Assignment 2

NAME: Nicholas Tolley
DUE DATE: February 28th, 6pm

Problem 1 (100 pts)

In the earnings dataset you can find salary (*earn*) and some socio-demographic characteristics of each subject, including variables such as *height*, *weight*, gender (*male*), *ethnicity*, *education*, mother's (*mother_education*) and father's education (*father_education*), *walk* (e.g. walking time), *exercise*, if they smoke or not (*smokenow*), *tense*, *angry* and *age*.

The dataset can be found in Canvas in the Data folder (file name: earnings.csv):

(a) (10 points) Subset the data and consider only the variables: *education*, *mother_education*, *father_education*, *walk*, *exercise*, *tense*, *angry*, *weight*, *height*. Check the correlation by performing a figure similar to Figure 1 below (make sure not to use the default colours but rather choose your own). Take special care to the labels and legend. What can you say about the results? What would you expect from a linear regression model (hint: there are some variables to be excluded/included in the model)? Perform a test statistic for the correlation between earn and education, write the hypothesis test and the results you will obtain.

```
library(ggplot2)
library(boot)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

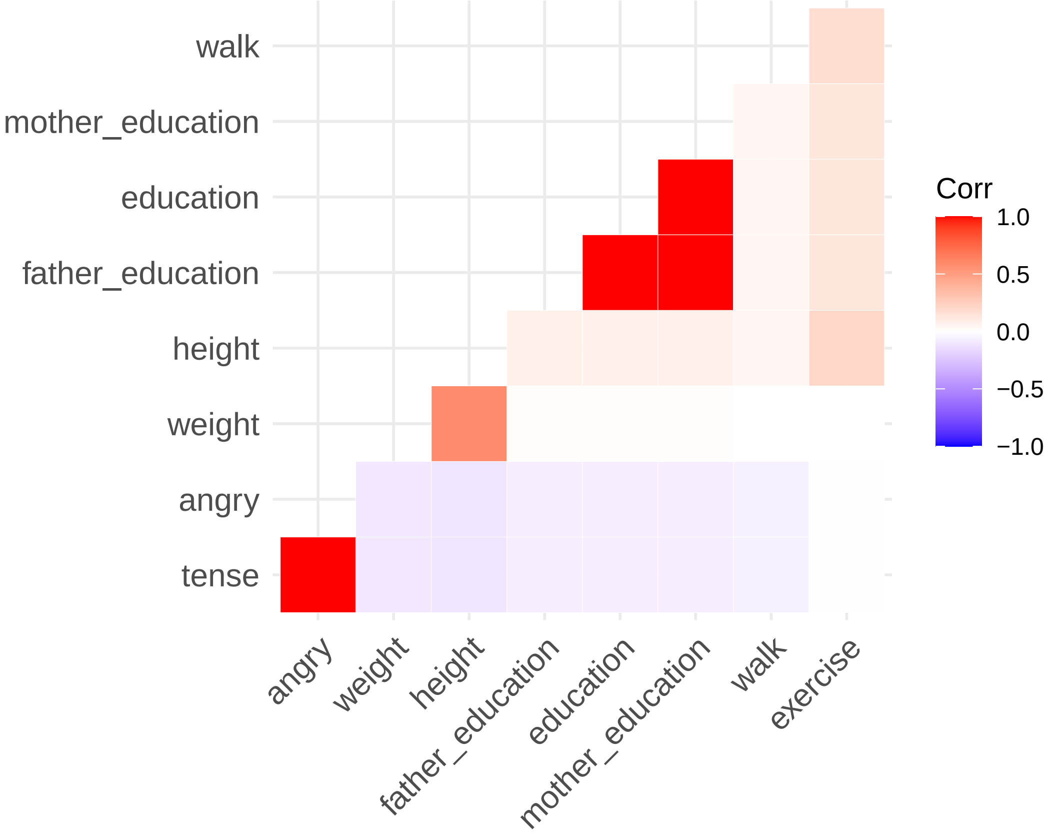
```
library(ggcorrplot)
library(tidyr)

df <- read.csv('earnings.csv')
```

The code below stores a subset of the dataset with the columns indicated above. Since we are calculating the correlation between columns, rows with missing values in any column are removed.

```
subset_cols <- c('education', 'mother_education', 'father_education',
                 'walk', 'exercise', 'tense', 'angry', 'weight', 'height')
df_subset <- subset(df, select=subset_cols)
df_subset <- drop_na(df_subset)
```

Next we can calculate the correlation between the columns contained in the subset, and visualize the result as a heatmap.

```
corr_matrix <- cor(df_subset)
ggcorrplot(corr_matrix, hc.order=TRUE, type="lower", outline.col = "white")
```



As we can see there are several variables that exhibit a near perfect correlation with one another. The most highly correlated columns include:

education <-> mother_education

education <-> father_education

father_education <-> mother_education

tense <-> angry

If we were to build a linear regression model, we would need to remove 2 of the education variables, and either the tense or angry variable. This is because the dataset exhibits what is known as multicollinearity, in other words there is redundant information in the columns. If we were to try and create a linear model on the full dataset, there would not be a unique combination of regression ("beta") coefficients that minimize the residual error. For example, the same coefficient could be assigned tot he angry or tense columns.

The code below calculates the correlation coefficient, and associated p-value, between the earn and education columns. The p-value refers to the probability of the null hypothesis that these two variables are uncorrelated (correlation=0).

```
cor.test(df$earn, df$education)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$earn and df$education
## t = 13.748, df = 1812, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2650564 0.3484265
## sample estimates:
##       cor
## 0.3073311
```

We can see that despite the correlation being relatively low (cor=0.3), the result is highly significant with a p-value < 2.2e-16. We can therefore reject the null hypothesis that the "earn" and "education" variables are uncorrelated with one another.

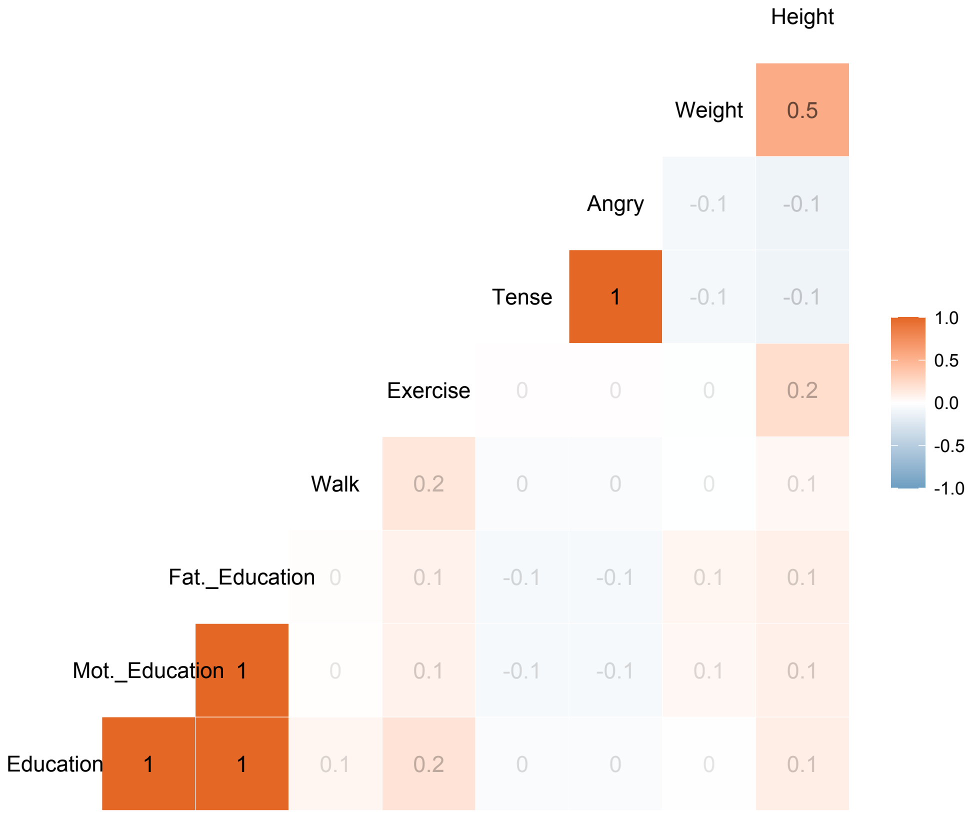


Figure 1: Correlation

(b) (10 points) Perform a linear regression model using the variable *earn* as the dependent variable and years of eductaion *education* as the independent variable. What can you say about this covariate? Is it significant? Write down the hypothesis test. Plot the linear regression you have obtained in ggplot by using a subset of the data. This subset is obtained by restricting the variable *earn* to be less than 2e+05 (similar to Figure 2 below)

The code below creates a linear model that predicts "earn" by the education covariate. From the linear model there are three potential hypothesis tests with the following null hypotheses:

- The intercept of the linear model is zero
- The beta coefficient of the linear model is zero
- The model has the same explanatory power (residual error) as constant (flat line) model

```
summary(lm(df$earn ~ df$education))
```

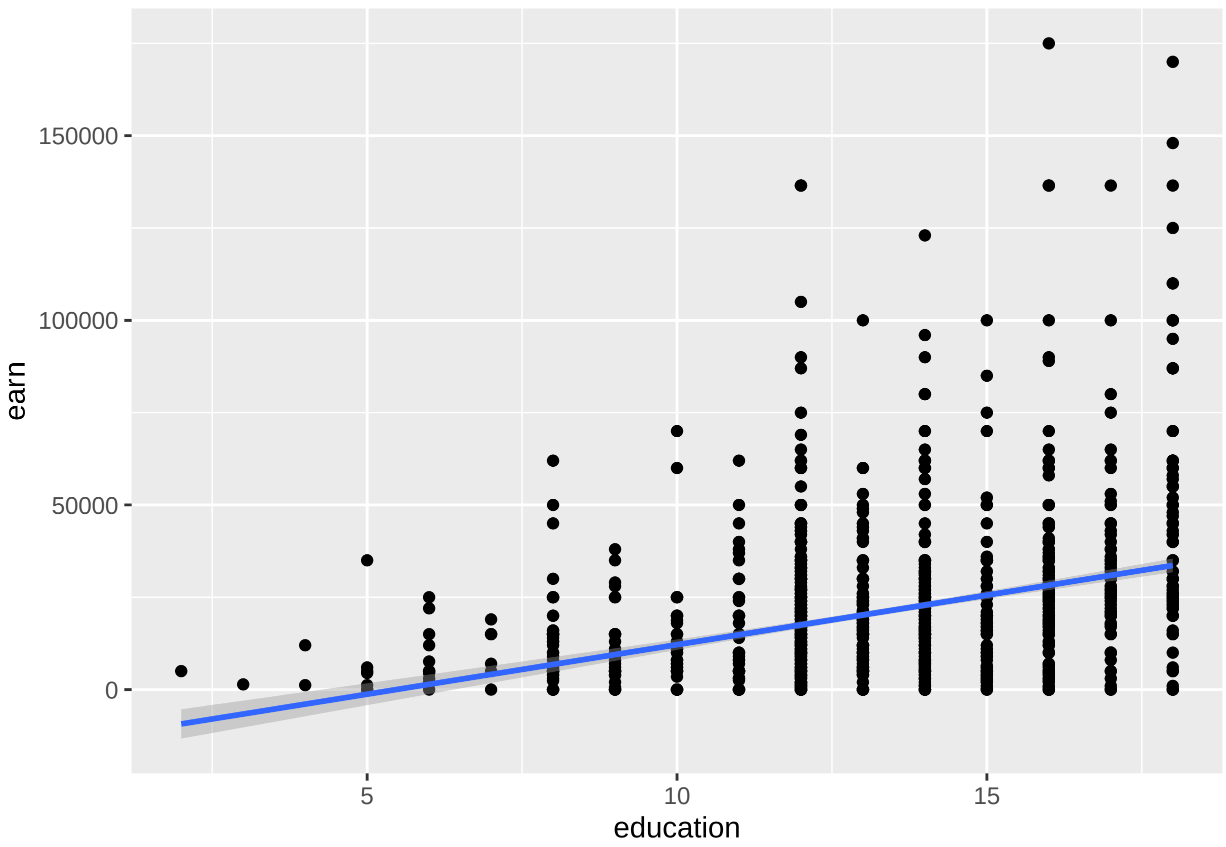
```
##
## Call:
## lm(formula = df$earn ~ df$education)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34051 -12373  -3212   7207 382207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -14724.1     2657.0  -5.542 3.43e-08 ***
## df$education   2709.7      197.1  13.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21460 on 1812 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.09445,    Adjusted R-squared:  0.09395
## F-statistic:   189 on 1 and 1812 DF,  p-value: < 2.2e-16
```

As we can see from the from the summary of the model output above, both the intercept and the slope are highly significant indicating that the *earn* variable is predicted well by the *education* variable.

Next using the subset of the with *earn* $<$ 2e+05, we can visualize how well our linear model explains *earn* using just the *education* covariate

```
earn_subset <- subset(df, df$earn < 2e5)

ggplot(earn_subset, aes(x=education, y=earn)) + geom_point(na.rm=TRUE) +   geom_smooth(method='lm', na.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(c) (20 points) Draw the qqplot by using the library ggplot for the model obtained in point b. Then perform the qqplot (using the library ggplot) for the two different groups of sex (similar to Figure 3 below). Take special care to the legend and the label. What can you say about this plot?

(d) (20 points) Perform in R the backward and forward procedure to select the covariates, remember to remove the rows with missing values. Did you obtain the same or different results from the two different procedures, please explain. Which procedure would you prefer? Comment what you discovered and the theoretical implications. Just for the backward solution compute the RSS and show the trend of RSS for beta1 in a plot by using ggplot in R (similar to Figure 4). (Hint: For RSS plot, set the range of x-axis to be [0,1000]).

(e) (20 points) Perform a bootstrap of 500 samples for beta 1 (*height*), beta 2 (*male*), and beta 3 (*education*) for the coefficient obtained in the backward procedure in point d. Plot the beta coefficients that you have obtained with histograms with ggplot (similar to Figure 5). Remember to use the data without missing values.

(f) (20 points) Compute the LOO and K-fold cross validation and write the results. Compute the mean square error for both the LOO and the K-fold cross validation. Then plot the prediction against the true value for LOO, using ggplot. Describe the results. Remember to use the data without missing values.

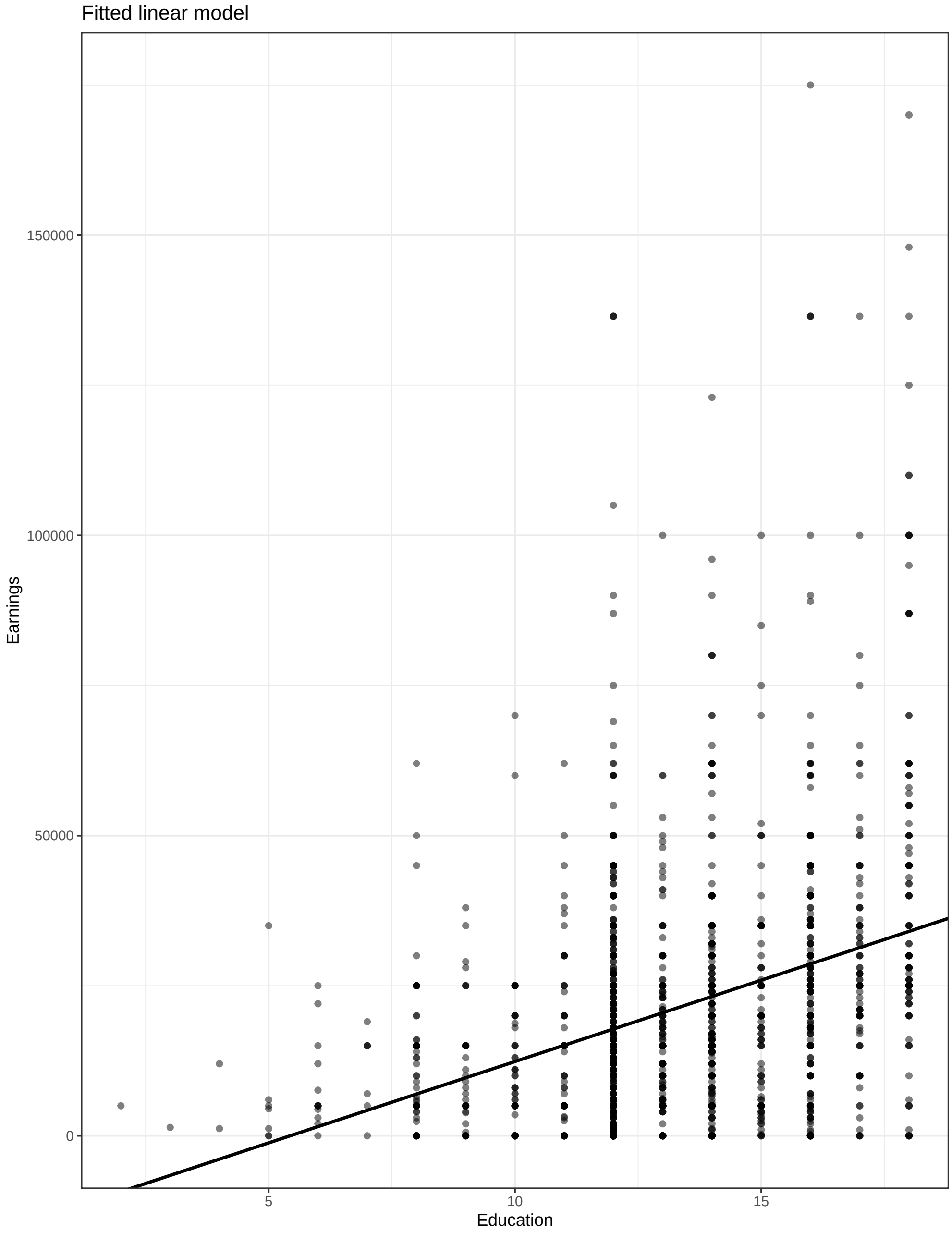


Figure 2: Linear Regression

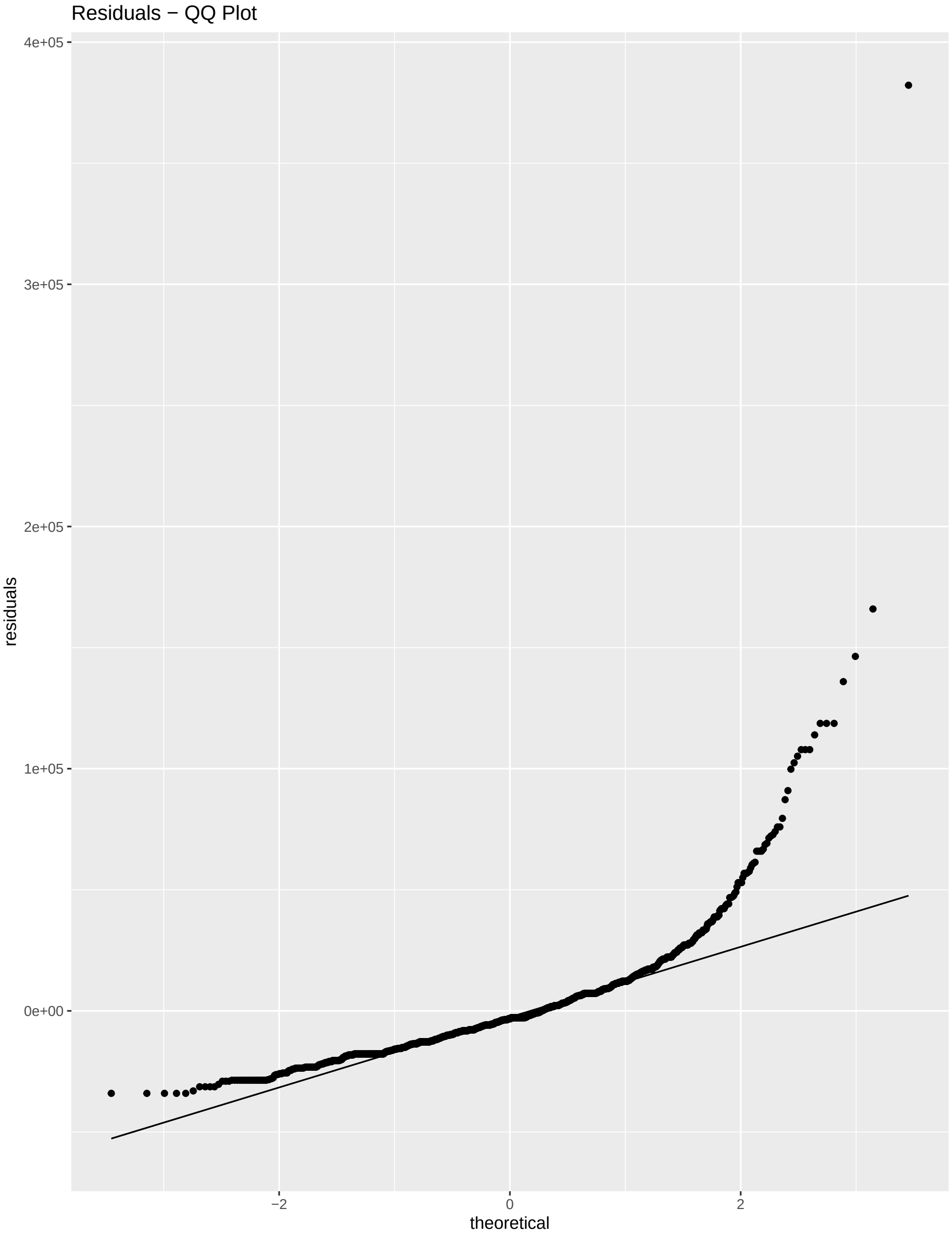


Figure 3: QQplot for different groups

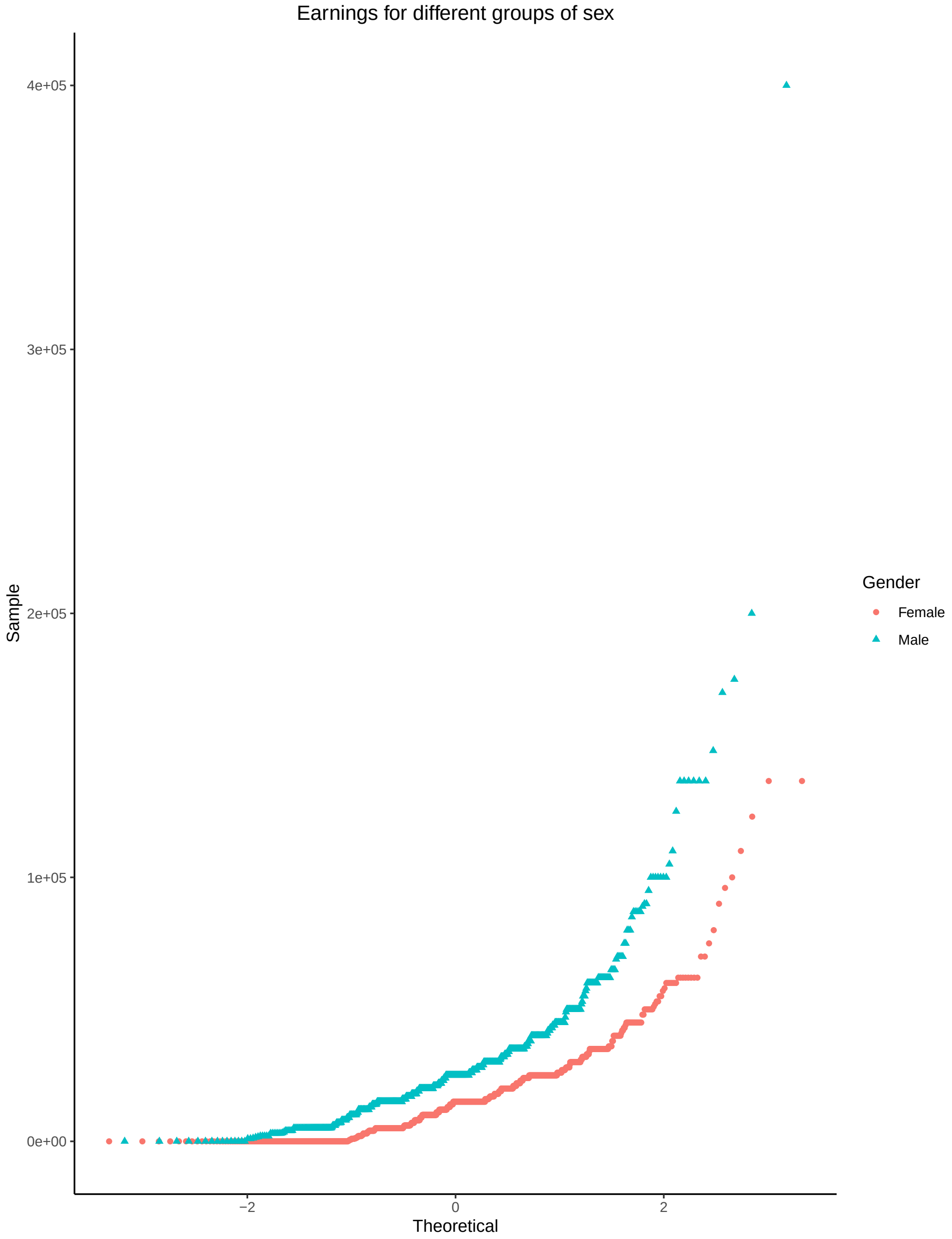


Figure 4: RSS for the backward procedure

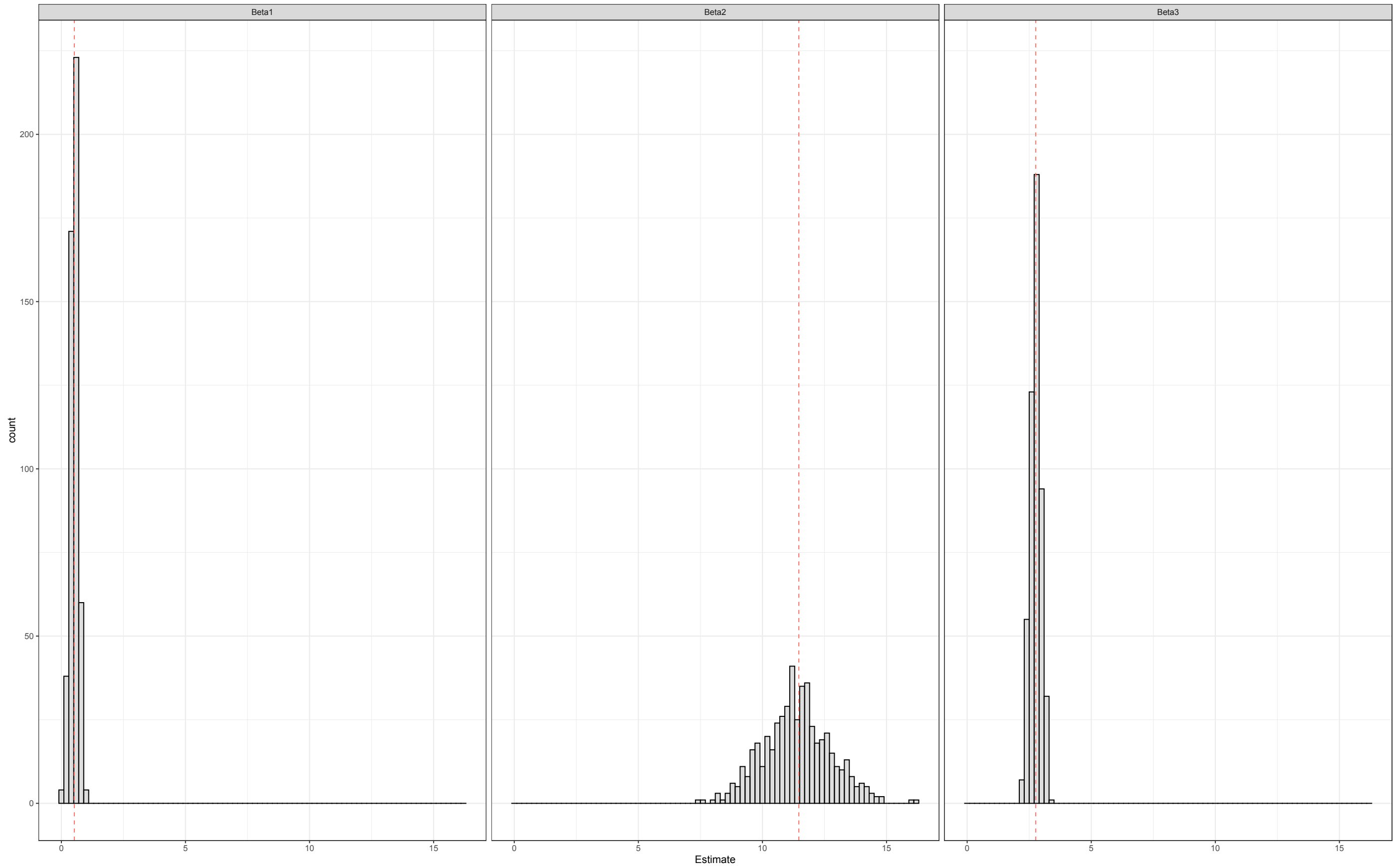


Figure 5: Bootstrap Results