
# 2nd-Year Master of Statistics and Data Science Computer Intensive Methods: Final projects (2024/2025)

**Project 3**

Mikita Bisliuk (2364811), Edmond Sacla Aide (2159278)

26 January, 2025

# Project 3

## Part 1

The data we use in this question is the Chicks dataset. The data contains information over an experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.

The question of primary interest is if there is a difference between the chicks weights across the diet groups. Let $Y_{ij}$ be the weight of a chick $i$ in diet group $j$.

### Question 1.1

#### Formulating a One-Way ANOVA Model

**Model:** Let $Y_{ij}$ represent the weight of chick $i$ in diet group $j$. The one-way ANOVA model can be written as:

$$Y_{ij} = \mu_j + \epsilon_{ij},$$

where:

- $\mu_j$ is the mean weight for diet group $j$,
- $\epsilon_{ij}$ is the random error term with $\epsilon_{ij} \sim N(0, \sigma^2)$.

**Hypotheses:**

- **Null Hypothesis ($H_0$)**: All diet groups have the same mean weight.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

- **Alternative Hypothesis ($H_A$)**: At least one diet group has a different mean weight.

$$H_A : \text{Not all } \mu_j \text{ are equal.}$$

**Test Statistic:** The test statistic for one-way ANOVA is given by:

$$F = \frac{\text{Between-group variability (MSB)}}{\text{Within-group variability (MSW)}},$$

where:

- $MSB = \frac{SSB}{k-1}$, the mean sum of squares between groups,
- $MSW = \frac{SSW}{N-k}$, the mean sum of squares within groups,
- $k$ is the number of groups, and $N$ is the total number of observations.

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## feed       5 231129   46226  15.365 5.936e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Observed F is: 15.36"
```

```
## [1] "Critical F is: 2.36"
```

- $F_{\text{Observed}} > F_{\text{critical}}$, reject $H_0$.
- (Also p_{value}= 5.936e-10 $< 0.05$), reject $H_0$.

Conclusion: there is a significant difference in mean weights between diet groups. Otherwise, no significant difference is detected.

**Question 1.2 : Testing the Null Hypothesis of No Diet Effect Using Semi-Parametric Bootstrap**

To test the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ (no diet effect) using a semi-parametric bootstrap, we follow these steps:

1. **Fit the Null Model**:
   - Under the null hypothesis, the response variable $Y$ is modeled without considering group differences. This gives the residuals.
2. **Resample Residuals**:
   - Randomly sample (with replacement) from the residuals of the null model to generate bootstrap datasets.
3. **Generate Bootstrap Datasets**:
   - Add the resampled residuals back to the fitted values under $H_0$ to create new bootstrap datasets.
4. **Compute Test Statistic for Each Bootstrap Sample**:
   - For each bootstrap dataset, compute the $F$-statistic from the ANOVA test.
5. **Bootstrap Distribution**:
   - Collect the $F$-statistics from all bootstrap samples to approximate their null distribution.
6. **p-value Calculation**:
   - Calculate the proportion of bootstrap $F$-statistics that are greater than or equal to the observed $F$-statistic from the original data.

**Hypotheses:**

- **Null Hypothesis ($H_0$)**: There is no diet effect (all groups have the same mean weight).
- **Alternative Hypothesis ($H_A$)**: There is a diet effect (at least one group's mean weight is different).

```
## [1] "Bootstrap p-value: 0.001"
```

The result indicates evidence against the null hypothesis, suggesting a diet effect on chick weights. Otherwise, fail to reject $H_0$.

**Question 1.3 : Testing the Null Hypothesis of No Diet Effect Using Permutation Test**

To test the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ (no diet effect) using a permutation test, we follow these steps:

1. **Compute Observed $F$-Statistic**:
   - Perform a one-way ANOVA on the original dataset to compute the observed $F$-statistic.
2. **Permute the Data**:
   - Shuffle the diet group labels randomly across all observations while keeping the response variable $Y$ unchanged. This breaks any relationship between the response and the groups under the null hypothesis.
3. **Generate Permuted Datasets**:
   - For each permutation, compute the $F$-statistic using the permuted dataset.
4. **Build Permutation Distribution**:
   - Collect the $F$-statistics from all permuted datasets to approximate their null distribution under $H_0$.
5. **Calculate p-value**:
   - The p-value is the proportion of permuted $F$-statistics that are greater than or equal to the observed $F$-statistic.

**Hypotheses:**

- **Null Hypothesis ($H_0$)**: There is no diet effect (all groups have the same mean weight).
- **Alternative Hypothesis ($H_A$)**: There is a diet effect (at least one group's mean weight is different).

```
## [1] "Bootstrap p-value: 0.001"
```

- p-value $< \alpha = 0.05$, reject $H_0$, suggesting evidence of a diet effect. The permutation test provides a non-parametric approach to testing group differences without relying on the assumptions of normality or equal variances.

**Question 1.4: Estimation of $\theta = \mu_{\text{Sunflower}} - \mu_{\text{Soybean}}$ and Construction of a 90% C.I. Using Parametric Bootstrap**

To estimate $\theta$ and construct a 90% confidence interval using a parametric bootstrap, we follow these steps:

1. **Define $\theta$:**
   - $\theta = \mu_{\text{Sunflower}} - \mu_{\text{Soybean}}$, the mean difference between the Sunflower and Soybean diet groups.
2. **Fit the Model:**
   - Fit a one-way ANOVA model to the data to estimate the group means ($\mu_{\text{Sunflower}}$ and $\mu_{\text{Soybean}}$).
3. **Compute Observed $\theta$:**
   - Calculate the observed value of $\theta$ as the difference between the means of the Sunflower and Soybean groups.
4. **Parametric Bootstrap:**

- Assume the residuals follow a normal distribution.
- For each bootstrap iteration:
  - Resample the residuals and add them to the fitted values for the Sunflower and Soybean groups to generate new datasets.
  - Compute $\theta$ for each bootstrap dataset.

5. **Theata Construct the Confidence Interval:**
   - calculate the bootstraap mean
   - Extract the 5th and 95th percentiles of the bootstrap distribution to form the 90% confidence interval.

```
## [1] "mean difference between the Sunflower and Soybean diet groups is: 82.927"
```

```
## [1] "CI of mean difference between the Sunflower and Soybean diet groups is: 51.358"
## [2] "CI of mean difference between the Sunflower and Soybean diet groups is: 114.064"
```

```
## $theta_observed
## [1] 82.92739
##
## $confidence_interval
##        5%       95%
##  51.35788 114.06360
```

## Part 2

### Question 2.1

In this question we focused on the Computers dataset that can be accessed via the R package Ecdat. Make sure you install the package Ecdat in order to access the data. This data shows the prices of Personal Computers from 1993 until 1995. It contains with 6259 observations on 10 variables. Visit https://rdrr.io/cran/Ecdat/man/Computers.html to read more about the data set. Use the code below to acsess the data.

let us focus on the variables "price in US dollars of 486 PCs" (the variable price in the dataset) and size of hard drive in MB (the variable hd in the dataset). Let $Y_i$ be the price and $X_i$ be the size of hard drive in MB.

We consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

We aim to estimate the following regression model using Ordinary Least Squares (OLS):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where:

- $y_i$: Response variable (e.g., price).
- $x_i$: Predictor variable (e.g., hard drive size).
- $\beta_0$: Intercept.
- $\beta_1$: Slope.
- $\epsilon_i$: Error term.

```
##
## Call:
## lm(formula = price ~ hd, data = Computers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1338.45  -382.23   -44.47   315.34  2674.65
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.817e+03  1.257e+01   144.6   <2e-16 ***
## hd          9.665e-01  2.564e-02    37.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 524.3 on 6257 degrees of freedom
## Multiple R-squared:  0.1851, Adjusted R-squared:  0.185
## F-statistic:  1421 on 1 and 6257 DF,  p-value: < 2.2e-16
```

**Question 2.2: Prediction of Price and Estimation of Prediction Error**

1. Fit the regression model using the classical OLS approach.
2. Use the `predict()` function to obtain predicted values.
3. Calculate the prediction error as the difference between actual and predicted values.
4. Compute metrics such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) to quantify the prediction error.

```
## [1] "The squared mean squared error is: 524.253"
```

**Question 2.3: 10-Fold Cross-Validation for Price Prediction and Error Estimation**

1. Divide the dataset into 10 approximately equal folds.
2. For each fold:
   - Use 9 folds as the training set to fit the model.
   - Use the remaining fold as the test set to predict prices and calculate prediction errors.
3. Calculate the Mean Squared Error (RMSE) for each fold.
4. Average the MSEs across all folds to obtain the cross-validated error.

```
## 10-Fold Cross-Validated RMSE: 524.4827
```

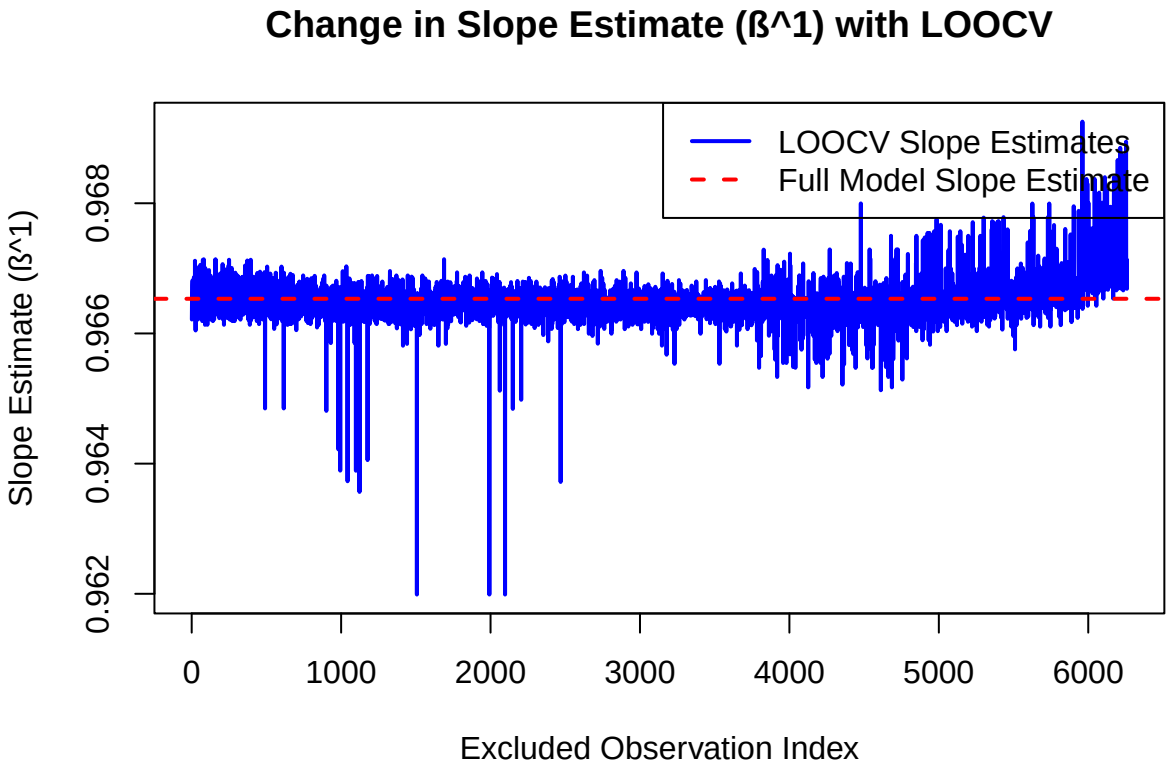
```
## Classical OLS RMSE: 524.2534
```

**Question 2.4: Leave-One-Out Cross-Validation (LOOCV) to Investigate Changes in $\hat{\beta}_1$**

Use LOOCV to analyze the variation in the slope estimate ($\hat{\beta}_1$) when each observation is excluded. Visualize the changes in $\hat{\beta}_1$ to assess its sensitivity to individual observations.

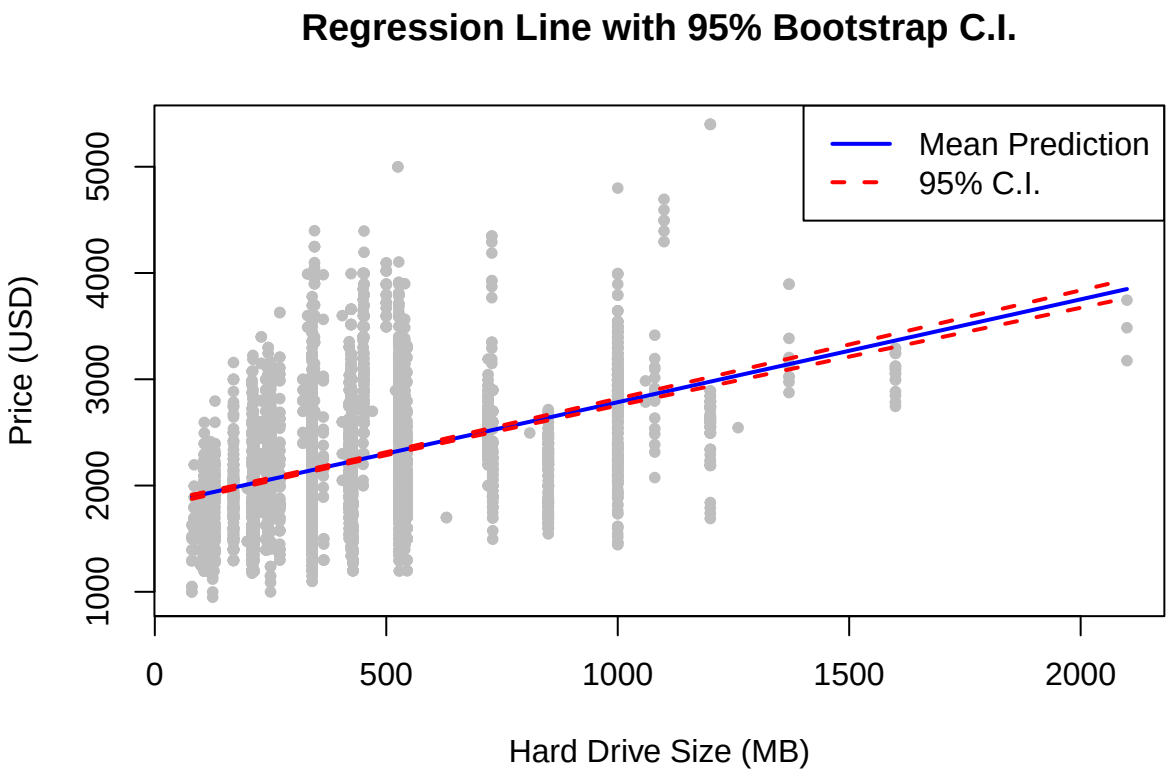
1. Exclude each observation one at a time.
2. Fit the regression model to the remaining $n-1$ observations.
3. Record the slope estimate ($\hat{\beta}_1$) for each iteration.

4. Visualize the changes in $\hat{\beta}_1$ across all iterations.




**Question 2.5:Non-Parametric Bootstrap for Constructing 95% CI for Predicted Values**

1. Fit the regression model to the original data.
2. Resample the data with replacement $B$ times to create bootstrap datasets.
3. For each bootstrap sample:
   - Fit the regression model.
   - Predict the response variable using the bootstrap regression line at fixed predictor values.
4. Calculate the 95% CI at each predictor value using the percentiles of the bootstrap predictions.




```
## C.I. for the predicted values for the slope: 1933.825 3811.436
```

## Part 3

In this question we use the same model formulated in Q2.

**Question 3.1: Non-Parametric Bootstrap for Constructing 95% CI for $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$**

1. Fit the original regression model to the data.
2. Resample the data with replacement $B$ times to create bootstrap datasets.
3. For each bootstrap sample:
   - Fit the regression model and estimate $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$.
4. Calculate the standard errors of $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ across all bootstrap samples.
5. Construct 95% confidence intervals for $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ using the percentiles of their bootstrap distributions.

```
## 95% C.I. for SE(^0): 12.2359 12.90604
```

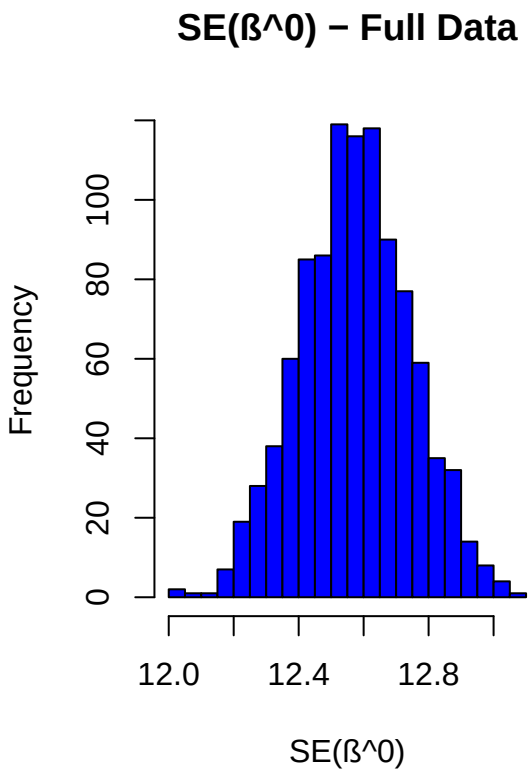
```
## 95% C.I. for SE(^1): 0.02482244 0.02654871
```

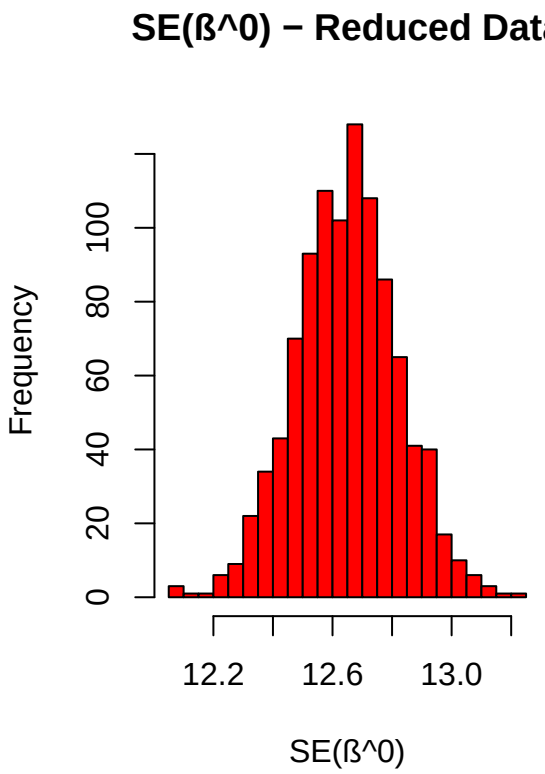
**Question 3.2: Investigating the Influence of Observations with Hard Drive Size > 2000 MB Using Bootstrap**

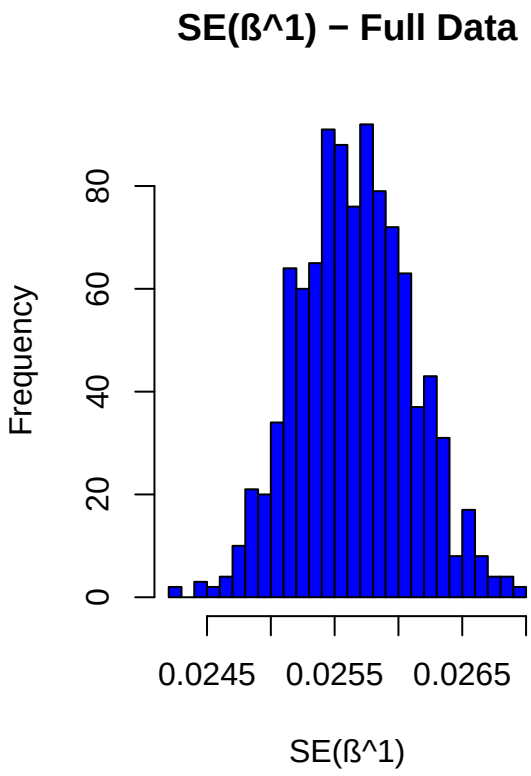
To assess the influence of these observations, the bootstrap procedure can be used to compare standard errors with and without these influential points in the dataset.

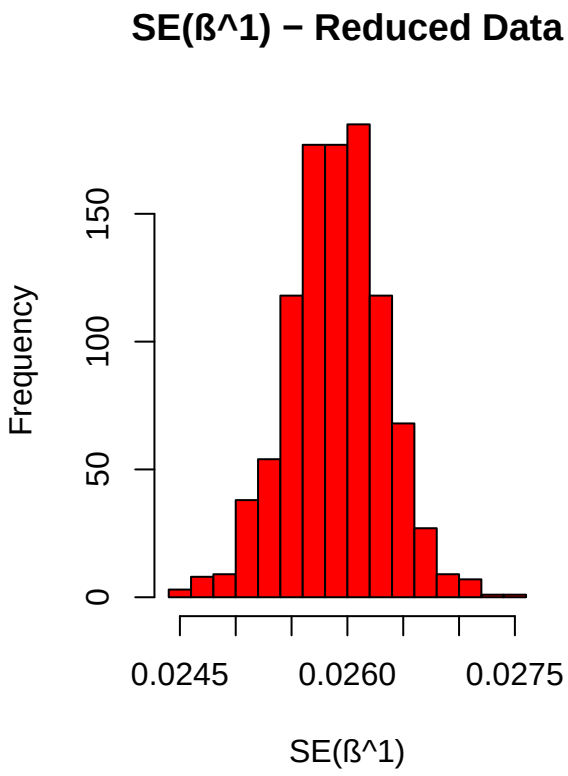
1. **Identify Observations**:
   - Subset the data into two groups: one with hard drive size $> 2000\,\text{MB}$, and another without these observations.
2. **Bootstrap Procedure**:
   - Perform bootstrap resampling separately for the two datasets:
     - Full dataset (all observations included).
     - Reduced dataset (excluding observations with hard drive size $> 2000\,\text{MB}$).
   - For each bootstrap sample:
     - Fit a linear regression model.
     - Estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.
   - Calculate the standard errors $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ for each dataset.
3. **Compare Results**:
   - Compare the empirical distributions of $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ between the full and reduced datasets.
4. **Visualization**:
   - Plot histograms or density plots of $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ for both datasets to visualize the influence of the observations.

```
## 95% C.I. for SE(^0) - Full Data: 12.2359 12.90604

## 95% C.I. for SE(^1) - Full Data: 0.02482244 0.02654871

## 95% C.I. for SE(^0) - Reduced Data: 12.31111 12.97453

## 95% C.I. for SE(^1) - Reduced Data: 0.02502753 0.02672845
```


SE(ß^0) − Full Data
Frequency
0
20
40
60
80
100
12.0
12.4
12.8
SE(ß^0)


SE(ß^0) − Reduced Data
Frequency
0
20
40
60
80
100
12.2
12.6
13.0
SE(ß^0)




## Part 4

Consider a sample of 20 observations from a population with mean $\mu$:

```
## [1] 20
```

**Question 4.1**

Estimate  using the mean and the median.

```
## mean of mu 0.2909559
```

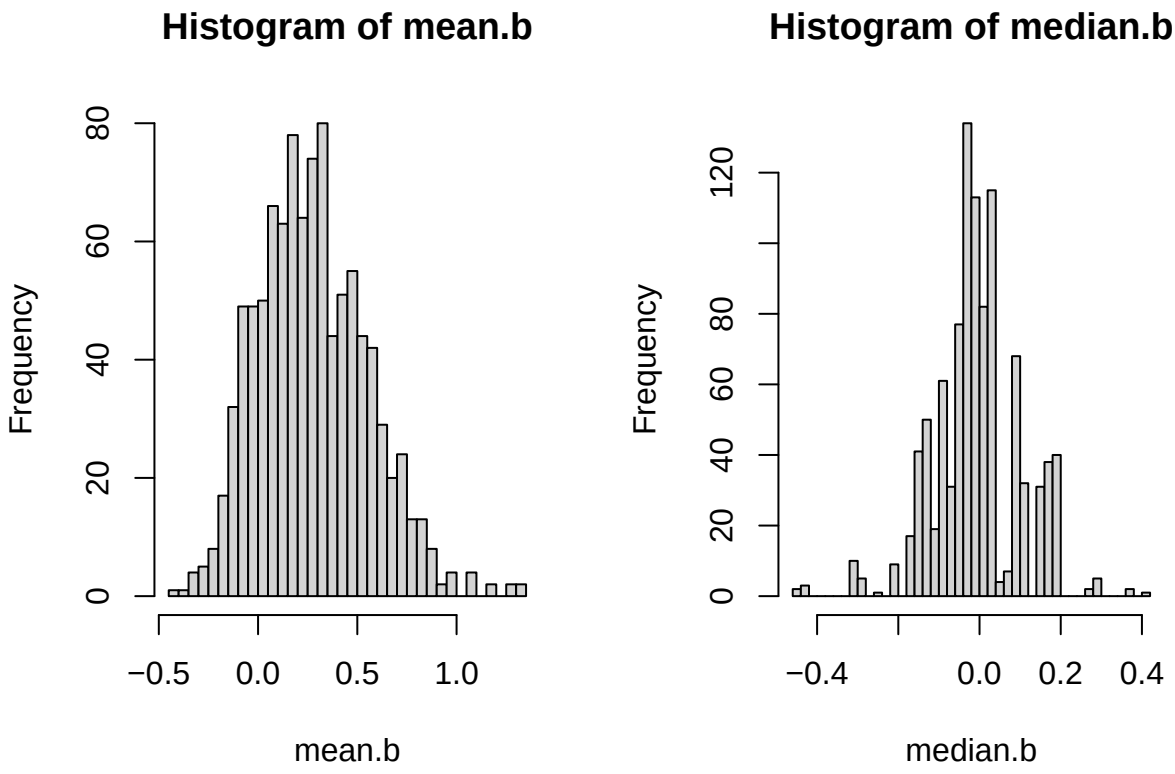
```
## mean of mu -0.006405105
```

**Question 4.2: Approximation of the Distribution of Sample Mean and Median Using Non-Parametric Bootstrap**

1. **Bootstrap Resampling**:
   - Generate $B = 1000$ bootstrap samples from the original data by sampling with replacement.
   - For each bootstrap sample, calculate the sample mean and the sample median.
2. **Approximation of Distributions**:

- Collect the $B$ bootstrap estimates for both the sample mean and the sample median.
- Plot histograms or density plots to visualize the approximated distributions.

3. **Summary Statistics**:
   - Compute descriptive statistics (mean, standard deviation) for the bootstrap distributions.




**Question 4.3: Estimation of Standard Error and 95% Confidence Intervals for Sample Mean and Median Using Semi-Parametric Bootstrap**

1. **Semi-Parametric Bootstrap**:
   - Decompose the data into a parametric component (mean and standard deviation) and a residual component.
   - Generate bootstrap samples by resampling residuals and adding them back to the parametric predictions.
   - For each bootstrap sample, calculate the sample mean and median.
2. **Standard Error and Confidence Interval**:
   - Compute the standard errors as the standard deviation of the bootstrap estimates.
   - Construct 95% CIs using the percentiles of the bootstrap distributions.

```
## The standard error of the sample mean is : 0.008622562
```

```
## The standard error of the sample median is : 0.003389343
```

```
## The C.I for the sample mean is : -0.169777 0.8449082
```

```
## The C.I for the sample mean is : -0.1798164 0.1930504
```

**Question 4.4: Estimation of Mean Squared Error (MSE) for the Mean and Median Using Jackknife**

1. **Jackknife Method**:
   - For a sample of size $n$, iteratively leave out one observation at a time to create $n$ subsamples, each of size $n-1$.
   - Calculate the statistic (mean or median) for each subsample.
2. **Estimate MSE**:
   - Compute the jackknife estimate of the parameter $\hat{\theta}_{(i)}$ for each subsample.
   - Calculate the jackknife estimate of the variance:

$$\text{Var}_{\text{jackknife}} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \bar{\theta})^2$$

     where $\bar{\theta}$ is the average of the jackknife estimates.
   - Use the variance to compute the MSE:

$$\text{MSE} = \text{Var}_{\text{jackknife}}$$

```
## MSE for Mean: 0.3508179
```

```
## MSE for Median: 0.007265759
```

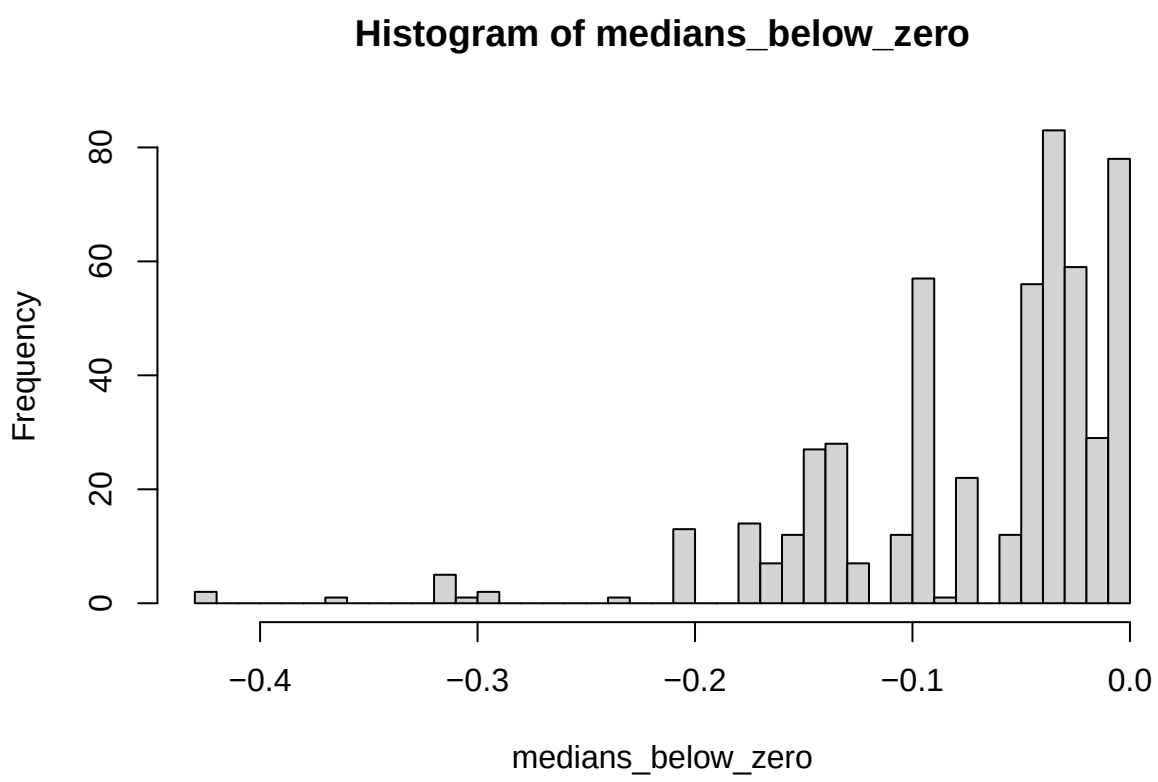
```
## The Median is preferred based on lower MSE.
```

**Question 4.5: Estimation of $\pi(M < 0)$, Its Distribution, and 95% Confidence Interval**

1. **Bootstrap Sampling**:
   - Resample the data $B$ times (with replacement).
   - Compute the median $M$ for each bootstrap sample.
2. **Estimate $\pi(M < 0)$**:
   - Count the number of bootstrap medians that are less than 0.
   - Calculate $\hat{\pi}(M < 0)$ as the proportion:

$$\hat{\pi}(M < 0) = \frac{\text{Count of } M < 0}{B}$$

3. **Distribution of $\hat{\pi}(M < 0)$**:
   - The bootstrap procedure generates a distribution of $M < 0$.
4. **Confidence Interval**:
   - Use the bootstrap distribution to construct a 95% confidence interval for $\pi(M < 0)$ based on the 2.5th and 97.5th percentiles.




```
## Estimated  (M < 0): 0.529

## 95% Confidence Interval for  (M < 0): -0.2037189 -0.006405105
```