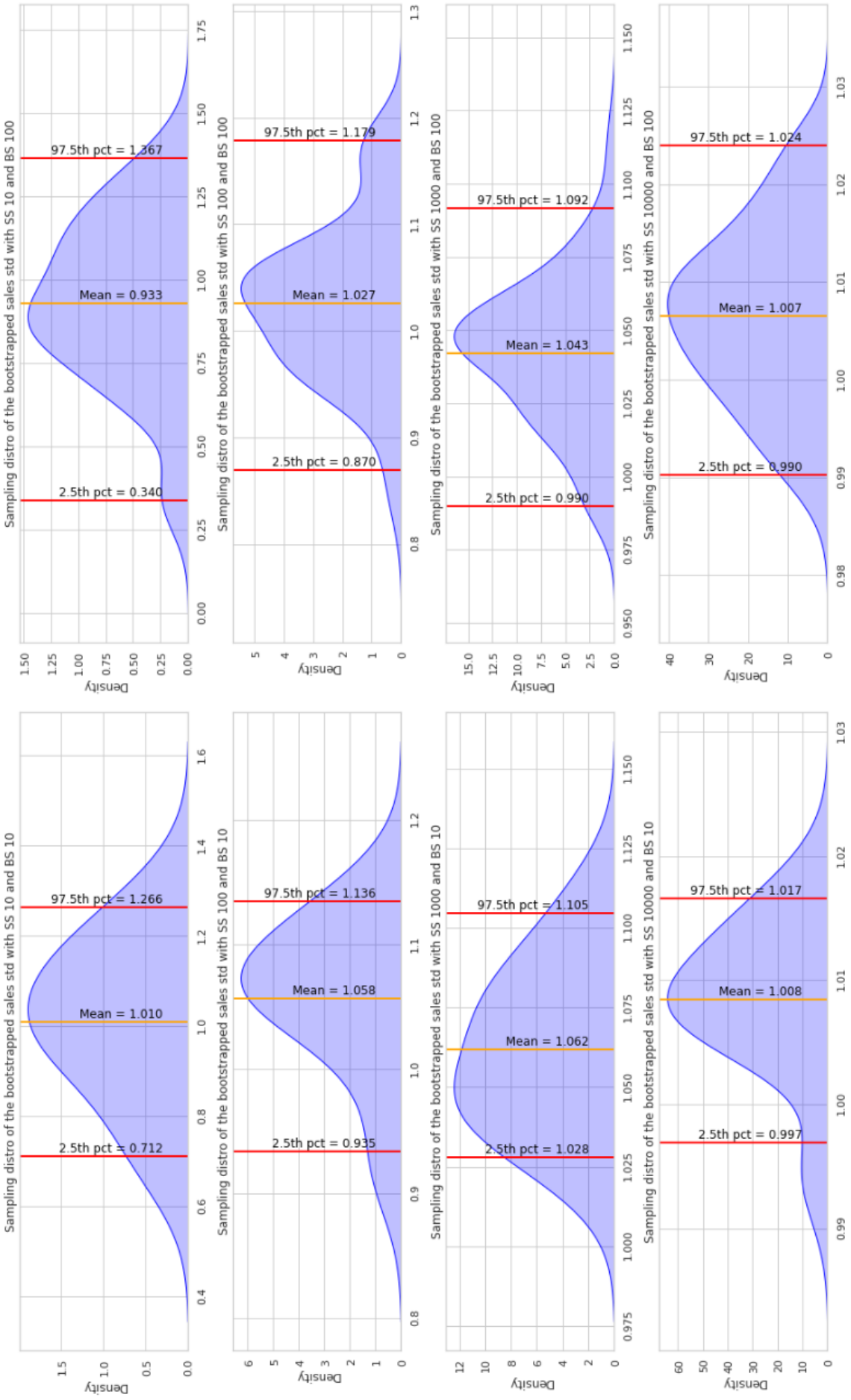


Figure 10: Sampling distributions of sales std Part 1

Appendix E Bootstrapped R-squared Plot

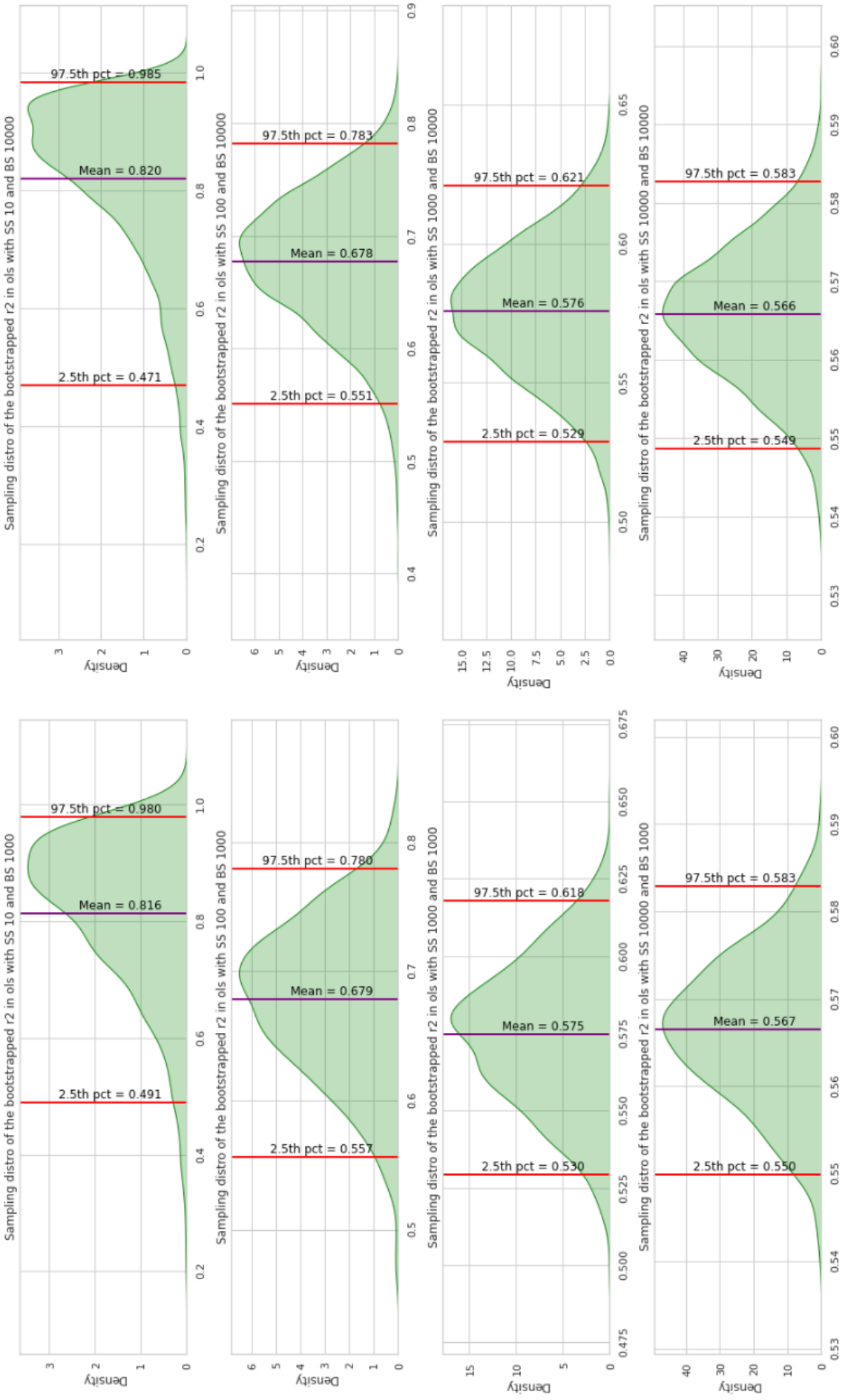


Figure 11: Sampling distributions of R-squared Part 2

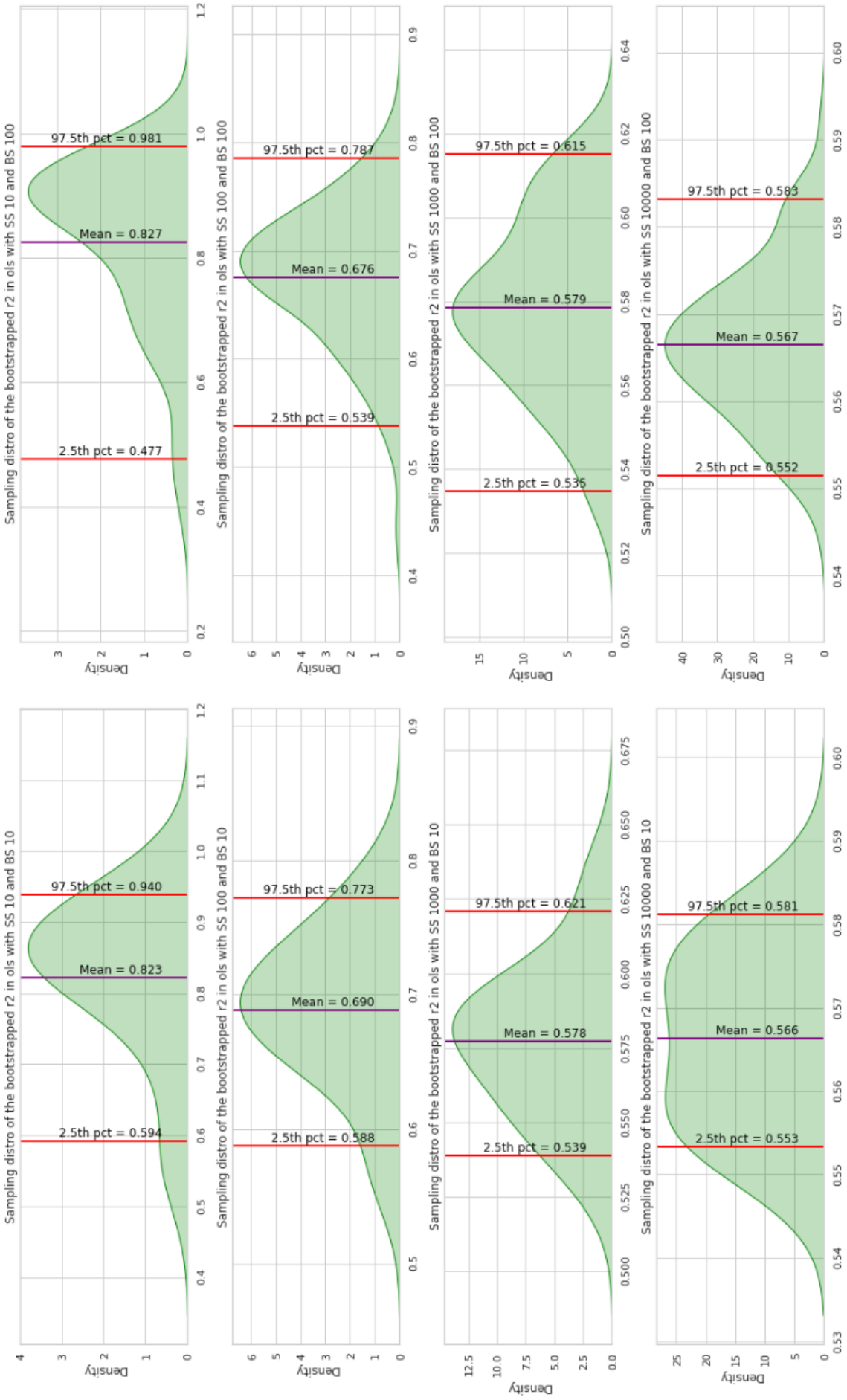


Figure 12: Sampling distributions of R-squared Part 1

Appendix F Bootstrapped OLS parameter plots

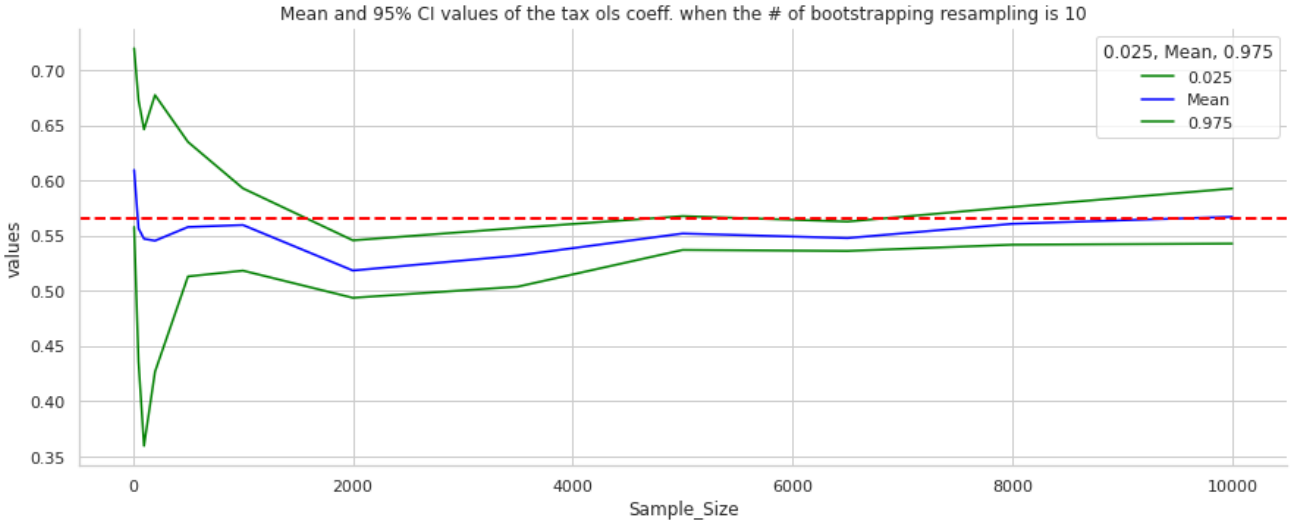


Figure 13: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 10

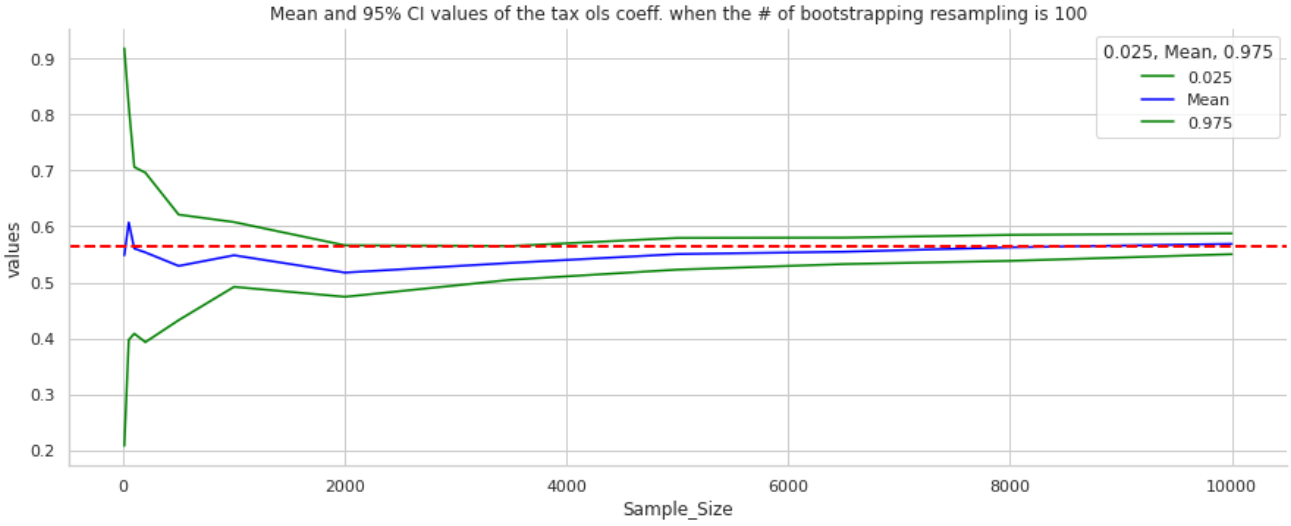


Figure 14: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 100

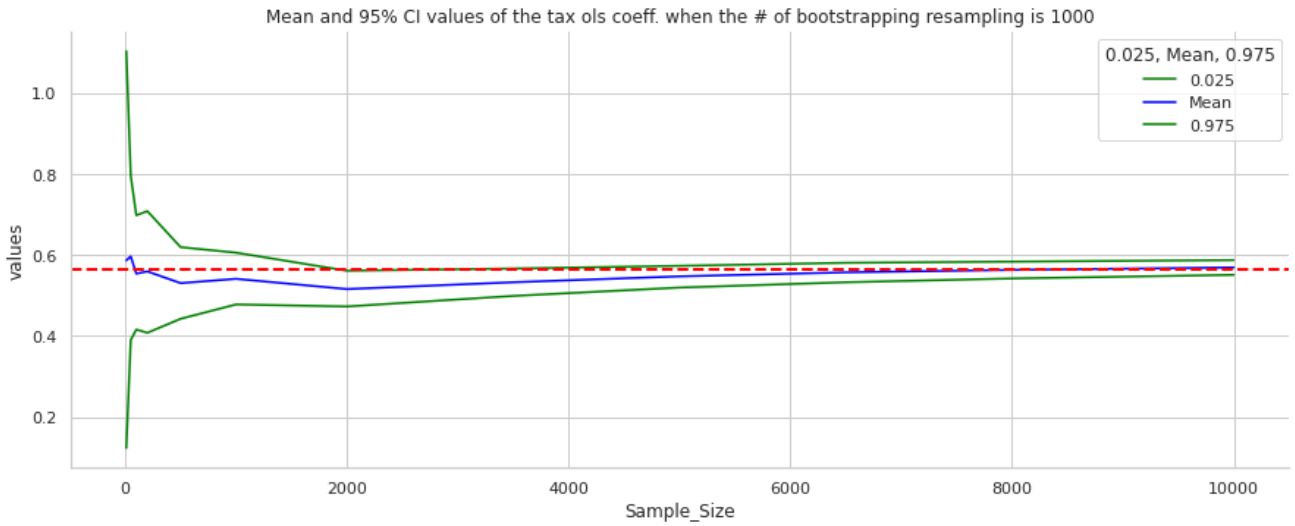


Figure 15: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 1000

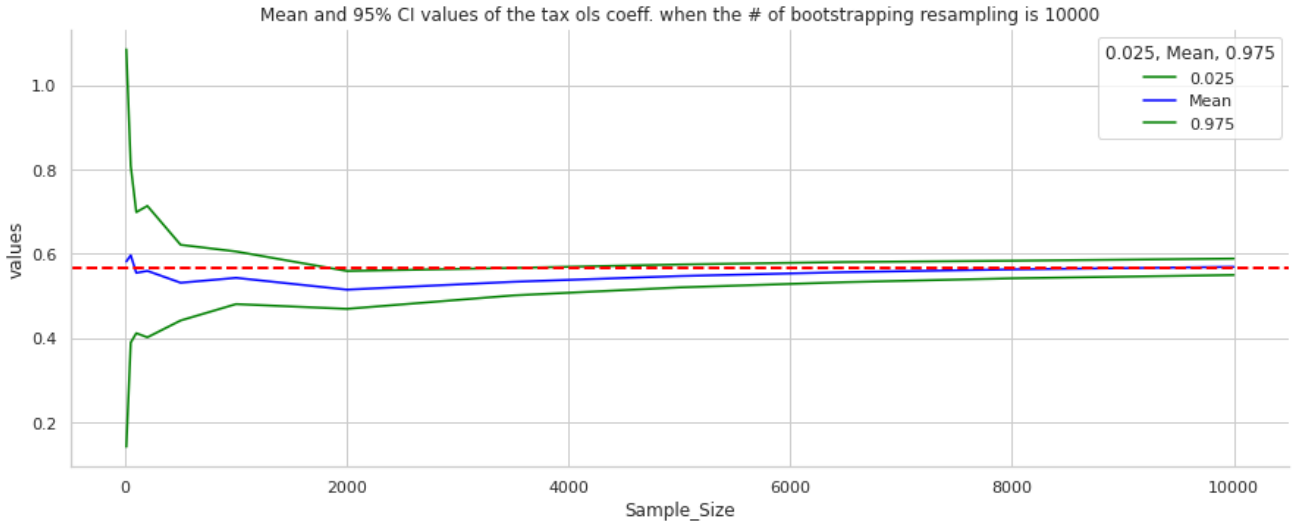


Figure 16: Mean and 95% CI of Bootstrapped tax OLS when the # of bootstrap resampling is 10000

References

[1] Wasserman, L. (2006). *All of nonparametric statistics*. New York, NY: Springer.

[2] Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

[3] Shao, J. & Tu, D. (1995). *The jackknife and bootstrap*. New York, NY, USA: Springer Verlag.

[4] Rao, C. R. (1989). *Statistics and Truth. Putting Chance to Work*. International Co-operative Publishing House, Burtonsville, Md.

[5] Hansen, Bruce E. *Econometrics*. `https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf`

[6] 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`

[7] 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`

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

[9] 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`