Computer Intensive Methods - Project 3

1 Chickwts dataset (Question 1)

In the first task, we will analyze the data set *chickwts*, which consists of two features: the weight of the chicks and their diet (*feed*). The second of them takes 6 categorical variables. Let's denote the set of categorical variables: $D = \{horsebean, linseed, soybean, sunflower, meatmeal, casein\}$. Below, we will check whether diet affects the weight of the chicks, i.e., using significance level $\alpha = 0.05$, we will test the hypotheses:

$$H_0 : \forall_{d \in D} \forall_{f \in D : f \neq d}\, \mu_d = \mu_f \qquad H_1 : \exists_{d \in D} \exists_{f \in D : f \neq d}\, \mu_d \neq \mu_f,$$

where μ_i means the average of weight of chickens with the i diet. By testing the above hypotheses we will check whether the average weights of the chickens for each class of the *feed* feature are the same.

1.1 One-way ANOVA

The one-way ANOVA model we can formulate as:

$$y_{i,j} = \mu + \tau_j + \varepsilon_{i,j},$$

where: $y_{i,j}$ is an observation, μ is the grand mean of the observations, $\tau_j = \mu_j - \mu$. μ_j is the weight mean for the j category from D and $\varepsilon_{i,j}$ are the random errors from the normal distribution with zero mean. The statistic that will help us decide which hypothesis is true is the F statistic:

$$F = \frac{\frac{1}{r-1} \sum_{i=1}^{r} n_i (\overline{x_i} - \tilde{x})^2}{\frac{1}{n-r} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{i,j} - \overline{x_i})^2},$$

where $\overline{x_i}$ and n_i is the average weight and number of chickens on the i diet, respectively. The $\tilde{x}$ is the average weight of the chickens, n is the number the chickens and $r = \#D$. In our case, under H_0, the F has Fisher-Snedecor distribution with 5 and 65 degrees of freedom. We reject the H_0 when F is equal or greater than $1 - \alpha$ quantile of the F-distribution with 5 and 65 df.

In our case $F = 15.365$ and value $p = 5.936 \cdot 10^{-10}$ (< 0.05), which indicates that the null hypothesis is false.

1.2 Test with semi-parametric bootstrap

Let's test the previously defined test problem using semi-parametric bootstrap. First, let us assume that the null hypothesis is true, then one-way ANOVA model has form:

$$y_{i,j} = \mu + \tau_j + \varepsilon_{i,j},$$

where $\tau_j = 0$ for all $j \in D$. The weight depends only on random errors, it doesn't depend on the diet(all means of weights depend on the diet are equal).

At the beginning of the iteration, let's resample the random errors and build a model on them. Then, let's calculate the F statistic. Let's repeat the iteration 1000 times. The p-value estimate will be the fraction of obtained statistics that are greater than the observed F statistic from Section 1.1. It is equal to 0.00099 (<0.05), so the null hypothesis is false.

If the null hypothesis were true, resampling the random errors would not change the model significantly, and the values of obtained statistics would oscillate around the observed F statistic.

1.3 Permutation test

In the permutation test, the main idea is that if H_0 is true, then a model based on the response vector Y and the explanatory vector X is not significantly different from a model built on Y and a permutation of X. That is, in our case the models should give similar F statistics if H_0 were true.

Let's perform a permutation test. Firstly, during one iteration, let's resample (with: replacement=FALSE) the vector X (i.e. the *feed* column in the data frame *chickwts*), thus obtaining a permutation of X. Then let's build an ANOVA model based on the permutation of X and the weights vector Y and calculate the F statistic. Let's repeat the iteration 1000 times and determine the fraction of obtained statistics that are greater than the observed F statistic from section 1.1 (this will be an approximation of the p-value for the permutation test).

As a result of the test, we got a p-value equal to 0.00099 (<0.05), which means that the null hypothesis is false.

1.4 Estimator of $\theta = \mu_{sunflower} - \mu_{soybean}$

Below we use parametric bootstrap to estimate the parameter $\theta = \mu_{sunflower} - \mu_{soybean}$ and determine its 95% confidence interval. We know that the distribution of the estimator $\hat{\theta}$ is $N\left(\mu_{sunflower} - \mu_{soybean}, \frac{\sigma^2_{sunflower}}{12} + \frac{\sigma^2_{soybean}}{14}\right)$ (12-size of the sunflower class; 14-size of the soybean class). We don't know the parameters of that distribution $\hat{\mu}_{sunflower}$ and the $\hat{\mu}_{soybean}$. For this reason, we will simulate sample of weights and estimate θ based on them.

During a single iteration, we will generate a sample of random variables from a normal distribution with mean equal to the sample mean of the class *sunflower*, and variance equal to the sample variance of the same class. We will repeat a similar operation for the class *soybean*. The generated samples will be of equal size as in the case of samples of the corresponding classes from the data set. Based on the generated samples, we will calculate the estimates of the means: $\hat{\mu}_{sunflower}$ and $\hat{\mu}_{sunflower}$, then we will calculate $\hat{\theta}^*$. Parametric bootstrap will consist in repeating this iteration 1000 times (we generate sample from a distribution).

As a result of the simulation, we obtained 1000 estimates values of the θ. The estimator of theta is the mean of the sample of the statistics $\hat{\theta}^*$. It equal to:

$$\hat{\theta} = 81.425.$$

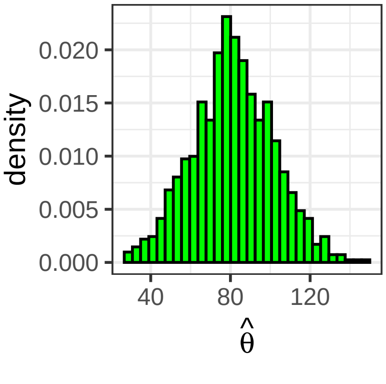


Figure 1: The histogram of $\hat{\theta}$

The theoretical 90% C.I has form:

$$\left[\hat{\theta} - t_{0.95,24}\sqrt{\frac{\hat{\sigma}^2_{sunflower}}{12} + \frac{\hat{\sigma}^2_{soybean}}{14}}; \hat{\theta} + t_{0.95,24}\sqrt{\frac{\hat{\sigma}^2_{sunflower}}{12} + \frac{\hat{\sigma}^2_{soybean}}{14}}\right] = [46.866; 115.985]$$

The bootstrap 90% C.I has form:

$$\left[\hat{\theta} - \hat{\beta} + C_{0.05}\sqrt{\hat{\nu}}; \hat{\theta} - \hat{\beta} + C_{0.95}\sqrt{\hat{\nu}}\right] = [47.942; 114.908],$$

where $\hat{\beta}$ is mean of $\hat{\theta}_b^* - \hat{\theta}$ and $\hat{\nu}$ is sample variance of the bootstrap sample of $\hat{\theta}^*$.

2 Computers dataset (Question 2)

In this section we will mainly focus on the bootstrap methods and their impact on statistical models based on *Computers* dataset. Firstly, let's consider the linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where Y_i is the price of the i-th. computer (he variable *price* in the dataset) in US dollars. The X_i is the size of hard drive in MB of the i-th. computer (the variable *hd* in the dataset).

2.1 The OLS Model

At the beginning, we will start by estimating the above model using the OLS classical OLS approach. The estimated model has form:

$$\hat{Y}_i = 1816.9176 + 0.9665 \cdot X_i + \varepsilon^*. \qquad (*)$$

2.2 Estimation of prediction error

We call the difference $Y_i - \hat{Y}_i$ the prediction error. However, in our dataset we have a vector $\underline{Y}$ with 6259 observations, so to measure the prediction error we will use the MSE measure (Mean Squared Error):

$$MSE = \frac{1}{6259} \sum_{i=1}^{6259} \left(Y_i - \hat{Y}_i\right)^2.$$

The second measure of the prediction error is RMSE:

$$RMSE = \sqrt{MSE}.$$

We will use both measures to measure the prediction error.

The MSE of the $(*)$ model is equal to 274841.6 and RMSE is equal to 524.2534.

2.3 Cross validation

Let's check what values of the above measures we will obtain using a 10 fold cross validation and compare them with the previous results.

First, let's divide our set of observations into 10 subsets of the same size. In one iteration, let's fit the OLS model to the set that is the sum of 9 subsets (train set) and determine the prediction of observations from the remaining 1 subset. We will repeat the iterations 10 times, but in such a way that each of the subsets is a test set. We will calculate MSE and RMSE based on the predicted values of Y_i obtained during cross validation.

The result of this experiment is:

- MSE=274964.1,

- RMSE=524.3702.

We can notice that the values of the measures are higher compared to the previous task (Q2.2). The reason for this is that when using cross validation, we trained and tested the models on two different datasets. In the previous task, we trained and tested the model on the same dataset. The idea of the cross validation is better because it avoids overfitting the model to the data.

2.4 Leave-one-out cross validdation

Let's take a closer look at how the selection of the training set affects the model. For this purpose, we will use the leave-one-out cross validation. This is cross validation in which the number of folds is equal to the number of observations, in our case it is equal to 6259.

In each iteration, we will build the model on 6258 observations and determine its slope. As a result, we will get 6259 slopes. We can see on the plot below, that each of them oscillates around the slope from task 1 (Q2.1).

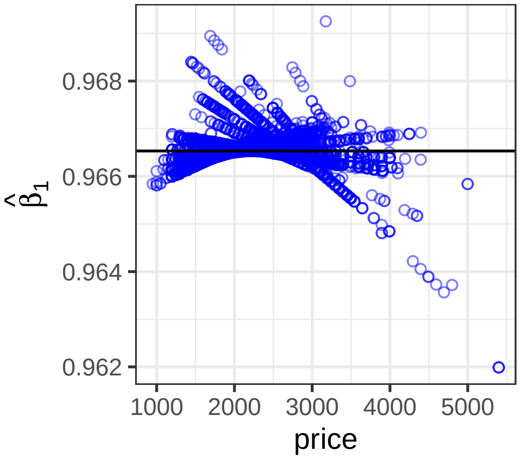


Figure 2: Slope values for the LOO CV

2.5 Bootstrap procedure for the predicted values

Using bootstrap methods we can also estimate the predicted values (regression line) of the model. However, our main goal will be to find the 95% confidence interval of the regression line.

As we know, non parametric bootstrap involves drawing data with replacement. During one iteration, we will draw a sample with replacement from the data set *Computers*. On the obtained sample, we will train the model, and then use it to predict the set of the entire set *Compuers*. We will repeat the iteration 1000 times, as a result of which we will receive 1000 predictions of each value of Y_i - price. For each of them, we will determine 95% C.I. Where the lower 95% bound of the estimated value of Y_i will be the 0.025 quantile of the obtained sample prediction of this value, and the upper bound of the interval will be the 0.975 quantile. The boundaries of the obtained intervals should create separate lines that will define the 95% prediction area of the *price* variable.

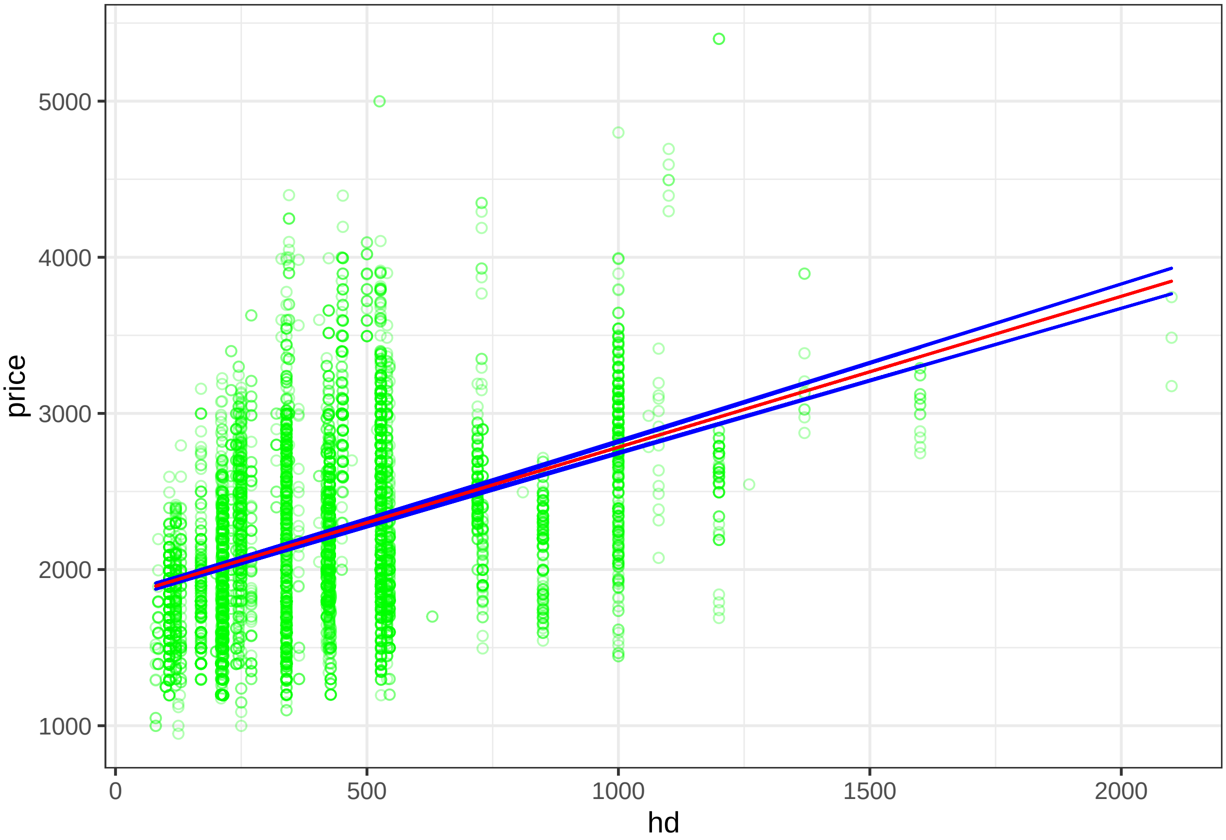


Figure 3: 95% C.I of the model regression line

The above graph shows the result of our bootstrap. The red line indicates the regression line from the first task (Q2.1), and the blue lines outline its 95% C.I.

3 Computers dataset - cont. (Question 3)

In this section, we will mainly analyze the standard deviations of the model coefficients. For this purpose, we will use the non parametric bootstrap method.

3.1 95% C.I for $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$

Let's use nonparametric bootstrap to estimate 95% confidence intervals of the standard deviations of the model coefficients. We will do this by repeating the iteration 1000 times. The iteration will consist of drawing a sample of size 6259, with replacement from the *Computers* dataset. Then we will fit a linear model to it (similarly to the previous section) and determine the standard deviations of the estimated coefficients: $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$. As a result of the experiment, we will obtain two samples of size 1000 of the sought statistics. The estimates of the 95% confidence intervals will be the 95% quantile C.I.

As a result of the above we get:

- $SE(\hat{\beta}_0) \in [12.2358, 12.9252]$;
- $SE(\hat{\beta}_1) \in [0.0248, 0.0266]$.

Their lengths are 0.6894 and 0.0018, respectively.

3.2 Influence of the observations, for which the hard drive size is larger than 2000 MB

Let's check what is the influence of the observations having the disk size larger than 2000 MB, on the $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.

First, perform the same experiment as in problem (Q3.1), but with the difference that we remove from the *Computers* dataset the observations that have a disk larger than 2000 MB. As a result, we will get two new samples of size 1000 of the $SE(\hat{\beta}_0^*)$ and $SE(\hat{\beta}_1^*)$. Additionally, we estimate the coefficients of the regression line.

If the observations we consider in this problem have a significant impact on the model, then the empirical distributions of the obtained statistics should be different.

We got estimated coefficients:

- $\hat{\beta}_0^* = 1815.1597$
- $\hat{\beta}_1^* = 0.9711$

As we can see, the coefficients have not changed significantly compared to the model from task Q2.2. We got quantile 95% C.I:

- $SE(\hat{\beta}_0^*) \in [12.3069, 12.9746]$, length: 0.6677;
- $SE(\hat{\beta}_1^*) \in [0.0251, 0.0267]$, length: 0.0016.

Removing the considered observations caused the confidence intervals to narrow slightly. The graphs below (pink corresponds to the results from Q3.1 and blue to the results from Q3.2) prove that the results from the two tasks are not significantly different, which proves that the considered observations do not have a significant impact on the models.

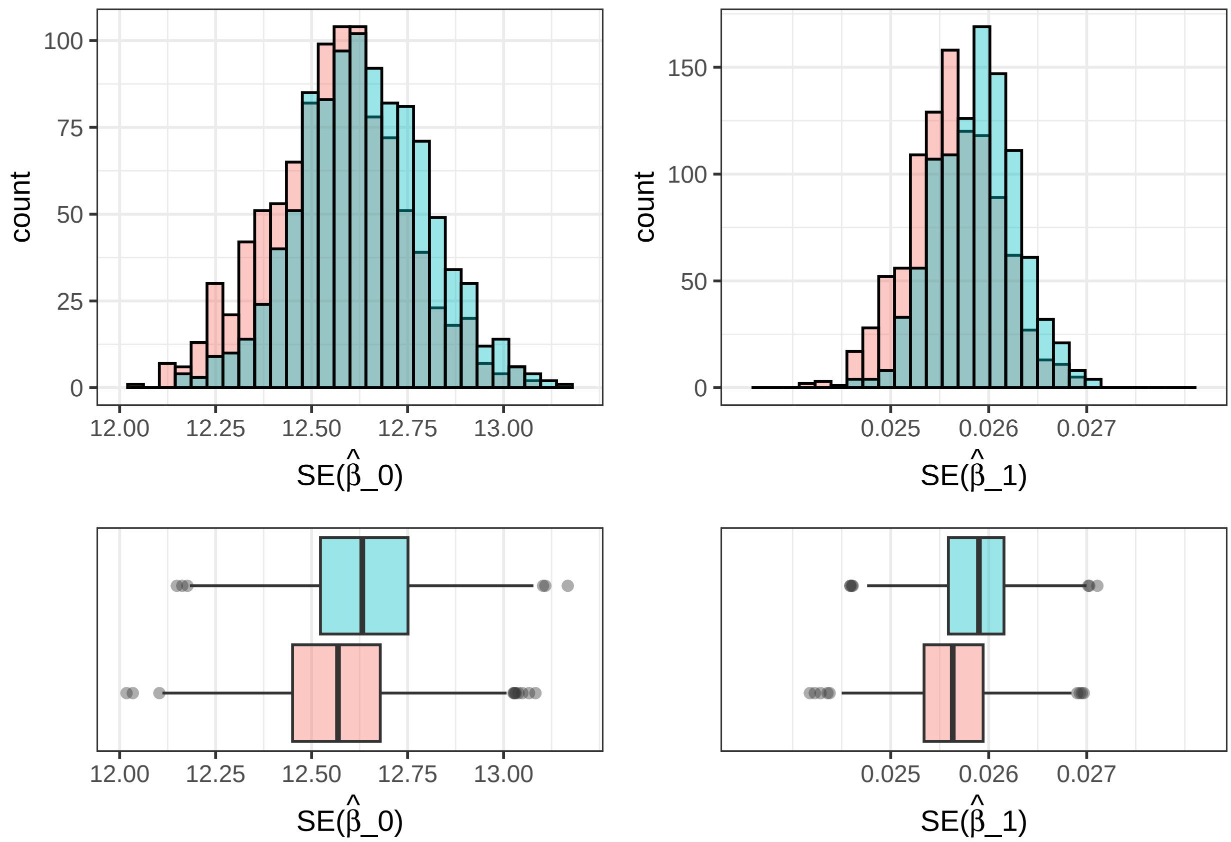


Figure 4: Histogram and boxplot of the $SE(\hat{\beta}_0)$

4

We consider a sample of 20 observation with mean μ.

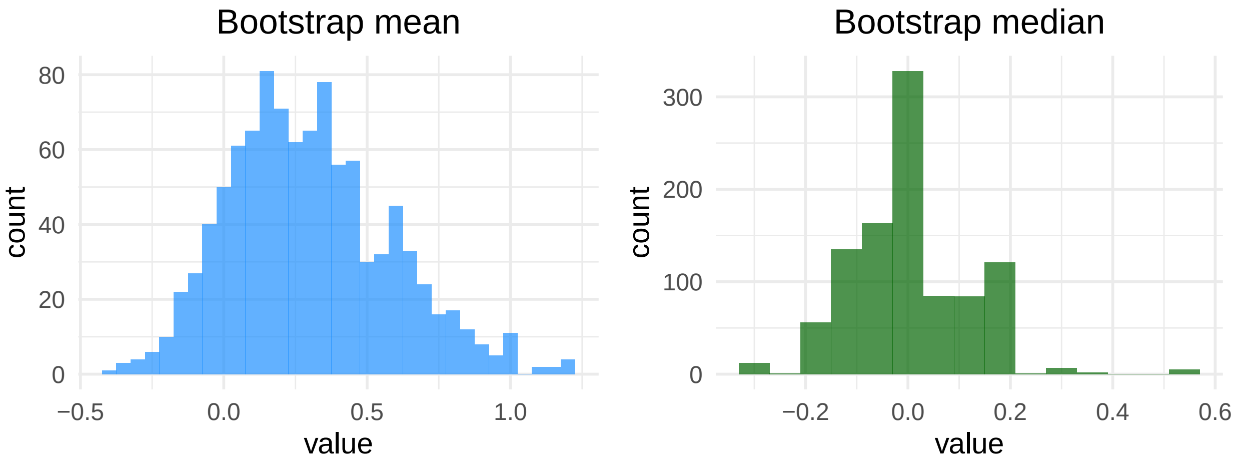
4.1

Below there are values of estimated mean and median.

	mean	median
value	0.291	-0.0064

4.2

To have a closer look at the data we want to approximate sample mean's and median's distribution using non parametric bootstrap.



By looking at above plots we can notice that the mean's distribution recalls normal distribution, that agrees with Central Limit Theorem. However, we cannot tell the same for median's distribution - it is more peaked due to the presence of outliers.

4.3

The next step to explore the mean and median is estimating the standard error of the sample and calculating 95% confidence intervals for the sample mean and median. We use a semi parametric bootstrap to do that.

With regard to use semi parametric bootstrap we need to estimate the error term by residuals and bootstrap them, next just fit the model. In our case we as a model consider the measure - mean or median and focus on their's residuals. The process was following: we sampled the residuals with replacement and added it to the mean of the vector. Next, we calculated mean/median of given vector. After replication of the experiment 10000 times we could calculate SE and 95% confidence intervals.

	mean	median
orig	0.291	-0.0064
SE	0.2719	0.1087
CI lower	-0.1714	-0.1792
CI upper	0.8953	0.1931

The SE is big for both measures, that indicates that confidence intervals are also wide. It is caused by the presence of outliers. The mean's CI is wider that median's due to the chosen method.

4.4

Now, we will estimate mean squared error for the mean and median using jackknife procedure. This method bases on leaving out one observation from the observed sample. Then the i-th jackknife sample is given by

$$\hat{\theta}^{-(i)} = \frac{1}{n-1} \sum_{j, j \neq i} x_j,$$

then the jackknife estimate for the mean is

$$\hat{\theta}^{(\cdot)} = \frac{1}{n} \sum_{i} \hat{\theta}^{(-i)}.$$

After calculating the mean's estimate we need to calculate it's bias and variance

$$MSE = Var + Bias^2.$$

We used following formulas

$$Bias(\hat{\theta}) = (n-1)(\hat{\theta}_{est} - \mu),$$

$$Var(\hat{\theta}) = \frac{(n-1)}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{est})^2.$$

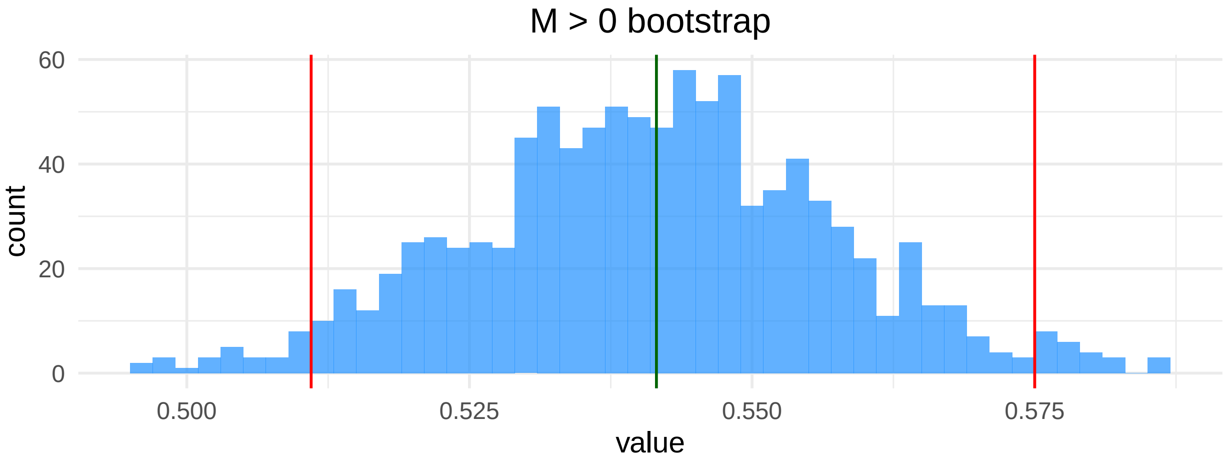
Similar calculations was made for the median.

	mean	median
MSE	31.0076	30.5955

Median has smaller MSE due to the existence of outliers in dataset. Then median is preferred estimator.

4.5

Let M be the median and $\pi_{M<0} = P(M < 0)$. We would like to estimate the distribution of $\hat{\pi}_{M<0}$ and constrict 95% confidence interval for $\pi_{M<0}$. To do that we will perform non parametric bootstrapping (sample with replacement) with 10000 replications, calculate probability $\pi_{M<0}$ and again replicate the procedure 10000 times to determine CI using quantiles.



The estimated $\pi_{M>0} = 0.5415$ with confidence interval [0.511, 0.575]. The probability of $M < 0$ is greater of 1/2, but close to it. That indicates that most of the observations in x is smaller than 0. The confidence interval includes 1/2 then we cannot say that the difference between positive and negative observations is significant.