Assignment 2

NAME: Enmin Zhou
DUE DATE: February 23rd, 11:59pm

Problem 1 (100 pts)

```
rm(list=ls())
library(foreign)
library(ggplot2)
library(boot)
library(MASS)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
library(dplyr)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

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) and a figure with ggpairs. Take special care to the labels and legend. What can you say about the results? What would you expect from a linear regression model?

a: I use green for positively high correlation and yellow for negatively high correlation and white for low correlation. "Tense" and "Angry", "mother education" and "education", "fother education" and "education", "father education" and "mother educatin" have correlation 1. "weight" have very low correlation with all of them except for "Heighet".

```
earnings = read.csv('earnings.csv')
attach(earnings)
d_1 <- earnings %>% select(education, mother_education, father_education, walk, exercise, tense, angry,
colnames(d_1) <-c("Education","Mot._education","Fat._education","Walk","Exercise","Tense","Angry","Weig
p1 <- ggcorr(d_1, label = TRUE, low = "yellow", mid = "white", high = "green")
p1
```

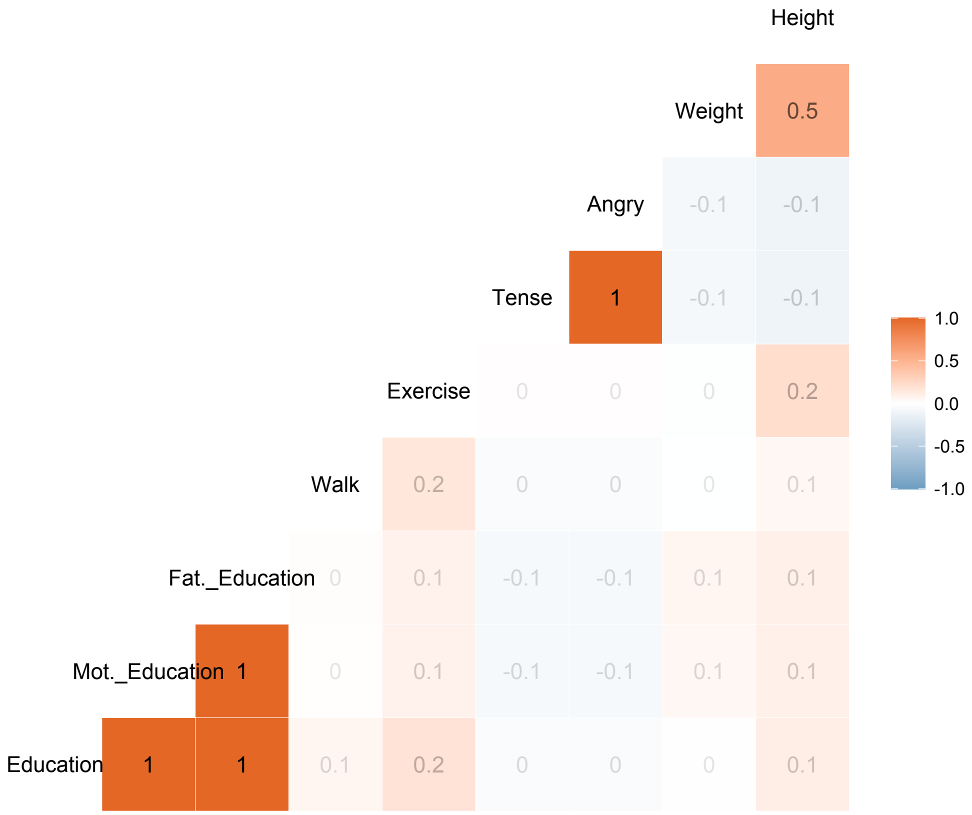
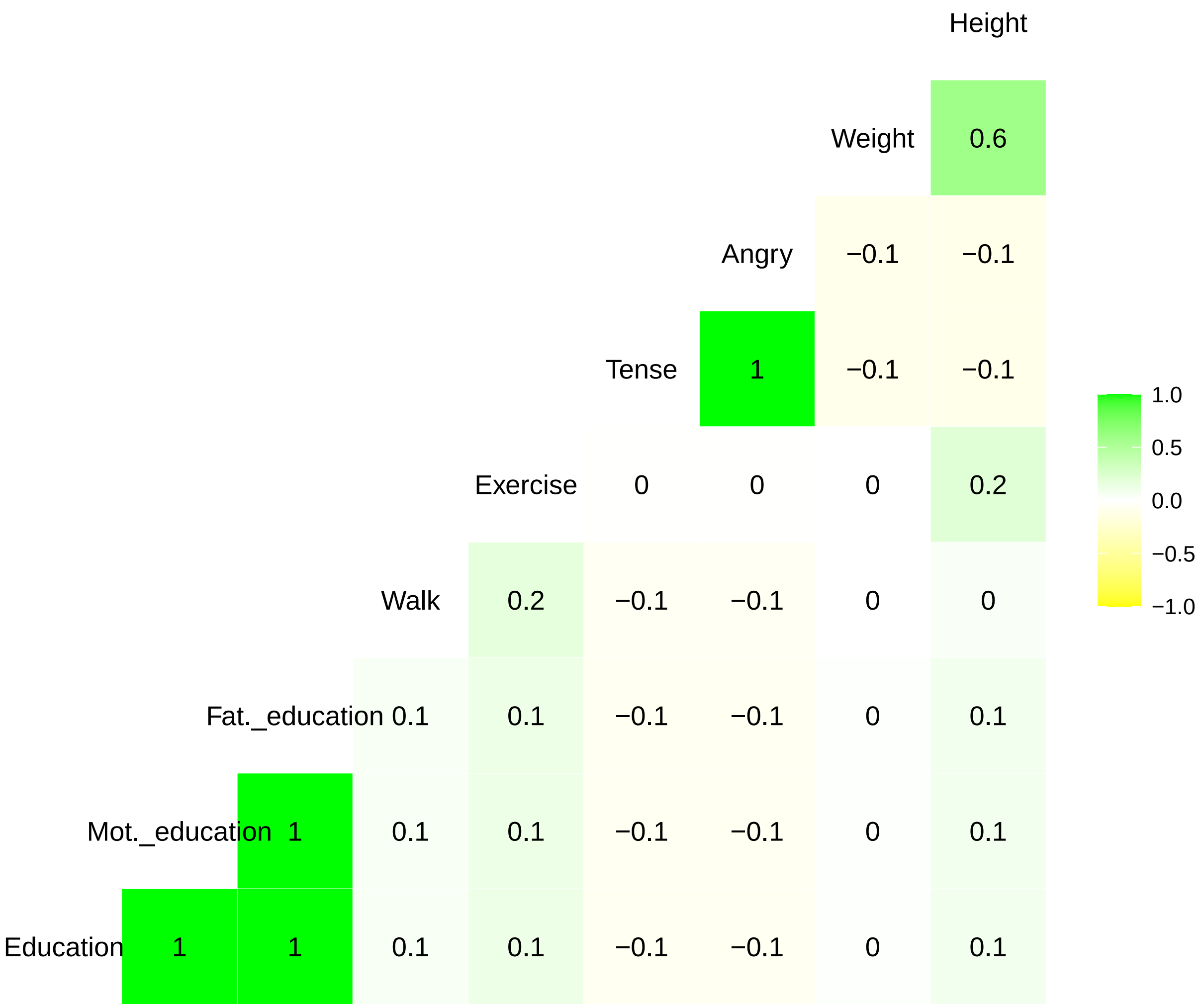
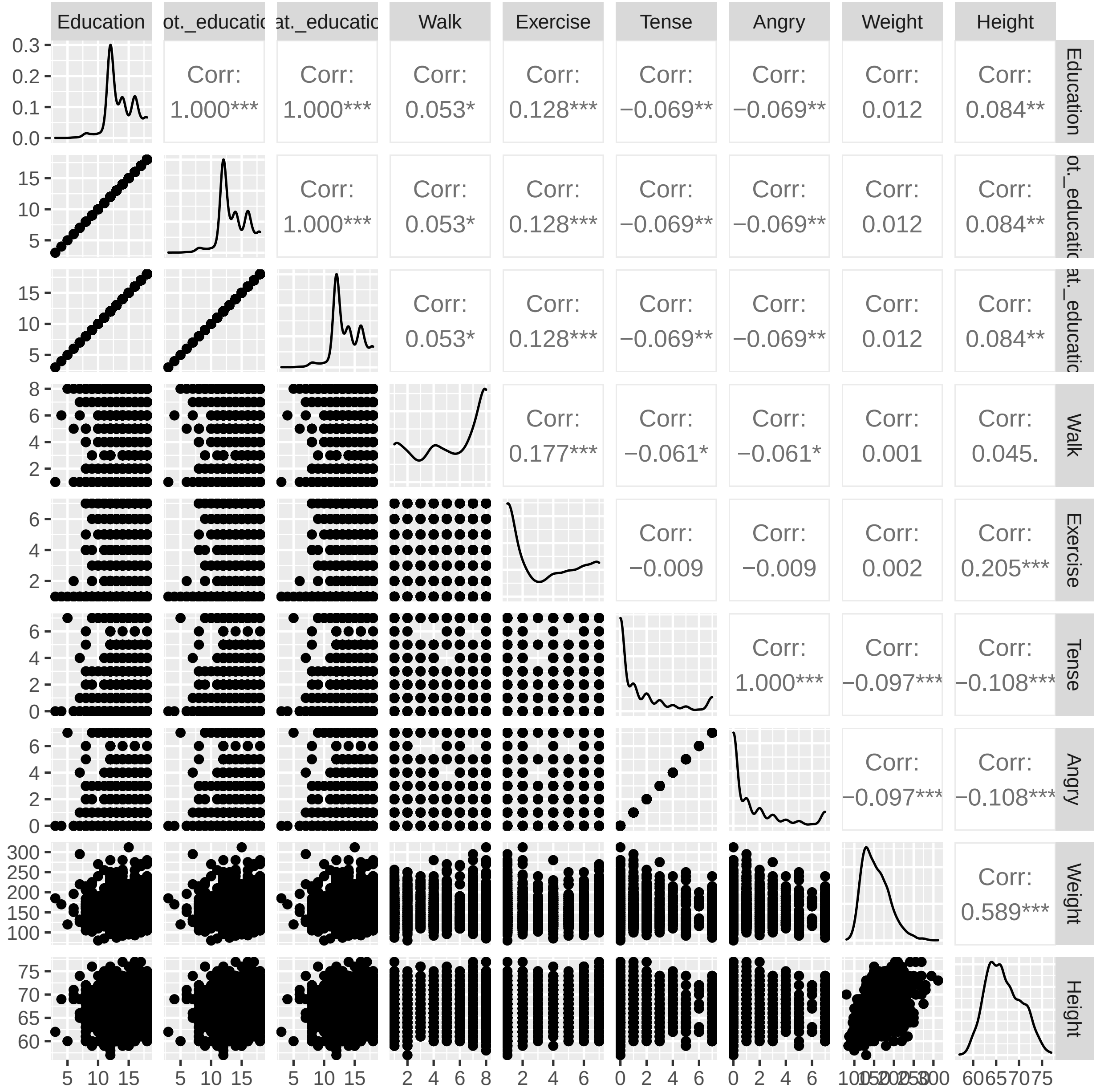


Figure 1: Correlation



```
p1_2 <- ggpairs(d_1)
p1_2
```



(b) (10 points) Perform a linear regression model using the variable *earn* as the dependent variable and *height* as the independent variable. What can you say about this covariate? Is it significant? 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)

b: The slope is 1595 and p-value is smaller than 2.2e-16 which means height is correlated to earn but the slope is not large as seen in graph. From the graph, we can also see that in each height rane, the earnings distributed widely over 0 - 200000.

```
attach(earnings)
```

```
## The following objects are masked from earnings (pos = 3):
##
##     age, angry, earn, education, ethnicity, exercise, father_education,
##     height, male, mother_education, smokenow, tense, walk, weight
```

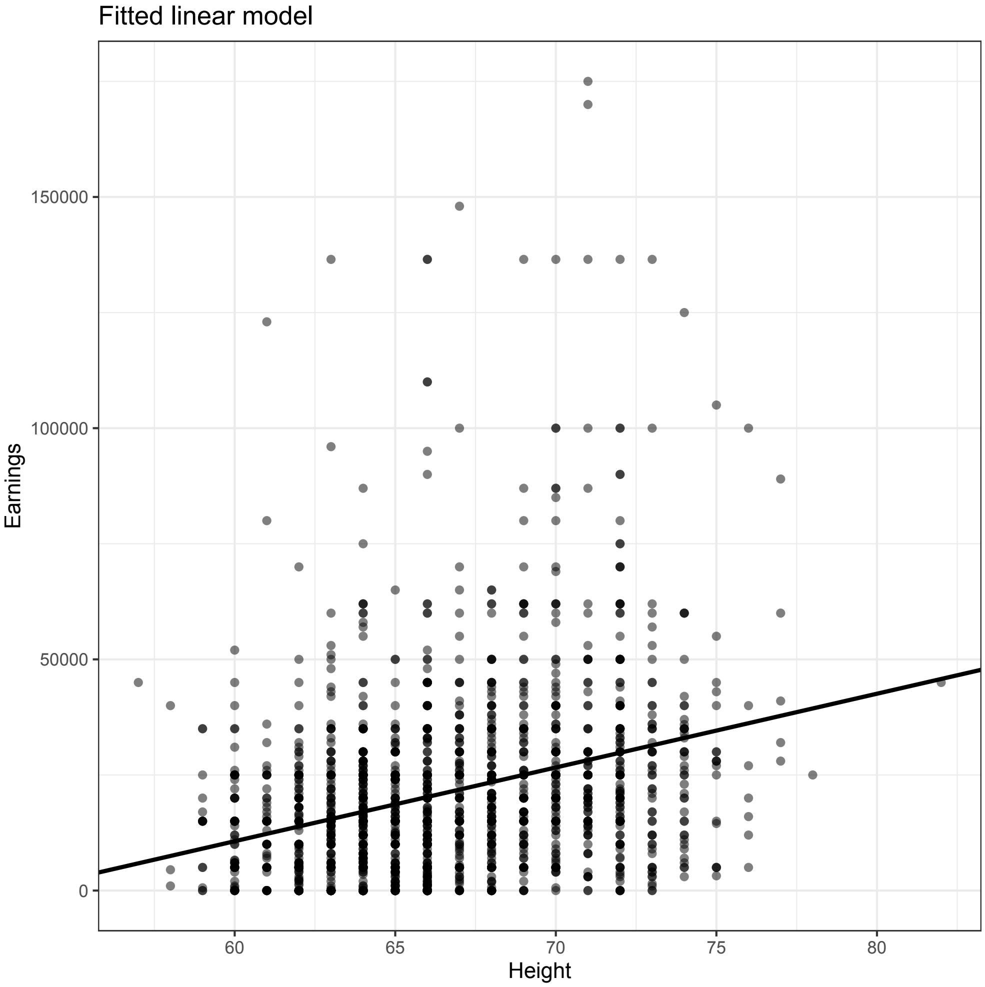


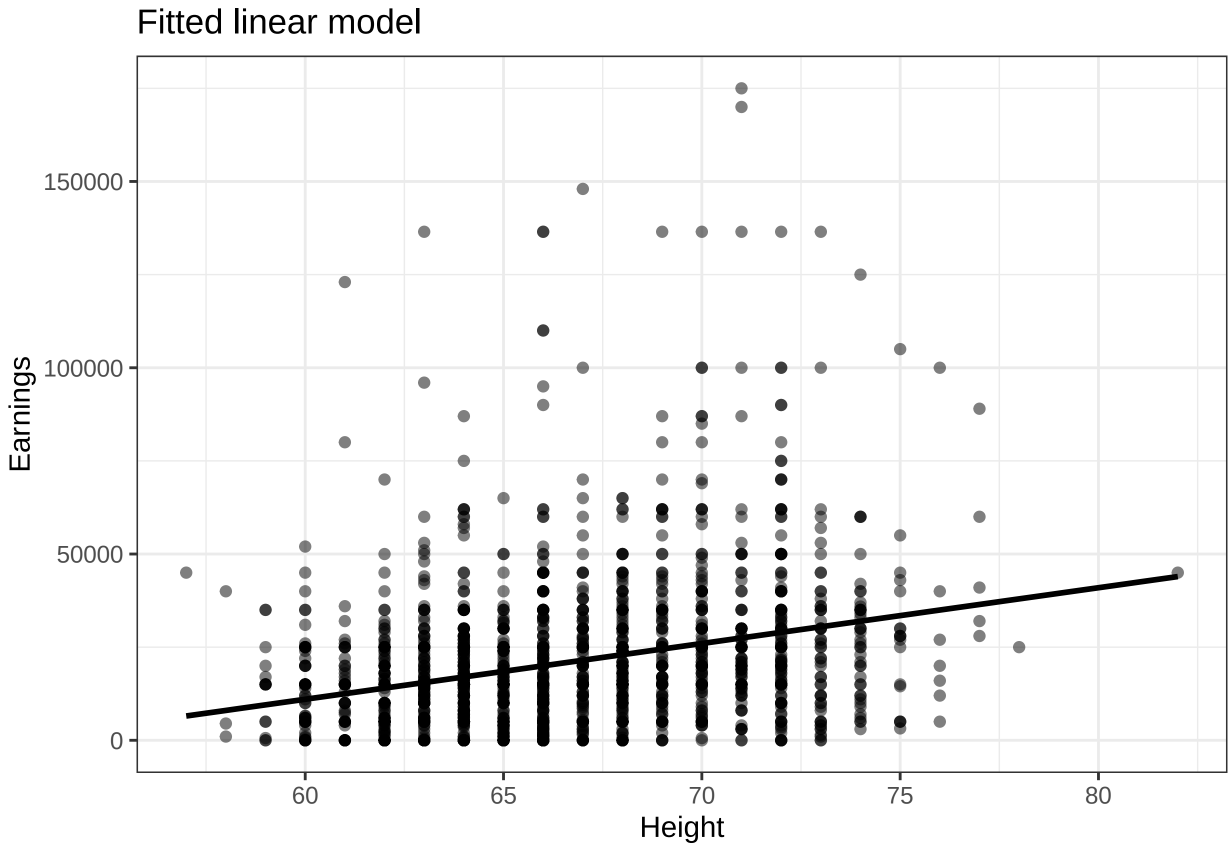
Figure 2: Linear Regression

```r
model_2 <- lm(earn~height)
summary(model_2)
```

```
##
## Call:
## lm(formula = earn ~ height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31405 -12456  -3645   6570 370190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -85027.3     8860.7  -9.596   <2e-16 ***
## height        1595.0      132.9  12.003   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21690 on 1814 degrees of freedom
## Multiple R-squared:  0.07357,    Adjusted R-squared:  0.07306
## F-statistic: 144.1 on 1 and 1814 DF,  p-value: < 2.2e-16
```

```r
df_2 <- earnings %>% filter(earn < 2*10^5)
p2 <- ggplot(data = df_2, aes(x = height, y = earn)) + geom_point(alpha = 0.5)
p2 <- p2 + geom_smooth(method = "lm", se = FALSE, color='black')
p2 <- p2 + theme_bw() +
  scale_x_continuous(name="Height") +
  scale_y_continuous(name="Earnings") + ggtitle("Fitted linear model")
p2
```

```
## `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?

c: We cann see that generally, male earns more than females. But both of the genders does not follow the normal distribution since the red line and the blue line are below the diagnoal line, which menas that the sample distribution does not match the theoretical normal distribution.

```
p3 <- qplot(sample = earn, shape=as.factor(earnings$male), color=as.factor(earnings$male))
p3 <- p3 + theme_classic() +
  scale_x_continuous(name = "Theoretical") +
  scale_y_continuous(name = "Sample",limits=c(0, 4e+05),breaks = c(0,1e+05,2e+05,3e+05,4e+05),
                   labels = function(x) format(x, scientific = TRUE)) +
  scale_colour_discrete(name = 'Gender',
                      breaks=c("0", "1"),
                      labels=c("Female", "Male")) +
  scale_shape_discrete(name = 'Gender',
                       breaks=c("0", "1"),
                       labels=c("Female", "Male")) +
  ggtitle("Earnings for different groups")
p3
```

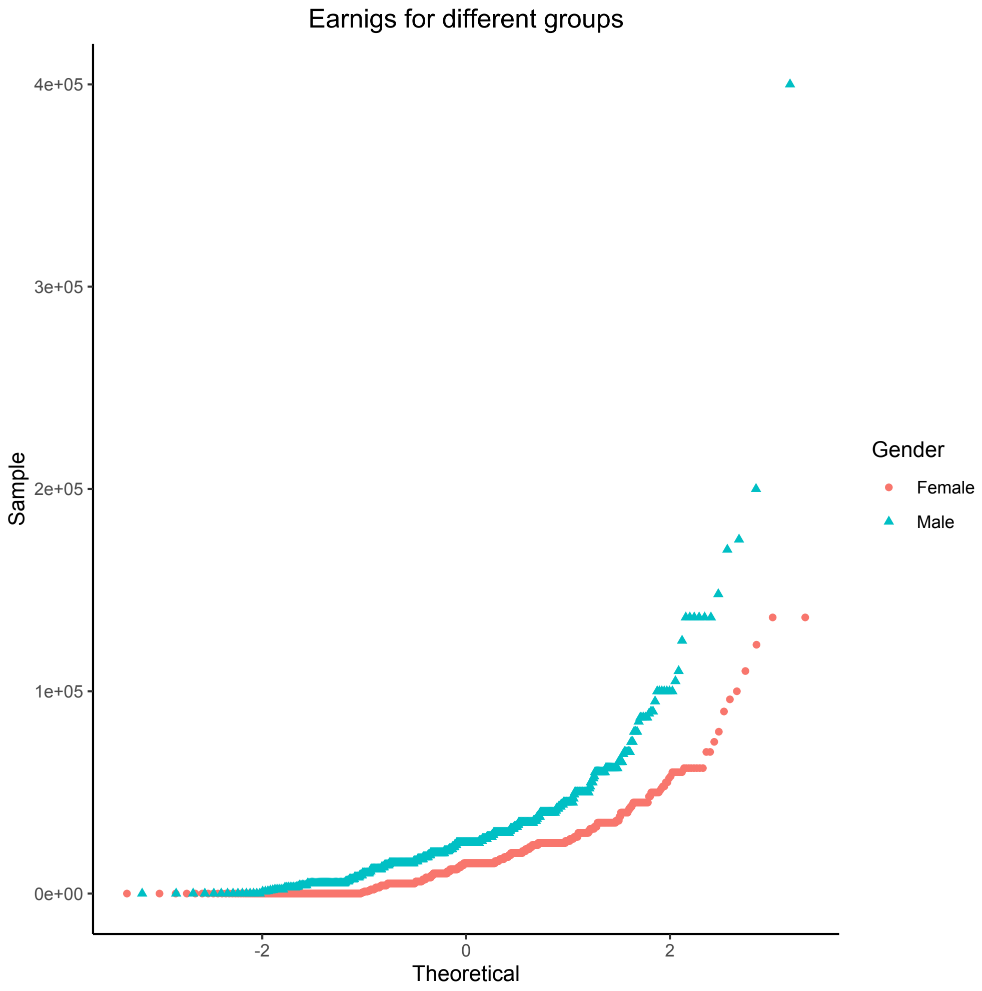
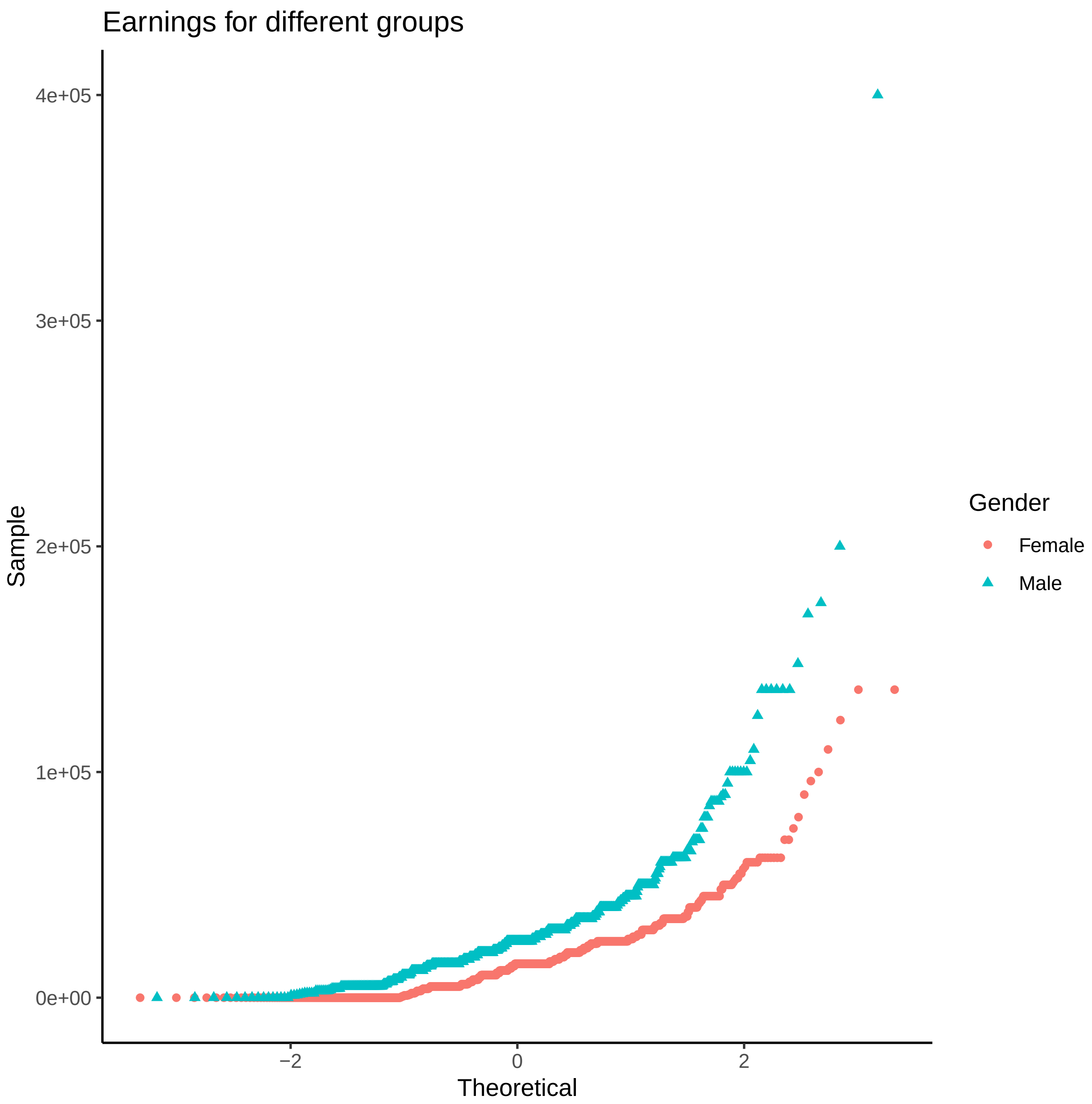


Figure 3: QQplot for different groups



(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]).

d: I prefered forward selection because both have 28699.37, but the forward selection reaches the best combination with fewer steps.

```
d_4 <- na.omit(earnings) # remove missing values
fit_for1 <- lm(earn~., data=d_4)
fit_for2 <- lm(earn ~ 1, data=d_4)
print("FORWARD SELECTION")
```

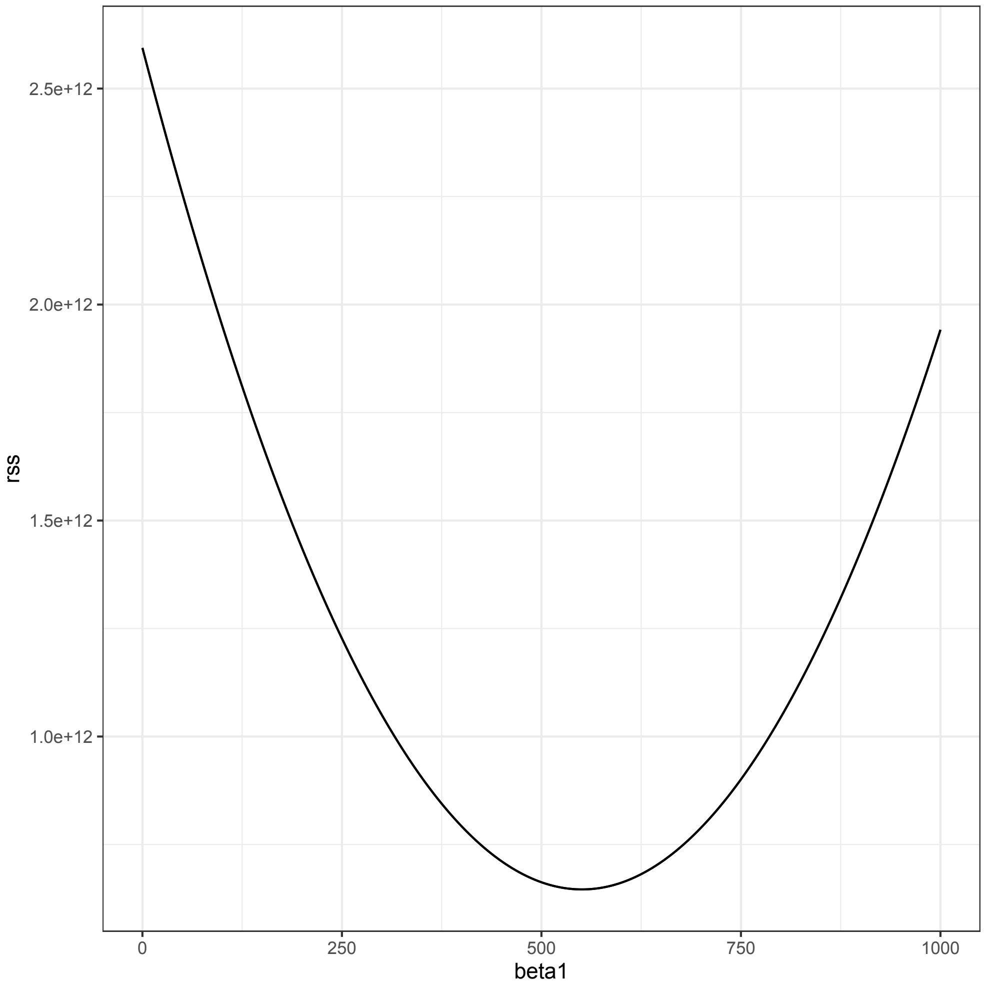


Figure 4: RSS for the backward procedure

```
## [1] "FORWARD SELECTION"
```

```
model_forward <- stepAIC(fit_for2,direction="forward",scope=list(upper=fit_for1,lower=fit_for2))
```

```
## Start:  AIC=29031.73
## earn ~ 1
## 
##                      Df  Sum of Sq        RSS   AIC
## + male                1 7.9739e+10 7.3974e+11 28886
## + education           1 7.4527e+10 7.4495e+11 28896
## + mother_education    1 7.4527e+10 7.4495e+11 28896
## + father_education    1 7.4527e+10 7.4495e+11 28896
## + height              1 5.9505e+10 7.5997e+11 28925
## + weight              1 2.9325e+10 7.9016e+11 28981
## + age                 1 1.3370e+10 8.0611e+11 29010
## + exercise            1 5.0787e+09 8.1440e+11 29025
## + tense               1 3.9410e+09 8.1554e+11 29027
## + angry               1 3.9410e+09 8.1554e+11 29027
## + ethnicity           3 4.8139e+09 8.1467e+11 29029
## <none>                           8.1948e+11 29032
## + smokenow            1 9.1854e+08 8.1856e+11 29032
## + walk                1 9.0848e+08 8.1857e+11 29032
## 
## Step:  AIC=28886.32
## earn ~ male
## 
##                      Df  Sum of Sq        RSS   AIC
## + education           1 6.9797e+10 6.6994e+11 28746
## + mother_education    1 6.9797e+10 6.6994e+11 28746
## + father_education    1 6.9797e+10 6.6994e+11 28746
## + age                 1 1.7756e+10 7.2198e+11 28853
## + height              1 3.6894e+09 7.3605e+11 28881
## + ethnicity           3 5.3068e+09 7.3443e+11 28882
## <none>                           7.3974e+11 28886
## + weight              1 9.4776e+08 7.3879e+11 28886
## + smokenow            1 9.1854e+08 7.3882e+11 28886
## + tense               1 3.4937e+08 7.3939e+11 28888
## + angry               1 3.4937e+08 7.3939e+11 28888
## + walk                1 2.3203e+08 7.3951e+11 28888
## + exercise            1 9.0739e+07 7.3965e+11 28888
## 
## Step:  AIC=28745.61
## earn ~ male + education
## 
##               Df  Sum of Sq        RSS   AIC
## + age          1 1.9518e+10 6.5043e+11 28705
## + ethnicity    3 3.5226e+09 6.6642e+11 28744
## + height       1 1.4225e+09 6.6852e+11 28744
## + weight       1 1.0396e+09 6.6890e+11 28745
## <none>                    6.6994e+11 28746
## + exercise     1 5.4813e+08 6.6940e+11 28746
## + smokenow     1 1.1241e+08 6.6983e+11 28747
## + walk         1 3.1565e+06 6.6994e+11 28748
## + tense        1 2.5823e+06 6.6994e+11 28748
## + angry        1 2.5823e+06 6.6994e+11 28748
```

```
##
## Step:  AIC=28705.03
## earn ~ male + education + age
##
##              Df  Sum of Sq        RSS   AIC
## + height      1 2921837319 6.4750e+11 28700
## + tense       1 1350718589 6.4908e+11 28704
## + angry       1 1350718589 6.4908e+11 28704
## <none>                     6.5043e+11 28705
## + smokenow    1  362030699 6.5006e+11 28706
## + ethnicity   3 2093480782 6.4833e+11 28706
## + exercise    1  287699226 6.5014e+11 28706
## + weight      1  137467965 6.5029e+11 28707
## + walk        1   33481537 6.5039e+11 28707
##
## Step:  AIC=28700.55
## earn ~ male + education + age + height
##
##              Df  Sum of Sq        RSS   AIC
## + tense       1 1430271509 6.4607e+11 28699
## + angry       1 1430271509 6.4607e+11 28699
## <none>                     6.4750e+11 28700
## + smokenow    1  357020068 6.4715e+11 28702
## + exercise    1  235648665 6.4727e+11 28702
## + weight      1  111208434 6.4739e+11 28702
## + walk        1   41917712 6.4746e+11 28702
## + ethnicity   3 1521906378 6.4598e+11 28703
##
## Step:  AIC=28699.37
## earn ~ male + education + age + height + tense
##
##              Df  Sum of Sq        RSS   AIC
## <none>                     6.4607e+11 28699
## + exercise    1  288700139 6.4579e+11 28701
## + smokenow    1  282249180 6.4579e+11 28701
## + weight      1  137259702 6.4594e+11 28701
## + walk        1   25908265 6.4605e+11 28701
## + ethnicity   3 1607572422 6.4447e+11 28702
```

```
summary(model_forward)
```

```
##
## Call:
## lm(formula = earn ~ male + education + age + height + tense,
##     data = d_4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43220 -10972  -2314   6306 371527
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70750.77   14208.16  -4.980 7.14e-07 ***
## male         12694.11    1658.43   7.654 3.55e-14 ***
## education     2952.42     237.43  12.435  < 2e-16 ***
## age            257.45      36.48   7.058 2.62e-12 ***
## height         550.82     213.41   2.581  0.00995 **
## tense          490.96     275.55   1.782  0.07500 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21230 on 1434 degrees of freedom
## Multiple R-squared:  0.2116, Adjusted R-squared:  0.2089
## F-statistic: 76.98 on 5 and 1434 DF,  p-value: < 2.2e-16
```

```
print("BACKWARD SELECTION")
```

```
## [1] "BACKWARD SELECTION"
```

```
## backward selection
fit_bac <- lm(earn~., data = d_4)
model_back <- step(fit_bac, direction= 'backward')
```

```
## Start:  AIC=28708.32
## earn ~ height + weight + male + ethnicity + education + mother_education +
##     father_education + walk + exercise + smokenow + tense + angry +
##     age
##
##
## Step:  AIC=28708.32
## earn ~ height + weight + male + ethnicity + education + mother_education +
##     father_education + walk + exercise + smokenow + tense + age
##
##
## Step:  AIC=28708.32
## earn ~ height + weight + male + ethnicity + education + mother_education +
##     walk + exercise + smokenow + tense + age
##
##
## Step:  AIC=28708.32
## earn ~ height + weight + male + ethnicity + education + walk +
##     exercise + smokenow + tense + age
##
##               Df  Sum of Sq        RSS   AIC
## - ethnicity  3 1.4940e+09 6.4531e+11 28706
## - walk       1 3.1861e+07 6.4385e+11 28706
## - weight     1 5.1199e+07 6.4387e+11 28706
## - smokenow   1 2.5692e+08 6.4407e+11 28707
## - exercise   1 3.2969e+08 6.4415e+11 28707
## <none>                    6.4382e+11 28708
## - tense      1 1.4955e+09 6.4531e+11 28710
## - height     1 2.2805e+09 6.4610e+11 28711
## - age        1 2.0675e+10 6.6449e+11 28752
## - male       1 2.5472e+10 6.6929e+11 28762
## - education  1 6.6279e+10 7.1009e+11 28847
##
## Step:  AIC=28705.66
## earn ~ height + weight + male + education + walk + exercise +
##     smokenow + tense + age
##
```

```
##             Df  Sum of Sq        RSS   AIC
## - walk       1 6.3440e+07 6.4537e+11 28704
## - weight     1 7.7849e+07 6.4539e+11 28704
## - smokenow   1 3.0394e+08 6.4561e+11 28704
## - exercise   1 3.4289e+08 6.4565e+11 28704
## <none>                    6.4531e+11 28706
## - tense      1 1.4071e+09 6.4672e+11 28707
## - height     1 2.8270e+09 6.4814e+11 28710
## - age        1 2.2170e+10 6.6748e+11 28752
## - male       1 2.4948e+10 6.7026e+11 28758
## - education  1 6.7532e+10 7.1284e+11 28847
##
## Step:  AIC=28703.8
## earn ~ height + weight + male + education + exercise + smokenow +
##     tense + age
##
##             Df  Sum of Sq        RSS   AIC
## - weight     1 7.4075e+07 6.4545e+11 28702
## - exercise   1 3.0181e+08 6.4568e+11 28702
## - smokenow   1 3.1560e+08 6.4569e+11 28702
## <none>                    6.4537e+11 28704
## - tense      1 1.4239e+09 6.4680e+11 28705
## - height     1 2.8131e+09 6.4819e+11 28708
## - age        1 2.2147e+10 6.6752e+11 28750
## - male       1 2.4931e+10 6.7030e+11 28756
## - education  1 6.7475e+10 7.1285e+11 28845
##
## Step:  AIC=28701.97
## earn ~ height + male + education + exercise + smokenow + tense +
##     age
##
##             Df  Sum of Sq        RSS   AIC
## - smokenow   1 3.3784e+08 6.4579e+11 28701
## - exercise   1 3.4429e+08 6.4579e+11 28701
## <none>                    6.4545e+11 28702
## - tense      1 1.4060e+09 6.4685e+11 28703
## - height     1 2.9325e+09 6.4838e+11 28706
## - age        1 2.2336e+10 6.6778e+11 28749
## - male       1 2.5233e+10 6.7068e+11 28755
## - education  1 6.7691e+10 7.1314e+11 28844
##
## Step:  AIC=28700.72
## earn ~ height + male + education + exercise + tense + age
##
##             Df  Sum of Sq        RSS   AIC
## - exercise   1 2.8870e+08 6.4607e+11 28699
## <none>                    6.4579e+11 28701
## - tense      1 1.4833e+09 6.4727e+11 28702
## - height     1 2.9449e+09 6.4873e+11 28705
## - age        1 2.2071e+10 6.6786e+11 28747
## - male       1 2.5335e+10 6.7112e+11 28754
## - education  1 6.7697e+10 7.1348e+11 28842
##
## Step:  AIC=28699.37
## earn ~ height + male + education + tense + age
##
##             Df  Sum of Sq        RSS   AIC
## <none>                    6.4607e+11 28699
## - tense      1 1.4303e+09 6.4750e+11 28700
## - height     1 3.0014e+09 6.4908e+11 28704
## - age        1 2.2443e+10 6.6852e+11 28746
## - male       1 2.6396e+10 6.7247e+11 28755
## - education  1 6.9666e+10 7.1574e+11 28845
```

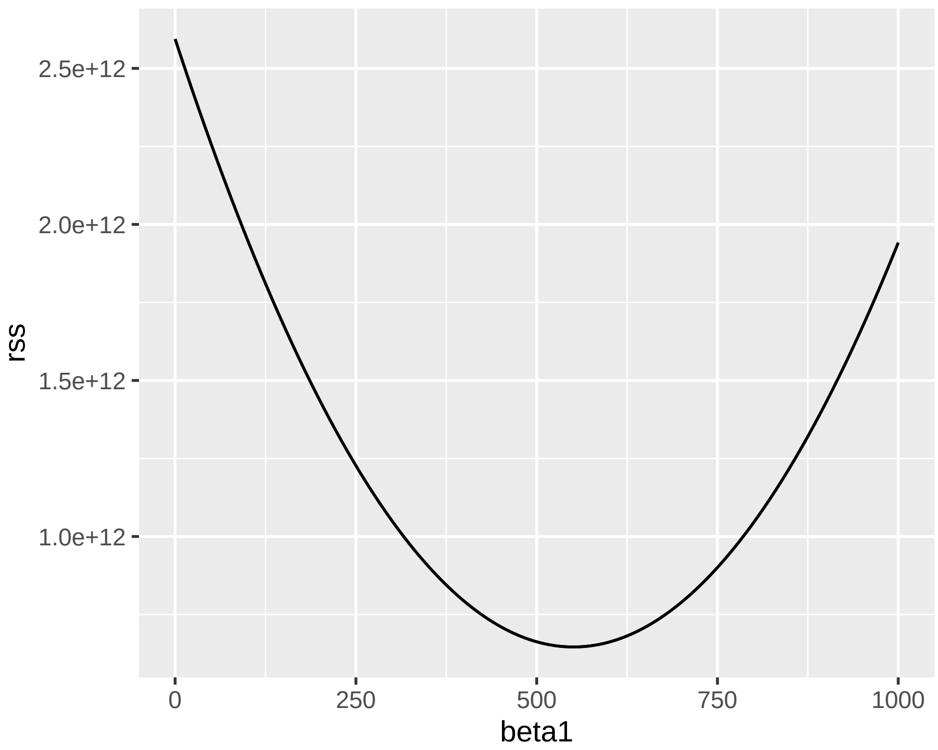
```
summary(model_back)
```

```
##
## Call:
## lm(formula = earn ~ height + male + education + tense + age,
##     data = d_4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43220 -10972  -2314   6306 371527
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70750.77   14208.16  -4.980 7.14e-07 ***
## height         550.82     213.41   2.581  0.00995 **
## male         12694.11    1658.43   7.654 3.55e-14 ***
## education     2952.42     237.43  12.435  < 2e-16 ***
## tense          490.96     275.55   1.782  0.07500 .
## age            257.45      36.48   7.058 2.62e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21230 on 1434 degrees of freedom
## Multiple R-squared:  0.2116, Adjusted R-squared:  0.2089
## F-statistic: 76.98 on 5 and 1434 DF,  p-value: < 2.2e-16
```

```
beta1 <- seq(0, 1000, 1)

rss <- function(beta, beta1, data){
  res <- d_4$earn - (beta[1]+beta1*d_4$height+beta[3]*d_4$male+beta[4]*d_4$education+beta[5]*d_4$tense
  return(sum(res^2))
}

results <- data.frame(beta1 = beta1,
                      rss = sapply(beta1, rss, beta = as.numeric(model_back$coefficients)))
results %>% ggplot(aes(beta1, rss)) + geom_line() +
  geom_line(aes(beta1, rss))
```



(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.

```
n <- 500
d_5 <- na.omit(earnings)
samples <- 500
coef_boot <- matrix(NA, n, 3)
for (i in 1:n){
  s_boot <- sample(c(1:dim(d_5)[1]), samples, replace=FALSE)
  data_boot <- d_5[s_boot,]
  fit3 <- lm(earn/1000 ~ height + male + education + tense + age, data = data_boot)
  coef_boot[i,] <- fit3$coefficients[2:4]
}
d_5 <- data.frame(value = c(coef_boot[,1], coef_boot[,2], coef_boot[,3]),
                        beta = rep(c("beta_1","beta_2","beta_3"), each = n))
df_model_backward <- data.frame(beta = c("beta_1","beta_2","beta_3"),value = as.numeric(model_back$coef
p5 <- ggplot(d_5, aes(x = value)) + geom_histogram(binwidth = 0.5,color="black", fill="grey") +
  facet_wrap(.~beta, ncol = 3) + geom_vline(data=df_model_backward, aes(xintercept=value, color="red"),
p5 <- p5 + scale_x_continuous(limits = c(0, 18)) + theme(legend.position="none")
p5
```

```
## Warning: Removed 18 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```

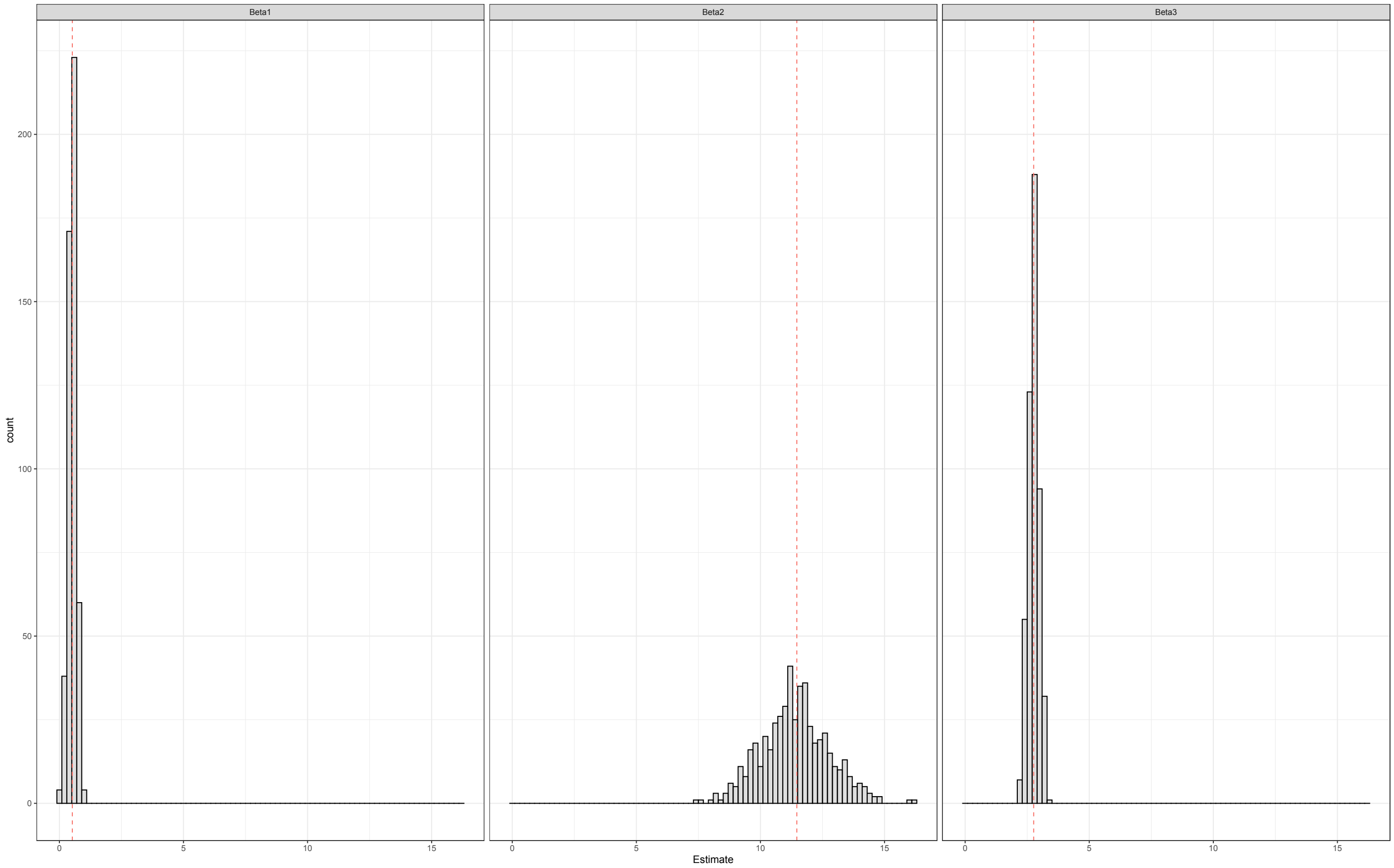
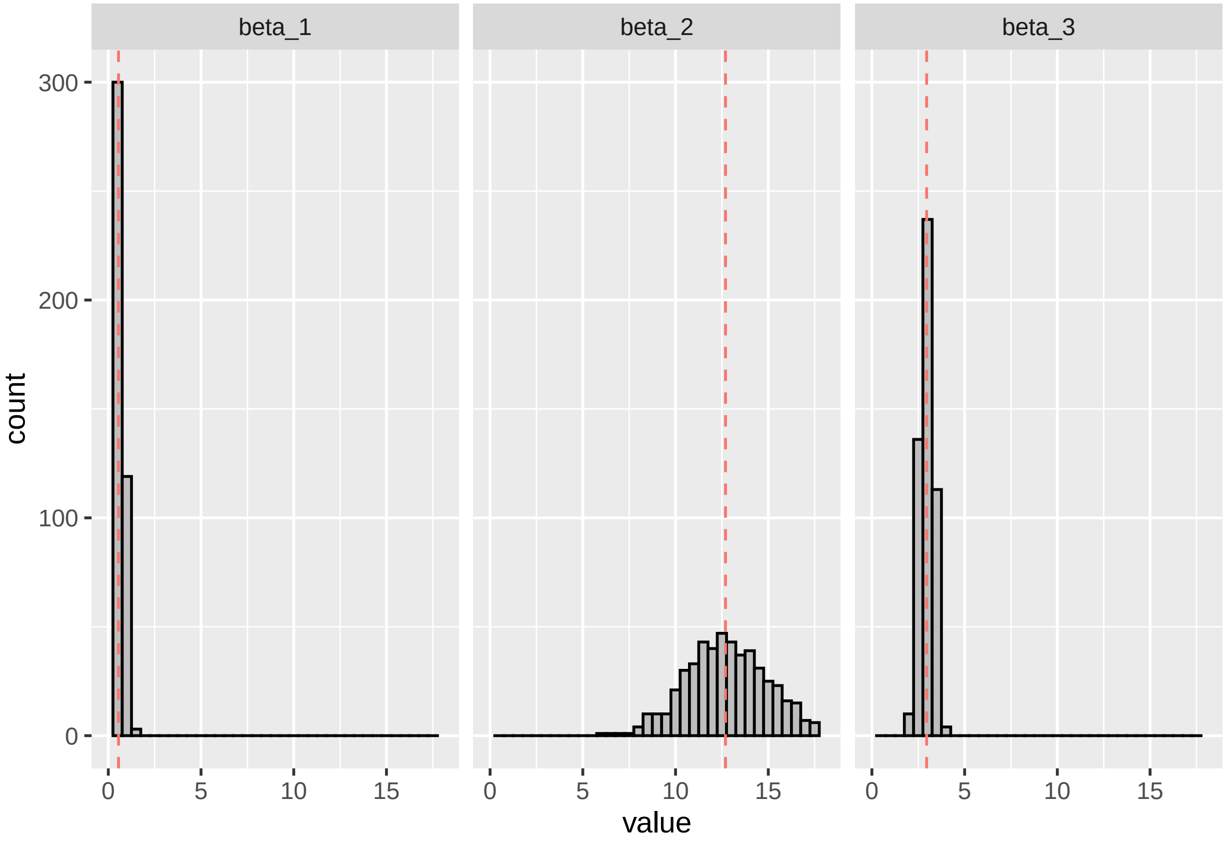


Figure 5: Bootstrap Results



(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.

f: mse for LOOCV is 453234294 while mse for K-fold CV is 425241313. Both are big since earnings have very big magnitude and the lm fitted model does not fit every point closely so that the MSE is big. After ploting the prediction, I find that the predictions are influenced some big outliers (large figures in earn). Only capturing the linear relationship does not fully capture the relationship between earn and other 5 features.

```
train.control_loocv <- trainControl(method = "LOOCV")
# Train the model
d_6 <- na.omit(earnings)
model_loocv <- train(earn~height + male + education + tense + age, data = d_6, method = "lm", trControl
# Summarize the results
print("LOOCV")
```

```
## [1] "LOOCV"
```

```
print(model_loocv)
```

```
## Linear Regression
##
## 1440 samples
##    5 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 1439, 1439, 1439, 1439, 1439, 1439, ...
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   21289.3  0.2036473  12870.92
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

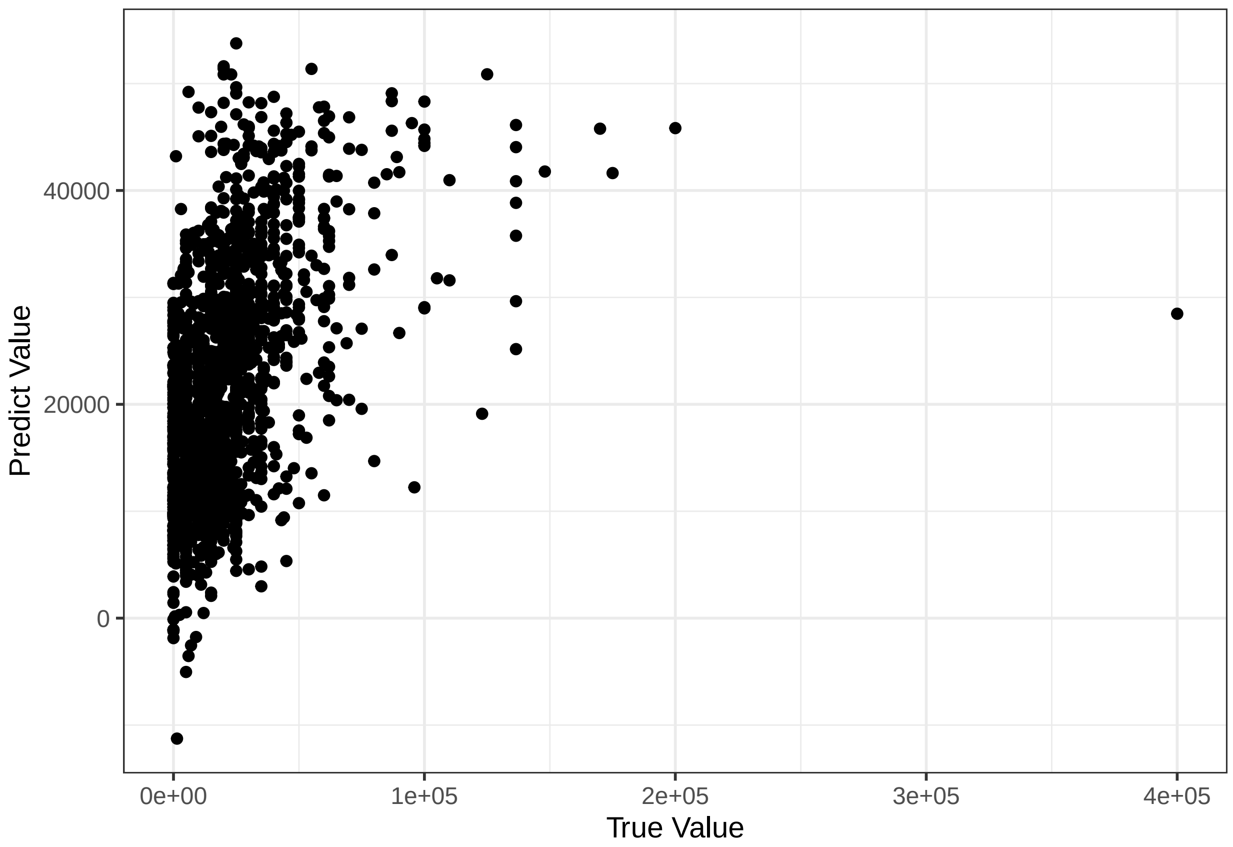
```
train.control_kcv <- trainControl(method = "cv", number = 10)
# Train the model
model_kcv <- train(earn~ height + male + education + tense + age, data = d_6, method = "lm", trControl
# Summarize the results
print("K-fold CV")
```

```
## [1] "K-fold CV"
```

```
print(model_kcv)
```

```
## Linear Regression
##
## 1440 samples
##    5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1295, 1295, 1295, 1297, 1296, 1298, ...
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   20580.1  0.2423434  12870.63
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
predictions <- predict(model_loocv, newdata = d_6)
res6 <- data.frame(Estimated=d_6$earn, True_value=predictions)
ggplot(aes(x=Estimated, y=True_value), data=res6) + geom_point() + labs(x="True Value", y="Predict Valu
```


Predict Value
40000
20000
0
0e+00
1e+05
2e+05
3e+05
4e+05
True Value