ECON 430: Project 1

Anshika Sharma, Cristian Martinez

November 17, 2020

Contents

Data Description and Directory

```
library(AER)
data(PSID1976)
survey = PSID1976

head(survey)
```

```
##   participation hours youngkids oldkids age education   wage repwage hhours
## 1           yes  1610         1       0  32        12 3.3540    2.65   2708
## 2           yes  1656         0       2  30        12 1.3889    2.65   2310
## 3           yes  1980         1       3  35        12 4.5455    4.04   3072
## 4           yes   456         0       3  34        12 1.0965    3.25   1920
## 5           yes  1568         1       2  31        14 4.5918    3.60   2000
## 6           yes  2032         0       0  54        12 4.7421    4.70   1040
##   hage heducation   hwage fincome    tax meducation feducation unemp city
## 1   34         12  4.0288   16310 0.7215         12          7   5.0   no
## 2   30          9  8.4416   21800 0.6615          7          7  11.0  yes
## 3   40         12  3.5807   21040 0.6915         12          7   5.0   no
## 4   53         10  3.5417    7300 0.7815          7          7   5.0   no
## 5   32         12 10.0000   27300 0.6215         12         14   9.5  yes
## 6   57         11  6.7106   19495 0.6915         14          7   7.5  yes
##   experience college hcollege
## 1         14      no       no
## 2          5      no       no
## 3         15      no       no
## 4          6      no       no
## 5          7     yes       no
## 6         33      no       no
```

```
#length(survey)
#names(survey)
```

Dataset: PSID1976: Labor Force Participation Data

Description: Cross-section data originating from the 1976 Panel Study of Income Dynamics (PSID), based on data for the previous year, 1975.

Purpose of the project: The following project explores family income dynamics in 1976. The subjects in the study were interviewed regarding the income generated in 1975. The focus of the analysis was to observe how wages of wives are affected by hours worked, years of education, wage of their husband, hours the husband worked, whether the husband attended college, and the marginal tax rate facing the wives. There were 753 observations with 428 participating the labor market. Although the data is over 40 years old, it continues to serve as an important reminder that family income dynamics still exists today, not only in the United States, but worldwide.

Data Directory: A data frame containing 753 observations on 21 variables.

1. participation: Did the individual participate in the labor force in 1975? (This is essentially wage > 0 or hours > 0.)
2. hours: Wife's hours of work in 1975.
3. youngkids: Number of children less than 6 years old in household.

4. ldkids: Number of children between ages 6 and 18 in household.
5. age: Wife's age in years.
6. education: Wife's education in years.
7. wage: Wife's average hourly wage, in 1975 dollars.
8. repwage: Wife's wage reported at the time of the 1976 interview (not the same as the 1975 estimated wage). To use the subsample with this wage, one needs to select 1975 workers with participation == "yes", then select only those women with non-zero wage. Only 325 women work in 1975 and have a non-zero wage in 1976.
9. hhours: Husband's hours worked in 1975.
10. hage: Husband's age in years.
11. heducation: Husband's education in years.
12. hwage: Husband's wage, in 1975 dollars.
13. fincome: Family income, in 1975 dollars. (This variable is used to construct the property income variable.)
14. tax: Marginal tax rate facing the wife, and is taken from published federal tax tables (state and local income taxes are excluded). The taxable income on which this tax rate is calculated includes Social Security, if applicable to wife.
15. medication: Wife's mother's educational attainment, in years.
16. feducation: Wife's father's educational attainment, in years.
17. unemp: Unemployment rate in county of residence, in percentage points. (This is taken from bracketed ranges.)
18. city: Factor. Does the individual live in a large city?
19. experience: Actual years of wife's previous labor market experience.
20. college: Factor. Did the individual attend college?
21. hcollege: Factor. Did the individual's husband attend college?

1: Variable Selection

a) Boruta Algorithm

```
library(leaps)
library(Boruta)

boruta.train <- Boruta(wage~., data = survey, doTrace = 2)
print(boruta.train)
```

```
## Boruta performed 99 iterations in 34.73892 secs.
##  10 attributes confirmed important: college, education, experience,
## fincome, hhours and 5 more;
##  6 attributes confirmed unimportant: city, feducation, hage,
## meducation, unemp and 1 more;
##  4 tentative attributes left: age, hcollege, heducation, oldkids;
```

```
cat(getSelectedAttributes(boruta.train), sep = "\n")
```

```
## participation
## hours
## education
## repwage
## hhours
## hwage
## fincome
## tax
## experience
## college
```

```
plot(boruta.train,las =2, cex.axis=0.75)
#The following variables were suggested by Boruta Algorithm
```

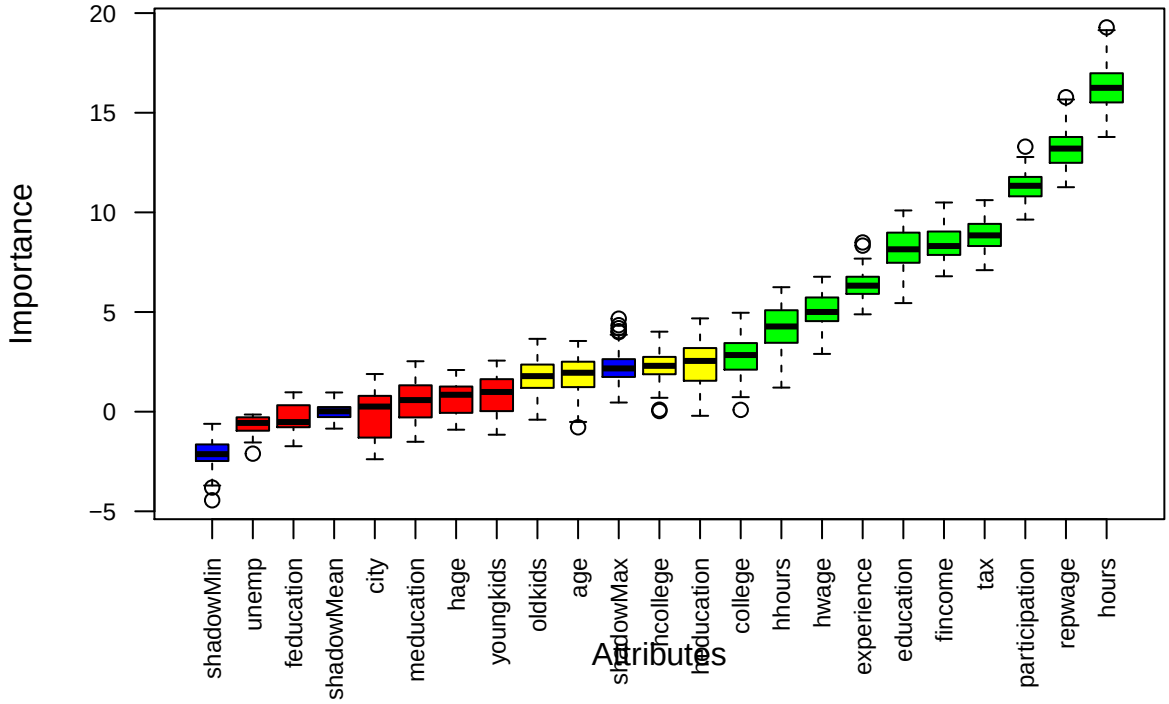


Figure 1: Results from the Boruta Algorithm

b) Mallows Cp

```
ss=regsubsets(wage~ hours + education + youngkids + oldkids + age +hhours + participation + education
+ repwage+ hage+heducation + fincome + meducation + feducation + unemp
+ city+ experience + college + hwage+ tax+ hcollege , method=c("exhaustive"),nbest=3,data=survey)
subsets(ss,statistic="cp",legend=F,main="Mallows CP",col="steelblue4", ylim=c(0,10))
```

```
##                   Abbreviation
## hours                       hr
## education                   ed
## youngkids                    y
## oldkids                      o
## age                          a
## hhours                      hh
## participationyes             p
## repwage                      r
## hage                        hg
## heducation                  hd
## fincome                     fn
## meducation                   m
## feducation                  fd
## unemp                        u
## cityyes                     ct
## experience                  ex
## collegeyes                  cl
## hwage                       hw
## tax                          t
## hcollegeyes                 hc
```

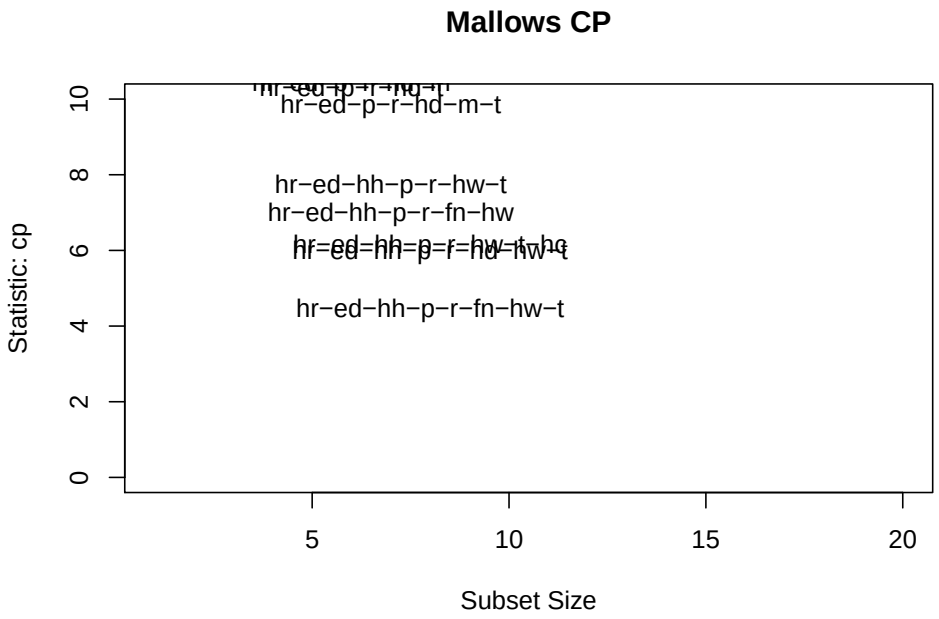


Figure 2: Results from Mallows Cp

Mallow Cp Chooses:

- hours
- education
- hhours
- participationyes
- repwage
- fincome
- hwage
- tax

c) Preferred choice of predictors

Based on the results in part (a) and (b), the following variables were selected (by with sub-setting the participation in the labor force in 1975 == yes)

- wage (dependent variable)
- hours
- education
- hhours

- hwage
- tax
- hcollege

```
#New data-frame with 6 selected variables:
wage = survey$wage[survey$participation=="yes"]
hours = survey$hours[survey$participation=="yes"]
education =survey$education[survey$participation=="yes"]
hwage = survey$hwage[survey$participation=="yes"]
tax = survey$tax[survey$participation=="yes"]
hhours = survey$hhours[survey$participation=="yes"]
hcollege = survey$hcollege[survey$participation=="yes"]

data = data.frame("wage" = wage,
                          "hours" = hours ,
                          "education" = education,
                          "hhours" = hhours,
                          "hwage" = hwage,
                          "tax" = tax,
                          "hcollege" =hcollege
                           )
```

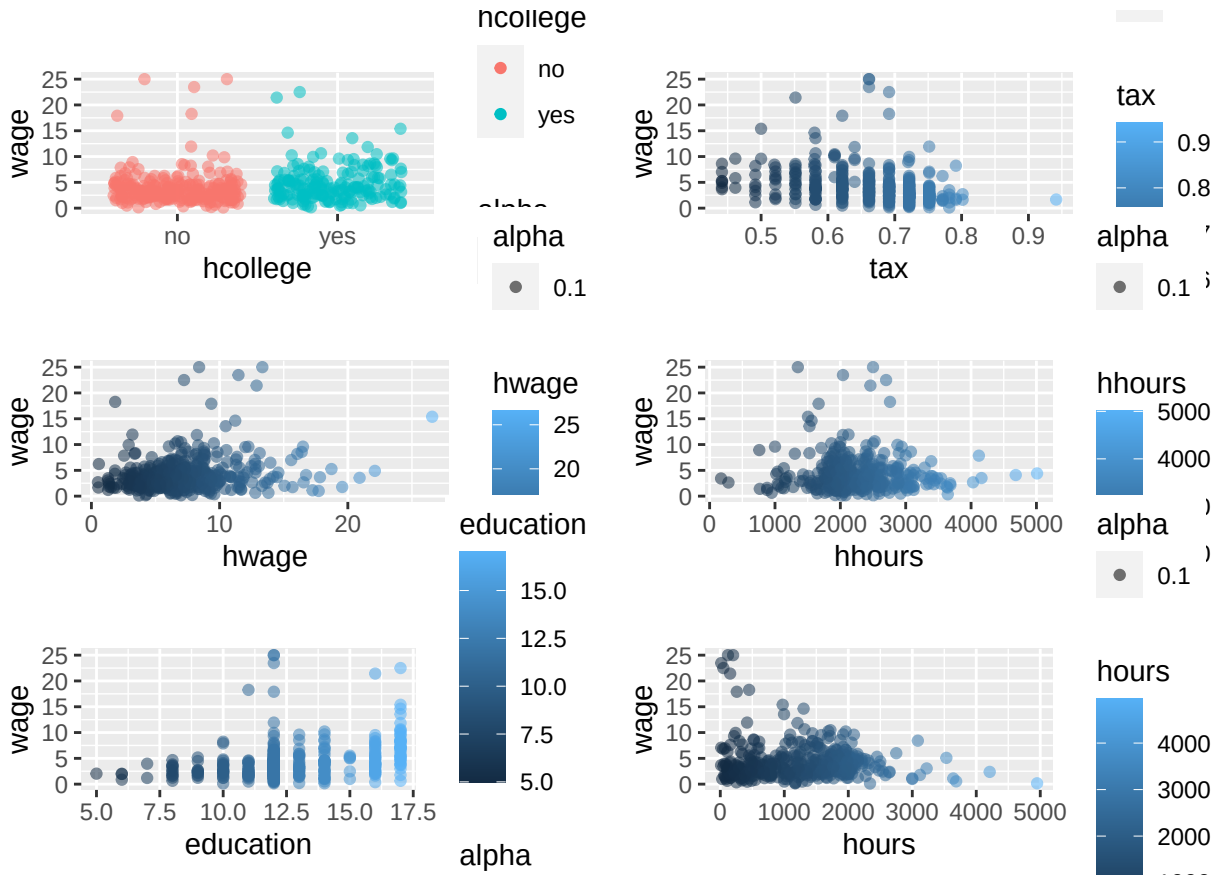
2: Descriptive Analysis

a) Univariate analysis

Quantile-Quantile plots, histograms and scatterplots have been produced for non-factor variables. For factor variables, we have used barplots.

Overview on Categorical and Continuous variables:

```
#hcollege (husband's attendance to college)
g1<- ggplot(data,aes(x=hcollege,y=wage,alpha=0.1, col=hcollege))+
labs(title=paste("", names(hcollege)))+
geom_jitter()

#tax (Marginal tax rate facing the wife)
g2<- ggplot(data,aes(x=tax,y=wage,alpha=0.1, col=tax))+
labs(title=paste("", names(tax)))+
 geom_point()

#hwage (Husband's wage)
g3<- ggplot(data,aes(x=hwage,y=wage,alpha=0.1, col=hwage))+
labs(title=paste("", names(hwage)))+
 geom_point()

#hhours (Husband's hours worked )
g4<- ggplot(data,aes(x=hhours,y=wage,alpha=0.1, col=hhours))+
labs(title=paste("", names(hhours)))+
 geom_point()

#education (Wife's education in years.)
g5<- ggplot(data,aes(x=education,y=wage,alpha=0.1, col=education))+
labs(title=paste("", names(education)))+
 geom_point()


#hours (Wife's hours of work)
g6<- ggplot(data,aes(x=hours,y=wage,alpha=0.1, col=hours))+
labs(title=paste("", names(hours)))+
geom_point()

require(gridExtra)
grid.arrange(g1, g2, g3, g4, g5, g6, ncol=2)
```



- hcollege: The barplot shows that the data consists of more wifes that their husband did not attend college.
- tax: From the plot, we observe that when wife's average hourly wage is the highest, they face a marginal tax rate around 0.67.
- hwage: We observe from the plot that there is positive relationship with husband's wage and wife's wage. But there are some observations that altyhough the husband wage is high, wife's wage is very low. As need to work for wife is likely to decrease when the husband wage is high, this might I explain why we see low wages as they might be working less.
- education: Education and wife's wages show positive relationship as seen in the graph but wifes that have 12 years education have the highest wage in our data.
- hhours: Husband's hours worked in 1976=5 and wife wages have a positive relationship but there are outliers as well.
- hours: Wife's hours of work in 1975 has almost a positive relationship with wages but there are some outliers as well. There are some observations where wife's wage is very low but hours of work is very high. If the wife does not have the option to not to work maybe because the husbands' wage is not adequate enough, the wife will work for long hours even with low wages.

Freedman's and Diaconis (FD) was used to determine the number of bins in the histograms and the Cullen-frey graph was used to get the best fitted distributions.

- Tax:

```r
#tax
par(mfrow=c(2,2))

#Histogram
hist(tax, breaks = "FD",  col = "skyblue3", main = "Histogram of tax ", freq = FALSE)
rug(tax)

#Density Functions using Cullen-Frey Graph
#descdist(tax, boot = 1000)
fln <- fitdist(tax, "lnorm")
plot.legend <- c("lnorm")
denscomp(list(fln), legendtext = plot.legend,main = "Fitted Dist of tax")

#Quantile Plots:
fit_beta = fitdist(tax, "beta")
fit_lnorm = fitdist(tax, "lnorm")
fit_norm  = fitdist(tax, "norm")
plot.legend = c("beta",  "lnorm", "norm")
qqcomp(list(fit_beta, fit_lnorm, fit_norm), legendtext = plot.legend,main = "Q-Q for tax")

#Boxplot
Boxplot(~tax, data=data, id=FALSE, main = "Boxplot for tax")
```

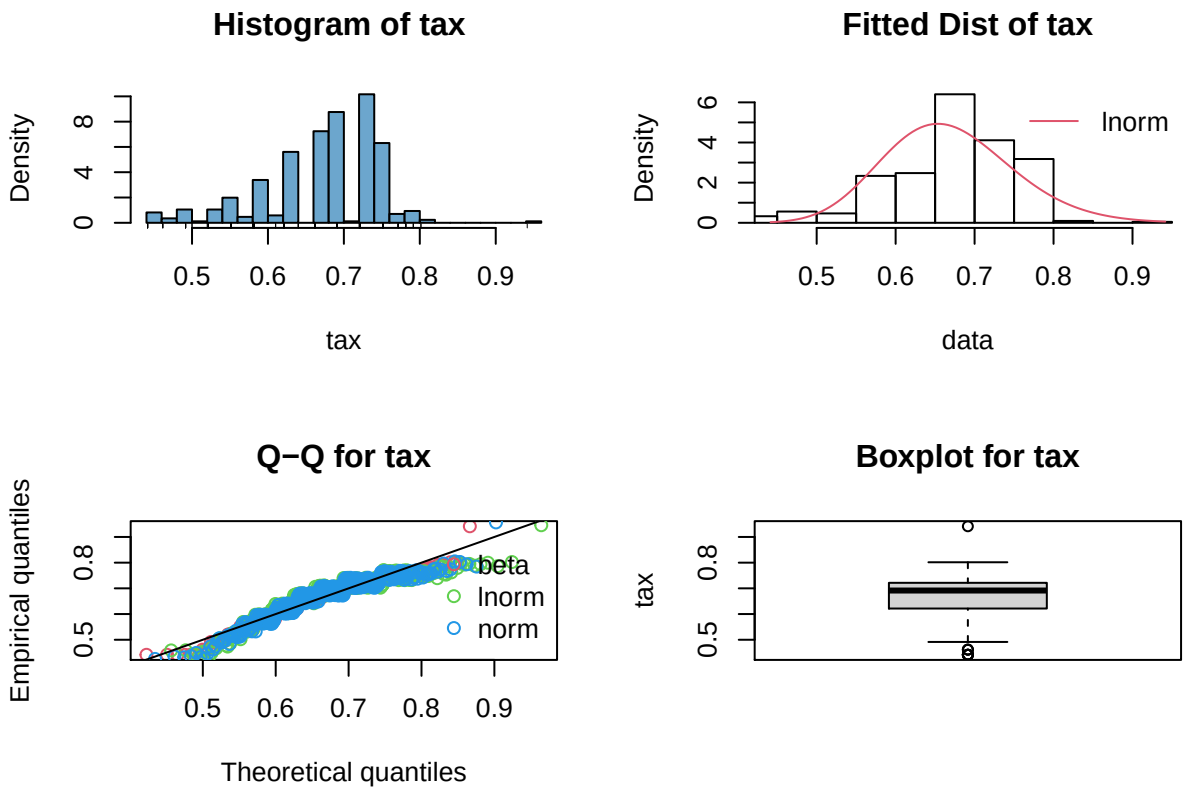


Figure 3: Descriptive Analysis for tax

From Fig.4, the distribution for the tax variable seems balanced and normalized for the most part. However, there is a presence of some outliers on both the upper and lower tails. Apart from that, most of the values concentrated on the center.

- Hours:

```
#hours
par(mfrow=c(2,2))

#Histogram
hist(hours, breaks = "FD",  col = "skyblue3", main = "Histogram of hours ", probability = TRUE)
lines(density(data$hours),lwd = 2, col ="navyblue")
rug(hours)

#Boxplot
Boxplot(~hours, data=data, id=FALSE, main = "Boxplot for hours")

#Density Functions using Cullen-Frey Graph
#descdist(hours, boot = 1000)
fln <- fitdist(hours, "logis")
plot.legend <- c("logis")
denscomp(list(fln), legendtext = plot.legend,main = "Fitted Dist of hours")

#Quantile Plots:
fit_logis = fitdist(hours, "logis")
fit_lnorm = fitdist(hours, "lnorm")
fit_norm  = fitdist(hours, "norm")
plot.legend = c("beta",  "lnorm", "norm")
qqcomp(list(fit_logis, fit_lnorm, fit_norm), legendtext = plot.legend, main = "Q-Q for hours")
```

From density histogram and boxplot in Fig.5, we notice that are positively skewed (right skewed) indicating the presence of some outliers on the upper tail.

- Education:

```
#education
par(mfrow=c(1,2))

#Density Functions using Cullen-Frey Graph
#descdist(education, boot = 1000)
fln <- fitdist(education, "norm")
plot.legend <- c("norm")
denscomp(list(fln), legendtext = plot.legend,main = "Fitted Dist of education")

#Quantile Plots:
fit_norm = fitdist(education, "norm")
fit_logis = fitdist(education, "logis")
fit_lnorm = fitdist(education, "lnorm")


plot.legend = c("norm",  "logis", "lnorm")
qqcomp(list(fit_norm, fit_logis, fit_lnorm), legendtext = plot.legend,main = "Q-Q for education")
```

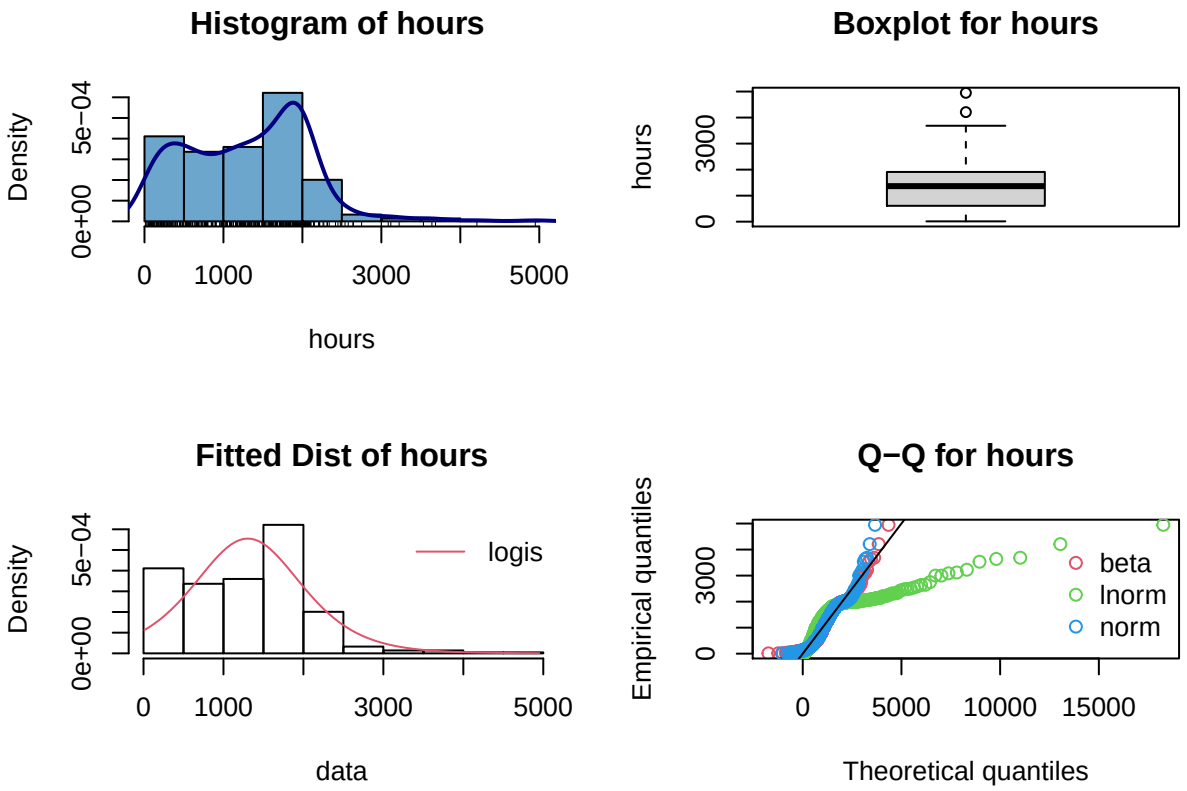


Figure 4: Descriptive Analysis for hours

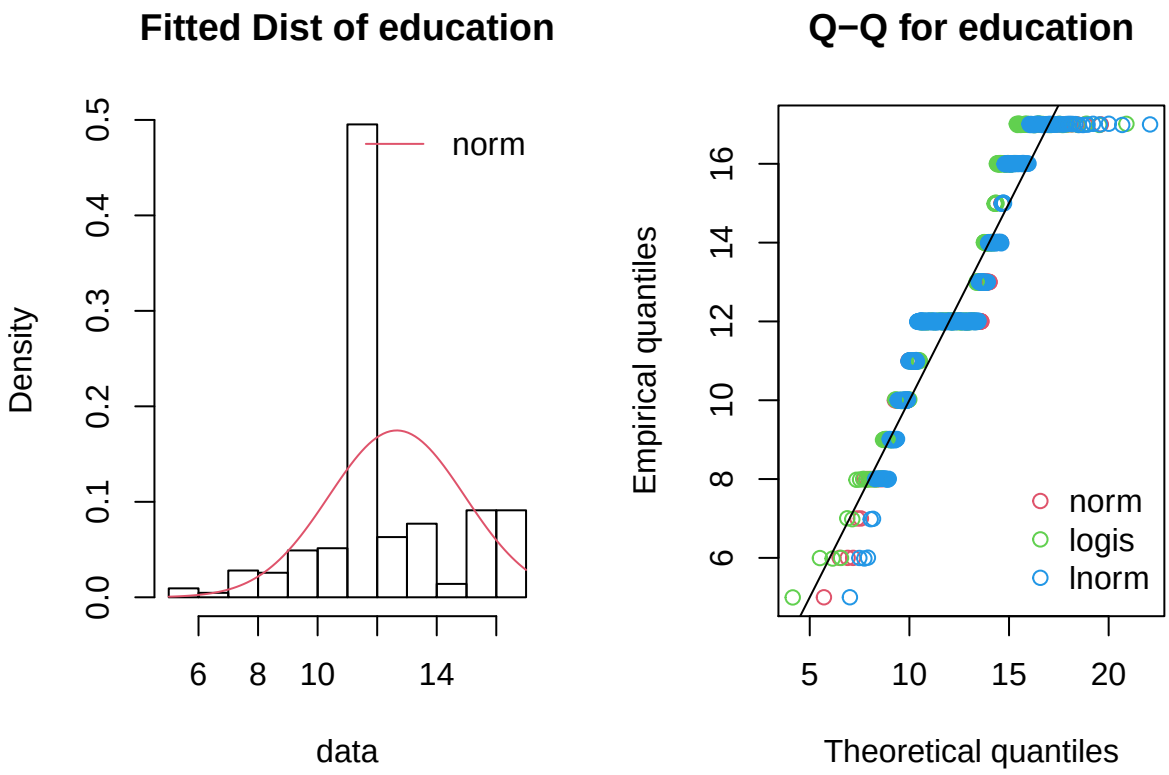


Figure 5: Descriptive Analysis for education

From fitted distribution and qqplot in Fig.5, it is seen that most of the observations are concentrated at 11-12 years of education.

- Hhours:

```
#hhours
par(mfrow=c(1,3))
hist(hhours, breaks = "FD",  col = "skyblue3", main = "Histogram of hhours", freq=FALSE)
lines(density(data$hhours),lwd = 2, col ="navyblue")
rug(hhours)

#Q-Q plot
#descdist(hhours, boot = 1000)
fit_norm = fitdist(hhours, "norm")
fit_logis = fitdist(hhours, "logis")
plot.legend = c("norm",  "logis")
qqcomp(list(fit_norm, fit_logis), legendtext = plot.legend,main = "Q-Q for hhours")

#Boxplot
Boxplot(~hhours, data=data, id=FALSE)
```

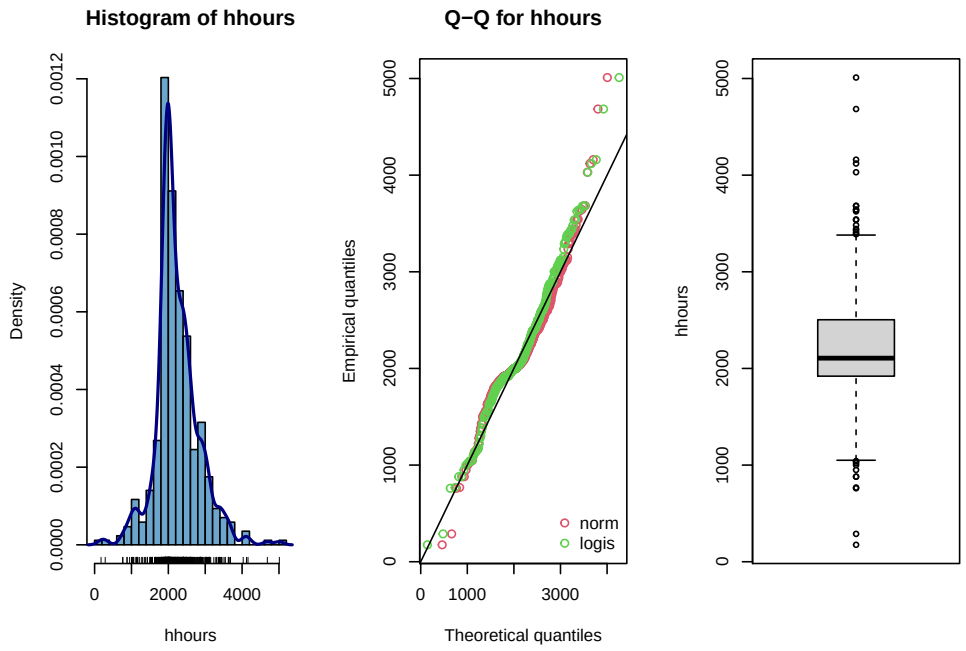


Figure 6: Descriptive Analysis for hhours

Figure 6 suggests that the variable hhours is positively skewed (right skewed) indicating the presence of some outliers, or influential observations on the upper tail. Apart from that, the histogram seems somewhat normalized with most of the values concentrated on the center.

- Hwage:

```r
#hwage
par(mfrow=c(1,3))
hist(hwage, breaks = "FD",  col = "skyblue3", main = "Histogram of hwage")
lines(density(data$hwage),lwd = 2, col ="navyblue")
rug(hwage)

#Density Functions using Cullen-Frey Graph
#descdist(hwage, boot = 1000)
fln <- fitdist(hwage, "lnorm")
plot.legend <- c("norm")
denscomp(list(fln), legendtext = plot.legend,main = "Fitted Dist of hwage")

fit_norm = fitdist(hwage, "norm")
fit_logis = fitdist(hwage, "logis")
fit_lnorm = fitdist(hwage, "lnorm")

plot.legend = c("norm",  "logis", "lnorm")
qqcomp(list(fit_norm, fit_logis, fit_lnorm), legendtext = plot.legend,main = "Q-Q for hwage")
```

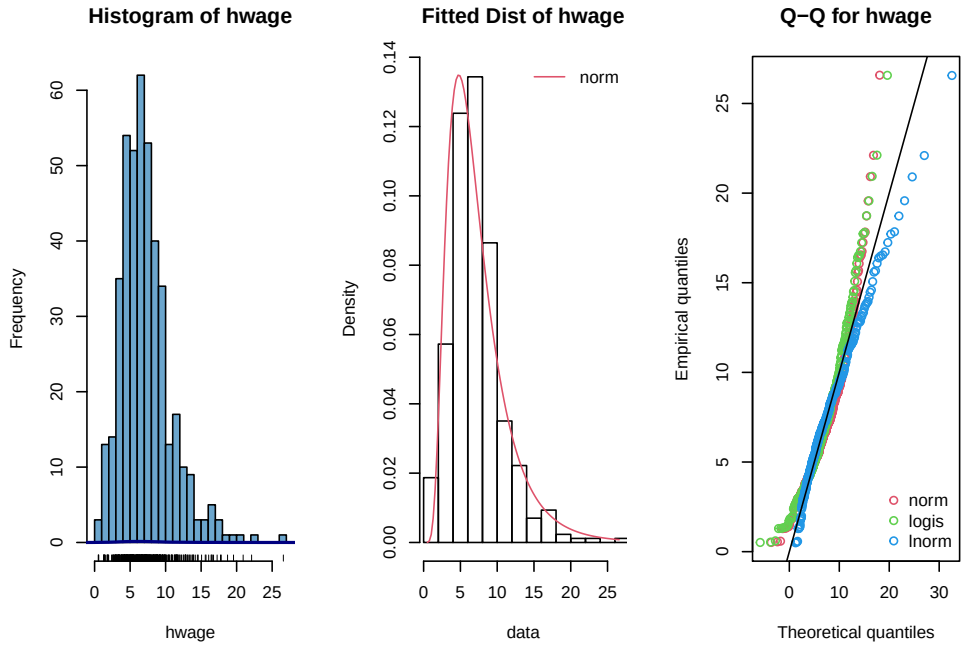


Figure 7: Descriptive Analysis for hwage

Figure 7 suggests that the variable hwage is positively skewed (right skewed) indicating the presence of some outliers, or influential observations on the upper tail. Apart from that, the histogram seems somewhat normalized with most of the values concentrated on the center.

- Correlation Plot:

```r
#Correlation Map
#Split data into two dataset,
#one contains category vairable only, one contains numeric data only
```

```
cat_var <- names(data)[which(sapply(data, is.factor))]
numeric_var <- names(data)[which(sapply(data, is.numeric))]
data_cat <- data[, cat_var]
data_cont <- data[,numeric_var]


corrplot(cor(data_cont), method="circle")
```

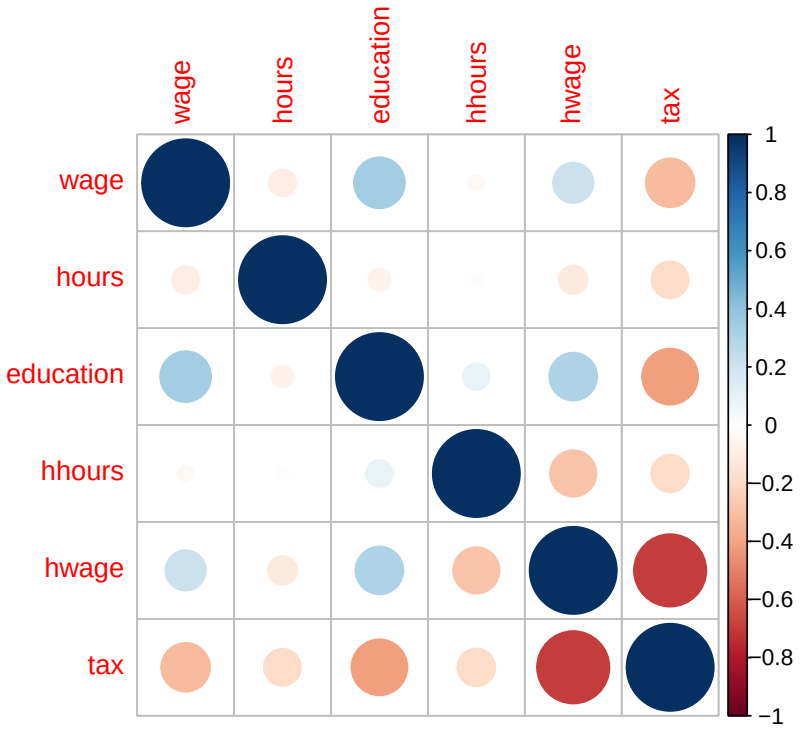


Figure 8: Correlation Plot

The variables hours and education seem to be important. Tax has a negative correlation with education, and with hwage. This could pose a problem of multi-collinearity.

b) Density Plots

```
library(car)
par(mfrow=c(3,2))
densityPlot(~ wage, data=data, xlab="wage", main="Density Plot for wage")
densityPlot(~ hours, data=data, xlab="hours", main="Density Plot for hours")
densityPlot(~ education, data=data, xlab="education", main="Density Plot for education")
densityPlot(~ hhours, data=data, xlab="hhours", main="Density Plot for hhours")
densityPlot(~ hwage, data=data, xlab="hwage", main="Density Plot for hwage")
densityPlot(~ tax   , data=data, xlab="tax   ", main="Density Plot for tax   ")
```

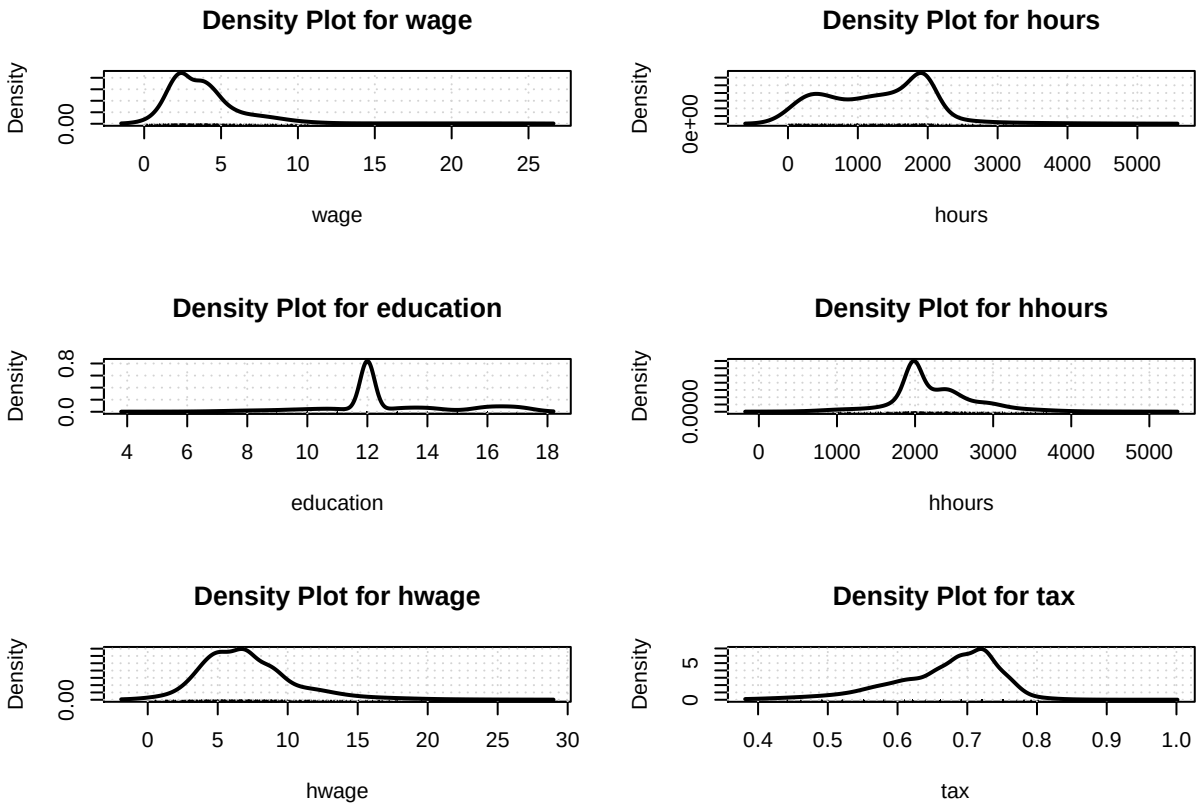


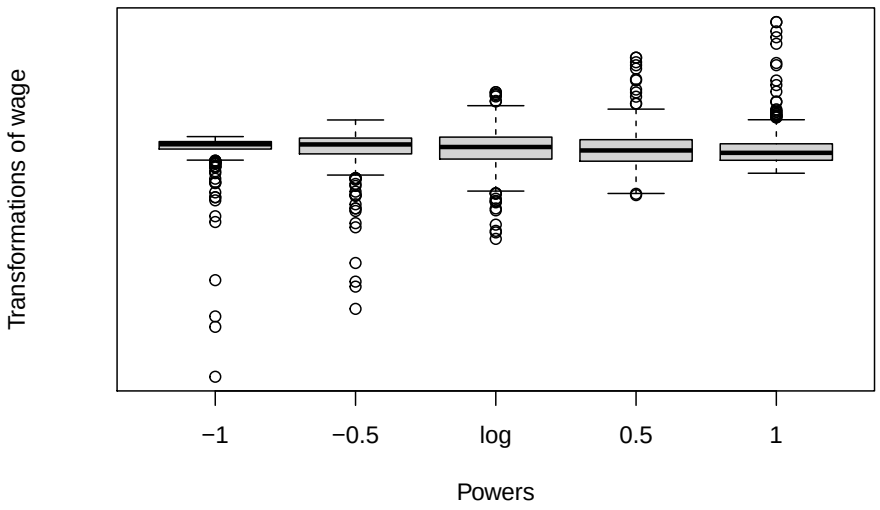
Figure 9: Density Plots

c) Transformations

Identify if there are any non-linearities within your variables. What transformations should you perform to make them linear? What would happen if you included non- linear variables in your regression models without transforming them first?

1. Dependent variable: wage

```
#Checking if there is a need for transforming the dependent variable
symbox(~wage, data=data)
```



The symbox seems to be favoring a log transformation for our dependent variable in Figure 11. Let's look at the histogram and see if transforming our dependent variable makes sense:

```
par(mfrow=c(1,2))
hist(data$wage,breaks = "FD", col = "skyblue3", main = "Histogram of wage", probability = TRUE)
lines(density(data$wage),col="red3", lwd=4)

hist(log(data$wage) ,breaks = "FD", col = "skyblue3", main = "Histogram of log wage",
probability = TRUE)
lines(density(log(data$wage)),col="red3", lwd=4)
```

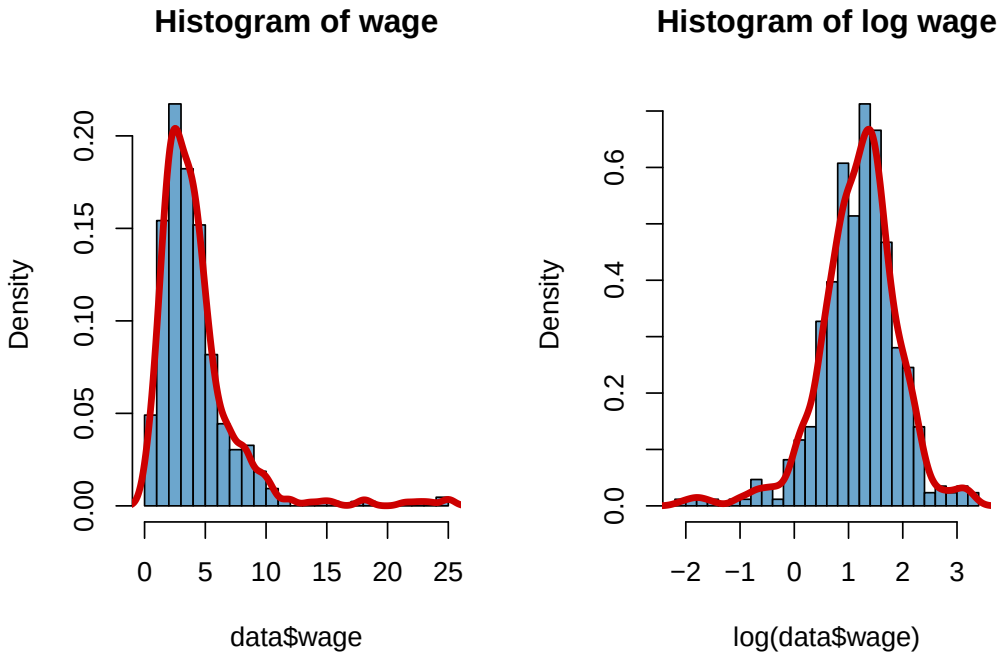
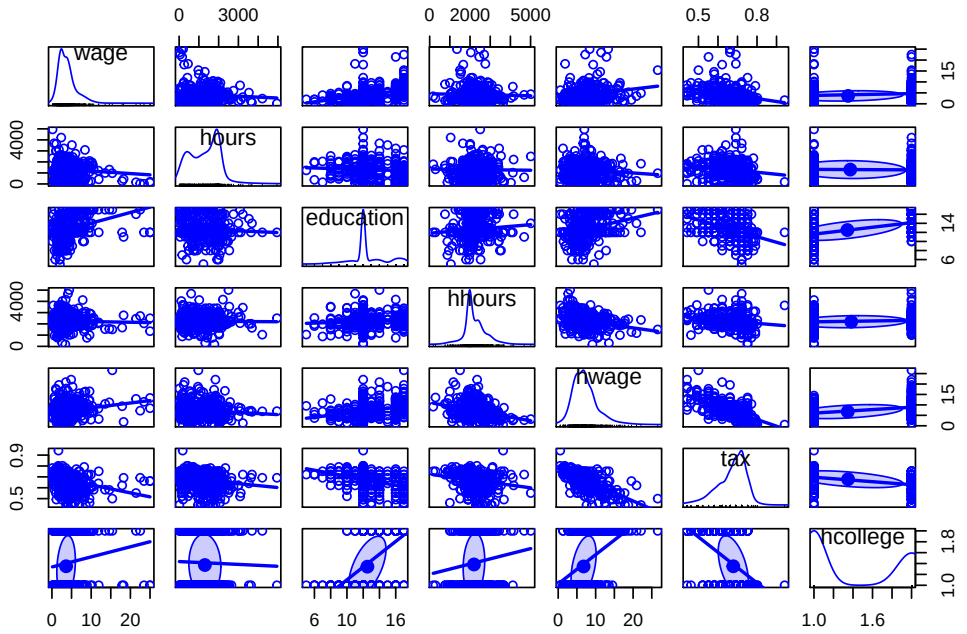
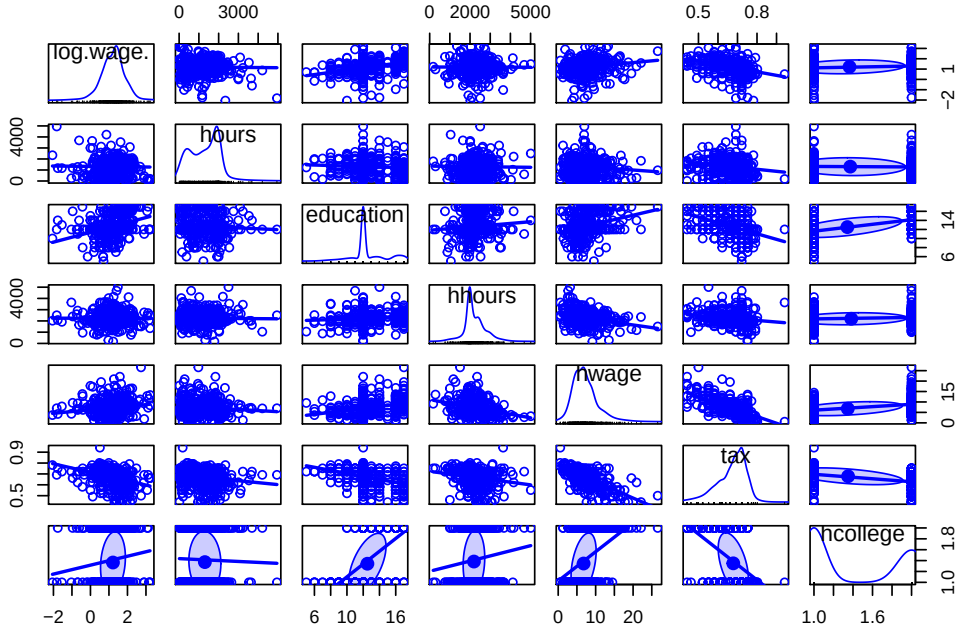


Figure 12 also suggests that Logarithm transformation is appropriate in this cases since our variable is skewed. The transformation spreads out the smaller values and compresses the larger ones, producing a more symmetric distribution.

Scatter Plots for comparing wage with transformation (log wage) vs wage without transformation:

```
#Wage
scatterplotMatrix(~ wage + hours + education
              +hhours + hwage+ tax+ hcollege ,smooth=FALSE, ellipse=list(levels=0.5))

#Log- wage
scatterplotMatrix(~ log(wage) + hours + education
              +hhours + hwage+ tax+ hcollege ,smooth=FALSE, ellipse=list(levels=0.5))
```





Scatterplots are also in favor of log transformation for our dependent variable.

Testing to see if a transformation is needed and if it is needed, transform it:

```
BoxCox = powerTransform(cbind(wage, hours, education, hwage,
tax, hhours)~ hcollege, data = data, family = "bcPower")
summary(BoxCox)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## wage         0.1998        0.20       0.1223       0.2773
## hours        0.6403        0.64       0.5409       0.7396
## education    1.0793        1.00       0.7493       1.4093
## hwage        0.4158        0.50       0.3254       0.5063
## tax          2.9275        2.93       2.5188       3.3362
## hhours       0.7016        0.70       0.5536       0.8497
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                     LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 733.536  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                      LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 572.9767  6 < 2.22e-16
```

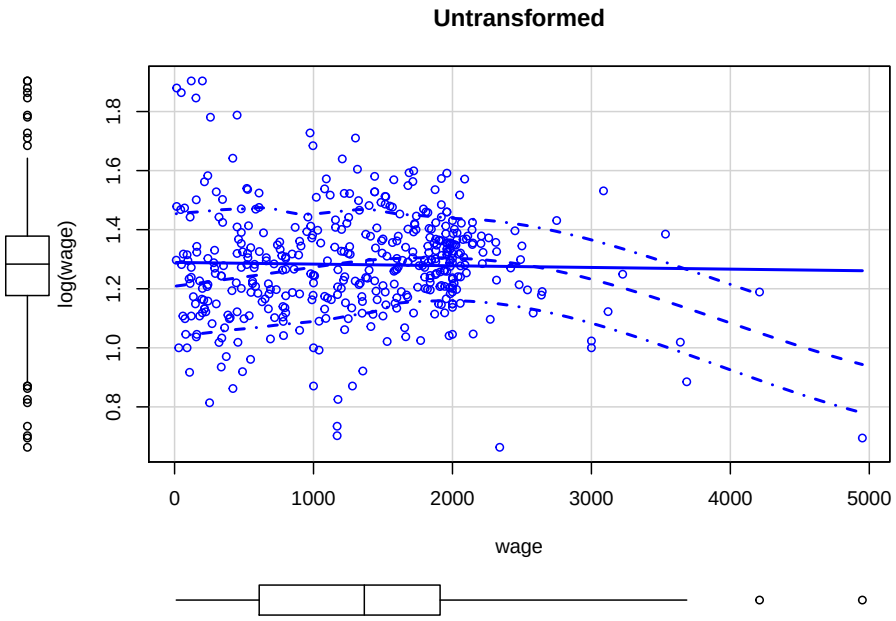
```
#Variables after transformation
Wage_B = (wage)^(0.20)
Hours_B = (hours)^(0.64)
Huswage_B = (hwage)^(0.5)
Tax_B = (tax)^(2.93)
HusHours_B = (hhours)^(0.70)
```

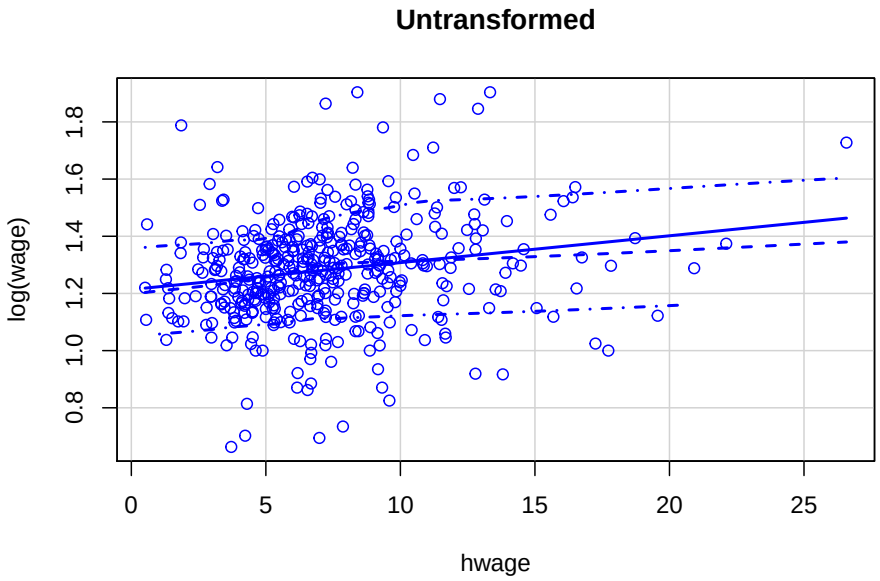
Based on the results from the powerTransform, we decided to transform the wage variable by raising it to the power of 0.20.

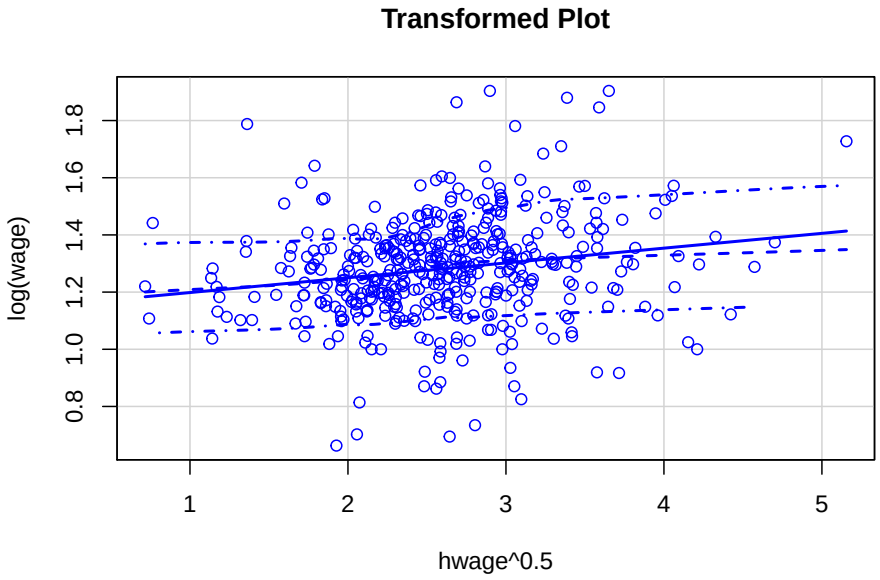
- Scatterplots for Untransformed and Transformed Variables

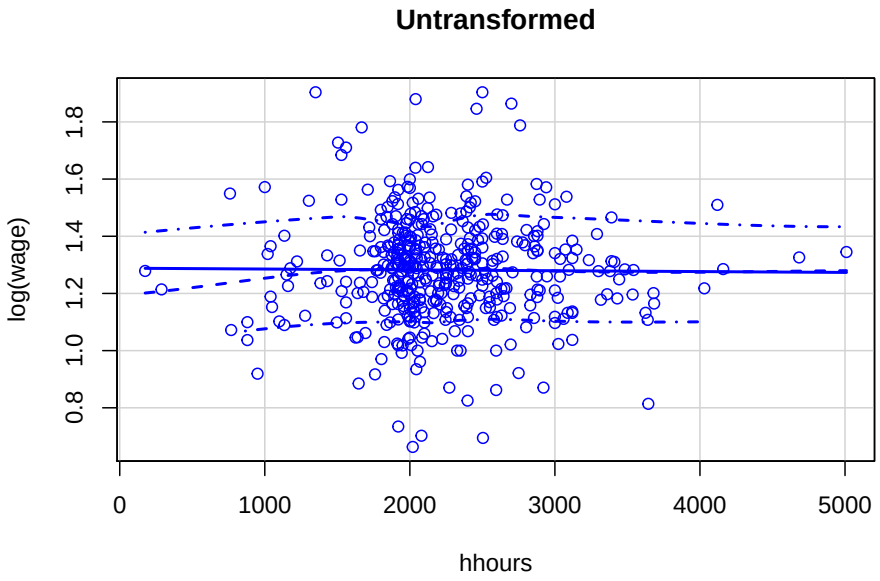
```
#hours
#Untransformed
scatterplot(Wage_B ~ hours, data=data, xlab="wage",ylab="log(wage)",
main="Untransformed")

#Transformed
scatterplot(Wage_B ~ Hours_B, data=data, xlab="hours", ylab="log(wage)",
main="Log-Log Plot",log="xy", boxplot=FALSE)

#hwage
#Untransformed
scatterplot(Wage_B ~ hwage, data=data, xlab="hwage",ylab="log(wage)",
main="Untransformed",boxplot=FALSE)

#Transformed
scatterplot(Wage_B ~ Huswage_B, data=data, xlab="hwage^0.5", ylab="log(wage)",
main="Transformed Plot", boxplot=FALSE)


#hhours
#Untransformed
scatterplot(Wage_B ~ hhours, data=data, xlab="hhours",ylab="log(wage)",
main="Untransformed", boxplot=FALSE)
```

```r
#Transformed
scatterplot(Wage_B ~ HusHours_B, data=data, xlab="hhours^0.70", ylab="log(wage)",
main="Transformed Plot", boxplot=FALSE)
```




Untransformed
log(wage)
0.8
1.0
1.2
1.4
1.6
1.8
0
5
10
15
20
25
hwage


Transformed Plot
log(wage)
0.8
1.0
1.2
1.4
1.6
1.8
1
2
3
4
5
hwage^0.5



From the given figures, it is evident that the transformations are effective in normalizing the distribution, and making it a better fit for the model.

Regression model including non-linear variables without transforming them first:

```
mod.1= lm(wage~ hours + education + hhours + hwage + tax + hcollege, data=data)
stargazer(mod.1, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
```

```
##                                 wage
## ------------------------------------------------
## hours                          -0.001***
##                                (0.0002)
##
## education                      0.424***
##                                 (0.078)
##
## hhours                         -0.001***
##                                (0.0003)
##
## hwage                          -0.236***
##                                 (0.078)
##
## tax                           -21.916***
##                                 (3.630)
##
## hcollegeyes                    -0.951***
##                                 (0.367)
##
## Constant                       19.440***
##                                 (3.835)
##
## ------------------------------------------------
## Observations                      428
## R2                               0.215
## Adjusted R2                      0.204
## Residual Std. Error       2.954 (df = 421)
## F Statistic           19.205*** (df = 6; 421)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The first model above is the model we all agreed on based on exploring all possible relationships within the choice of explanatory variables we had available. As it can be seen, all our explanatory variables are statistically significant at the 1% level. For hours, it is based on total hours worked in 1975, thus, for every incremental increase in hours worked, we expect the wives' average hourly wage to negatively decrease by 0.001, holding all else constant, which could entail substitution and income effects related to the wives' incentive to supply hours to the labor force. For education, for every year of education, we expect the average hourly wage of wives to increase by 42 cents, holding all else constant. For the wage of husbands, as their wages incrementally increase, we expect the wives' average hourly wage to decrease by 0.236 cents, holding all else constant. This makes economic sense because as the husband earns a higher wage, the wife can supply less hours to the labor market and not be required to get a higher paying job. For Tax, as wives earned a higher wage, thus, a higher income in most cases, the marginal tax rate negatively impacted the wage of wives by 21.916. Similar to total hours the wives worked, the total hours that the husband worked also negatively impacted the average hourly wage of their wives by 0.001, holding all else constant. If the husband attended college, we expect the average hourly wage of their wives to decrease by .95 cents. The rationale is since the husband attended college, they may have a higher paying job, thus the wife doesn't have to work as much, thus reducing their desire for a higher wage and supply less hours to the market. The constant in this model is deemed irrelevant, but serves as a model stabilizer. It should be noted that based on the variables selected, an adjusted R^2 of 0.204 was derived.

d) Removing Outliers

```
# Update after Box Cox
mod.2 = lm(Wage_B ~ Hours_B  + education + Huswage_B + Tax_B +
HusHours_B + hcollege, data = data)
stargazer(mod.2, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Wage_B
## -----------------------------------------------
## Hours_B                      -0.001***
##                              (0.0002)
##
## education                    0.022***
##                               (0.004)
##
## Huswage_B                    -0.144***
##                               (0.023)
##
## Tax_B                        -1.502***
##                               (0.156)
##
## HusHours_B                   -0.001***
##                              (0.0003)
##
## hcollegeyes                  -0.056***
##                               (0.019)
##
## Constant                     2.298***
##                               (0.175)
##
## -----------------------------------------------
## Observations                    428
## R2                             0.306
## Adjusted R2                    0.296
## Residual Std. Error      0.151 (df = 421)
## F Statistic           30.942*** (df = 6; 421)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
#qqplot
qqPlot(mod.2$residuals, id=list(n=3))
```

```
## [1] 408 185 349
```

```
#Bonferroni-corrected t-test:
outlierTest(mod.2)
```

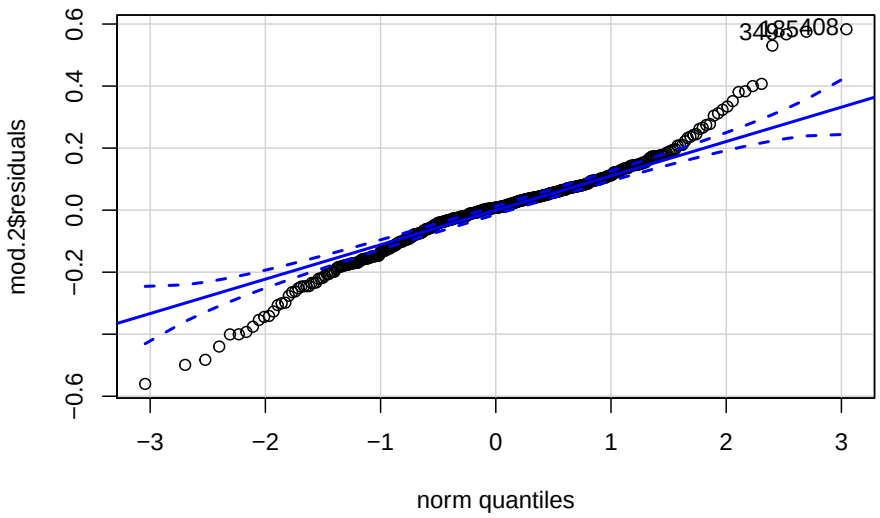


Figure 10: QQ Plot for outliers

```
##     rstudent unadjusted p-value Bonferroni p
## 408  3.97041         8.4397e-05     0.036122
## 185  3.92651         1.0071e-04     0.043105
```

```
#Residual vs fitted
plot(mod.2, 1)
```

```
#"bubble" plot of Studentized residuals
influencePlot(mod.2, id=list(n=3))
```

```
##        StudRes        Hat        CookD
## 126 -2.7049484 0.05630007 0.0614365963
## 157  2.9159585 0.15547062 0.2196982520
## 185  3.9265101 0.02096093 0.0455934171
## 211 -0.1645365 0.05197730 0.0002125329
## 349  3.8674983 0.02092382 0.0441999599
## 369  2.6121719 0.05012774 0.0507402048
## 408  3.9704101 0.01529885 0.0338031833
```

```
#Cooks Distance plot
influenceIndexPlot(mod.2, id=list(n=3),vars="Cook")
```

From the outliers test, the q-q plot, the residual plot and cook's distance plot, observations 126, 157, 185 and 408 seem to be the outliers, however, 157 seems to be the most influential outlier which could affect the accuracy of the predictive model. It has been removed from the model (mod.2), and the regression model is updated to mod.3:

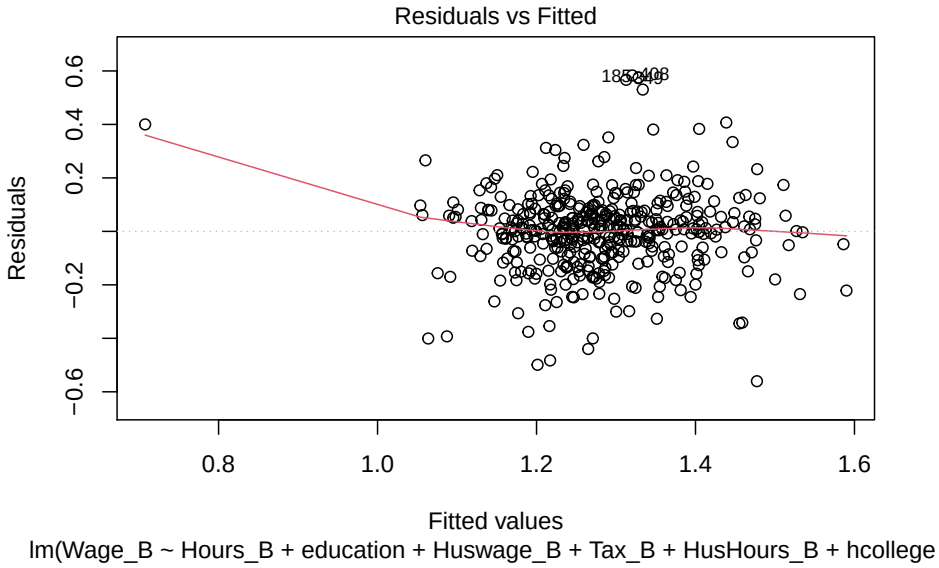


Figure 11: Residuals Plot

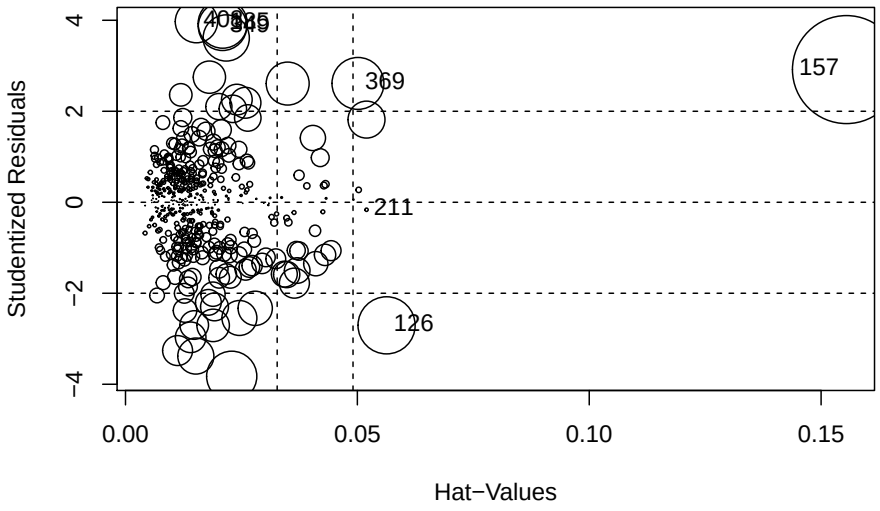


Figure 12: Bubble Plot for Studentized Residuals

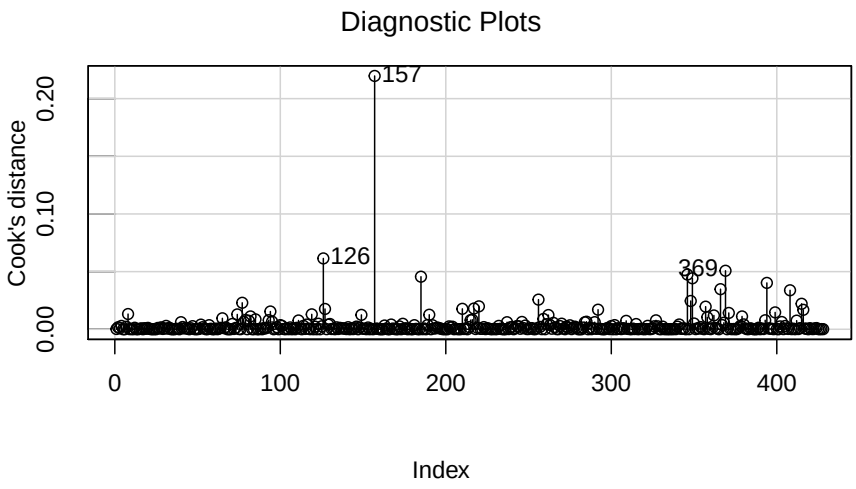


Figure 13: Cooks Distance Plot

```
mod.3<- update(mod.2, subset= -c(157))
#, 185, 408126,
stargazer(mod.3, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                Wage_B
## ------------------------------------------------
## Hours_B                       -0.001***
##                               (0.0002)
##
## education                     0.021***
##                                (0.004)
##
## Huswage_B                     -0.161***
##                                (0.023)
##
## Tax_B                         -1.672***
##                                (0.166)
##
## HusHours_B                    -0.002***
##                               (0.0003)
##
## hcollegeyes                   -0.056***
##                                (0.019)
##
## Constant                      2.463***
```

```
##                                (0.182)
##
## ------------------------------------------------
## Observations                     427
## R2                              0.318
## Adjusted R2                     0.309
## Residual Std. Error       0.149 (df = 420)
## F Statistic           32.683*** (df = 6; 420)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Comparing all three regression models:

```
stargazer(mod.1, mod.2, mod.3, column.labels = c("Base Model", "Transformed Model",
"No Outliers Model"), type = "text", float= FALSE)
```

```
##
## ================================================================================================
##                                              Dependent variable:
##                          ------------------------------------------------------------------------
##                                    wage                                  Wage_B
##                                Base Model           Transformed Model       No Outliers Model
##                                   (1)                    (2)                     (3)
## ------------------------------------------------------------------------------------------------
## hours                           -0.001***
##                                 (0.0002)
##
## Hours_B                                                -0.001***               -0.001***
##                                                        (0.0002)                (0.0002)
##
## education                        0.424***               0.022***                0.021***
##                                   (0.078)                (0.004)                 (0.004)
##
## hhours                          -0.001***
##                                 (0.0003)
##
## hwage                           -0.236***
##                                   (0.078)
##
## tax                             -21.916***
##                                   (3.630)
##
## Huswage_B                                              -0.144***               -0.161***
##                                                         (0.023)                 (0.023)
##
## Tax_B                                                  -1.502***               -1.672***
##                                                         (0.156)                 (0.166)
##
## HusHours_B                                             -0.001***               -0.002***
##                                                        (0.0003)                (0.0003)
##
## hcollegeyes                      -0.951***              -0.056***               -0.056***
##                                   (0.367)                (0.019)                 (0.019)
##
## Constant                         19.440***              2.298***                2.463***
##                                   (3.835)                (0.175)                 (0.182)
##
## ------------------------------------------------------------------------------------------------
## Observations                        428                    428                     427
## R2                                 0.215                  0.306                   0.318
## Adjusted R2                        0.204                  0.296                   0.309
## Residual Std. Error         2.954 (df = 421)       0.151 (df = 421)        0.149 (df = 420)
## F Statistic             19.205*** (df = 6; 421) 30.942*** (df = 6; 421) 32.683*** (df = 6; 420)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

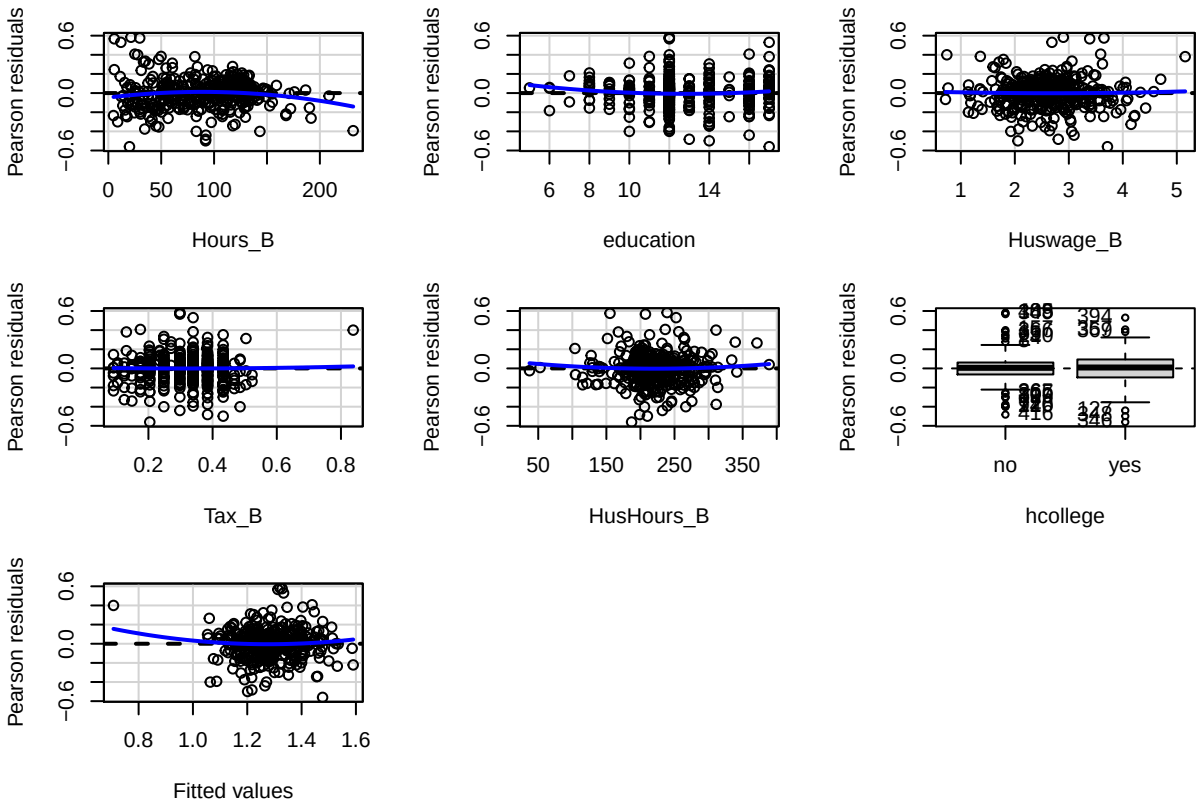
Removing the outliers yield different coefficient results for some of the variables. The significance levels for the both models stayed the same. The R^2 is highest, and AIC and BIC are the lowest in the mod.3.

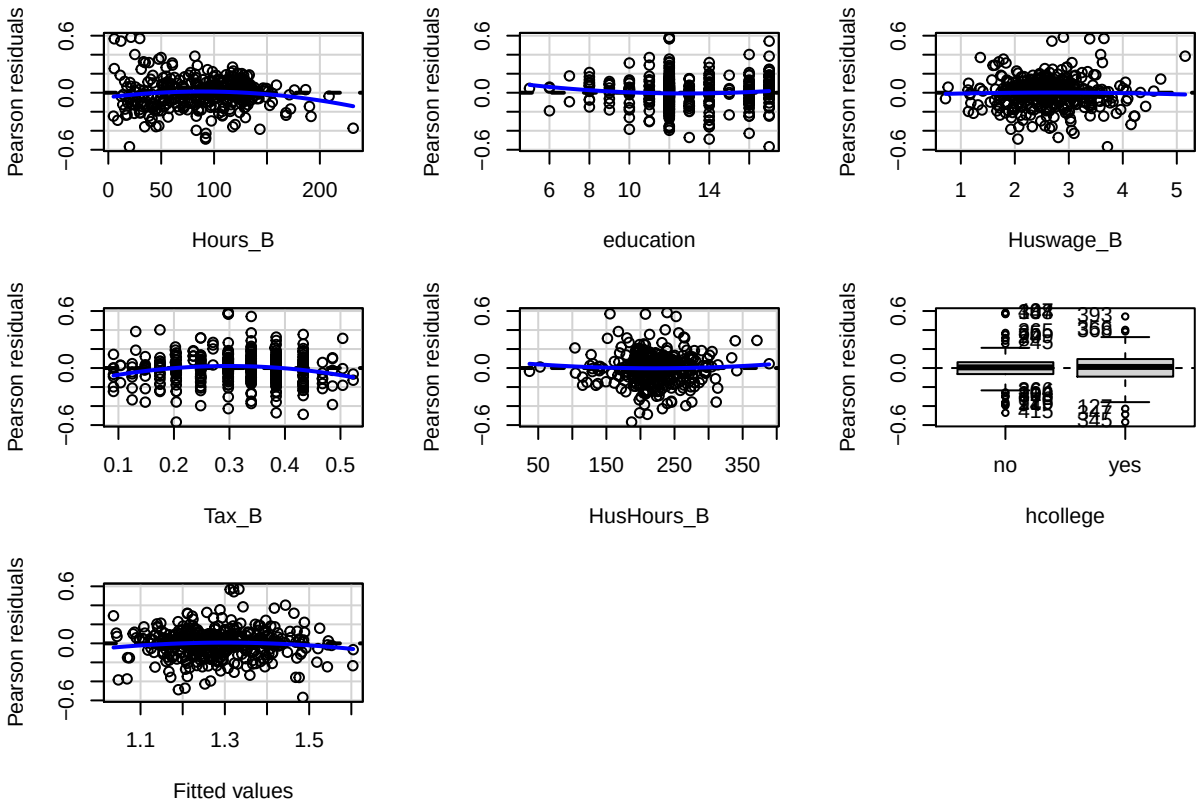
```
#Comparing residuals graphs to understand if taking outliers yielded to an improvement
###Pearson Residuals
residualPlots(mod.2)
```

```
##            Test stat Pr(>|Test stat|)
## Hours_B      -2.1039          0.03598 *
## education     1.6659          0.09648 .
## Huswage_B     0.3105          0.75635
## Tax_B         0.1873          0.85153
## HusHours_B    0.8337          0.40494
## hcollege
## Tukey test    1.4112          0.15818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residualPlots(mod.3)
```

```
##            Test stat Pr(>|Test stat|)
## Hours_B      -2.1422        0.0327493 *
## education     1.6482        0.1000613
## Huswage_B    -0.3253        0.7451206
## Tax_B        -3.4812        0.0005515 ***
## HusHours_B    0.6944        0.4878243
## hcollege
## Tukey test   -1.4577        0.1449108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Pearson residuals
Hours_B
education
Huswage_B
Tax_B
HusHours_B
hcollege
no
yes
Fitted values



Pearson Residuals: The p-value for Tukey Test is more than 5%. So we fail to reject the null that there is some pattern in the residuals. . Looking at the residuals plots, the mod.3 seems to have the best residual plot.

e) Checking for NAs

```
#Looking at the data to see if we have any missing variables
missmap(data[, -c(1, 11)], col = c("black", "red"), y.cex = 0.5, x.cex = 0.5)
```

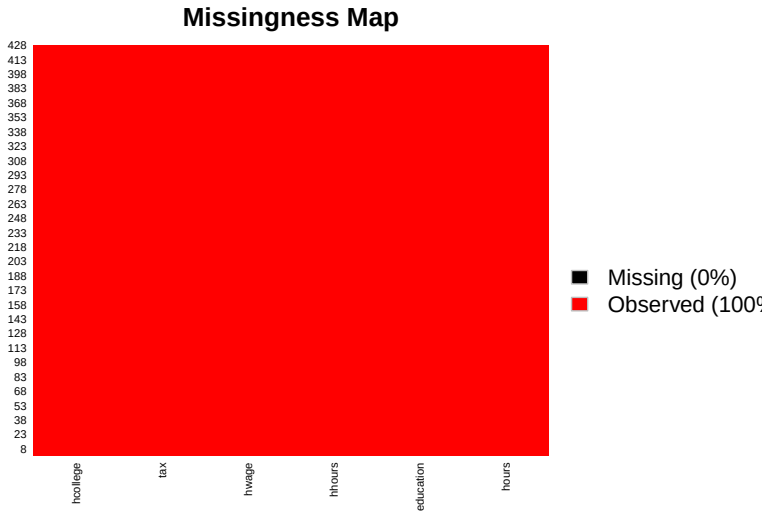


Figure 14: Missingness Map

Since there are no missing values in the variables selected, there is no need to omit any of them. The analysis can be continued.

3. Model Building

- Evaluate transformations of variables:

In part 2(c) boxcox transformation was used to transform the variables: tax, hours, hwage and hhours. The model (mod.1) was revised using new variables to create mod.2. From the results in part 2, it is seen that after the transformation, the R^2 of the model increases and AIC and BIC decrease.

- Test for multicollinearity

```
vif(mod.2)
```

```
##    Hours_B  education  Huswage_B      Tax_B HusHours_B   hcollege
##   1.421003   1.554675   4.179081   4.359533   1.975218   1.606168
```

```
vif(mod.3)
```

```
##    Hours_B  education  Huswage_B      Tax_B HusHours_B   hcollege
##   1.433067   1.572891   4.335895   4.648755   2.117543   1.603516
```

The VIF test was used to test both models for Multicollinearity. The models did not produce any VIF values greater than five, thus, the models do not have a Multicollinearity problem.

- Test for heteroskedasticity

```
#Plotting Residuals & Testing for Heteroskedasticity (Passed for mod.3)
plot(mod.3,1)
abline(h=0, col="red", lwd=2)
ncvTest(mod.3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.8796, Df = 1, p = 0.17038
```

```
bptest(mod.3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod.3
## BP = 34.858, df = 6, p-value = 4.591e-06
```

The next logical step is to test the model if Heteroskedasticity exists. The model passes the Non-constant Variance Test, but fails the Breusch-Pagan Test. However, since the model passed the Non-Constant Variance Test, we can continue with using Model 2 (mod.2) since the BP test is more restrictive than the NCV test. We acknowledge that the BPtest is the desired test to observe if Heteroskedasticity exists, but after careful consideration, we are going to proceed with model 2 (mod.2) and based on the results that will be observed below, Model 2 is the better model.

- Test for model mis-specification

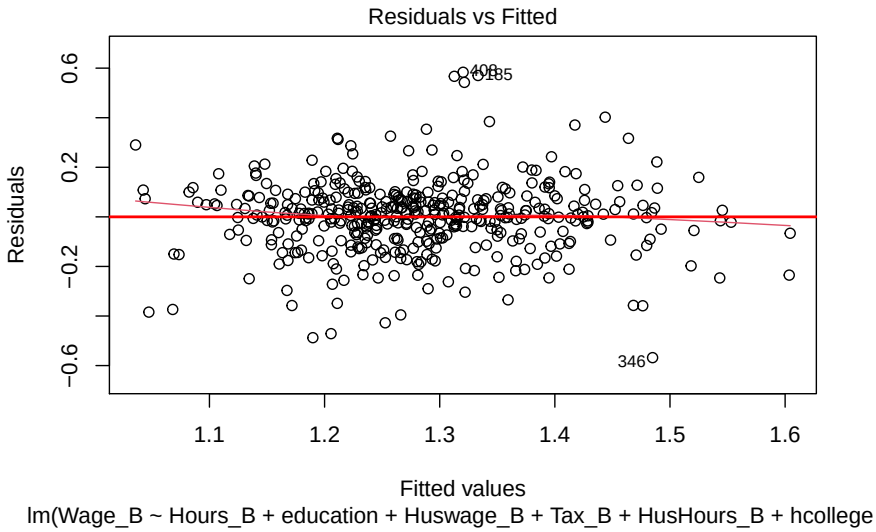


Figure 15: Residual Plots

```
#Test For Model Misspecification (Model is Fine)

resettest(mod.2, power= 2, type = "regressor") #Passed
```

```
##
##  RESET test
##
## data:  mod.2
## RESET = 1.5182, df1 = 5, df2 = 416, p-value = 0.1829
```

```
resettest(mod.3, power= 2, type = "regressor") #Failed
```

```
##
##  RESET test
##
## data:  mod.3
## RESET = 5.1571, df1 = 5, df2 = 415, p-value = 0.0001329
```

The RAMSEY RESET test was performed on mod.2 and mod.3. In mod.2, with a p-value (0.1829) greater than 5%, we failed to reject the null hypothesis and conclude that the model is not misspecified. Thus, no further improvement is required in the model. However, the RESET test fails in mod.3, with a p-vale close to 0.

- Look at Cook's distance Plot, Residuals Plot

Both the residuals plot and cook's distance plot have been plotted in part 2 (d)

We acknowledge that observation 157 is an outlier. However, its removal in model 3 causes the model to fail the Ramsey RESET Test. So we do not remove it.

- Use AIC and BIC for model selection

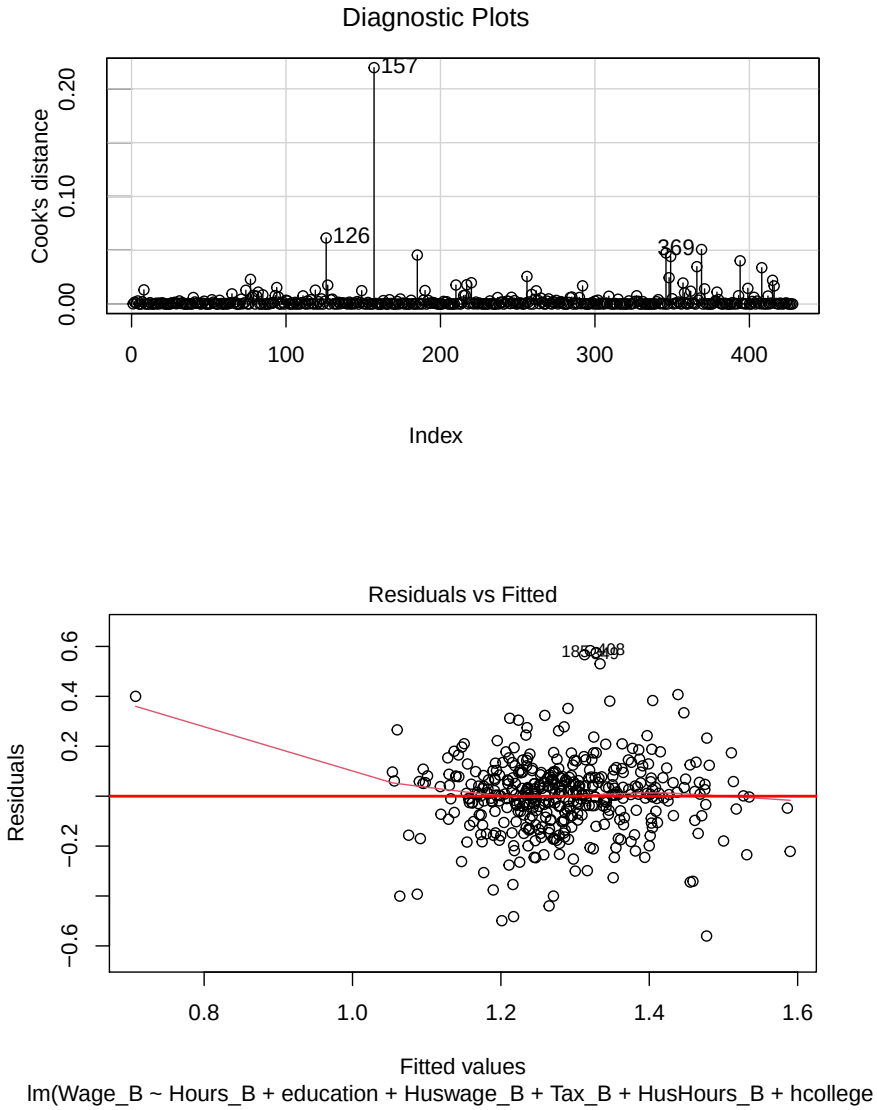


Figure 16: Outlier plots

```
AIC(mod.1,mod.2,mod.3)
```

```
##       df       AIC
## mod.1  8 2150.7246
## mod.2  8 -397.0534
## mod.3  8 -403.6477
```

```
BIC(mod.1,mod.2,mod.3)
```

```
##       df       BIC
## mod.1  8 2183.1975
## mod.2  8 -364.5804
## mod.3  8 -371.1934
```

Based on the results from both the AIC and BIC, Model 2 (mod.2) is the better model and will be confirmed as our finalized model for the remaining last steps of testing the robustness of the designated model.

- Evaluate the robustness of your coefficient estimates by bootstrapping your model. Provide a histogram of the bootstrapped estimates, and comment on the findings.

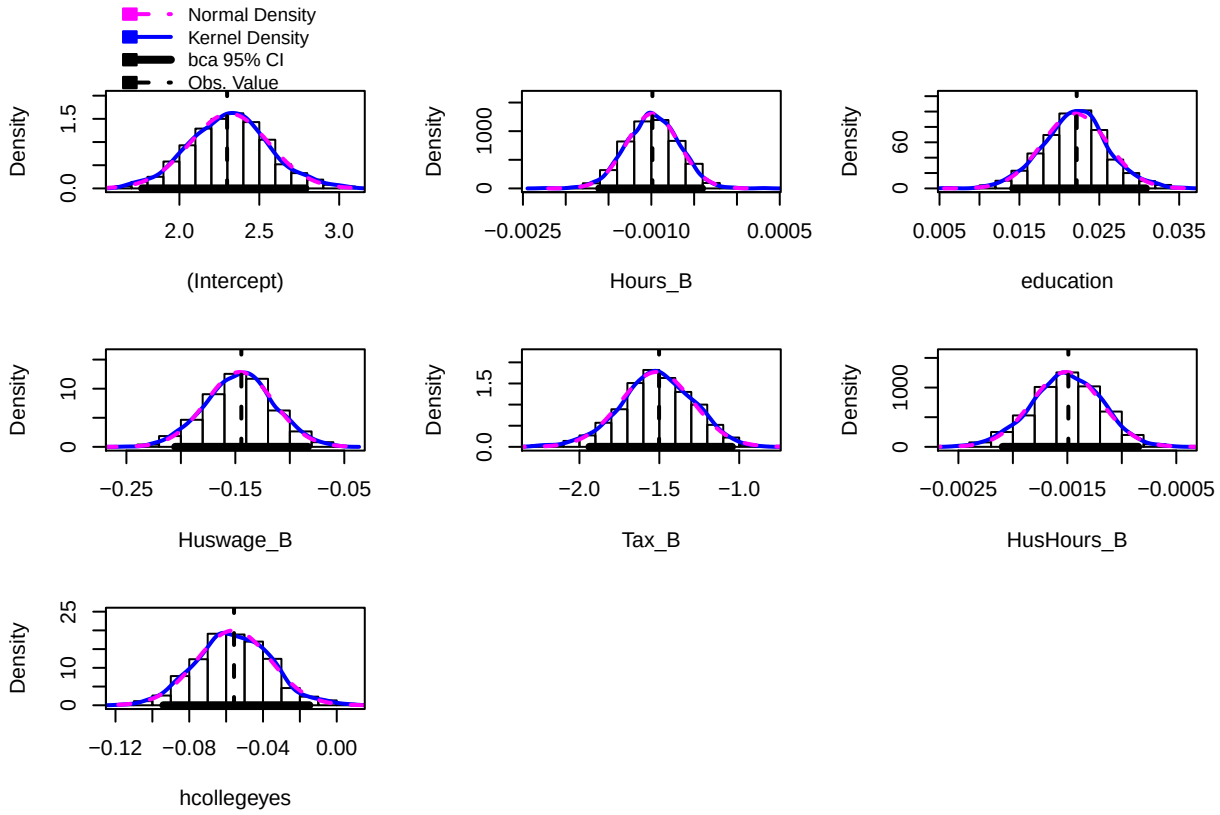
```
# Boostrapping the Model
set.seed(2425)
Betahat.Model = Boot(mod.2, R = 1000)
summary(Betahat.Model)
```

```
##
## Number of bootstrap replications R = 1000
##                 original    bootBias     bootSE     bootMed
## (Intercept)  2.29830220  1.4906e-02 0.24719527  2.31013663
## Hours_B     -0.00098781  3.5405e-06 0.00030516 -0.00099036
## education    0.02216864 -2.0613e-04 0.00409698  0.02200243
## Huswage_B   -0.14449577 -1.4269e-03 0.03100522 -0.14489900
## Tax_B       -1.50164671 -1.4587e-02 0.22518953 -1.51448713
## HusHours_B  -0.00149091 -1.9414e-05 0.00031608 -0.00150977
## hcollegeyes -0.05571652 -2.4616e-04 0.01996260 -0.05622078
```

```
confint(Betahat.Model)
```

```
## Bootstrap bca confidence intervals
##
##                    2.5 %        97.5 %
## (Intercept)  1.766658363  2.7888847539
## Hours_B     -0.001608809 -0.0004073239
## education    0.014213619  0.0308294639
## Huswage_B   -0.205214844 -0.0833690247
## Tax_B       -1.937489009 -1.0456521547
## HusHours_B  -0.002093303 -0.0008496827
## hcollegeyes -0.094027099 -0.0148531608
```

```
hist(Betahat.Model)
```



After Bootstrapping the selected model with 1000 replications, histograms for each explanatory variable were derived. The explanatory variables exhibit a normal distribution with some explanatory variables having better fits than others.

In particular, hours worked by husbands appears to be shaped like a cone at the top. If the husband attended college also appears to be slightly off of truly being normally distributed. As the number of replications increase, we expect every explanatory variable to be close to be or be normally distributed based on the Strong Law of Large Numbers.

- Use cross-validation to evaluate your model performance

```
#Cross Validation Using 5 Fold
#Data Prep for Five Fold & Cross Validation
BC = data.frame(
  Wage_B = (wage)^(0.20),
  Hours_B = (hours)^(0.64),
  Huswage_B = (hwage)^(0.5),
  Tax_B = (tax)^(2.93),
  HusHours_B = (hhours)^(0.70),
  HCollege = data$hcollege)


#Model 2
set.seed(1111)
```

```
train.control = trainControl(method = "cv", number = 5)

T_Model = train(Wage_B ~ + Hours_B + Huswage_B + HusHours_B + Tax_B
+ HCollege, data = BC, method = "lm", trControl = train.control)

print(T_Model)
```

```
## Linear Regression
##
## 428 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 343, 343, 340, 343, 343
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1570496  0.2468694  0.1127037
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Since we agreed to use Model 2 (mod.2) for our finalized model, we performed a 5-Fold Cross Validation to observe how our model performed and based on the output above, the model derived a RMSE of 0.1570 and a R^2 of 0.2469.

• Evaluate your model's out of sample performance by splitting the data into testing and training sets, and predicting on the testing set

```
# Cross Validation using Training & Testing Sets#
set.seed(1010)

training.samples = wage%>%
  createDataPartition( p = 0.8, list = FALSE)
train.data = BC[training.samples, ]
test.data = BC[-training.samples, ]

Cross_Model = lm(Wage_B ~ Hours_B + Huswage_B + HusHours_B + Tax_B
+ HCollege, data = BC)


predictions = Cross_Model %>% predict(test.data)

data.frame(
  RMSE = RMSE(predictions, test.data$Wage),
  R2 = R2(predictions, test.data$Wage)
)
```

```
##        RMSE        R2
## 1 0.1444621 0.3276932
```

We separated Model 2 (mod.2) into a testing and training data set and performed a cross validation on the testing data set and a RMSE of 0.1445 and a R^2 of 0.3280 was derived. Overall, we are pleased with the results.

We also wanted to observe how the model perfomed when the data was split evenly :

```
# Cross Validation using Training & Testing Sets#
set.seed(1010)

training.samples = wage%>%
  createDataPartition( p = 0.5, list = FALSE)
train.data = BC[training.samples, ]
test.data = BC[-training.samples, ]

Cross_Model = lm(Wage_B ~ Hours_B + Huswage_B + HusHours_B + Tax_B
+ HCollege, data = BC)


predictions = Cross_Model %>% predict(test.data)

data.frame(
  RMSE = RMSE(predictions, test.data$Wage),
  R2 = R2(predictions, test.data$Wage)
)
```

```
##        RMSE        R2
## 1 0.1573751 0.2586944
```

Upon evenly splitting Model 2 (mod.2) into a testing and training data set and performing a cross validation on the testing data set, we observed a RMSE of 0.15737 and a R^2 of 0.2586 was derived.

- Note: Make sure to also discuss any relevant marginal effects estimated

In conclusion, we elected to go with Model 2 (mod.2) due to the tests that the model passed that were performed above. Our final interpretation for Model 2, goes as follows: For hours, it is based on total hours worked in 1975, thus, for every incremental increase in hours worked, we expect the wives' average hourly wage to negatively decrease by 0.001, holding all else constant, which again could entail substitution and income effects related to the wives' incentive to supply hours to the labor supply. For education, for every year of education, we expect the average hourly wage of wives to increase by .02 cents, holding all else constant. For the wage of husbands, as their wages incrementally increase, we expect the wives' average hourly wage to decrease by 0.14 cents, holding all else constant. This makes economic sense because as the husband earns a higher wage, the wife can supply less hours to the labor market and not be required to get a higher paying job. For Tax, as wives earned a higher wage, thus, a higher income in most cases, the marginal tax rate negatively impacted the wage of wives by decreasing their wage by 1.50, measured in dollars. Similar to total hours the wives worked, the total hours that the husband worked also negatively impacted the average hourly wage of their wives by 0.001, holding all else constant. If the husband attended college, we expect the average hourly wage of their wives to decrease by .05 cents. The rationale is since the husband attended college, they may have a higher paying job, thus the wife doesn't have to work as much, thus reducing their desire for a higher wage and supply less hours to the market. The constant in this model is deemed irrelevant, but serves as a model stabilizer. It should be noted that based on the variables selected, an adjusted R^2 of 0.296 was derived. This model appears to be more aligned with true economic effects because the tax rate faced by the wives in the first model was in magnitude larger, but after the transformation, the interpretation makes more logical sense. Our finalized model improved based on an adjusted R^2 comparison from 0.204 to 0.296.

All the marginal effects the explanatory variables derived, all provided a statistical support for topics in labor economic theory when it comes to predicting and analyzing effects on wage determinants. Again, we acknowledge that the data is over 40 years old and it is plausible that the marginal effects that were derived in this analysis may not hold in today's society, but it serves an important reminder that income dynamics do exist and the topic is still evolving till this day.