Economics 403A: Homework 3

Pujan Thakrar

November 18, 2018

Question 1

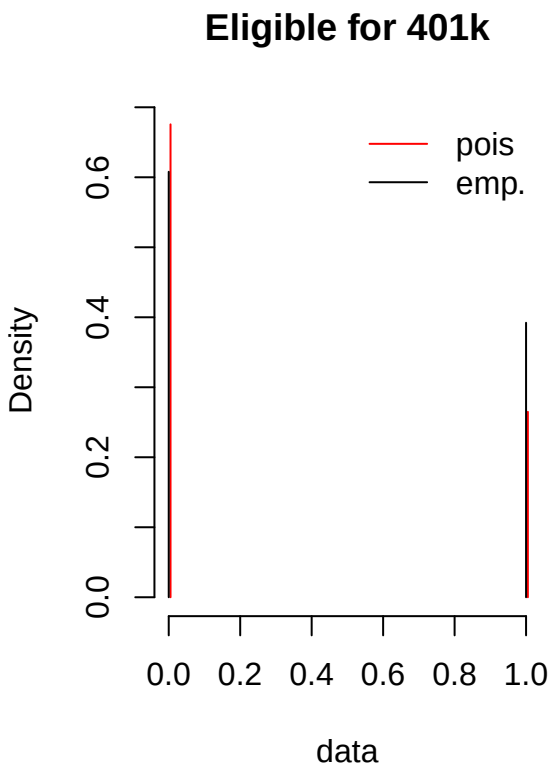
The dataset 401KSUBS contains information on net financial wealth (nettfa), age of the survey (age), annual family income (inc), family size (fsize), and a binary variable for eligibility in a 401(k) plan (e401k) among other variables. The wealth and income variables are both recorded in thousands of dollars. Our response variable for this problem is nettfa. Note: The complete the dataset includes 10 predictors (please refer to the file description for details).

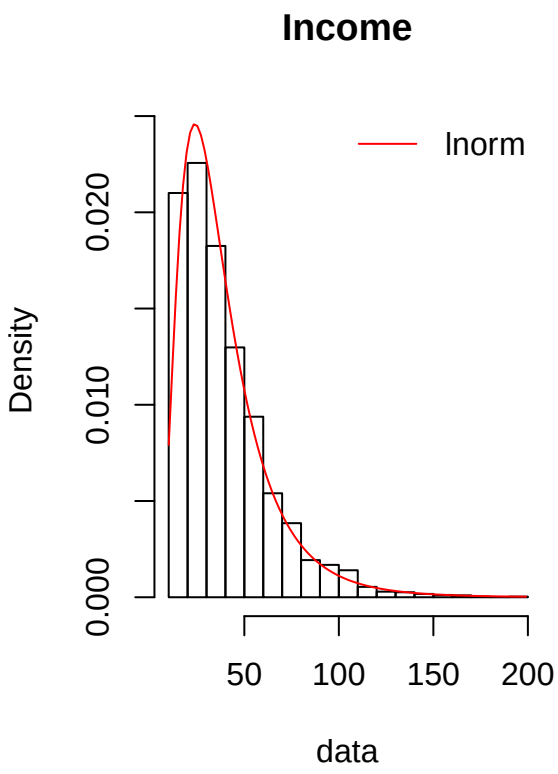
(a) Provide a descriptive analysis of your variables. This should include, histograms and fitted distributions, quantile plots, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

Histograms and Fitted Distributions:

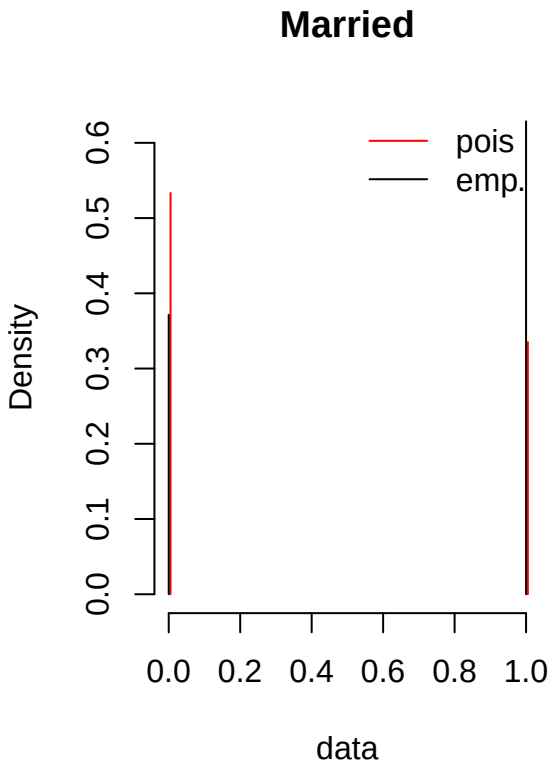
After using a Cullen-Frey graph to determine best distributions, I made the following histograms

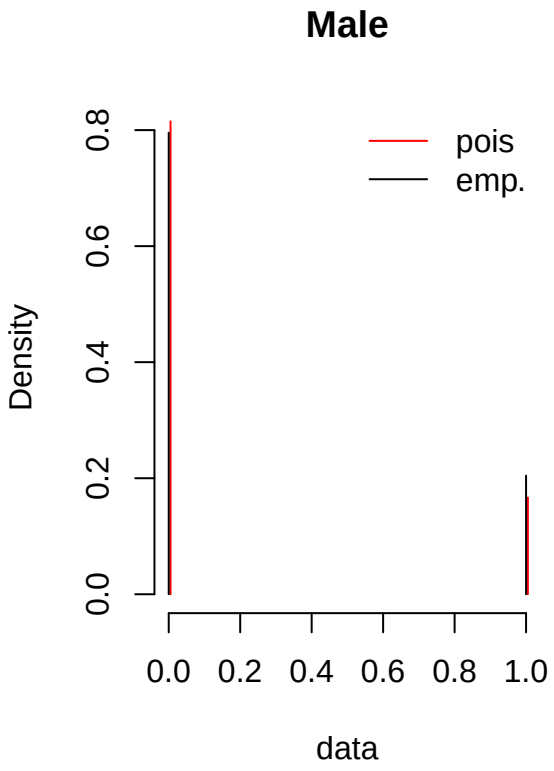
```
attach(data)
n= nrow(data)
k=1+log2(n)                          #determine number of bins to include

par(mfrow=c(1,2))
fn <- fitdist(as.numeric(e401k), "pois")
plot.legend <- c("pois")
denscomp(list(fn), legendtext = plot.legend,main = "Eligible for 401k")

fn <- fitdist(as.numeric(inc), "lnorm")
plot.legend <- c("lnorm")
denscomp(list(fn), legendtext = plot.legend ,main ="Income")
```



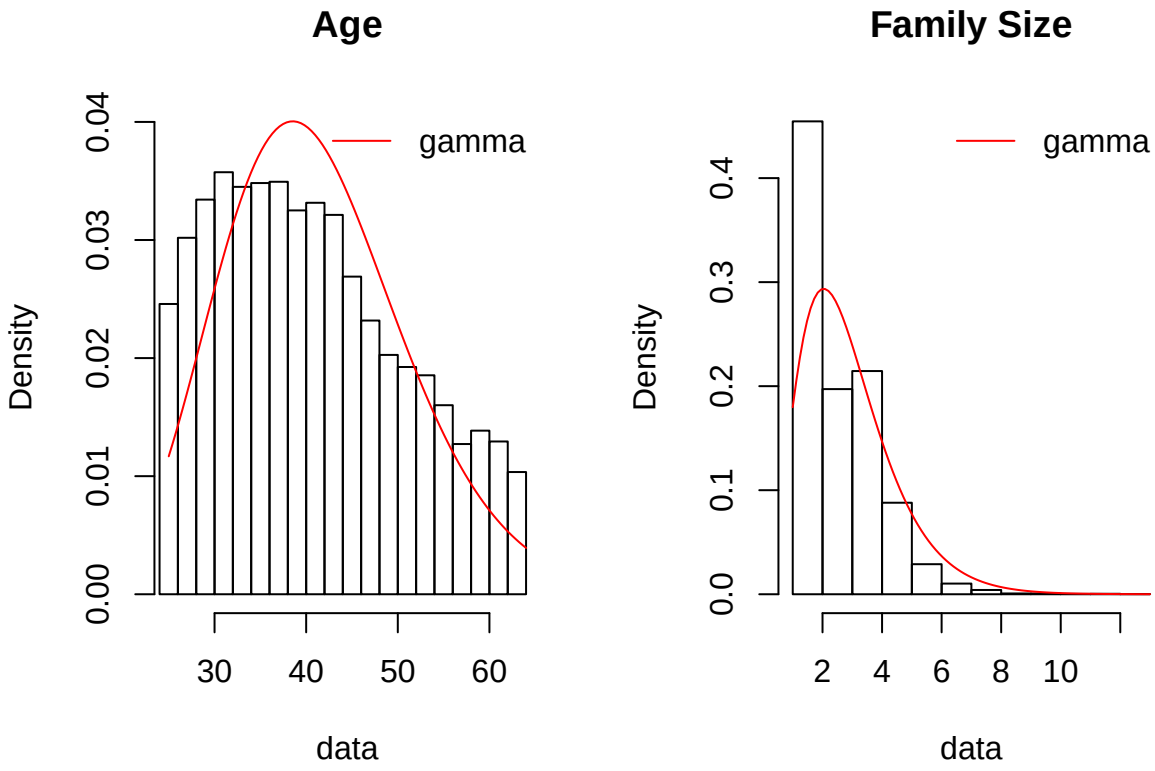



```
fn <- fitdist(as.numeric(marr), "pois")
plot.legend <- c("pois")
denscomp(list(fn), legendtext = plot.legend, main = "Married")

fn <- fitdist(as.numeric(male), "pois")
plot.legend <- c("pois")
denscomp(list(fn), legendtext = plot.legend, main="Male")
```

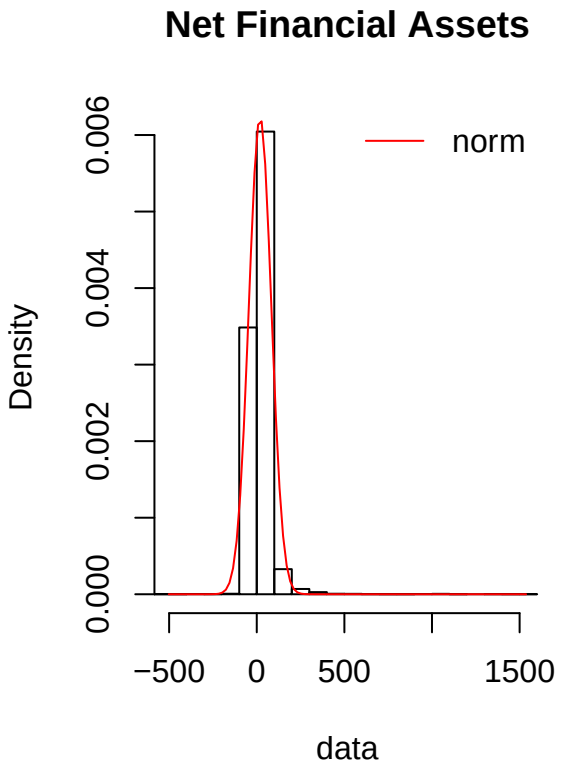


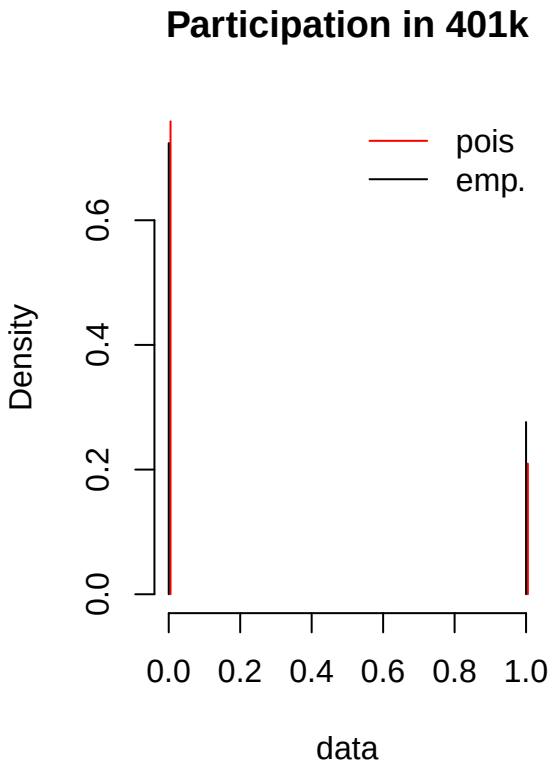



```
fm <- fitdist(as.numeric(age), "gamma")
plot.legend <- c("gamma")
denscomp(list(fm), legendtext = plot.legend,main="Age")

fl <- fitdist(as.numeric(fsize), "gamma")
plot.legend <- c("gamma")
denscomp(list(fl), legendtext = plot.legend, main="Family Size")
```

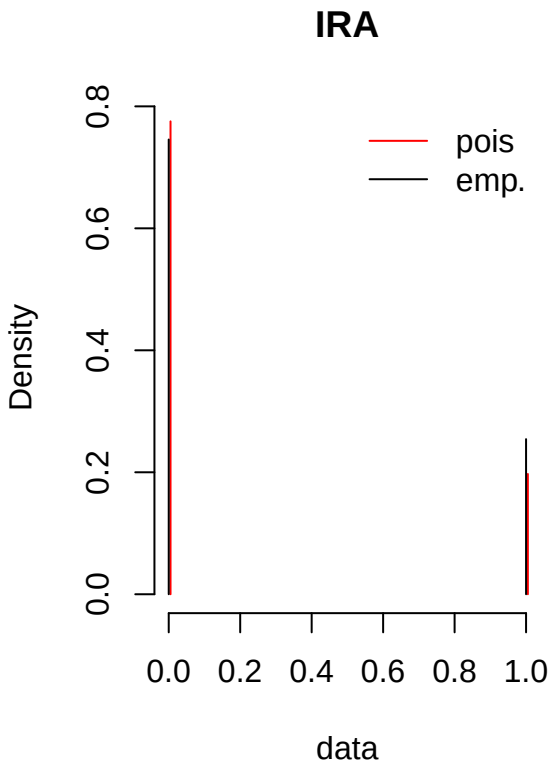


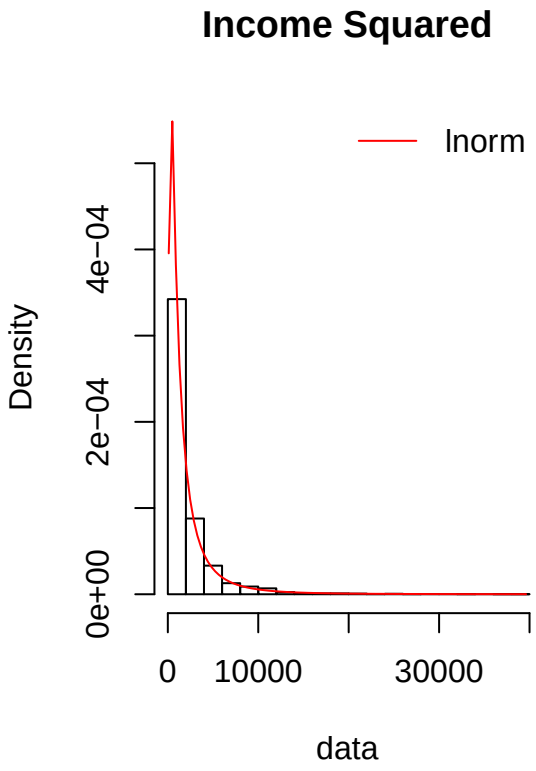

```
fn <- fitdist(as.numeric(nettfa), "norm")
plot.legend <- c("norm")
denscomp(list(fn), legendtext = plot.legend, main="Net Financial Assets")

fn <- fitdist(as.numeric(p401k), "pois")
plot.legend <- c("pois")
denscomp(list(fn), legendtext = plot.legend, main="Participation in 401k")
```



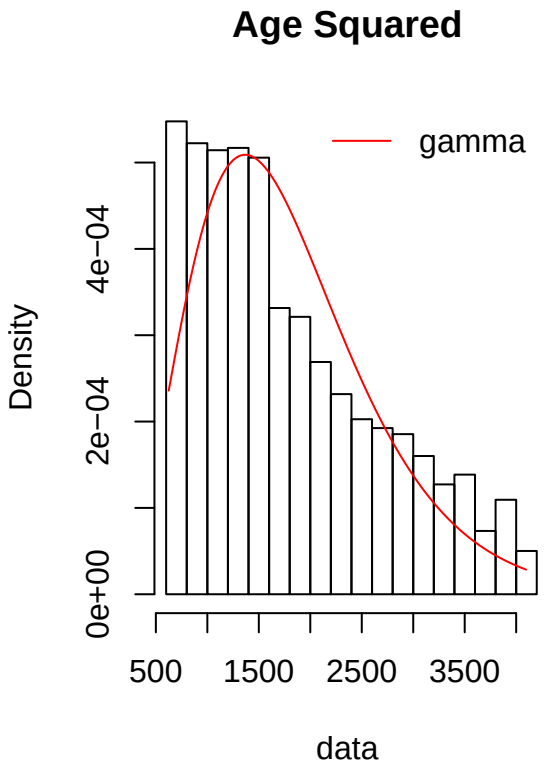



```
fn <- fitdist(as.numeric(pira), "pois")
plot.legend <- c("pois")
denscomp(list(fn), legendtext = plot.legend, main="IRA")

fn <- fitdist(as.numeric(incsq), "lnorm")
plot.legend <- c("lnorm")
denscomp(list(fn), legendtext = plot.legend, main="Income Squared")
```






```
fn <- fitdist(as.numeric(agesq), "gamma")
plot.legend <- c("gamma")
denscomp(list(fn), legendtext = plot.legend, main="Age Squared")
```



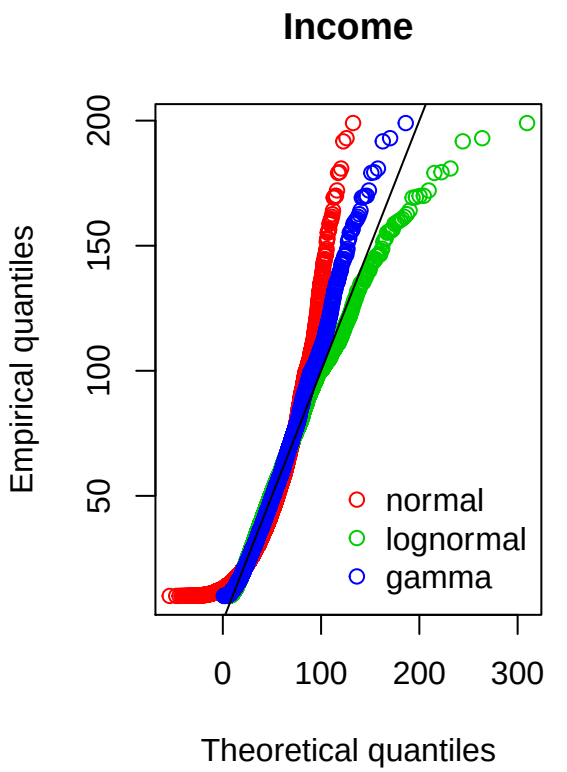
Observations:
-The sample has more individuals that are not eligible for a 401k.
-Most of the individuals in the sample have a lower income.
-We have more married individuals in the sample than non-married individuals.
-We have more females in the sample than males.
-Most of the individuals in the sample is within the working age however probabaly more well-established in their careers (i.e. less 20 year olds and more 30-40 year olds).
-Most individuals in the sample have smaller family sizes.
-Net financial assets are fairly normally distributed.Meaning that most individuals do not have very high or very low net financial assets.
-In the sample there are more individuals not participating in the 401k than individuals that are participating.
-Most individuals in the sample do not have IRA.
-Income squared shows more individuals with a lower level of income (this is helpful because we may want to give more weight to those with lower income levels).
-Age squared similarly shows us that our sample size is made up of younger individuals.

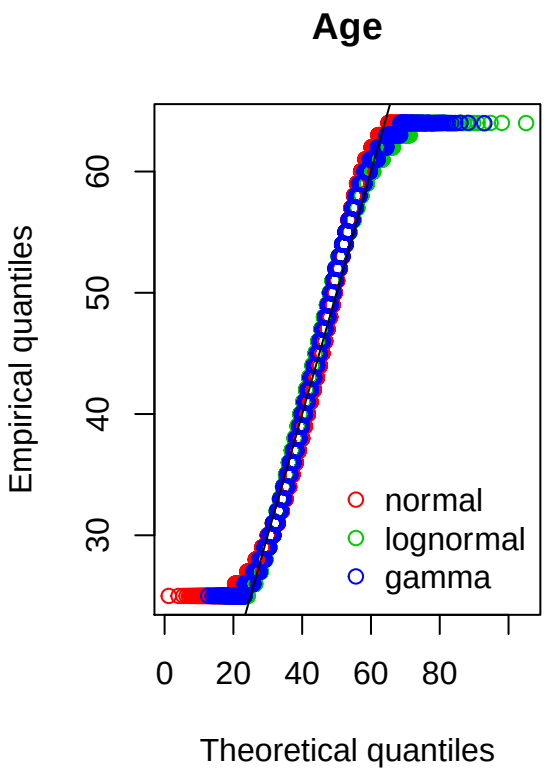
Q-Q Plots

For all of the non-factor variables, the distributions seem to be close to either a normal or gamma distribution. To confirm my estimates from the Cullen-Frey graphs, I will now produce QQ plots. I have also included the normal distribution for comparison.

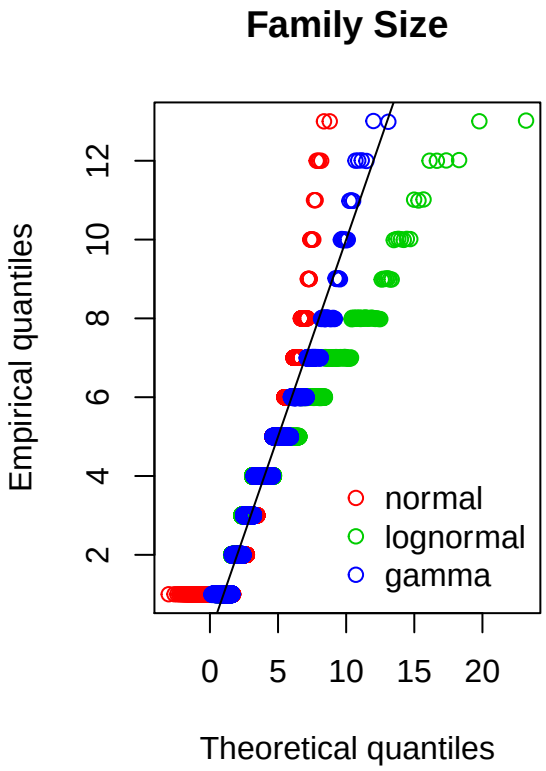
```
par(mfrow=c(1,2))

fit_norm = fitdist(as.numeric(inc), "norm")   #fit selected distributions
fit_lnorm = fitdist(as.numeric(inc), "lnorm")
fit_gamm = fitdist(as.numeric(inc), "gamma")
```

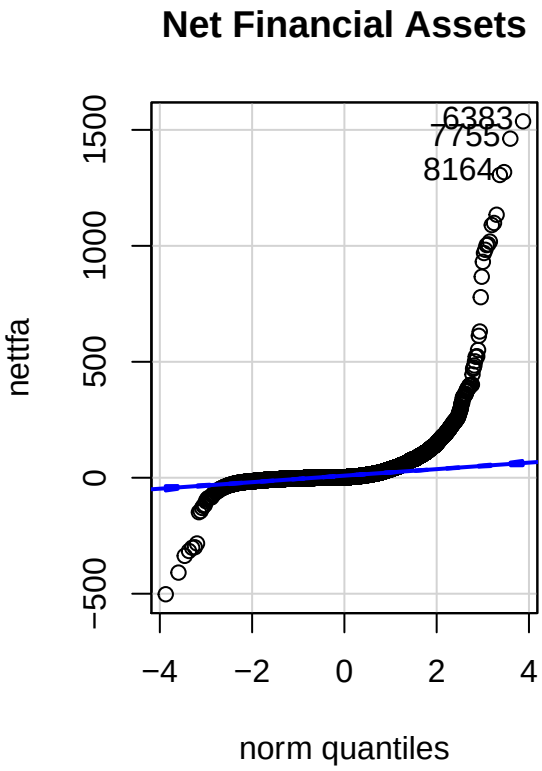
```
plot.legend = c("normal", "lognormal", "gamma")
qqcomp(list(fit_norm, fit_lnorm, fit_gamm), legendtext = plot.legend,main = "Income")

fit_norm = fitdist(as.numeric(age), "norm")
fit_lnorm = fitdist(as.numeric(age), "lnorm")
fit_gamm = fitdist(as.numeric(age), "gamma")
plot.legend = c("normal", "lognormal", "gamma")
qqcomp(list(fit_norm, fit_lnorm, fit_gamm), legendtext = plot.legend,main = "Age")
```





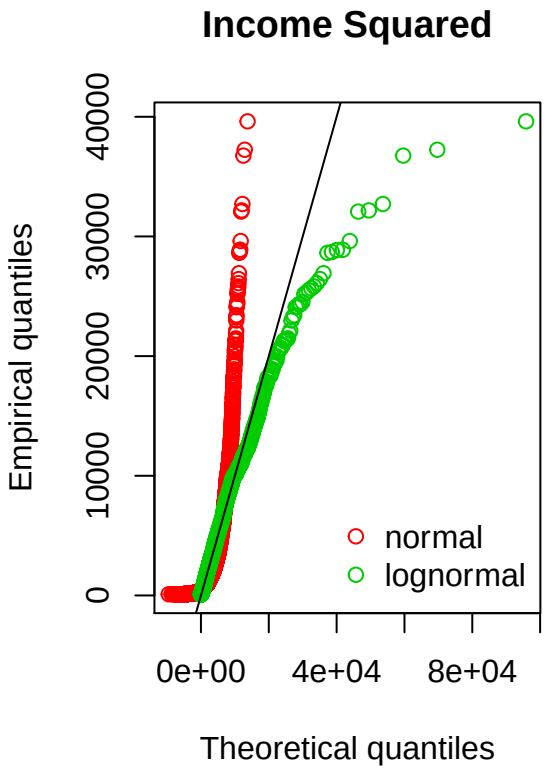

```
fit_norm = fitdist(as.numeric(fsize), "norm")
fit_lnorm = fitdist(as.numeric(fsize), "lnorm")
fit_gamm = fitdist(as.numeric(fsize), "gamma")
plot.legend = c("normal", "lognormal", "gamma")
qqcomp(list(fit_norm, fit_lnorm, fit_gamm), legendtext = plot.legend,main = "Family Size")

qqPlot(~ nettfa, data = data, id = list(n=3),main = "Net Financial Assets")
```

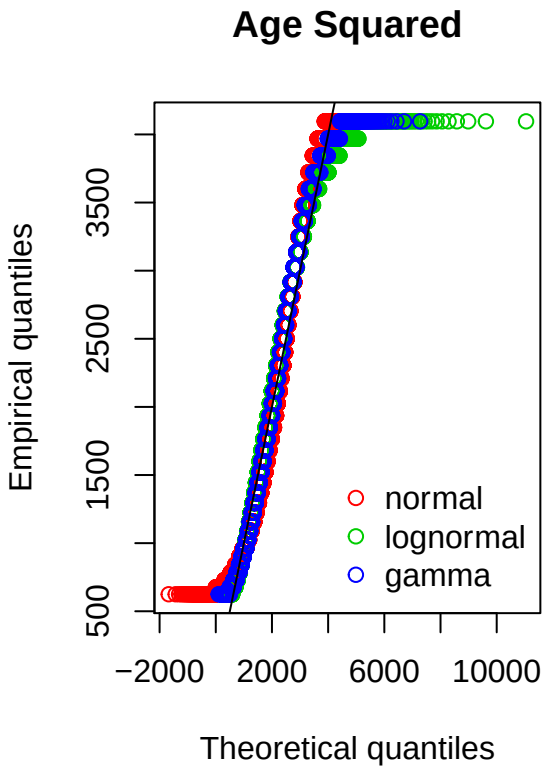





```
## [1] 6383 7755 8164
```

```
fit_norm = fitdist(as.numeric(incsq), "norm")
fit_lnorm = fitdist(as.numeric(incsq), "lnorm")
plot.legend = c("normal", "lognormal")
qqcomp(list(fit_norm, fit_lnorm), legendtext = plot.legend,main = "Income Squared")

fit_norm = fitdist(as.numeric(agesq), "norm")
fit_lnorm = fitdist(as.numeric(agesq), "lnorm")
fit_gamm = fitdist(as.numeric(agesq), "gamma")
plot.legend = c("normal", "lognormal", "gamma")
qqcomp(list(fit_norm, fit_lnorm, fit_gamm), legendtext = plot.legend,
       main = "Age Squared")
```

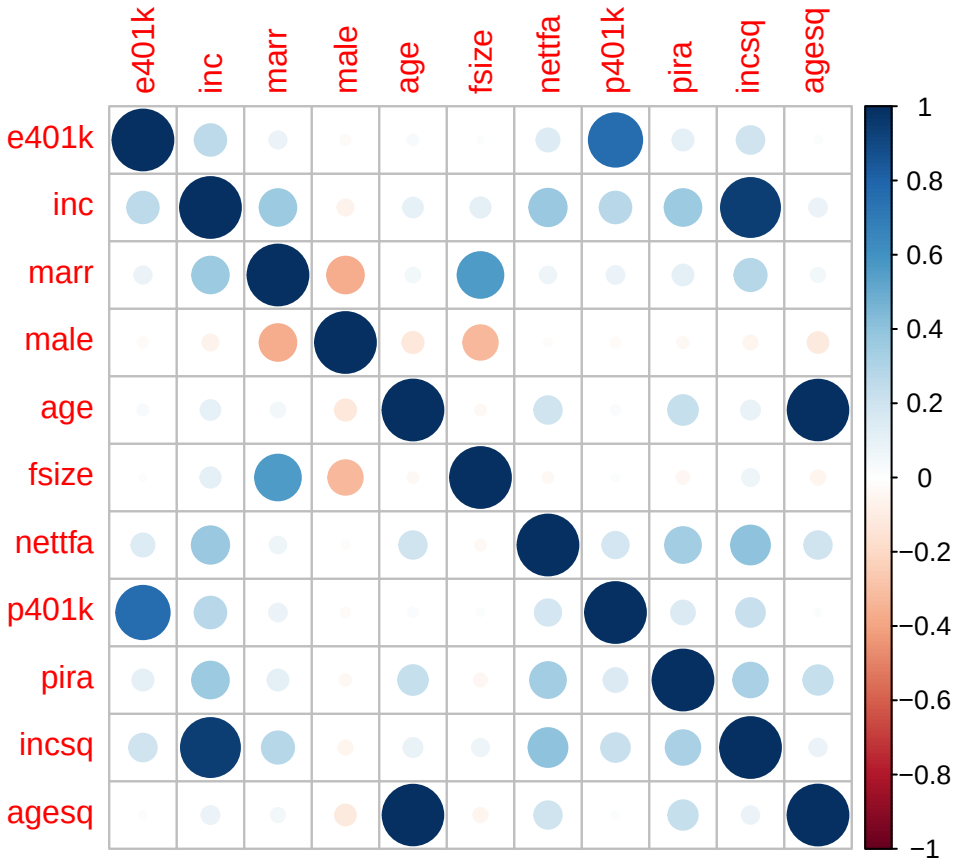




Since the Quantile-Quantile plots don't provide much information for factor variables, I excluded them. In general the variables tend to be mainly following a gamma distribution and in some case a lognormal distribution. Net financial assets follows a normal distribution as it has negative values and is fairly symmetric. These distributions make sense as gamma and lognormal distributions are right skewed and so are most of our non-factor variable. As previously mentioned, net financial assets is fairly symmetric with the majority of the sample not holding extreme levels of net financial assets. This would make it normal. In all cases we can see outliers that we may want to watch out for later.

Correlation plot

```
par(mfrow=c(1,1))

corrplot(cor(data))
```

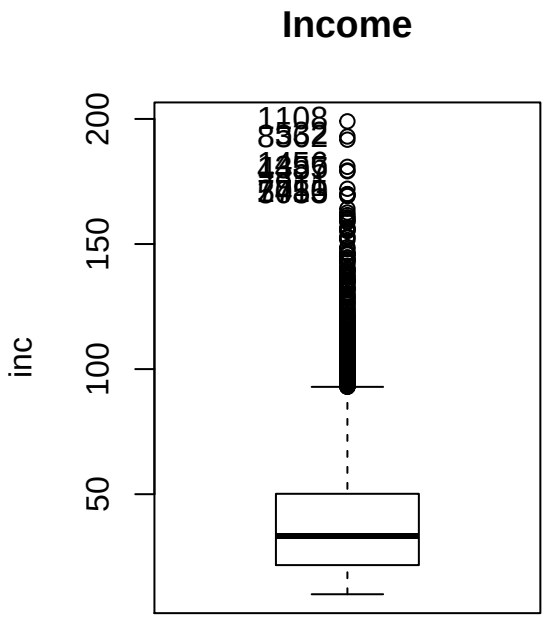


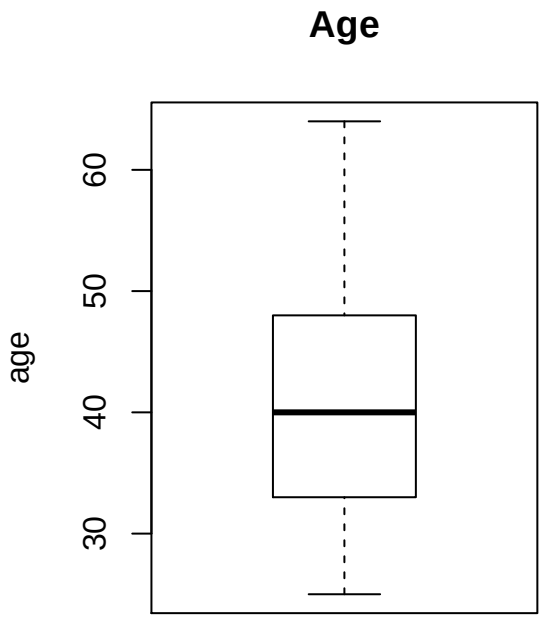
As we can see, net financial assets is positively correlated with all the other variables except for male and family size, in which case it is slightly negatively correlated. This makes sense intuitively. As you get older and have a higher income, you are more likely to have higher net financial assets. You are also likely to have a higher level of net financial assets if you are eligible and if you participate in a 401k. Similarly, if you participate in IRA, you will likely have higher net financial assets. If you are married, intuitively this must mean you are more likely to have a stable job, maybe even a house and therefore your net financial assets would be positively correlated as well. On the other hand, if you have a large family size, you are probably spending more money to support them and therefore have less net financial assets. A slight negative correlation also exist between the net financial assets and male dummy variable. This indicates that if you are male, you will likely have lower net financial assets, however this may not be significant as the the red dot is extremely faint.

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

Boxplot(~ inc, data=data,main="Income")
```

```
##  [1] "1108" "532"  "8362" "1456" "1355" "4487" "7911" "2410" "7755" "5083"
```

```
Boxplot(~ age, data=data,main="Age")
```

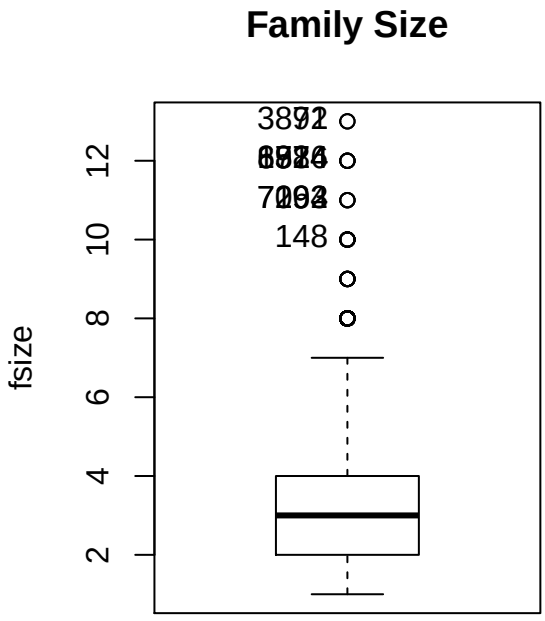


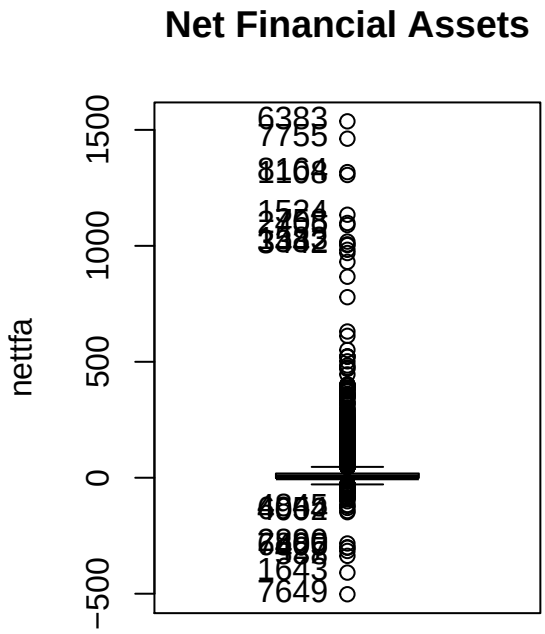



```
Boxplot(~ fsize, data=data,main="Family Size")
```

```
##  [1] "92"   "3871" "514"  "1774" "6984" "8826" "293"  "7094" "7162" "148"
```

```
Boxplot(~ nettfa, data=data,main="Net Financial Assets")
```



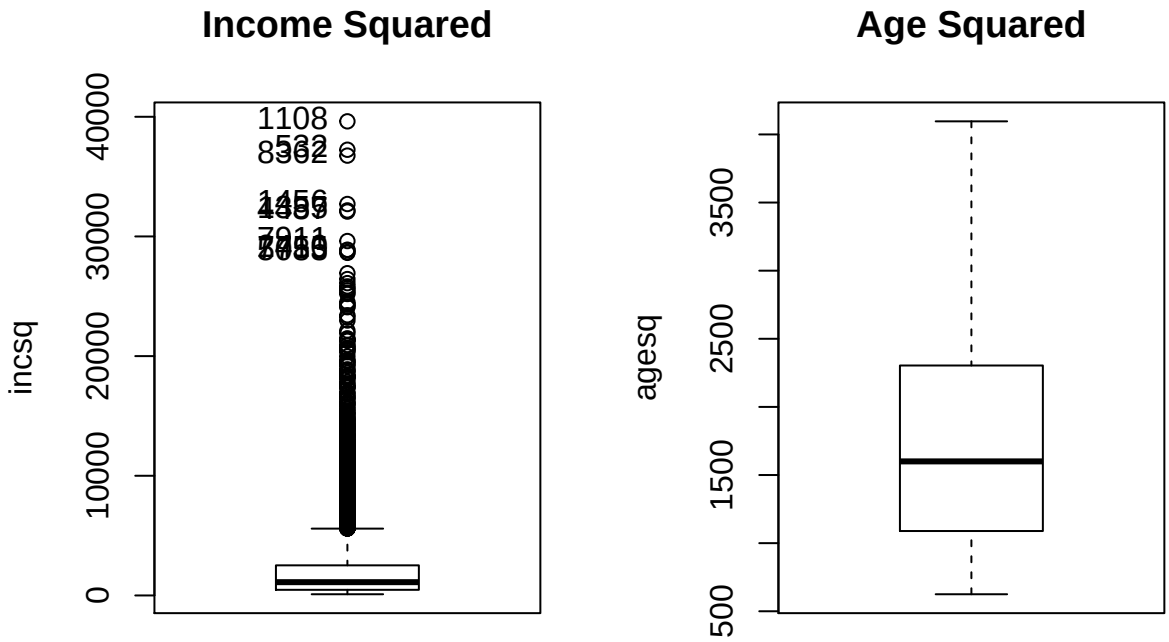



```
##  [1] "7649" "1643" "538"  "427"  "6400" "7805" "2890" "4002" "6954" "4845"
## [11] "6383" "7755" "8164" "1108" "1524" "758"  "2405" "1583" "3332" "3442"
```

```
Boxplot(~ incsq, data=data,main="Income Squared")
```

```
##  [1] "1108" "532"  "8362" "1456" "1355" "4487" "7911" "2410" "7755" "5083"
```

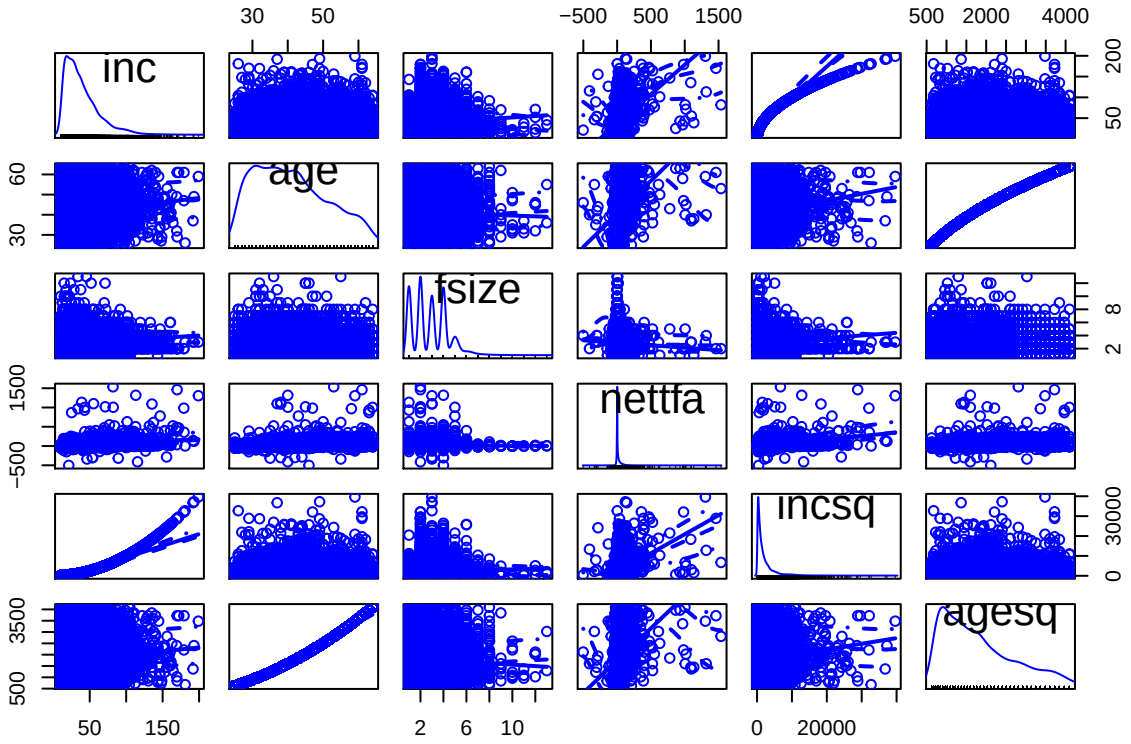
```
Boxplot(~ agesq, data=data,main="Age Squared")
```



For these boxplots we can make the following observations. Income, income squared and net financial assets have a lot of outliers and have a large spread. This may be due to variability of the measurement. It is very understandable that there would be a lot of variability in these three datasets because income and net financial assets are based on several different factors. Family size also shows variability. There are several family sizes that are greater than five. One explanation for this is if the data was self-reported. People may have not been clear on if the family size was based on family members they were supporting, immediate family, or extended family. Age and age squared both behave fairly well with slight skewness .In general, we see that all of the non-factor variables are right skewed to varying degrees.

Scatterplot Matrix

```
par(mfrow=c(1,1))
scatterplotMatrix(data[c('inc','age','fsize','nettfa','incsq','agesq')])
```



Only using the factor variables, we can see relationships between them from the scatterplot matrix. Some of the relationships we notice are that income tends to go down as family size increases and goes up with net financial assets. This makes sense because as your family size increases, it is likely that you are using your income to support more people. As your income increase it is likely that your net financial assets will also increase because you have more wealth. Age has a positive relationship with both net financial assets and income squared. From this we can reason that as your age increases it is likely that you have a better job and are therefore earning a higher income and have higher net financial assets. Family size has a negative relationship with both net financial assets and income squared. This is similar to the relationship between income and family size. As the size your family increases, there are more people to support and therefore less income to spend on financial assets.

Statisical Summaries

```
summary(data)
```

```
##      e401k             inc              marr             male       
##  Min.   :0.0000   Min.   : 10.01   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.: 21.66   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median : 33.29   Median :1.0000   Median :0.0000  
##  Mean   :0.3921   Mean   : 39.25   Mean   :0.6286   Mean   :0.2044  
##  3rd Qu.:1.0000   3rd Qu.: 50.16   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :199.04   Max.   :1.0000   Max.   :1.0000  
##       age              fsize            nettfa            p401k       
##  Min.   :25.00   Min.   : 1.000   Min.   :-502.30   Min.   :0.0000  
##  1st Qu.:33.00   1st Qu.: 2.000   1st Qu.:  -0.50   1st Qu.:0.0000  
##  Median :40.00   Median : 3.000   Median :   2.00   Median :0.0000
```

```
##  Mean   :41.08   Mean   : 2.885   Mean   :  19.07   Mean   :0.2762
##  3rd Qu.:48.00   3rd Qu.: 4.000   3rd Qu.:  18.45   3rd Qu.:1.0000
##  Max.   :64.00   Max.   :13.000   Max.   :1536.80   Max.   :1.0000
##       pira            incsq             agesq
##  Min.   :0.0000   Min.   :  100.2   Min.   : 625
##  1st Qu.:0.0000   1st Qu.:  469.2   1st Qu.:1089
##  Median :0.0000   Median : 1108.1   Median :1600
##  Mean   :0.2543   Mean   : 2121.2   Mean   :1794
##  3rd Qu.:1.0000   3rd Qu.: 2516.0   3rd Qu.:2304
##  Max.   :1.0000   Max.   :39617.3   Max.   :4096
```

```
describe(data)
```

```
##        vars    n    mean      sd  median trimmed     mad     min      max
## e401k     1 9275    0.39    0.49    0.00    0.37    0.00    0.00     1.00
## inc       2 9275   39.25   24.09   33.29   35.82   19.76   10.01   199.04
## marr      3 9275    0.63    0.48    1.00    0.66    0.00    0.00     1.00
## male      4 9275    0.20    0.40    0.00    0.13    0.00    0.00     1.00
## age       5 9275   41.08   10.30   40.00   40.50   11.86   25.00    64.00
## fsize     6 9275    2.89    1.53    3.00    2.76    1.48    1.00    13.00
## nettfa    7 9275   19.07   63.96    2.00    8.60    8.01 -502.30  1536.80
## p401k     8 9275    0.28    0.45    0.00    0.22    0.00    0.00     1.00
## pira      9 9275    0.25    0.44    0.00    0.19    0.00    0.00     1.00
## incsq    10 9275 2121.19 3001.47 1108.09 1487.52 1156.23  100.16 39617.32
## agesq    11 9275 1793.65  895.65 1600.00 1699.17  902.90  625.00  4096.00
##            range  skew kurtosis    se
## e401k       1.00  0.44    -1.80  0.01
## inc       189.03  1.60     3.68  0.25
## marr        1.00 -0.53    -1.72  0.01
## male        1.00  1.47     0.15  0.00
## age        39.00  0.40    -0.79  0.11
## fsize      12.00  0.79     1.26  0.02
## nettfa   2039.10 10.16   171.71  0.66
## p401k       1.00  1.00    -1.00  0.00
## pira        1.00  1.13    -0.73  0.00
## incsq   39517.16  4.02    24.92 31.17
## agesq    3471.00  0.78    -0.30  9.30
```

Again, we will only consider the non-factor variables here. Looking at the skewness,we can see that our non-factor variables are right skewed. Now if we look and kurtosis, we notice that income, income squared and nettfa are all leptokurtic. This means that there is more data that falls within the tails and will therefore be rejected. On the other hand, age and age squared are both platykurtic. Meaning that they have thinner tails. Range of the data is also high for all the non-factor variables, which means we have a lot of variablity in our data. The means tell us that on average the individuals in the sample are about 41 years of age with an income of about $39250, a family size of 3 and net financial assets being close to $19070.

(b) For each variable, test if a transformation to linearity is appropriate, and if so, apply the respective transformation. We will use these results later in part (m).

I will again only be looking at the non-factor variables for transformations.

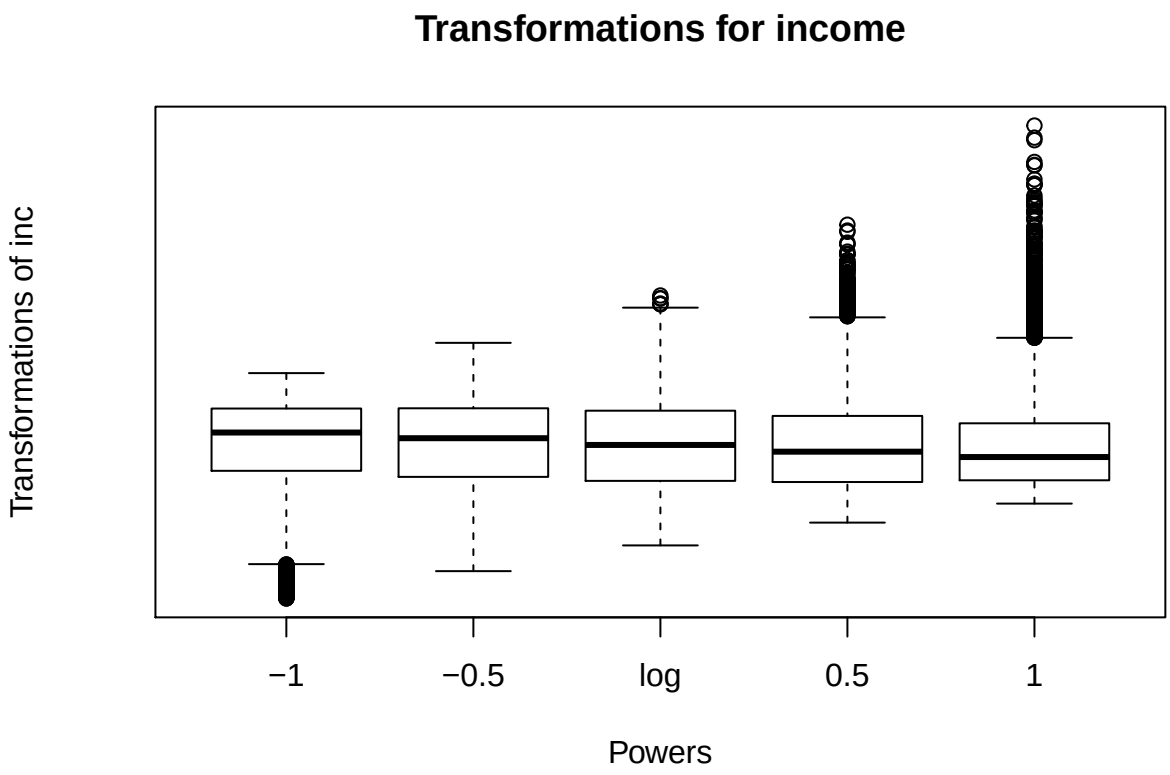
I start with testing for income by first doing a Box-Cox power test and then using symbox.

```
p1 = powerTransform(inc~1,data=data,family="bcPower")       #Box-Cox Test
summary(p1)
```

```
## bcPower Transformation to Normality
```

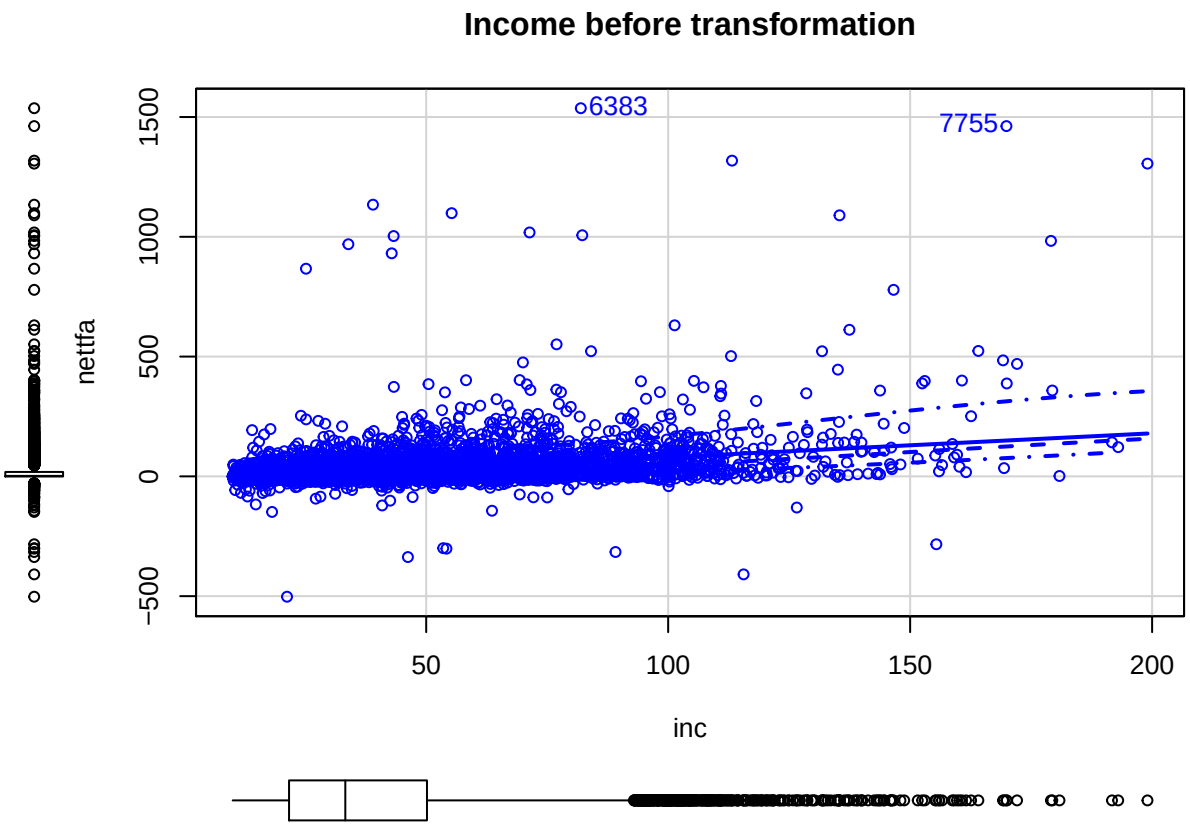
```
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    -0.067       -0.07      -0.0996      -0.0344
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 16.2473  1 5.5589e-05
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 4269.986  1 < 2.22e-16
```

```
symbox(~inc,data=data, main = "Transformations for income")       #Symbox
```



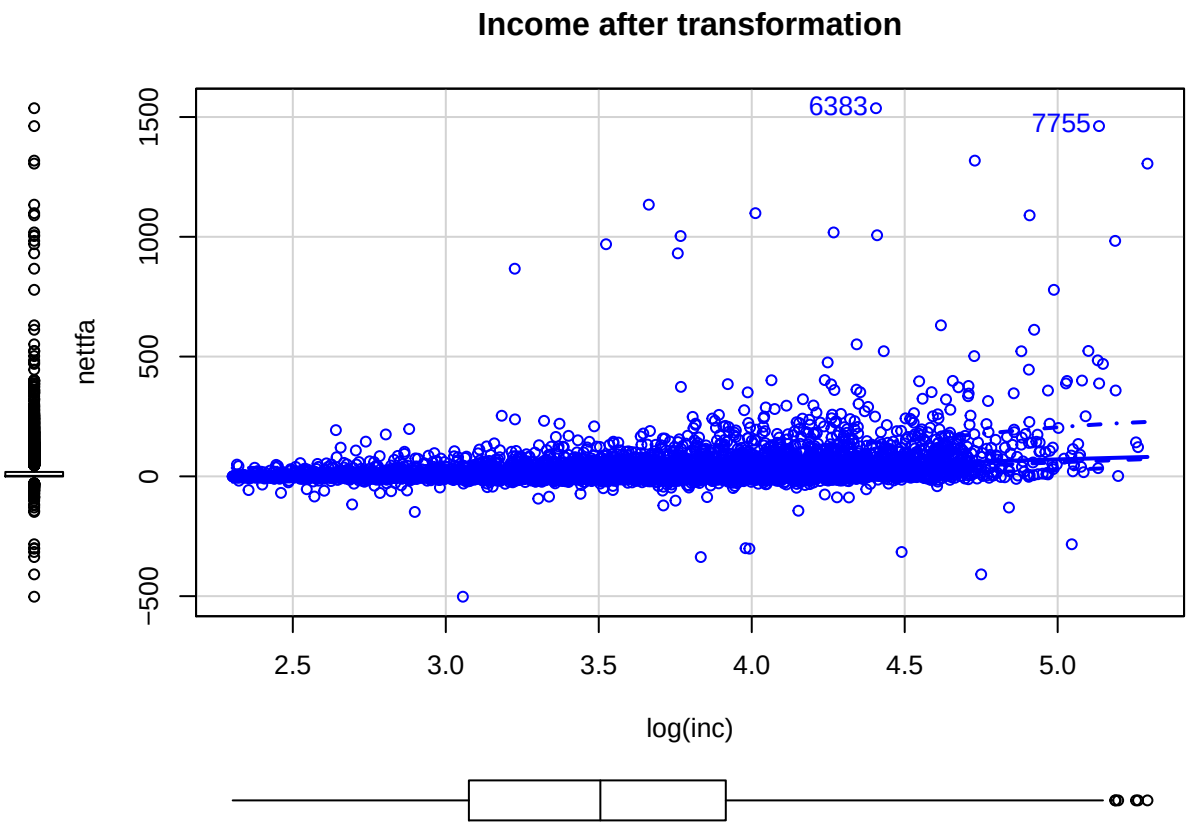
both the Box-Cox and the Symbox seem to be favouring a log transformation. We can look at the before and after scatterplots and histograms.

```
par(mfrow=c(1,2))
scatterplot(nettfa~inc,lwd=3,id=TRUE, main="Income before transformation")
```



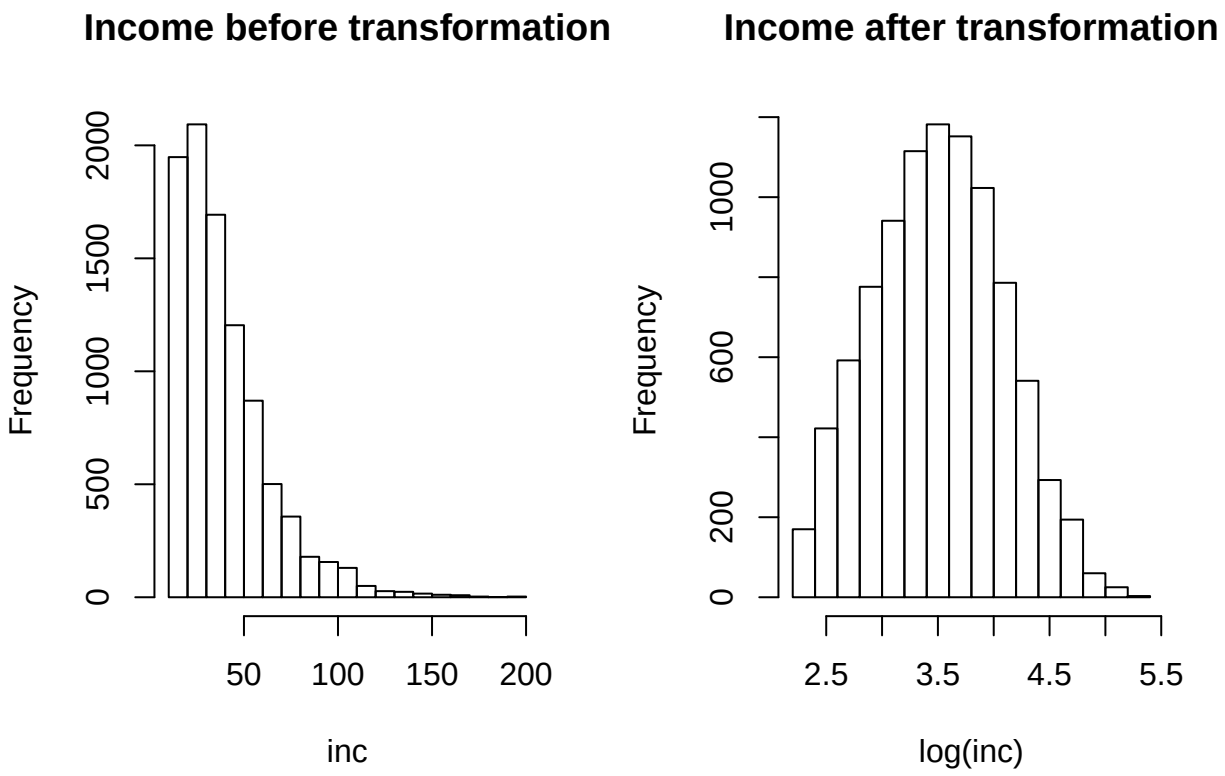

```
## [1] 6383 7755
```

```
scatterplot(nettfa~log(inc),lwd=3,id=TRUE,main="Income after transformation")
```




```
## [1] 6383 7755
```

```
hist(inc,main="Income before transformation" )
hist(log(inc), main= "Income after transformation")
```



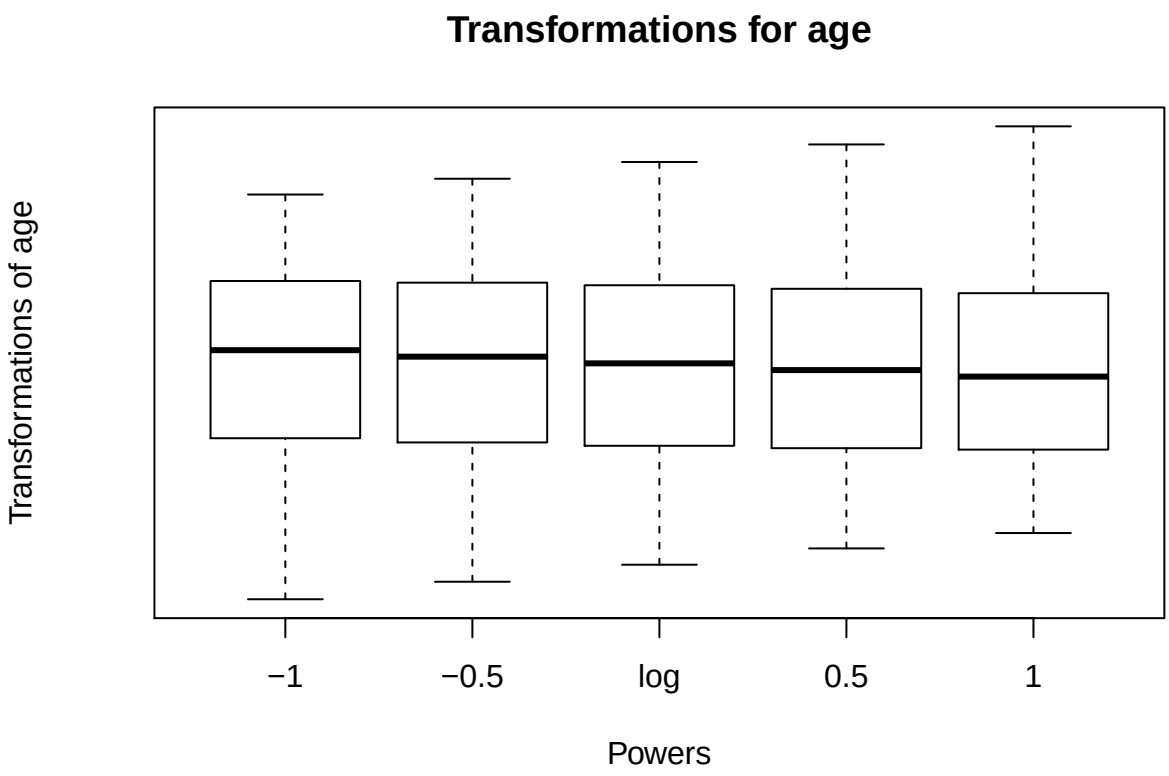
As we can see, the data ends up looking much more normal and spread out. Therefore I will keep this transformation.

Now I will repeat the procedure for the other factor variables.

```
p1 = powerTransform(age~1,data=data,family="bcPower")
summary(p1)
```

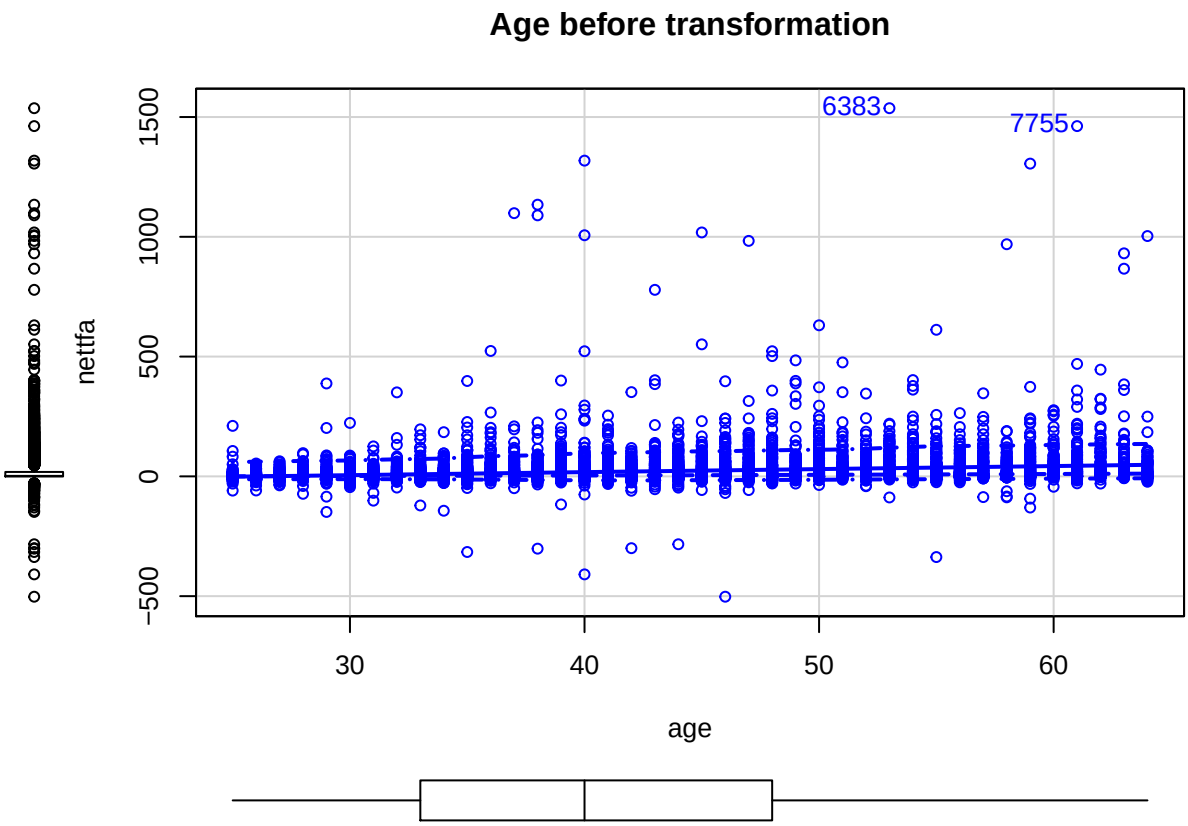
```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0374           0      -0.1214       0.0465
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df    pval
## LR test, lambda = (0) 0.7643395  1 0.38197
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 583.9085  1 < 2.22e-16
```

```
symbox(~age,data=data, main = "Transformations for age")
```



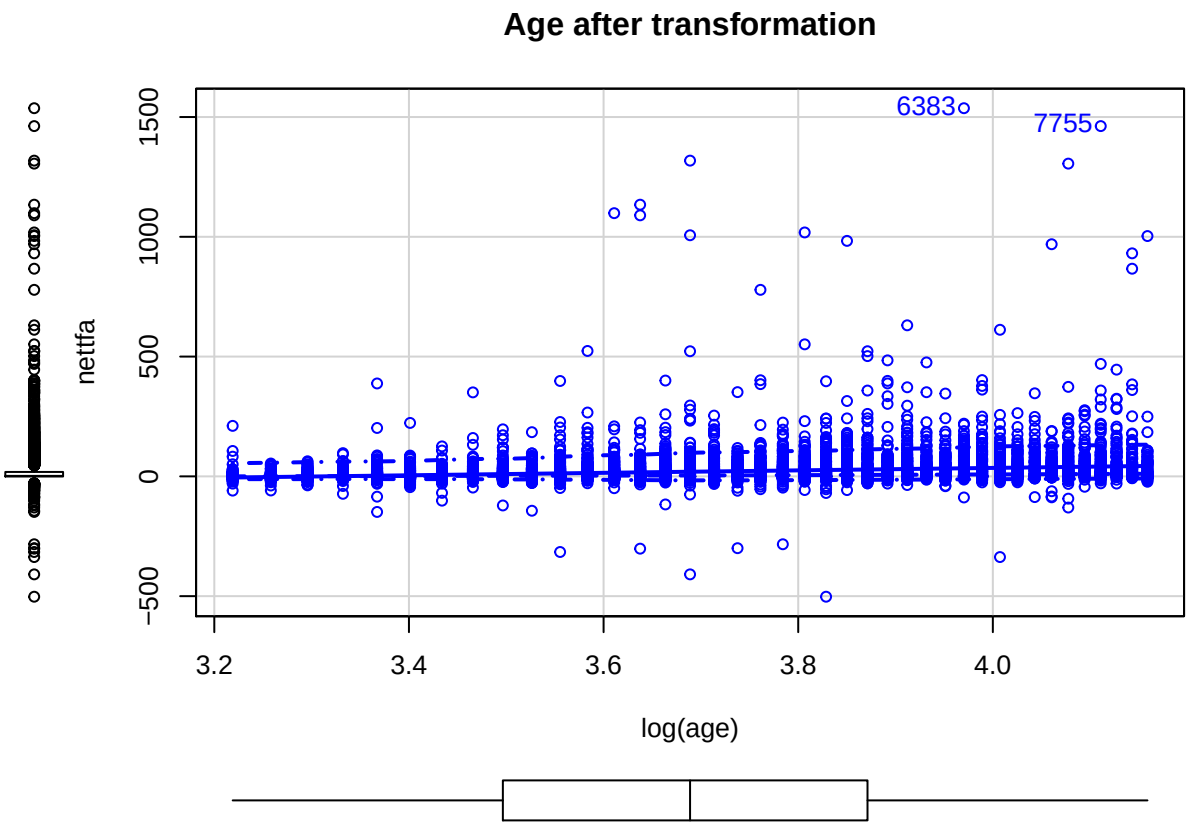
While the Box-Cox test suggests a log transformation, the marginal effect of the transformation won't be that great. We can however look at the scatterplots and histograms to make sure.

```
par(mfrow=c(1,2))
scatterplot(nettfa~age,lwd=3,id=TRUE, main="Age before transformation")
```




```
## [1] 6383 7755
```

```
scatterplot(nettfa~log(age),lwd=3,id=TRUE,main="Age after transformation")
```




```
## [1] 6383 7755
```

```
hist(age,main="Age before transformation" )
hist(log(age), main= "Age after transformation")
```

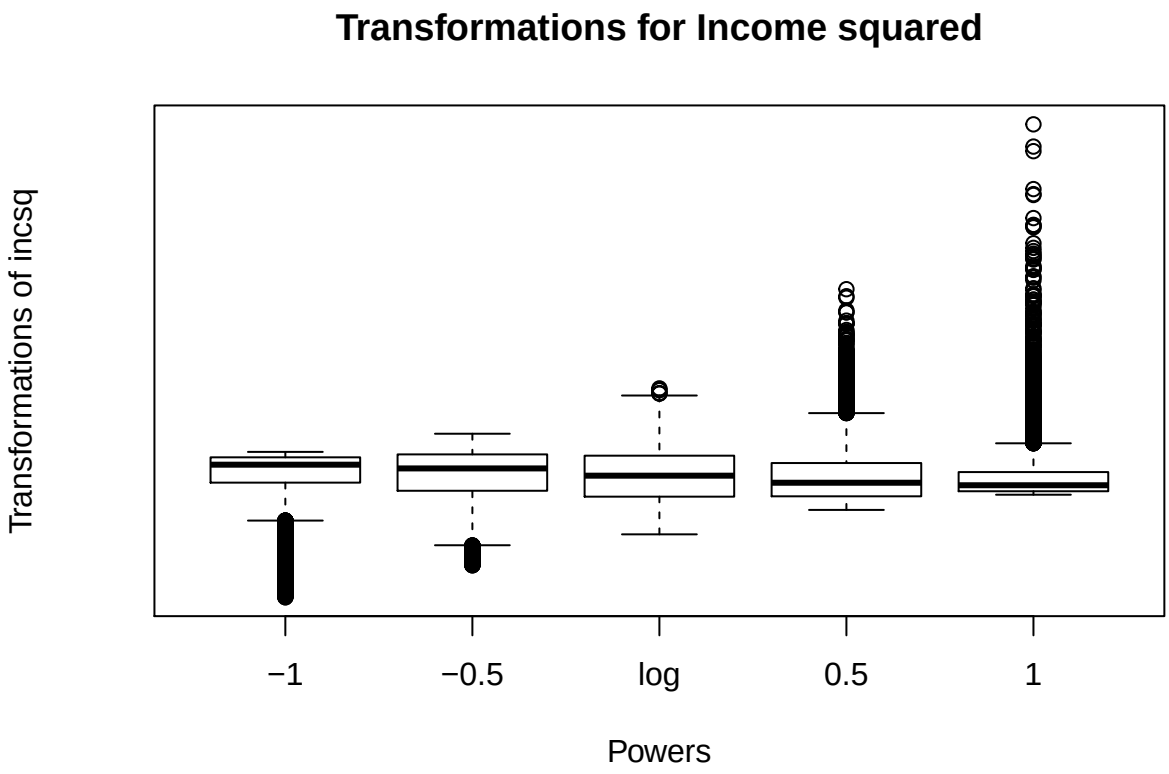
After looking at the histograms and the scatterplots I have also decided to keep the log transformation of the age variables. It removes the skewness and make age follow a normal distribution which can make running the analysis much easier.

Now, I will do this for age squared, income squared,family size, and net financial assets.

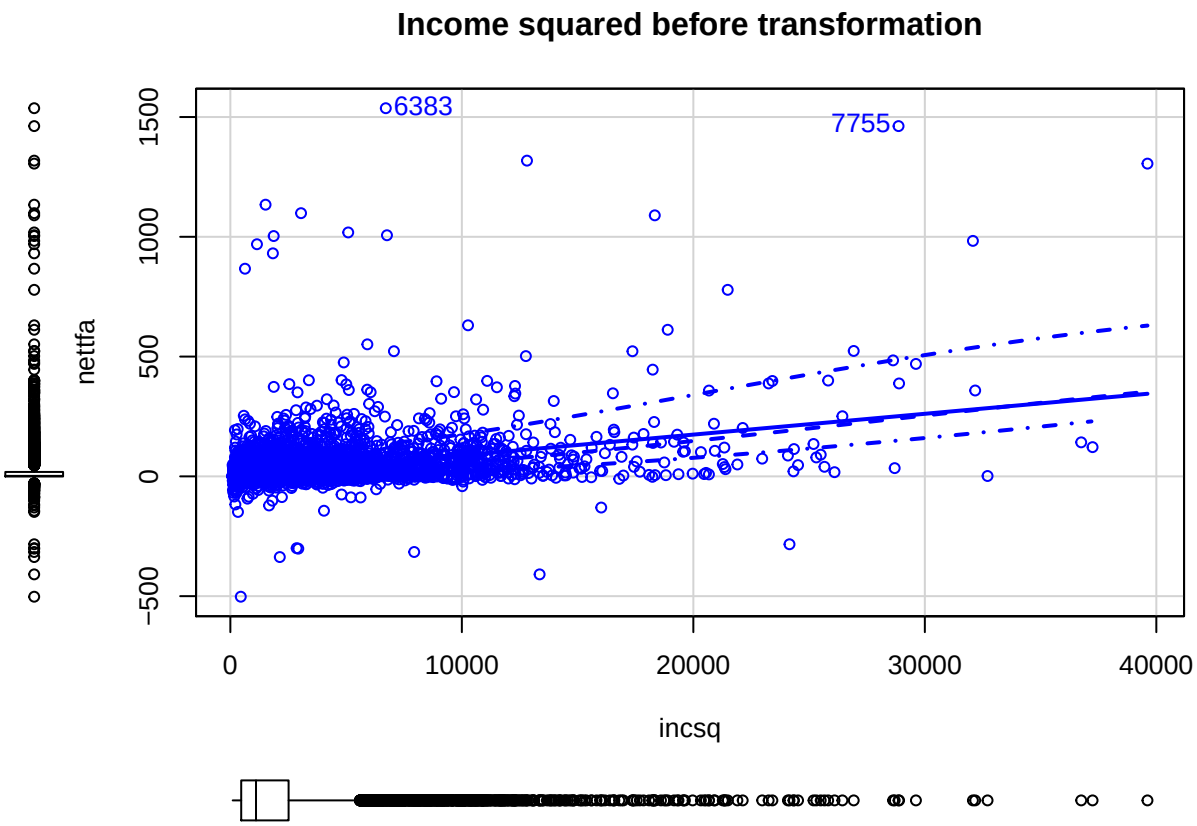
```
p1 = powerTransform(incsq~1,data=data,family="bcPower")
summary(p1)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0335       -0.03      -0.0498      -0.0172
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 16.2473  1 5.5589e-05
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 15928.15  1 < 2.22e-16
```

```
symbox(~incsq,data=data, main = "Transformations for Income squared")
```

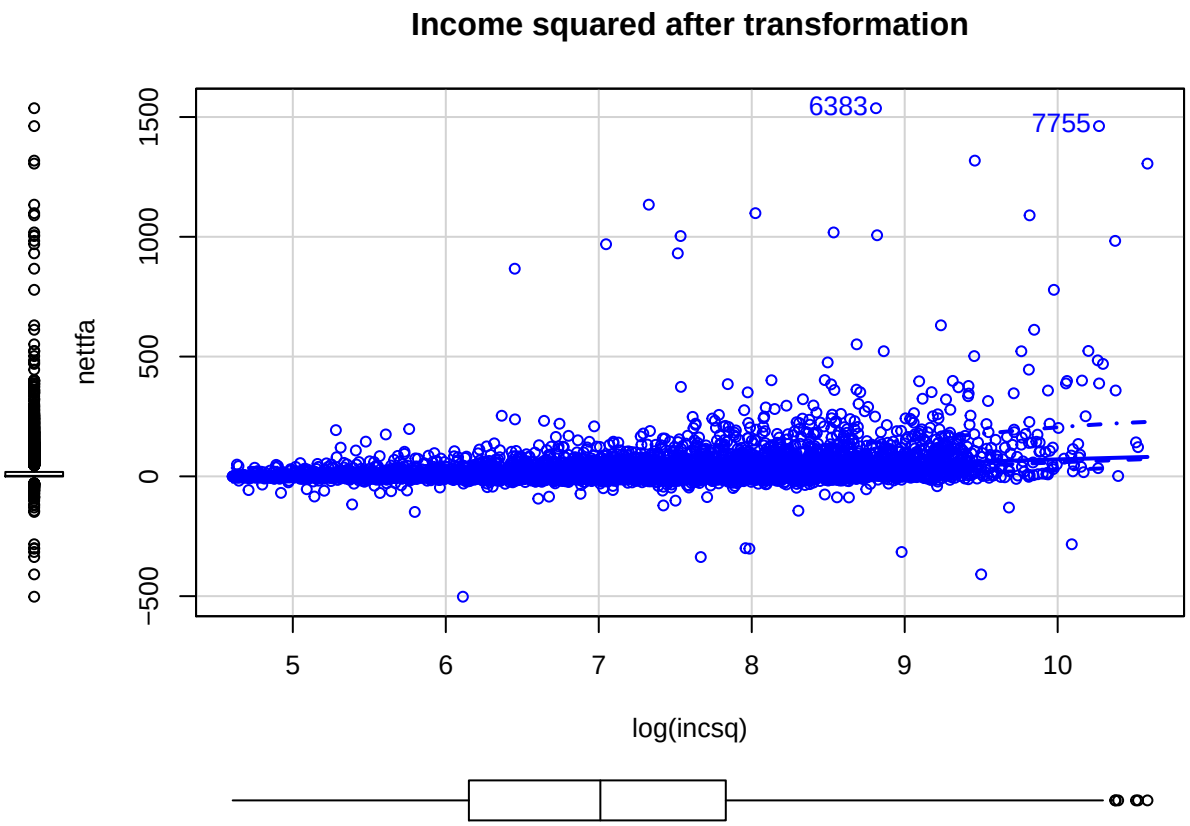



```
par(mfrow=c(1,2))
scatterplot(nettfa~incsq,lwd=3,id=TRUE, main="Income squared before transformation")
```



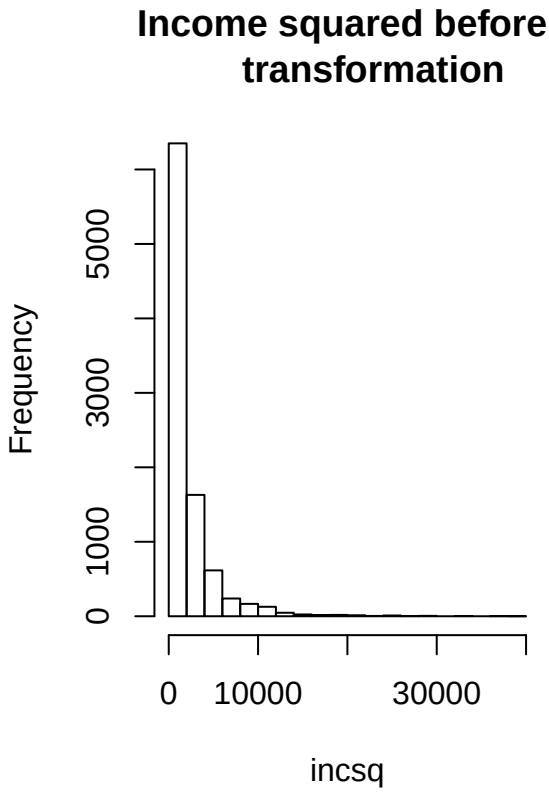

```
## [1] 6383 7755
```

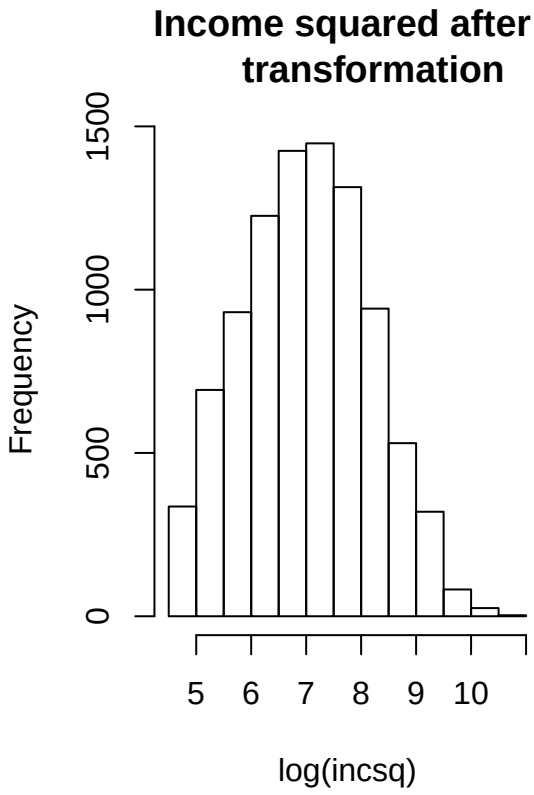
```
scatterplot(nettfa~log(incsq),lwd=3,id=TRUE,main="Income squared after transformation")
```




```
## [1] 6383 7755
```

```
hist(incsq,main="Income squared before
     transformation" )
hist(log(incsq), main= "Income squared after
     transformation")
```






```
p1 = powerTransform(agesq~1,data=data,family="bcPower")
summary(p1)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0187           0      -0.0607       0.0233
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df    pval
## LR test, lambda = (0) 0.7643395  1 0.38197
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 2217.854  1 < 2.22e-16
```

```
symbox(~agesq,data=data, main = "Transformations for Age squared")
```

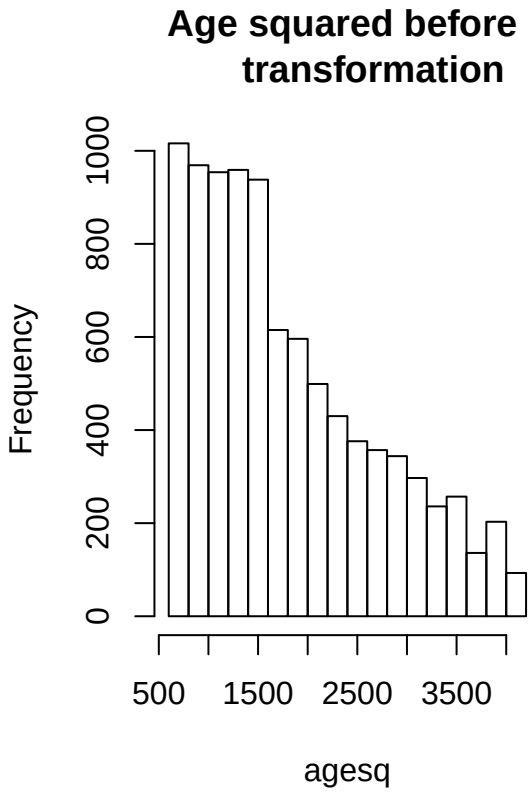

```
par(mfrow=c(1,2))
scatterplot(nettfa~agesq,lwd=3,id=TRUE, main="Age squared before transformation")
```

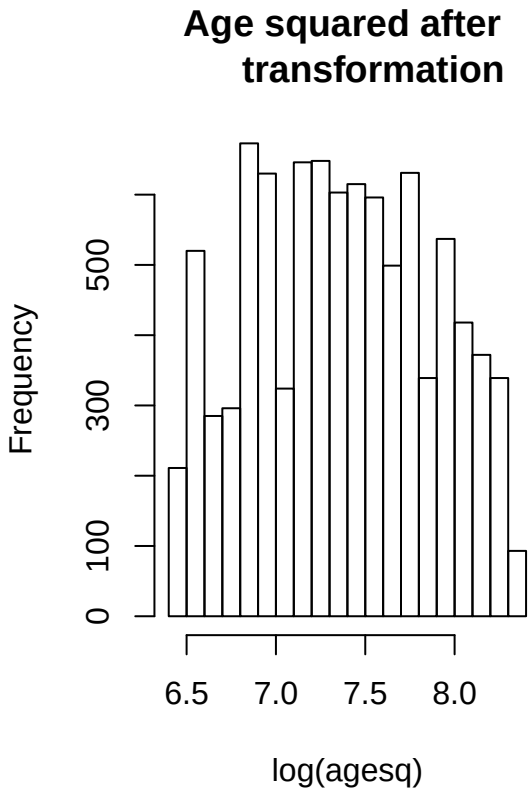

```
## [1] 6383 7755
```

```
scatterplot(nettfa~log(agesq),lwd=3,id=TRUE,main="Age squared after transformation")
```


```
## [1] 6383 7755
```

```
hist(agesq,main="Age squared before
     transformation" )
hist(log(agesq), main= "Age squared after
     transformation")
```

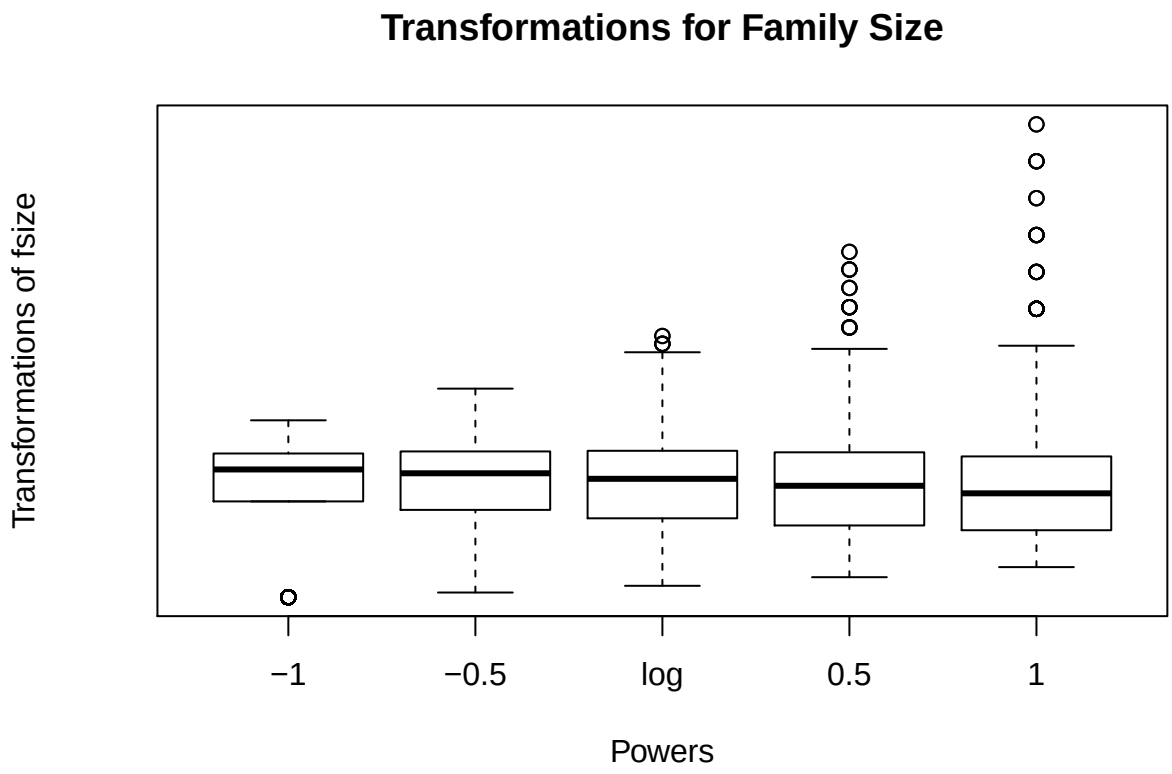




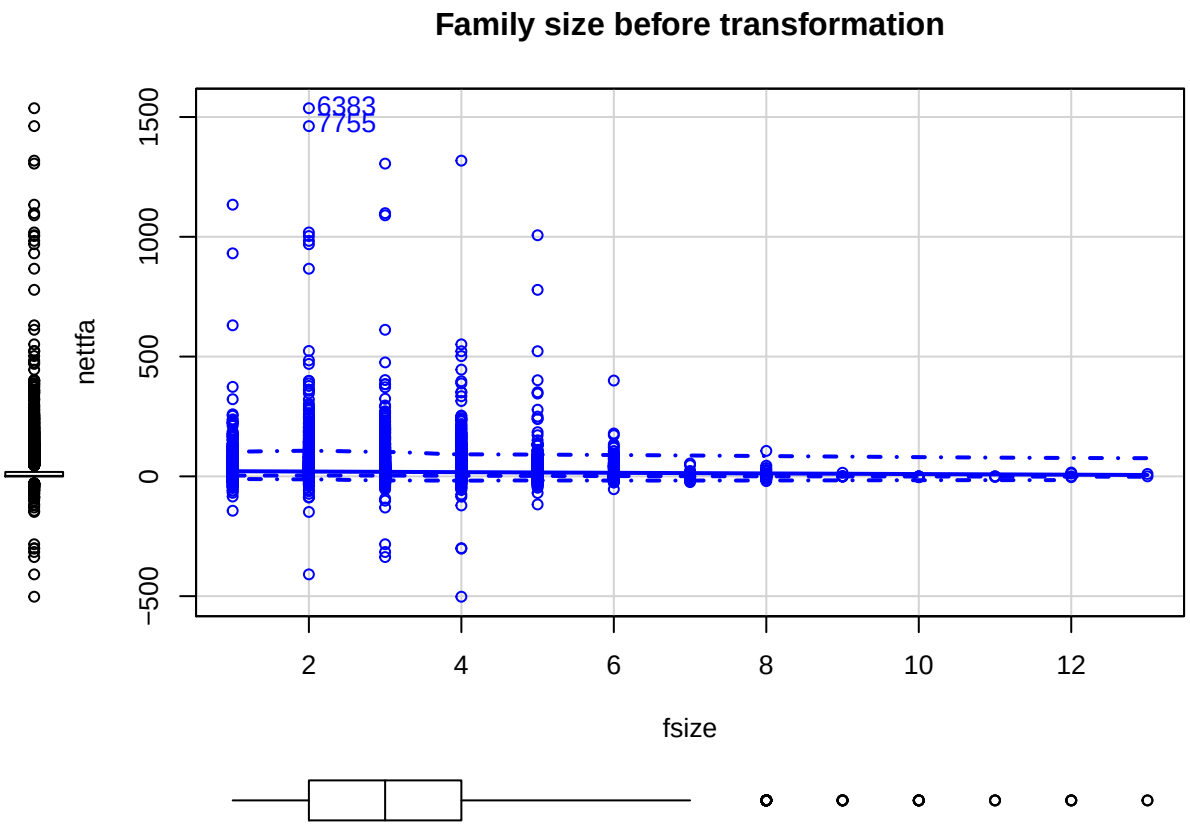

```
p1 = powerTransform(fsize~1,data=data,family="bcPower")
summary(p1)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.3042        0.33        0.268       0.3404
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 266.9023  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 1465.565  1 < 2.22e-16
```

```
symbox(~fsize,data=data, main = "Transformations for Family Size")
```

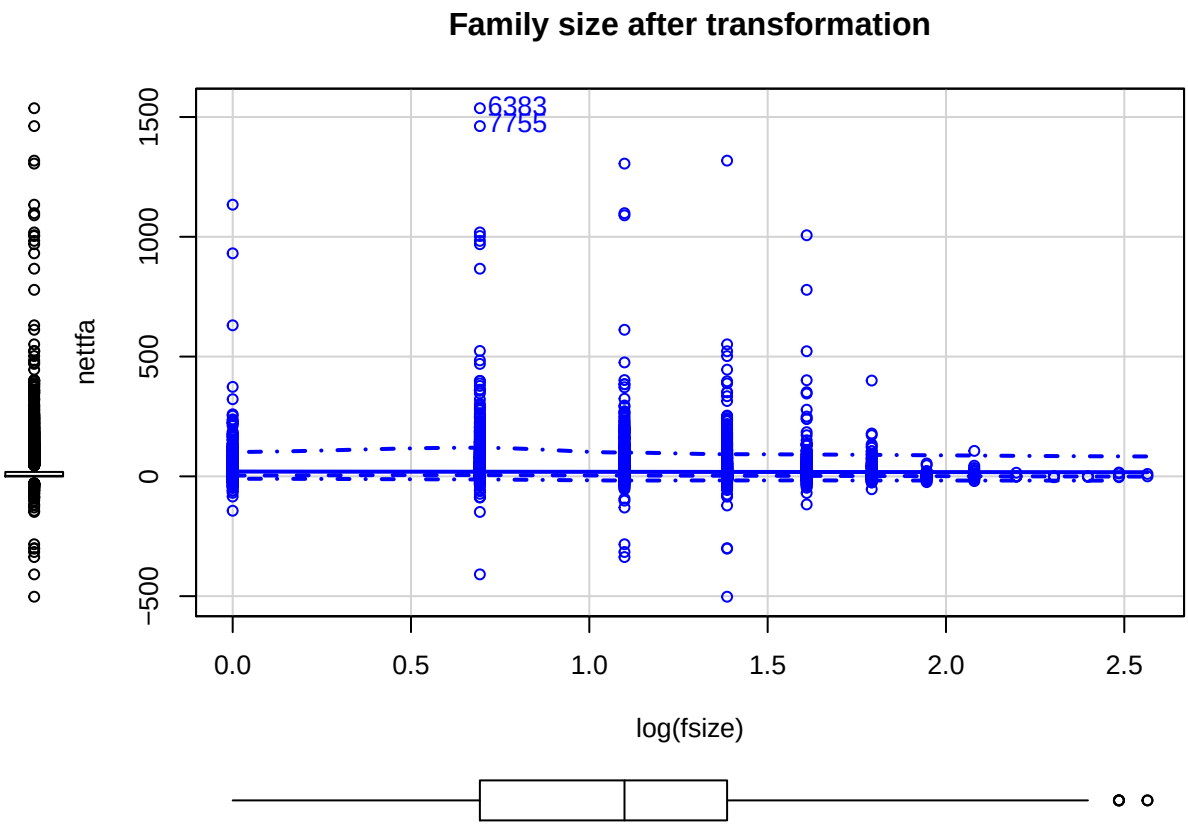



```
fsize_pow = fsize^(-0.5)

par(mfrow=c(1,2))
scatterplot(nettfa~fsize,lwd=3,id=TRUE, main="Family size before transformation")
```



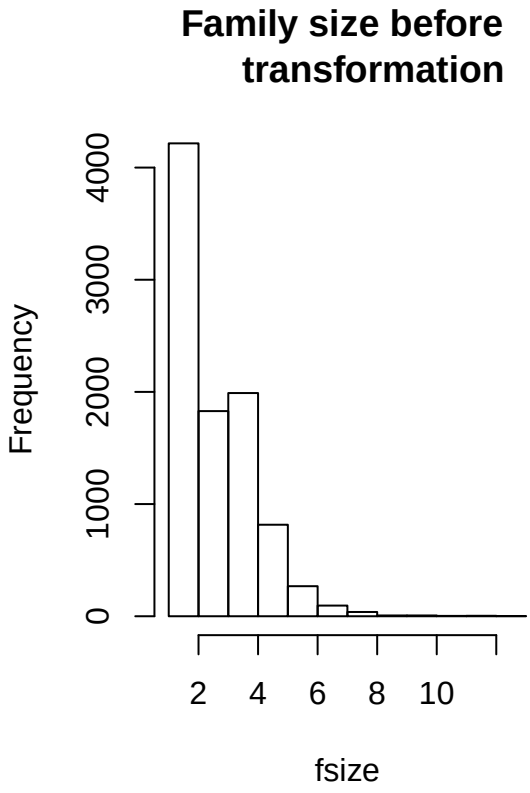

```
## [1] 6383 7755
```

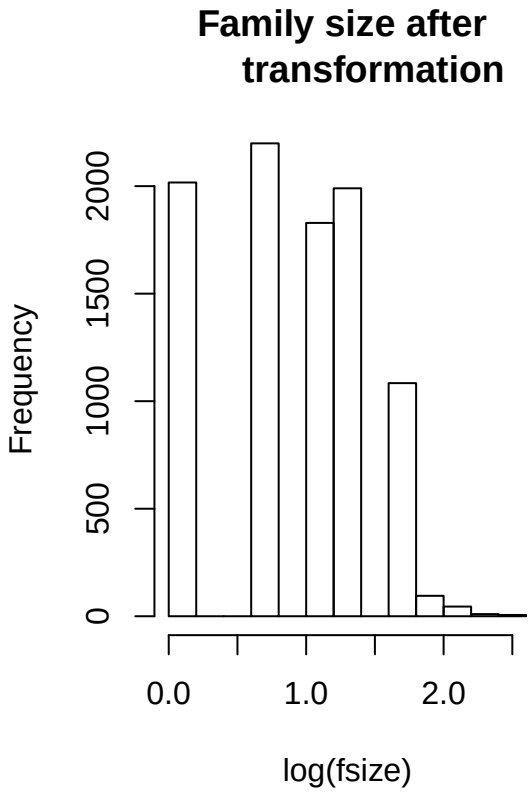
```
scatterplot(nettfa~log(fsize),lwd=3,id=TRUE,main="Family size after transformation")
```




```
## [1] 6383 7755
```

```
hist(fsize,main="Family size before
     transformation" )

hist(log(fsize), main= "Family size after
     transformation")
```



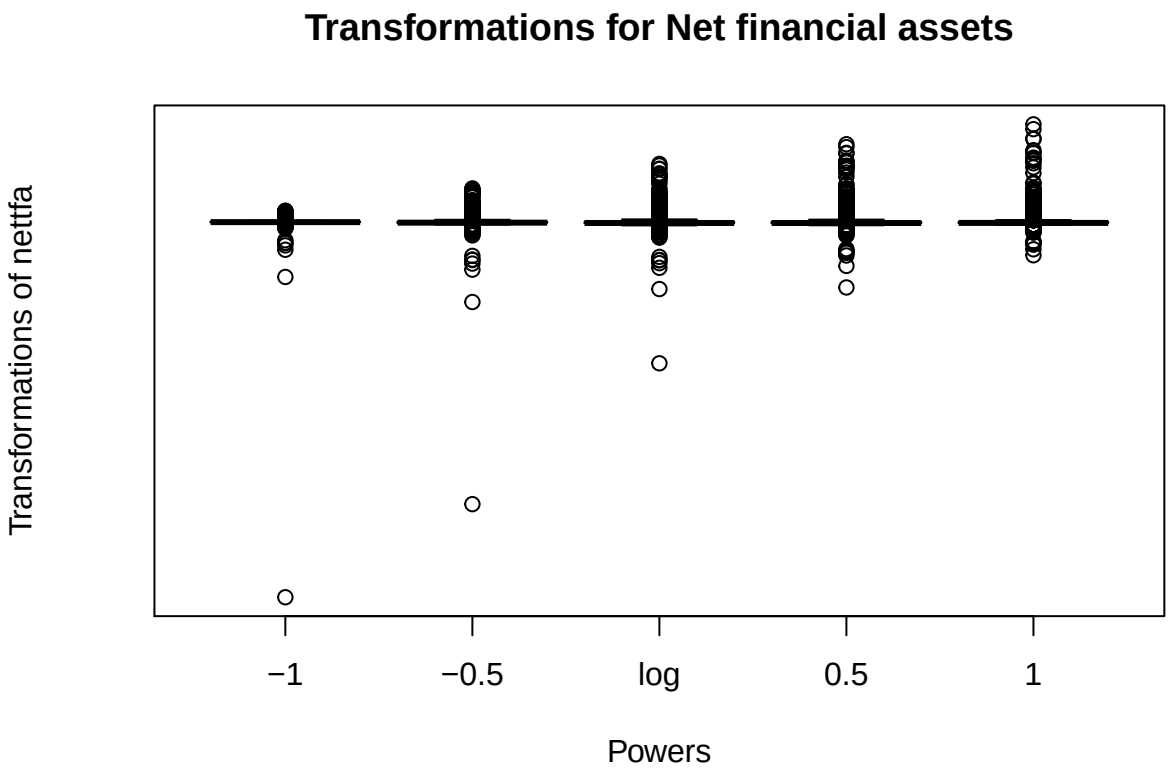



```
p1 = powerTransform(nettfa~1,data=data,family="bcnPower")
summary(p1)
```

```
## bcnPower transformation to Normality
##
## Estimated power, lambda
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.0131       0.013       0.0035       0.0228
##
## Estimated location, gamma
##    Est gamma Std Err. Wald Lower Bound Wald Upper Bound
## Y1    1.3736   0.0583           1.2593           1.4879
##
## Likelihood ratio tests about transformation parameters
##                                LRT df        pval
## LR test, lambda = (0)     7.308683  1 0.006862222
## LR test, lambda = (1) 43738.527907  1 0.000000000
```

```
symbox(~nettfa,data=data, main = "Transformations for Net financial assets")
```

```
## Warning in symbox.default(as.vector(mf[[1]]), ylab = ylab, ...): start set
## to 522.693
```



From these charts I have decided to apply the log transformation to the following variables: income, income squared, age, age squared, and family size.

(c) Estimate a multiple linear regression model for nettfa that includes income, age, and e401k as explanatory variables. We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates. If there are any outliers worth removing, remove them before proceeding with the next steps.

```
base_mod = lm(nettfa~inc+age+e401k,data=data)
S(base_mod)
```

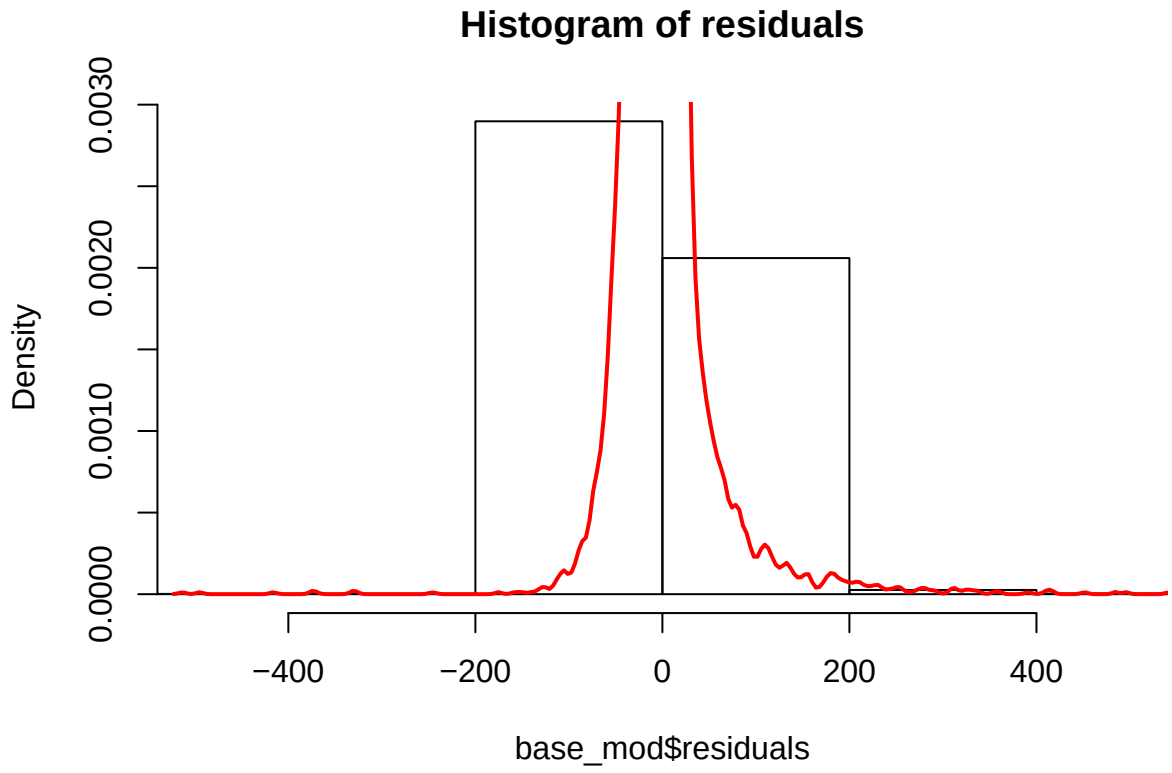
```
## Call: lm(formula = nettfa ~ inc + age + e401k, data = data)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.73111    2.60294 -23.716  < 2e-16 ***
## inc           0.92085    0.02620  35.151  < 2e-16 ***
## age           1.02986    0.05906  17.438  < 2e-16 ***
## e401k         5.98908    1.28592   4.657 3.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 58.25 on 9271 degrees of freedom
## Multiple R-squared: 0.171
## F-statistic: 637.4 on 3 and 9271 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 101727.8 101763.4
```

All of the variables are statistically significant at the 99% confidence level which we can see from the p-values. Economically, we can see that an increase in income, age and being eligible for a 401k can all make an individual's net financial assets increase. This makes sense given that when you are older it is likely you will be more established in your career and therefore earning a higher income and maybe more eligible for a 401k. These factors will likely increase your wealth and therefore, net financial assets will increase.

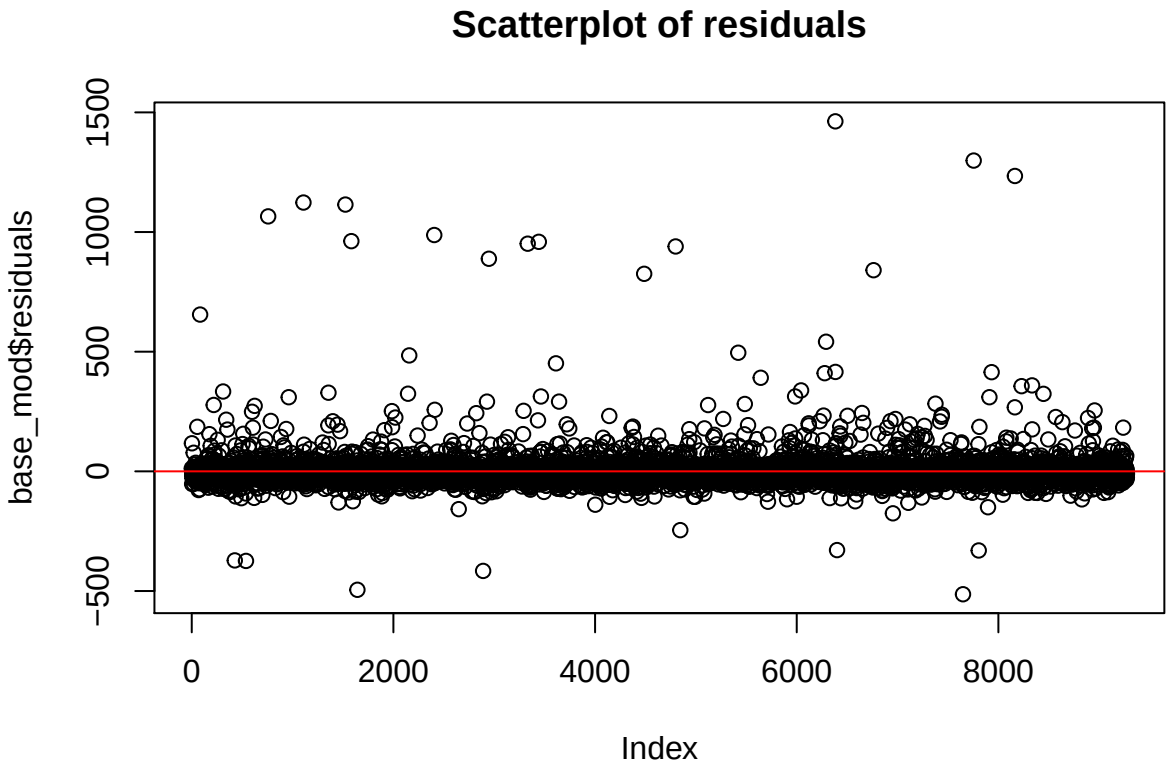
The estimates given by this model express the following. A $1000 increase in annual income will increase net financial assets by $920.85. A one year increase in age will raise net financial assets by $1029.86. Finally, the approximate difference in net financial assets between individuals who are eligible for a 401k and those who are not is approximately $5989.08.

Now we want to remove the outliers. First we can look at a scatterplot and histogram of the residuals.

```
hist(base_mod$residuals, breaks = k, freq = FALSE, ylab="Density", main = "Histogram of residuals", xli
lines(density(base_mod$residuals),lwd=2,col="red")
```




```
plot(base_mod$residuals, main="Scatterplot of residuals")
abline(h=0,col="red")
```



W can see from this that there are outliers, and we want to see how much of an impact these outliers make and which ones to remove.

Performing a Jarque-Bera test allows us determine whether or not there is skewness and excess kurtosis (there are influential outliers)

```
jarque.bera.test(base_mod$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  base_mod$residuals
## X-squared = 14843000, df = 2, p-value < 2.2e-16
```

We reject the null meaning that there are influential outliers in the dataset. To figure out what the outliers are we run the following code and subset the data.

```
car::outlierTest(base_mod)
```

```
##      rstudent unadjusted p-value Bonferonni p
## 6383 26.01501        4.5427e-144  4.2134e-140
## 7755 22.95787        1.7262e-113  1.6011e-109
## 8164 21.73732        3.1587e-102   2.9297e-98
## 1108 19.73952         5.4100e-85   5.0178e-81
## 1524 19.53029         2.8221e-83   2.6175e-79
## 758  18.63142         4.3372e-76   4.0228e-72
## 2405 17.24233         1.3395e-65   1.2424e-61
## 1583 16.76518         3.5945e-62   3.3339e-58
## 3442 16.71726         7.8528e-62   7.2834e-58
```

```
## 3332 16.57654         7.6984e-61   7.1403e-57
```

```
base_rm_outliers = update(base_mod, subset =
                            -c(6383,7755,8164,1108,1524,758,2405,1583,3442,3332))
S(base_rm_outliers)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k, data = data, subset =
##          -c(6383, 7755, 8164, 1108, 1524, 758, 2405, 1583, 3442, 3332))
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.2038     2.0071 -28.002  < 2e-16 ***
## inc           0.7869     0.0203  38.768  < 2e-16 ***
## age           0.9831     0.0455  21.605  < 2e-16 ***
## e401k         7.1026     0.9914   7.164 8.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.86 on 9261 degrees of freedom
## Multiple R-squared: 0.2139
## F-statistic: 839.9 on 3 and 9261 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96777.20 96812.87
```

Now we can compare the slopes of our two models to see the difference.

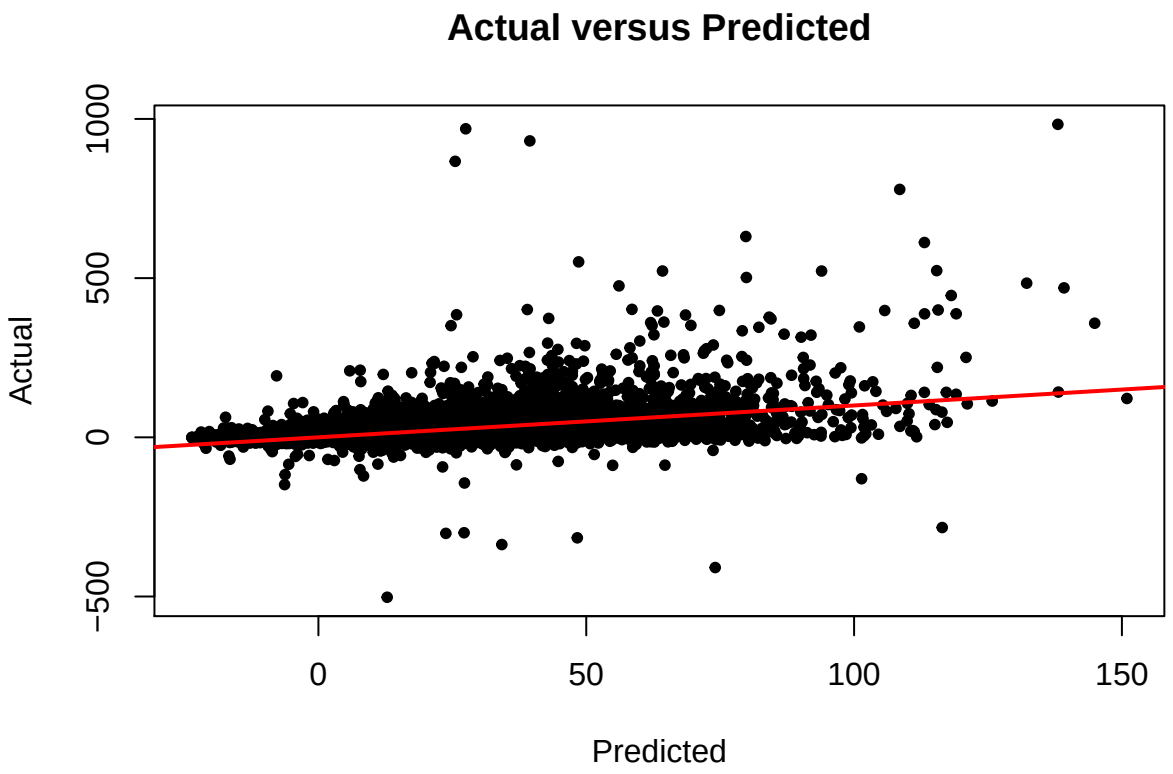
```
cbind(Original=coef(base_mod),NoOutliers = coef(base_rm_outliers))
```

```
##                Original NoOutliers
## (Intercept) -61.7311085 -56.203766
## inc           0.9208451   0.786927
## age           1.0298594   0.983114
## e401k         5.9890819   7.102636
```

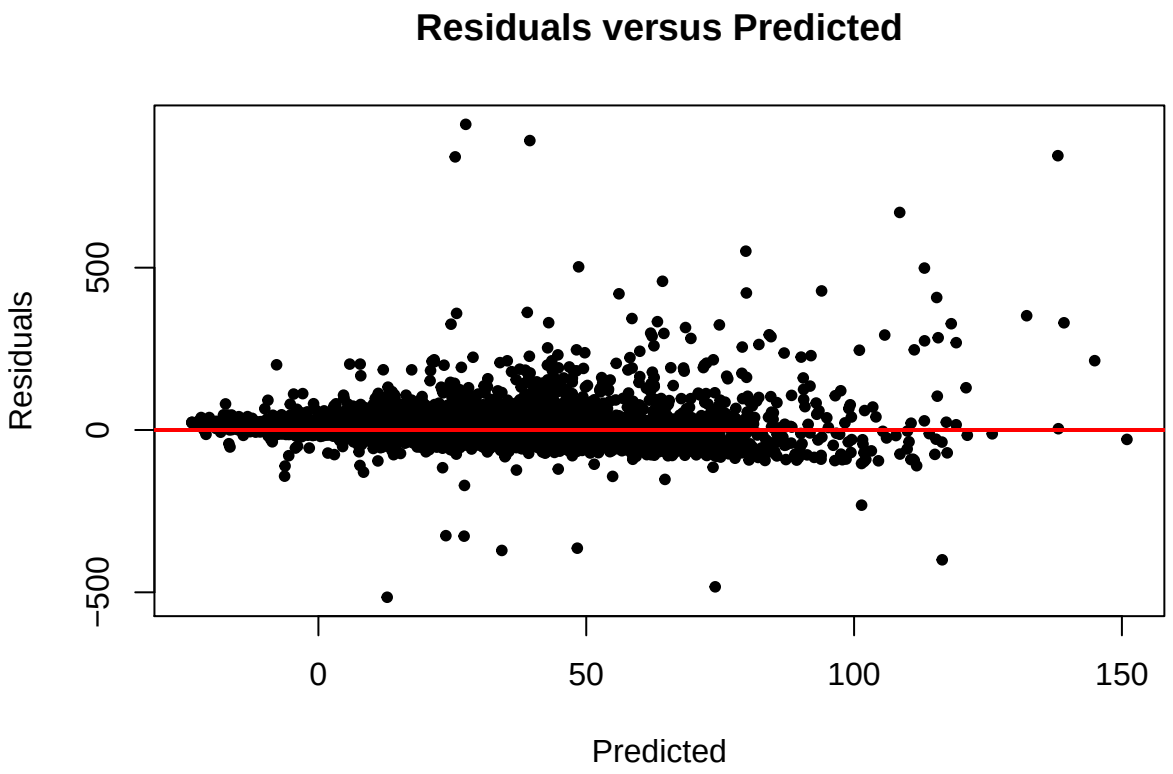
We can see that removing the outliers made a difference in the regression coefficients, while keeping all of the estimates at the same level of significance.

(d) For your model in part (c) plot the respective residuals vs. $\hat{y}$, and y vs. $\hat{y}$, and comment on your results.

```
outliers =c(6383,7755,8164,1108,1524,758,2405,1583,3442,3332)
data_new = data[-outliers,]
nettfa_rm_outliers = data_new$nettfa
plot(predict(base_rm_outliers),nettfa_rm_outliers,
     xlab = "Predicted",ylab="Actual",pch=20, main= "Actual versus Predicted")
abline(a=0,b=1,col="red",lwd=2)
```




```
plot(predict(base_rm_outliers),resid(base_rm_outliers),
     xlab = "Predicted",ylab="Residuals",pch=20,, main="Residuals versus Predicted")
abline(h=0,col="red",lwd=2)
```



In the plot for actual versus predicted values we can see the data follows the 45 degree line pretty closely with the exception of some outliers. degree line. Therefore there is high correlation between our actual and predicted values indicating that we have a model that fits the data well.

For the residuals versus predicted values, we can see that the residuals are randomly distributed around zero with fairly constant variance, one again indicating that our model fits the data well.

(e) For a more economically realistic model, the income and age variables should appear as quadratics. Re-estimate your model from part (c) including these two quadratic terms. Now, what is the estimated dollar effect of 401(k) eligibility?

```
attach(data_new)
```

```
## The following objects are masked from data:
##
##     age, agesq, e401k, fsize, inc, incsq, male, marr, nettfa,
##     p401k, pira
```

```
sq_mod = lm(nettfa~inc+age+e401k+I(age^2)+I(inc^2),data=data_new)
S(sq_mod)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2), data
##          = data_new)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.4533895  7.7190798   1.872   0.0612 .
## inc          0.0390611  0.0586660   0.666   0.5055
## age         -1.7768807  0.3744244  -4.746 2.11e-06 ***
```

```
## e401k        9.5321731  0.9893787   9.635  < 2e-16 ***
## I(age^2)     0.0319773  0.0042987   7.439 1.11e-13 ***
## I(inc^2)     0.0065045  0.0004666  13.939  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.24 on 9259 degrees of freedom
## Multiple R-squared: 0.2354
## F-statistic: 570.1 on 5 and 9259 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96524.29 96574.22
```

From this model, we can see that being eligible for a 401k will actually increase net financial assets by about $9532.17. Note that income is no longer statistically significant.

(f) For the model estimated in part (e), add the interactions e401k(age-41) and e401k(age-41)^2. Note that the average age in the sample is about 41, so that in the new model, the coefficient on e401k is the estimated effect of 401(k) eligibility at the average age. Are the interaction terms significant. Would you suggest keeping one of the interactions (or both)? Explain.

```
int_mod = lm(nettfa~inc+age+e401k+I(age^2)+I(inc^2)+e401k:I(age-41)+e401k:I((age-41)^2),
             data=data_new)
S(int_mod)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##          e401k:I(age - 41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           26.6571843  9.4331544   2.826  0.00472 **
## inc                    0.0391406  0.0585426   0.669  0.50378
## age                   -2.1667573  0.4573238  -4.738 2.19e-06 ***
## e401k                  9.9840167  1.3301614   7.506 6.67e-14 ***
## I(age^2)               0.0340862  0.0052495   6.493 8.83e-11 ***
## I(inc^2)               0.0064739  0.0004657  13.901  < 2e-16 ***
## e401k:I(age - 41)      0.6349960  0.1016030   6.250 4.29e-10 ***
## e401k:I((age - 41)^2) -0.0056861  0.0090016  -0.632  0.52762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.15 on 9257 degrees of freedom
## Multiple R-squared: 0.2389
## F-statistic: 415.1 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96485.77 96549.98
```

Only the interaction term for effect of 401k eligibility at the average age is significant and I would therefore keep that term in the model. Economically, it makes sense to keep the interaction term for the effect of 401k eligibility at the average age because people will have higher wealth and net financial assets as they age if they are eligible for 401k than those that are not eligible.

(g) Comparing the estimates from parts (e) and (f), do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.

```
int_mod_avg=lm(nettfa~inc+I(age-41)+e401k+I((age-41)^2)+I(inc^2)+e401k:I(age-41)+e401k:I((age-41)^2),da
S(int_mod)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##          e401k:I(age - 41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           26.6571843  9.4331544   2.826  0.00472 **
## inc                    0.0391406  0.0585426   0.669  0.50378
## age                   -2.1667573  0.4573238  -4.738 2.19e-06 ***
## e401k                  9.9840167  1.3301614   7.506 6.67e-14 ***
## I(age^2)               0.0340862  0.0052495   6.493 8.83e-11 ***
## I(inc^2)               0.0064739  0.0004657  13.901  < 2e-16 ***
## e401k:I(age - 41)      0.6349960  0.1016030   6.250 4.29e-10 ***
## e401k:I((age - 41)^2) -0.0056861  0.0090016  -0.632  0.52762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.15 on 9257 degrees of freedom
## Multiple R-squared: 0.2389
## F-statistic: 415.1 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96485.77 96549.98
```

```
S(int_mod_avg)
```

```
## Call: lm(formula = nettfa ~ inc + I(age - 41) + e401k + I((age - 41)^2) +
##          I(inc^2) + e401k:I(age - 41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -4.8809217  1.6064426  -3.038  0.00239 **
## inc                    0.0391406  0.0585426   0.669  0.50378
## I(age - 41)            0.6283131  0.0598533  10.498  < 2e-16 ***
## e401k                  9.9840167  1.3301614   7.506 6.67e-14 ***
## I((age - 41)^2)        0.0340862  0.0052495   6.493 8.83e-11 ***
## I(inc^2)               0.0064739  0.0004657  13.901  < 2e-16 ***
## I(age - 41):e401k      0.6349960  0.1016030   6.250 4.29e-10 ***
## e401k:I((age - 41)^2) -0.0056861  0.0090016  -0.632  0.52762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.15 on 9257 degrees of freedom
## Multiple R-squared: 0.2389
## F-statistic: 415.1 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96485.77 96549.98
```

```
sq_mod_avg = lm(nettfa~inc+I(age-41)+e401k+I((age-41)^2)+I(inc^2),data=data_new)
S(sq_mod)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2), data
##          = data_new)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.4533895  7.7190798   1.872   0.0612 .
```

```
## inc          0.0390611  0.0586660   0.666   0.5055
## age         -1.7768807  0.3744244  -4.746 2.11e-06 ***
## e401k        9.5321731  0.9893787   9.635  < 2e-16 ***
## I(age^2)     0.0319773  0.0042987   7.439 1.11e-13 ***
## I(inc^2)     0.0065045  0.0004666  13.939  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.24 on 9259 degrees of freedom
## Multiple R-squared: 0.2354
## F-statistic: 570.1 on 5 and 9259 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96524.29 96574.22
```

```
S(sq_mod_avg)
```

```
## Call: lm(formula = nettfa ~ inc + I(age - 41) + e401k + I((age - 41)^2) +
##          I(inc^2), data = data_new)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4.6448272  1.5686946  -2.961  0.00307 **
## inc              0.0390611  0.0586660   0.666  0.50554
## I(age - 41)      0.8452604  0.0488290  17.311  < 2e-16 ***
## e401k            9.5321731  0.9893787   9.635  < 2e-16 ***
## I((age - 41)^2)  0.0319773  0.0042987   7.439 1.11e-13 ***
## I(inc^2)         0.0065045  0.0004666  13.939  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.24 on 9259 degrees of freedom
## Multiple R-squared: 0.2354
## F-statistic: 570.1 on 5 and 9259 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96524.29 96574.22
```

The estimates are nearly identical. This means that using average age does not actually impact the estimates for 401k eligibility.

(h) Now, drop the interaction terms from the model in part (f), but define five family size dummy variables: fsize1, fsize2, fsize3, fsize4, and fsize5. The variable fsize5 is unity for families with five or more members. Include the family size dummies in the model estimated in part (e) and make sure to choose the base group as fsize2. Comment on your estimates.

```
data_new$fsize1= 0
data_new$fsize2=0
data_new$fsize3=0
data_new$fsize4=0
data_new$fsize5=0

nn= 9265

for (i in 1:nn){
  if (data_new[i,6]==1){
    data_new[i,12]= 1
  }else if (data_new[i,6]==2){
```

```
    data_new[i,13]= 1
  }else if (data_new[i,6]==3){
    data_new[i,14]= 1
  }else if (data_new[i,6]==4){
    data_new[i,15]= 1
  }else{
    data_new[i,16]= 1
  }

}

attach(data_new)
```

```
## The following objects are masked from data_new (pos = 3):
##
##     age, agesq, e401k, fsize, inc, incsq, male, marr, nettfa,
##     p401k, pira
```

```
## The following objects are masked from data:
##
##     age, agesq, e401k, fsize, inc, incsq, male, marr, nettfa,
##     p401k, pira
```

```
dumm_mod = lm(nettfa~inc+age+e401k+I(age^2)+I(inc^2)+fsize1+fsize3+fsize4+fsize5,
              data=data_new)
S(dumm_mod)
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##          fsize1 + fsize3 + fsize4 + fsize5, data = data_new)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.1619047  7.8387305   0.914 0.360921
## inc          0.0773316  0.0593707   1.303 0.192771
## age         -1.3325053  0.3829096  -3.480 0.000504 ***
## e401k        9.2831590  0.9890674   9.386  < 2e-16 ***
## I(age^2)     0.0266316  0.0044124   6.036 1.64e-09 ***
## I(inc^2)     0.0062908  0.0004683  13.433  < 2e-16 ***
## fsize1       1.1876093  1.4068070   0.844 0.398586
## fsize3      -3.8789581  1.4169224  -2.738 0.006201 **
## fsize4      -4.6649908  1.4111462  -3.306 0.000951 ***
## fsize5      -6.1255141  1.6099004  -3.805 0.000143 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.17 on 9255 degrees of freedom
## Multiple R-squared: 0.2381
## F-statistic: 321.4 on 9 and 9255 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96499.06 96577.53
```

We see from this that the following are not statistically significant: family size of 2, income, and family size of 1. From this model we can see that the larger your family size, the lower your net financial assets will be. In this model, an increase in age will decrease net financial assets, this may be a result of as you age, you may have more expenses or enter retirement and therefore your net financial assets begin to decrease. In our

case it is estimated that a one year increase in age will lead to a $1332.51 decrease in net financial assets. An eligibility in 401k will lead to a increase in net financial assets by about $9283.16 . Meanwhile age square will increase net financial assets by $26.63 while income squared will increase it by $6.29 . A family size of 3 will increase net financial assets by about $3000 while a family size of 4 and 5 will increase net financial assets by about $2000 and $1000 respectively.

The F-statistic is significant and the R-squared is not very high.

(i) Now, do a Chow test for the unrestricted model vs. restricted models for fsize (one for each fsize) variable. For example, the first restricted model tested would be $\theta_1 = 0$, and so on. Also, perform a Chow test but this time on all the fsize variables jointly, i.e., $H_0 : \theta_1 = 0, \theta_2 = 0, \theta_3 = 0, \theta_4 = 0, \theta_5 = 0$.

```
unr_mod = lm(nettfa~inc+I(inc^2)+age+I(age^2)+e401k+e401k:I(age-41)+e401k:I((age-41)^2)+
               fsize1+fsize2+fsize3+fsize4+fsize5,data=data_new)
summary(unr_mod)
```

```
##
## Call:
## lm(formula = nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k +
##     e401k:I(age - 41) + e401k:I((age - 41)^2) + fsize1 + fsize2 +
##     fsize3 + fsize4 + fsize5, data = data_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -515.89  -15.06   -2.63    6.01  942.99
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           13.4547029  9.7420931   1.381 0.167285
## inc                    0.0771527  0.0592442   1.302 0.192852
## I(inc^2)               0.0062607  0.0004674  13.395  < 2e-16 ***
## age                   -1.7361162  0.4631453  -3.749 0.000179 ***
## I(age^2)               0.0288917  0.0053282   5.422 6.03e-08 ***
## e401k                  9.7886701  1.3293912   7.363 1.95e-13 ***
## fsize1                 7.3819618  1.6218706   4.552 5.39e-06 ***
## fsize2                 6.2385977  1.6066452   3.883 0.000104 ***
## fsize3                 2.3479071  1.6289320   1.441 0.149512
## fsize4                 1.5725482  1.6008758   0.982 0.325975
## fsize5                        NA         NA      NA       NA
## e401k:I(age - 41)      0.6386806  0.1014540   6.295 3.21e-10 ***
## e401k:I((age - 41)^2) -0.0062400  0.0089914  -0.694 0.487703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.08 on 9253 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2407
## F-statistic:   268 on 11 and 9253 DF,  p-value: < 2.2e-16
```

We see from this that we fall into the dummy variable trap. And therefore I will not include fsize5. Our new unrestricted model is therefore the following.

```
unr_mod = lm(nettfa~inc+I(inc^2)+age+I(age^2)+e401k+e401k:I(age-41)+e401k:I((age-41)^2)+
               fsize1+fsize2+fsize3+fsize4,data=data_new)
summary(unr_mod)
```

```
##
## Call:
```

```
## lm(formula = nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k +
##     e401k:I(age - 41) + e401k:I((age - 41)^2) + fsize1 + fsize2 +
##     fsize3 + fsize4, data = data_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -515.89  -15.06   -2.63    6.01  942.99
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           13.4547029  9.7420931   1.381 0.167285
## inc                    0.0771527  0.0592442   1.302 0.192852
## I(inc^2)               0.0062607  0.0004674  13.395  < 2e-16 ***
## age                   -1.7361162  0.4631453  -3.749 0.000179 ***
## I(age^2)               0.0288917  0.0053282   5.422 6.03e-08 ***
## e401k                  9.7886701  1.3293912   7.363 1.95e-13 ***
## fsize1                 7.3819618  1.6218706   4.552 5.39e-06 ***
## fsize2                 6.2385977  1.6066452   3.883 0.000104 ***
## fsize3                 2.3479071  1.6289320   1.441 0.149512
## fsize4                 1.5725482  1.6008758   0.982 0.325975
## e401k:I(age - 41)      0.6386806  0.1014540   6.295 3.21e-10 ***
## e401k:I((age - 41)^2) -0.0062400  0.0089914  -0.694 0.487703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.08 on 9253 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2407
## F-statistic:   268 on 11 and 9253 DF,  p-value: < 2.2e-16
```

Now to perform the chow test:

```
hyp = c("fsize1=0")
linearHypothesis(unr_mod,hyp)
```

```
## Linear hypothesis test
##
## Hypothesis:
## fsize1 = 0
##
## Model 1: restricted model
## Model 2: nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k + e401k:I(age -
##     41) + e401k:I((age - 41)^2) + fsize1 + fsize2 + fsize3 +
##     fsize4
##
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   9254 18016371
## 2   9253 17976125  1     40246 20.716 5.394e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hyp = c("fsize2=0")
linearHypothesis(unr_mod,hyp)
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## fsize2 = 0
##
## Model 1: restricted model
## Model 2: nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k + e401k:I(age -
##     41) + e401k:I((age - 41)^2) + fsize1 + fsize2 + fsize3 +
##     fsize4
##
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   9254 18005417
## 2   9253 17976125  1     29292 15.078 0.0001039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hyp = c("fsize3=0")
linearHypothesis(unr_mod,hyp)
```

```
## Linear hypothesis test
##
## Hypothesis:
## fsize3 = 0
##
## Model 1: restricted model
## Model 2: nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k + e401k:I(age -
##     41) + e401k:I((age - 41)^2) + fsize1 + fsize2 + fsize3 +
##     fsize4
##
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1   9254 17980161
## 2   9253 17976125  1    4036.2 2.0776 0.1495
```

```
hyp = c("fsize4=0")
linearHypothesis(unr_mod,hyp)
```

```
## Linear hypothesis test
##
## Hypothesis:
## fsize4 = 0
##
## Model 1: restricted model
## Model 2: nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k + e401k:I(age -
##     41) + e401k:I((age - 41)^2) + fsize1 + fsize2 + fsize3 +
##     fsize4
##
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1   9254 17977999
## 2   9253 17976125  1    1874.6 0.9649  0.326
```

```
hyp = c("fsize1=0","fsize2=0","fsize3=0","fsize4=0")
linearHypothesis(unr_mod,hyp)
```

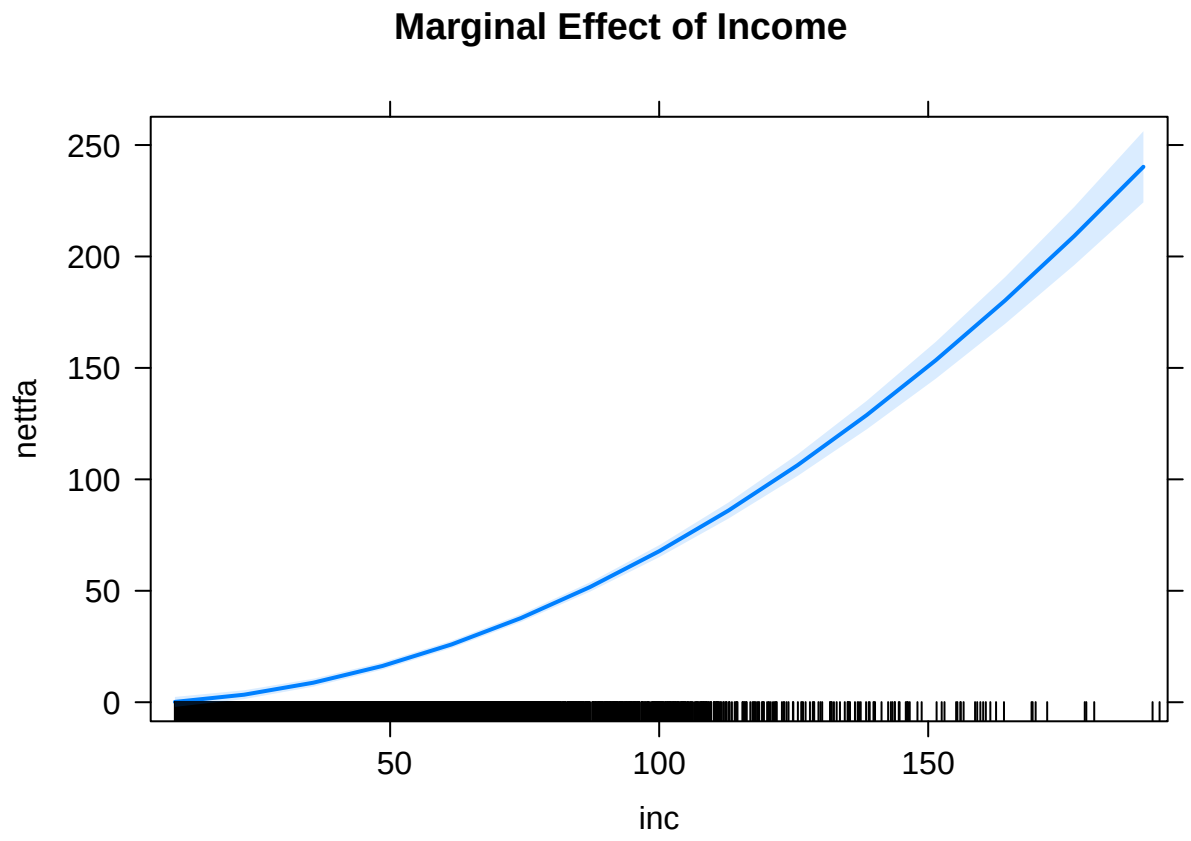
```
## Linear hypothesis test
##
## Hypothesis:
## fsize1 = 0
## fsize2 = 0
## fsize3 = 0
```

```
## fsize4 = 0
##
## Model 1: restricted model
## Model 2: nettfa ~ inc + I(inc^2) + age + I(age^2) + e401k + e401k:I(age -
##     41) + e401k:I((age - 41)^2) + fsize1 + fsize2 + fsize3 +
##     fsize4
##
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1   9257 18041509
## 2   9253 17976125  4     65384 8.4139 9.008e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

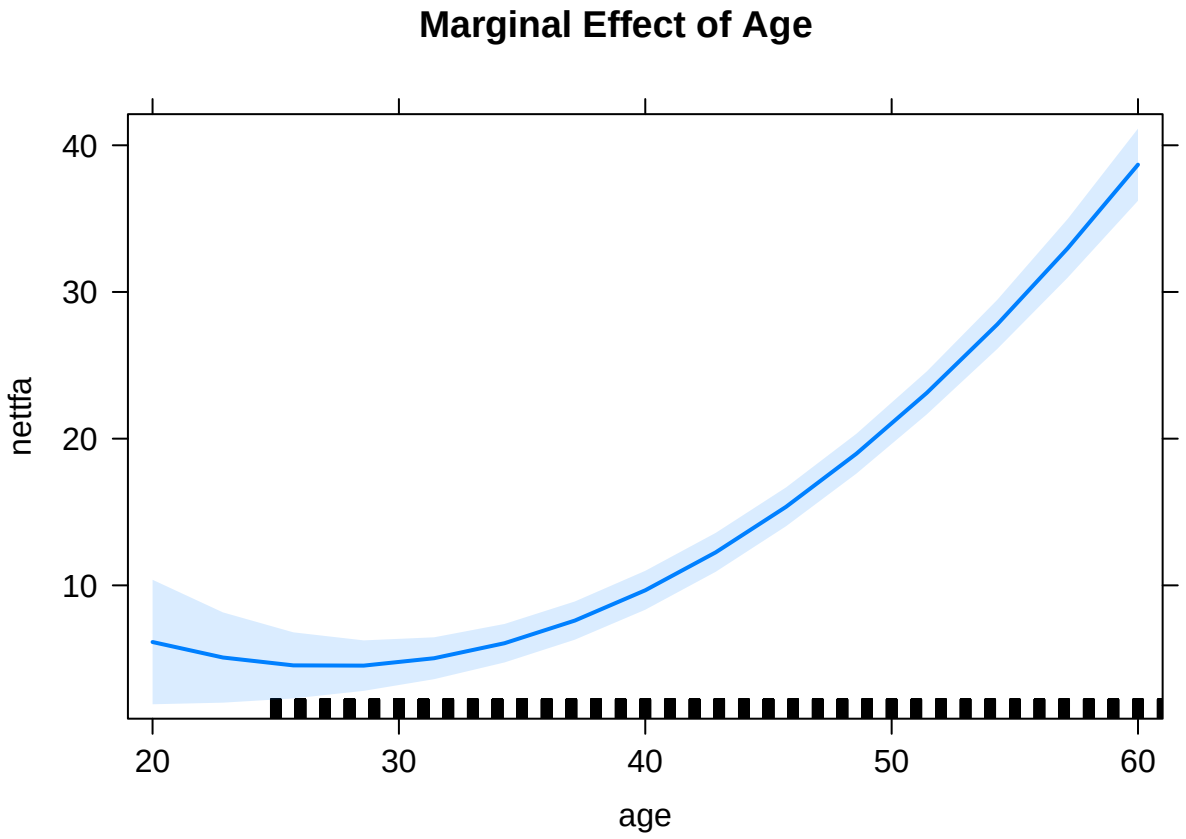
fsize1 and fsize2 are not zero but fsize3 and fsize 4 are zero. The final Chow test tells us that there are one or more of the null hypotheses is rejected.

(j) Based on your model in part (f) plot and discuss the marginal effects plots across your predictors.

```
plot(effect(mod=int_mod,"inc"),main="Marginal Effect of Income")
```

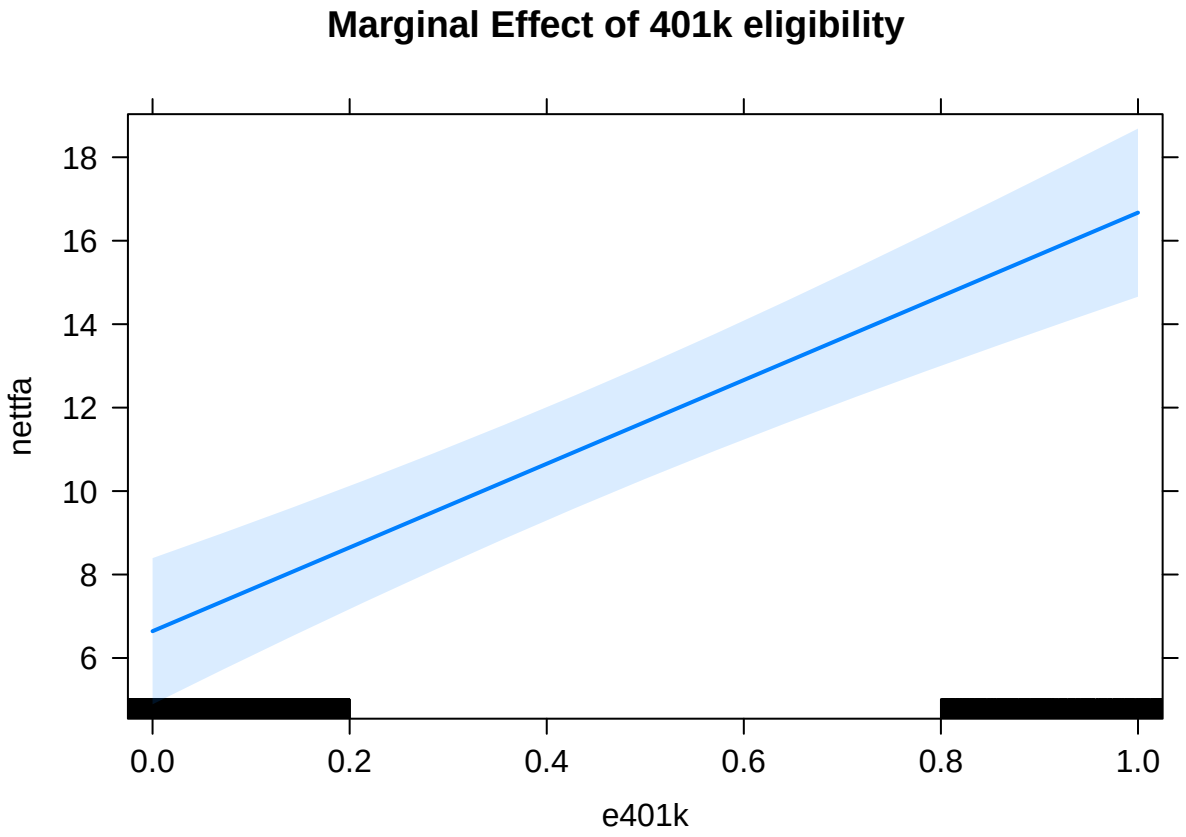



```
plot(effect(mod=int_mod,"age"), main = "Marginal Effect of Age")
```

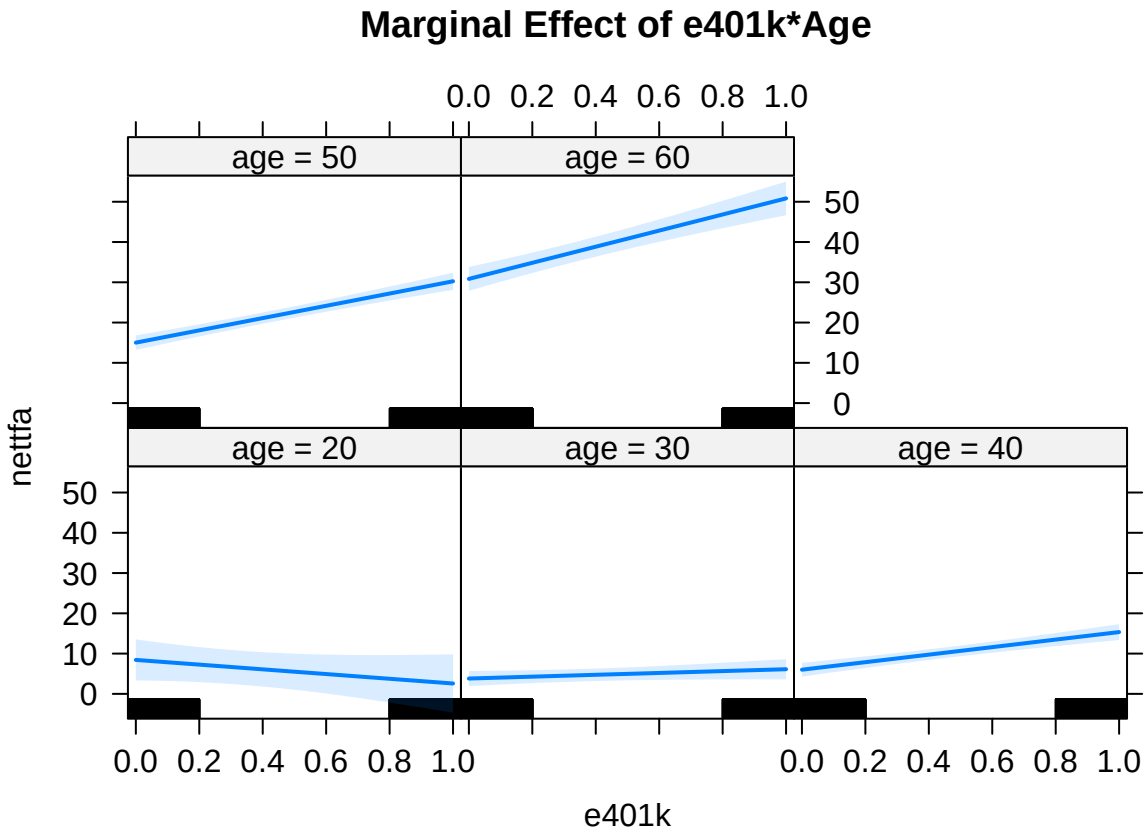



```
plot(effect(mod = int_mod,"e401k"),main="Marginal Effect of 401k eligibility")
```

```
## NOTE: e401k is not a high-order term in the model
```




```
plot(effect(mod=int_mod,"e401k:I(age-41)"),main="Marginal Effect of e401k*Age")
```



The income effect plot shows that marginal effects of income on net financial assets is increasing. Therefore a higher income is more likely to increase net financial assets than a lower income. THis is very similar to the marginal effects of age. The older an individual gets, the more likely net financial assets will increase. We can see that there is some variability at an age of around 20. This is probably because there is much more variation in spending behaviour in your 20s.

For 401k eligibility, we will look at the ends only. you can see that an eligibility in 401k will have a greater effect on net financial assets than not being eligible for a 401k

For the interaction effects plots, we can see that at different age levels, 401k eligibility will have different impacts. We can see that the older you are, the greater the effect of eligibility for 401k is on net financial assets. The difference between 20 and 30 is not very significant but the difference between 50 and 60 is very significant.

(k) Is there an optimal level of net financial wealth based on income and age? If so, compute this level and show the respective perspective and image plots.

```
deltaMethod(int_mod,"-b2/(2*b6)",parameterNames = paste("b",1:8,sep = ""))
```

```
##                Estimate       SE     2.5 %   97.5 %
## -b2/(2 * b6) -3.022955 4.726215 -12.28617 6.240255
```

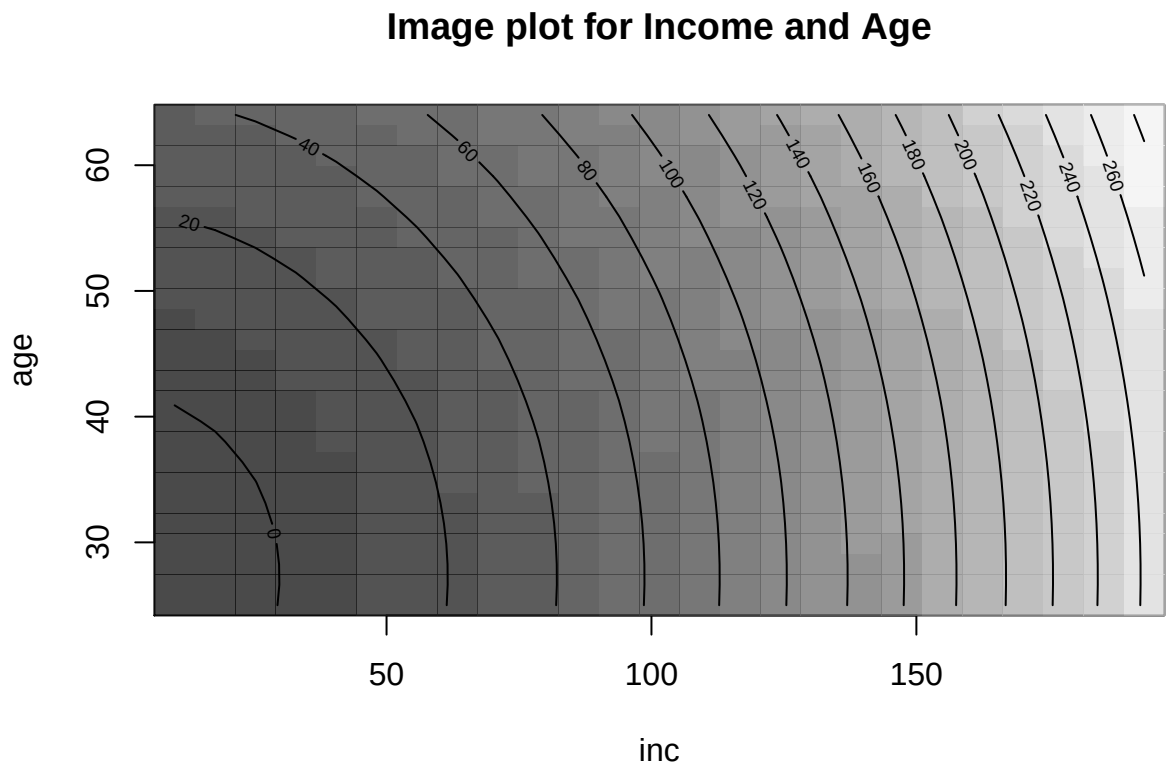
```
deltaMethod(int_mod,"-b3/(2*b5)",parameterNames = paste("b",1:8,sep = ""))
```

```
##               Estimate       SE    2.5 % 97.5 %
## -b3/(2 * b5) 31.78347 1.940611 27.97995 35.587
```

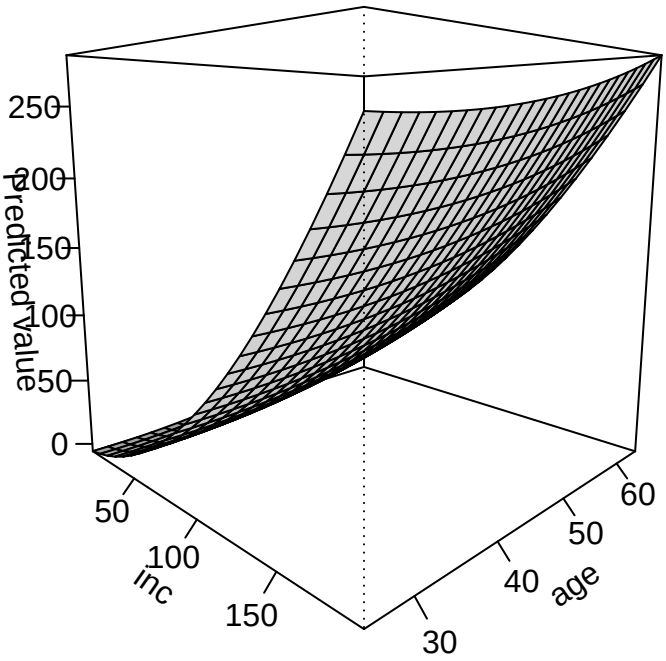
The delta method is telling us that the optimal level of income is around $3022.95 and the optimal age is about 32. From our marginal effects plots we know this is where the minimum point is. Howevever both

income and age are increasing. Therefor optimal level would be much higher than this. We can see this from the image and perspective plots.

```
image(int_mod,"inc","age",main="Image plot for Income and Age")
```




```
persp(int_mod,"inc", "age", main="Perspective plot for Income and Age")
```



From this we can see that our optimal levels the highest level of income will be achieved around the age of 60 wit an income around $200000.

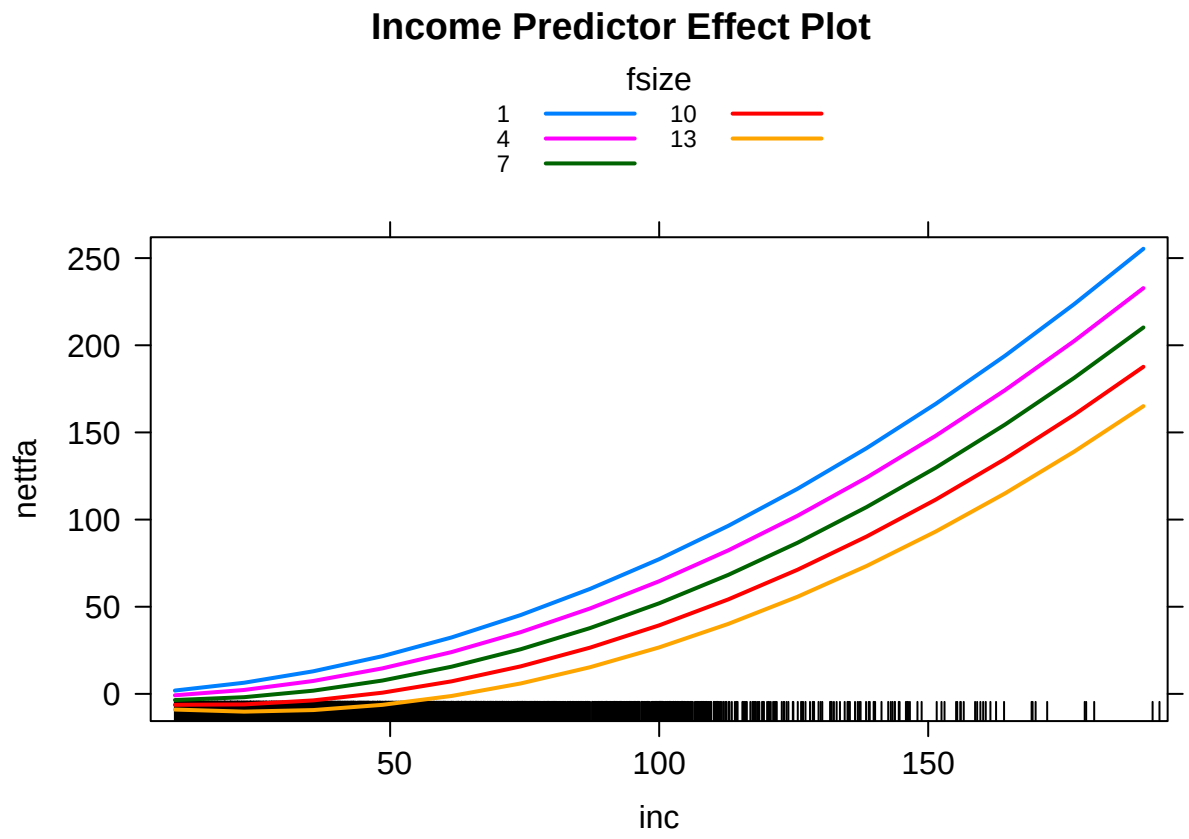
(l) For each predictor, plot the predictor effect plot by family size.

```
pred_mod <- lm(nettfa~inc+age+e401k+I(age^2)+I(inc^2)+
                 e401k:I(age-41)+e401k:I((age-41)^2)+inc:fsize + age:fsize + e401k*fsize, data = data_ne
summary(pred_mod)
```

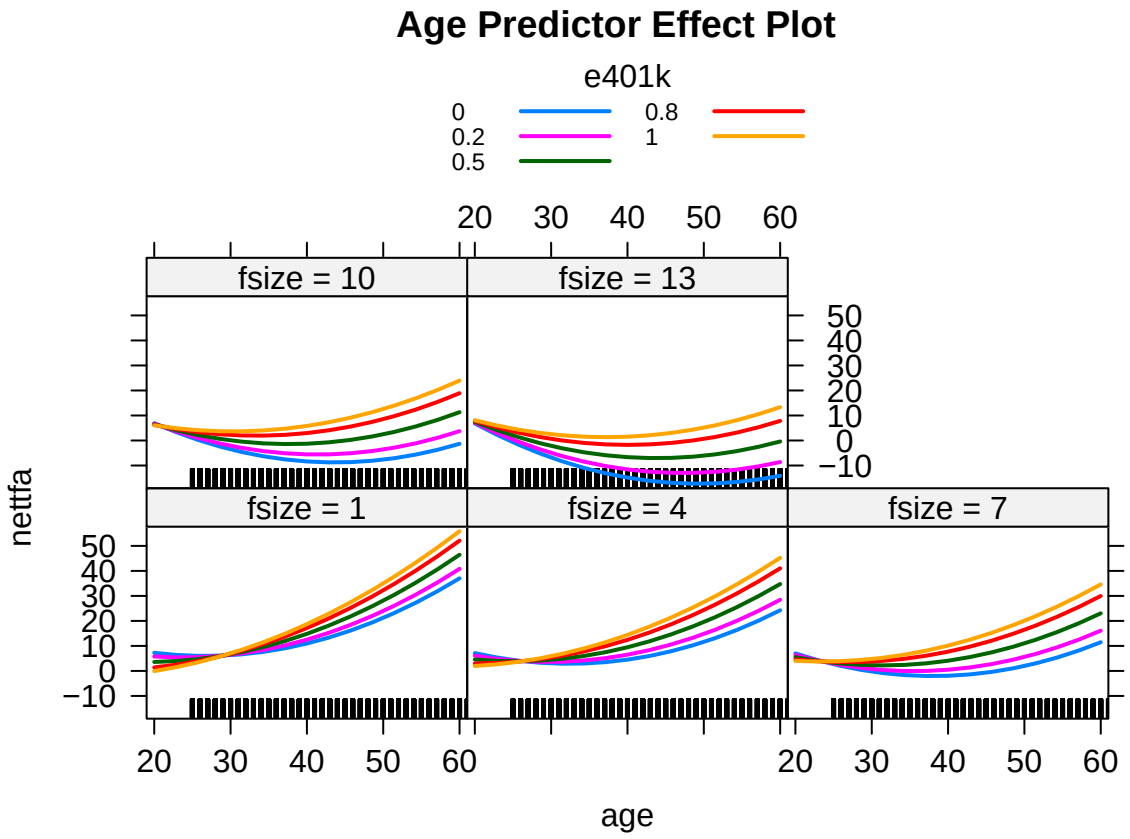
```
##
## Call:
## lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##     e401k:I(age - 41) + e401k:I((age - 41)^2) + inc:fsize + age:fsize +
##     e401k * fsize, data = data_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -516.62  -14.76   -2.53    5.64  942.62
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.1474633 10.1665567   0.703  0.48205
## inc                     0.1763061  0.0722041   2.442  0.01463 *
## age                    -1.3848408  0.4757402  -2.911  0.00361 **
## e401k                   7.6759230  2.4678567   3.110  0.00187 **
## I(age^2)                0.0279418  0.0053558   5.217 1.86e-07 ***
## I(inc^2)                0.0063423  0.0004656  13.620  < 2e-16 ***
## fsize                   3.5148244  1.3686091   2.568  0.01024 *
```

```
## e401k:I(age - 41)      0.6429682  0.1015204   6.333 2.51e-10 ***
## e401k:I((age - 41)^2) -0.0049507  0.0091867  -0.539  0.58997
## inc:fsize             -0.0367453  0.0146120  -2.515  0.01193 *
## age:fsize             -0.1056094  0.0324092  -3.259  0.00112 **
## e401k:fsize            0.7170729  0.6594177   1.087  0.27687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.03 on 9253 degrees of freedom
## Multiple R-squared:  0.2431, Adjusted R-squared:  0.2422
## F-statistic: 270.2 on 11 and 9253 DF,  p-value: < 2.2e-16
```

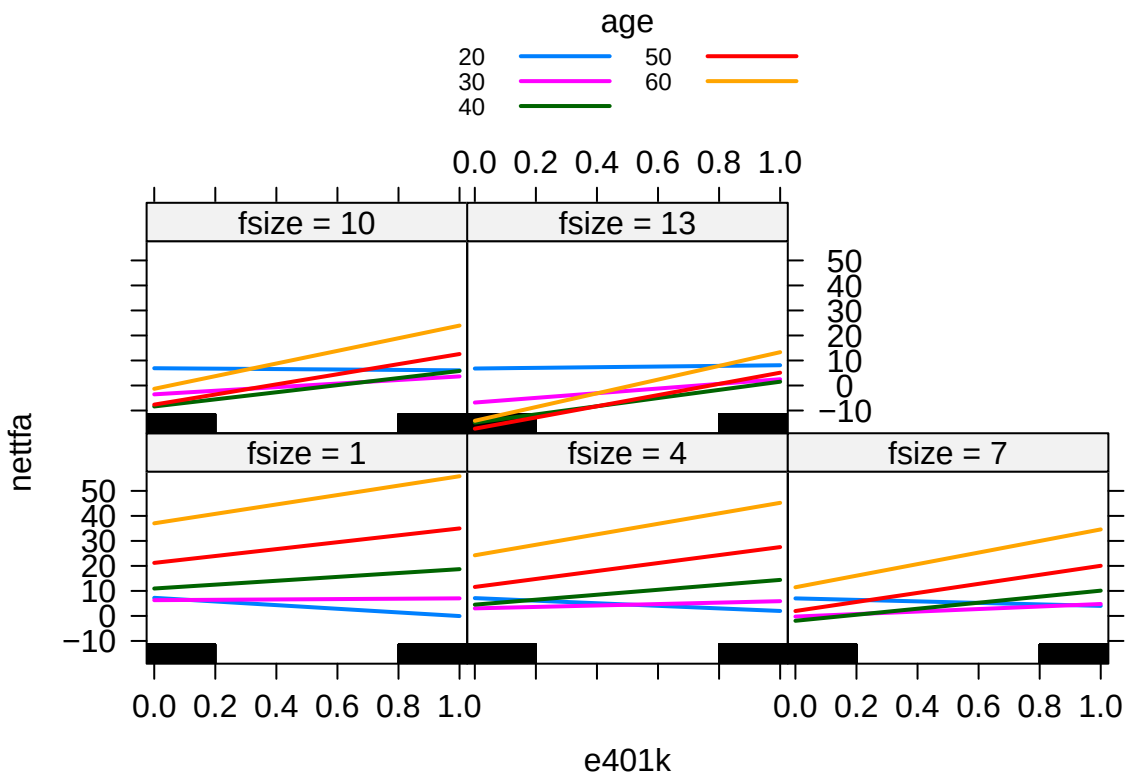
```
plot(predictorEffects(pred_mod, ~ inc), lines = list(multiline = TRUE),
     main="Income Predictor Effect Plot")
```




```
plot(predictorEffects(pred_mod, ~ age), lines = list(multiline = TRUE),
     main="Age Predictor Effect Plot")
```




```
plot(predictorEffects(pred_mod, ~ e401k), lines = list(multiline = TRUE),
     main="401k Eligibility Predictor Effect Plot")
```



For the income effect plot, we can see that a larger family size translates to a lower predictor effect from income on net financial assets.

For family sizes of 7 or less, an eligibility of 401k for ages around 30 or higher has the highest level of net financial assets. While for family sizes of greater than 7, an eligibility for 401k consistantly has a higher level of net financial assets.

For family sizes of 7 or less, the age of 60 or higher consistently has the highest level of net financial assets. However, With a family size of greater than 7, this is only true when the individuals have 401k eligibility. With large families and no 401k eligibility, 20 year-olds typically have the highest net financial assets.

(m) Estimate a multiple regression model using your transformed predictor variables from part (b) using the same model as in part (f) and compare the two models. Which one do you prefer?

```
S(int_mod)                                    #original model summary
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##          e401k:I(age - 41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            26.6571843  9.4331544   2.826  0.00472 **
## inc                     0.0391406  0.0585426   0.669  0.50378
## age                    -2.1667573  0.4573238  -4.738 2.19e-06 ***
## e401k                   9.9840167  1.3301614   7.506 6.67e-14 ***
## I(age^2)                0.0340862  0.0052495   6.493 8.83e-11 ***
## I(inc^2)                0.0064739  0.0004657  13.901  < 2e-16 ***
## e401k:I(age - 41)       0.6349960  0.1016030   6.250 4.29e-10 ***
## e401k:I((age - 41)^2) -0.0056861  0.0090016  -0.632  0.52762
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.15 on 9257 degrees of freedom
## Multiple R-squared: 0.2389
## F-statistic: 415.1 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96485.77 96549.98
```

```
trans_mod= lm(nettfa~log(inc)+log(age)+
                e401k+log(I(age^2))+log(I(inc^2))+e401k:I(age-41)+e401k:I((age-41)^2),data=data_new)
S(trans_mod)
```

```
## Call: lm(formula = nettfa ~ log(inc) + log(age) + e401k + log(I(age^2)) +
##          log(I(inc^2)) + e401k:I(age - 41) + e401k:I((age - 41)^2), data =
##          data_new)
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.932e+02  8.940e+00 -21.615  < 2e-16 ***
## log(inc)               2.720e+01  8.766e-01  31.029  < 2e-16 ***
## log(age)               3.059e+01  2.353e+00  12.999  < 2e-16 ***
## e401k                  4.537e+00  1.252e+00   3.625 0.000291 ***
## log(I(age^2))                 NA         NA      NA       NA
## log(I(inc^2))                 NA         NA      NA       NA
## e401k:I(age - 41)      5.876e-01  1.042e-01   5.638 1.77e-08 ***
## e401k:I((age - 41)^2)  3.290e-02  7.678e-03   4.284 1.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 45.95 on 9259 degrees of freedom
## Multiple R-squared: 0.1753
## F-statistic: 393.7 on 5 and 9259 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 97224.87 97274.81
```

When we use the log transformations, we get multicollinearity, so we will remove log(age squared) and log(income squared) terms.

```
trans_mod= lm(nettfa~log(inc)+log(age)+e401k+e401k:I(age-41)+e401k:I((age-41)^2),
              data=data_new)
S(trans_mod)
```

```
## Call: lm(formula = nettfa ~ log(inc) + log(age) + e401k + e401k:I(age -
##          41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.932e+02  8.940e+00 -21.615  < 2e-16 ***
## log(inc)               2.720e+01  8.766e-01  31.029  < 2e-16 ***
## log(age)               3.059e+01  2.353e+00  12.999  < 2e-16 ***
## e401k                  4.537e+00  1.252e+00   3.625 0.000291 ***
## e401k:I(age - 41)      5.876e-01  1.042e-01   5.638 1.77e-08 ***
## e401k:I((age - 41)^2)  3.290e-02  7.678e-03   4.284 1.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard deviation: 45.95 on 9259 degrees of freedom
## Multiple R-squared: 0.1753
## F-statistic: 393.7 on 5 and 9259 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 97224.87 97274.81
```

From lower AIC and BIC as well as a higer R-squared. I would still prefer the original model.

(n) Based on all the available predictors, estimate a model with additive and interactions terms, and compare it to your model in part (f).

```
S(int_mod)                          #recall the original model from part f
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) +
##          e401k:I(age - 41) + e401k:I((age - 41)^2), data = data_new)
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           26.6571843  9.4331544   2.826  0.00472 **
## inc                    0.0391406  0.0585426   0.669  0.50378
## age                   -2.1667573  0.4573238  -4.738 2.19e-06 ***
## e401k                  9.9840167  1.3301614   7.506 6.67e-14 ***
## I(age^2)               0.0340862  0.0052495   6.493 8.83e-11 ***
## I(inc^2)               0.0064739  0.0004657  13.901  < 2e-16 ***
## e401k:I(age - 41)      0.6349960  0.1016030   6.250 4.29e-10 ***
## e401k:I((age - 41)^2) -0.0056861  0.0090016  -0.632  0.52762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.15 on 9257 degrees of freedom
## Multiple R-squared: 0.2389
## F-statistic: 415.1 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96485.77 96549.98
```

```
est_mod = lm(nettfa~inc+age+e401k+I(age^2)+I(inc^2)+marr+male+p401k+pira+fsize+
               e401k:I((age-41)^2)+e401k:I(age-41),data=data_new)
S(est_mod)                                  #model with all predictors used
```

```
## Call: lm(formula = nettfa ~ inc + age + e401k + I(age^2) + I(inc^2) + marr
##          + male + p401k + pira + fsize + e401k:I((age - 41)^2) + e401k:I(age - 41),
##          data = data_new)
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           31.7729942  9.0885375   3.496 0.000475 ***
## inc                   -0.1818787  0.0599674  -3.033 0.002428 **
## age                   -1.9982517  0.4459383  -4.481 7.52e-06 ***
## e401k                 -1.6492778  1.6505907  -0.999 0.317722
## I(age^2)               0.0294317  0.0051285   5.739 9.83e-09 ***
## I(inc^2)               0.0068190  0.0004554  14.975  < 2e-16 ***
## marr                  -2.8357075  1.2549595  -2.260 0.023869 *
## male                   0.8509337  1.1909453   0.715 0.474934
## p401k                 16.3609953  1.5456522  10.585  < 2e-16 ***
## pira                  29.6974015  1.1157217  26.617  < 2e-16 ***
```

```
## fsize                 -0.5376169  0.3691500  -1.456 0.145326
## e401k:I((age - 41)^2) -0.0080442  0.0085954  -0.936 0.349363
## e401k:I(age - 41)      0.6067882  0.0970291   6.254 4.19e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 42.15 on 9252 degrees of freedom
## Multiple R-squared: 0.3066
## F-statistic: 340.9 on 12 and 9252 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 95632.82 95732.69
```

```
est_mod1 = lm(nettfa~inc+age+I(age^2)+I(inc^2)+marr+p401k+pira+e401k:I(age-41),data=data_new)
S(est_mod1)                                    #removed insignificant terms
```

```
## Call: lm(formula = nettfa ~ inc + age + I(age^2) + I(inc^2) + marr + p401k
##          + pira + e401k:I(age - 41), data = data_new)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       29.765294   7.469032   3.985 6.80e-05 ***
## inc               -0.184494   0.059130  -3.120  0.00181 **
## age               -1.947260   0.357826  -5.442 5.41e-08 ***
## I(age^2)           0.028835   0.004093   7.045 1.99e-12 ***
## I(inc^2)           0.006856   0.000452  15.168  < 2e-16 ***
## marr              -4.112107   0.992427  -4.143 3.45e-05 ***
## p401k             14.370749   1.025838  14.009  < 2e-16 ***
## pira              29.900373   1.111516  26.901  < 2e-16 ***
## e401k:I(age - 41)  0.569257   0.089796   6.339 2.41e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 42.16 on 9256 degrees of freedom
## Multiple R-squared: 0.3061
## F-statistic: 510.3 on 8 and 9256 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 95631.56 95702.90
```

```
est_mod2 = lm(nettfa~inc+age+I(age^2)+I(inc^2)+marr+e401k+p401k:I(age-41),data=data_new)
S(est_mod2)             #testing if p401k has more of an effect than e401k
```

```
## Call: lm(formula = nettfa ~ inc + age + I(age^2) + I(inc^2) + marr + e401k
##          + p401k:I(age - 41), data = data_new)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       25.3818367  7.7524692   3.274  0.00106 **
## inc                0.1327499  0.0612264   2.168  0.03017 *
## age               -2.0580322  0.3733589  -5.512 3.64e-08 ***
## I(age^2)           0.0326704  0.0042765   7.639 2.40e-14 ***
## I(inc^2)           0.0059857  0.0004732  12.650  < 2e-16 ***
## marr              -5.0888304  1.0356085  -4.914 9.08e-07 ***
## e401k              9.1909739  0.9848858   9.332  < 2e-16 ***
## p401k:I(age - 41)  0.9345125  0.1035718   9.023  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44 on 9257 degrees of freedom
## Multiple R-squared: 0.244
## F-statistic: 426.9 on 7 and 9257 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96422.72 96486.93
```

```
est_mod3 = lm(nettfa~inc+age+I(age^2)+I(inc^2)+marr:I(age^2)+p401k+e401k:I(age-41),data=data_new)
S(est_mod3)                        #testing interaction of married and age
```

```
## Call: lm(formula = nettfa ~ inc + age + I(age^2) + I(inc^2) +
##          marr:I(age^2) + p401k + e401k:I(age - 41), data = data_new)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       23.4400823  7.7522204   3.024 0.002504 **
## inc                0.0576413  0.0598375   0.963 0.335424
## age               -2.0594162  0.3726489  -5.526 3.36e-08 ***
## I(age^2)           0.0340004  0.0042963   7.914 2.78e-15 ***
## I(inc^2)           0.0062658  0.0004659  13.448  < 2e-16 ***
## p401k             15.9874369  1.0640459  15.025  < 2e-16 ***
## I(age^2):marr     -0.0017616  0.0005135  -3.431 0.000604 ***
## e401k:I(age - 41)  0.6154259  0.0933196   6.595 4.49e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 43.8 on 9257 degrees of freedom
## Multiple R-squared: 0.2508
## F-statistic: 442.8 on 7 and 9257 DF,  p-value: < 2.2e-16
##     AIC     BIC
## 96339.0 96403.2
```

```
est_mod4 = lm(nettfa~age+I(age^2)+I(inc^2)+marr:I(age^2)+p401k+e401k:I(age-41),data=data_new)
S(est_mod4)        #testing by removing income because it is insignificant
```

```
## Call: lm(formula = nettfa ~ age + I(age^2) + I(inc^2) + marr:I(age^2) +
##          p401k + e401k:I(age - 41), data = data_new)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       23.8775008  7.7388799   3.085 0.002039 **
## age               -2.0177487  0.3701285  -5.451 5.12e-08 ***
## I(age^2)           0.0334554  0.0042589   7.855 4.43e-15 ***
## I(inc^2)           0.0066855  0.0001652  40.472  < 2e-16 ***
## p401k             16.1688965  1.0472347  15.440  < 2e-16 ***
## I(age^2):marr     -0.0016297  0.0004949  -3.293 0.000994 ***
## e401k:I(age - 41)  0.6149519  0.0933180   6.590 4.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 43.8 on 9258 degrees of freedom
## Multiple R-squared: 0.2508
## F-statistic: 516.5 on 6 and 9258 DF,  p-value: < 2.2e-16
##      AIC      BIC
```

```
## 96337.93 96395.00
```

```
est_mod5 = lm(nettfa~I(inc^2)+marr:I(age-41)+p401k+e401k:I(age-41),data=data_new)
S(est_mod5)      #testing with the interaction of marriage on average age
```

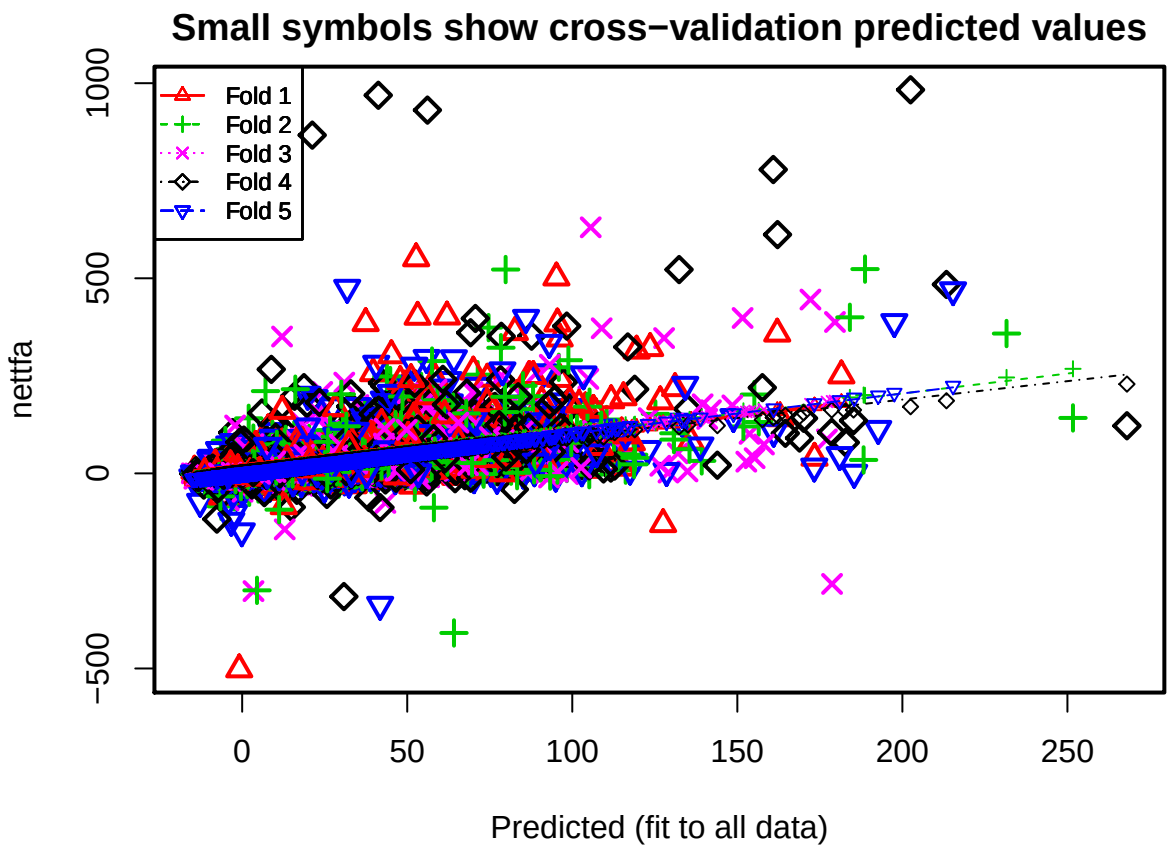
```
## Call: lm(formula = nettfa ~ I(inc^2) + marr:I(age - 41) + p401k +
##          e401k:I(age - 41), data = data_new)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.5249544  0.6024171  -0.871    0.384
## I(inc^2)           0.0064276  0.0001599  40.192   <2e-16 ***
## p401k             15.7103861  1.0524798  14.927   <2e-16 ***
## marr:I(age - 41)   0.7839734  0.0652253  12.019   <2e-16 ***
## I(age - 41):e401k  0.8585822  0.0879404   9.763   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 44.1 on 9260 degrees of freedom
## Multiple R-squared: 0.2404
## F-statistic: 732.6 on 4 and 9260 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 96461.41 96504.21
```

Therefore est_mod, est_mod1, est_mod3, and est_mod4 are all better than my model from part (f) using multiple r-squared, AIC, and BIC.

(o) Lastly, choose you favorite model from all the ones estimated and perform a five-fold cross validation test on it.

Since it seems like est_mod1 was the best model overall, I will perform the five-fold cros validation test on it.

```
cv.lm(data = data_new, form.lm = est_mod1, m = 5, plotit = TRUE, printit = FALSE)
```



Overall, it seems like Fold 5 did well but in general, this model would not do a good job given a different sample.

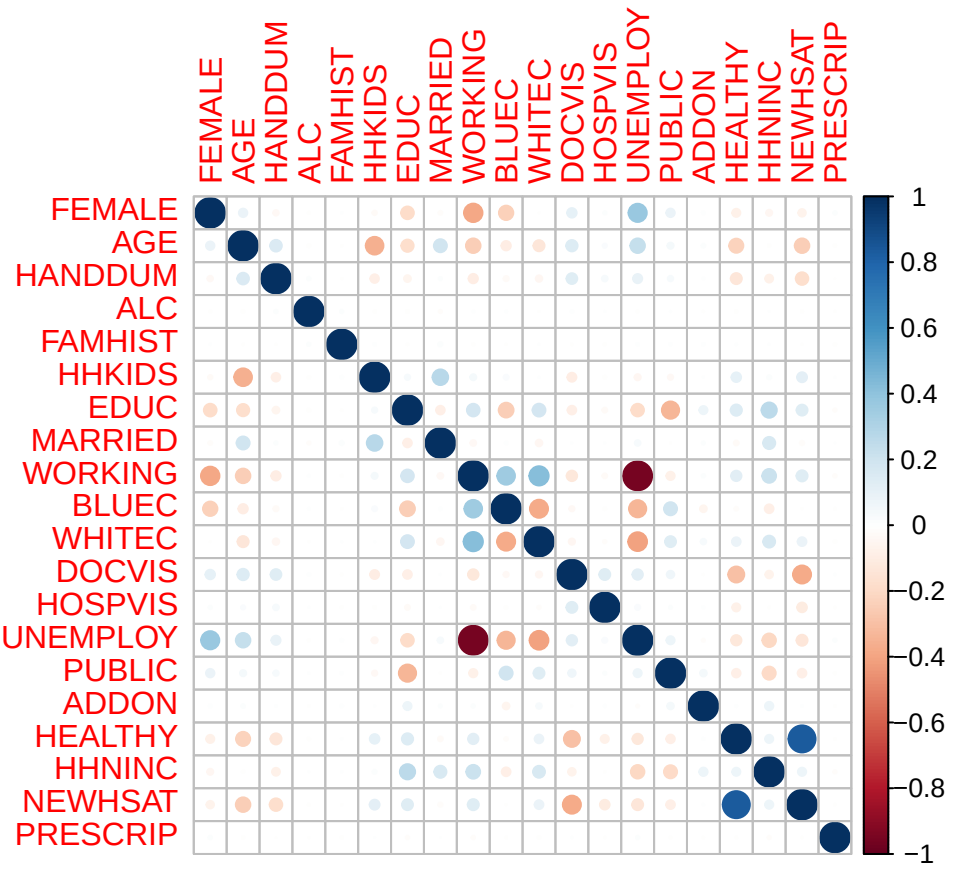
Question 2

Assume a healthcare insurance company hired you as a consultant to develop an econometric model to estimate the number of doctor visits a patient has over a 3 month period. The rational behind this study is that patients with a higher number of doctors visits wold pose a higher liability in terms of insurance expenses, and therefore, this may be mitigated via a higher insurance premium. The panel data are from the German Health Care Usage Dataset, and consist of 7,293 Individuals across varying numbers of periods with a total of 27, 326 observations.

(a) Build a multiple regression model with a subset of 10 predictors (at most), including interaction and non-linear transformations if appropriate. For this part you only need to briefly discuss a justification for the model chosen, and discuss the respective regression output.

Since there are NA values, I first omit the NA and then continue with a correlation plot to determine which variables are correlated with doctor visits in the last three months. There were too many variables to include for the document so I included a subset which I thought might be the most interesting.

```
attach(healthData)

healthData1 = na.omit(healthData)
x=c(2,4,5,6,7,9,10,11,17,18,19,22,23,24,25,26,30,41,42,43)
corrplot(cor(healthData1[,x]))
```



Based on the correlation plot, I take some of the variables I thought should be included and run a regression.

```
my.mod = lm(DOCVIS~HANDDUM+HOSPVIS+UNEMPLOY+HHNINC+EDUC+NEWHSAT+FEMALE:AGE
            +FEMALE+AGE, data=healthData1)
S(my.mod)
```

```
## Call: lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + UNEMPLOY + HHNINC + EDUC +
##           NEWHSAT + FEMALE:AGE + FEMALE + AGE, data = healthData1)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.315231   0.271346  26.959  < 2e-16 ***
## HANDDUM      0.824128   0.079299  10.393  < 2e-16 ***
## HOSPVIS      0.619944   0.035886  17.275  < 2e-16 ***
## UNEMPLOY     0.333197   0.076949   4.330 1.50e-05 ***
## HHNINC      -0.852422   0.189235  -4.505 6.68e-06 ***
## EDUC        -0.007675   0.014580  -0.526 0.598621
## NEWHSAT     -0.814910   0.014508 -56.171  < 2e-16 ***
## FEMALE       0.962192   0.252793   3.806 0.000141 ***
## AGE          0.024173   0.004012   6.025 1.71e-09 ***
## FEMALE:AGE  -0.004589   0.005628  -0.815 0.414870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.213 on 27287 degrees of freedom
## Multiple R-squared: 0.1612
## F-statistic: 582.5 on 9 and 27287 DF,  p-value: < 2.2e-16
```

```
##      AIC      BIC
## 167622.2 167712.5
```

Notice the UNEMPLOY and WORKING variables should be opposite from each other. However, Unemployment is actually much harder to measure and so I will use WORKING instead, which as we can see below does mildly improve the regression (lower AIC and BIC).

```
my.mod = lm(DOCVIS~HANDDUM+HOSPVIS+WORKING+HHNINC+EDUC+NEWHSAT+FEMALE:AGE
            +FEMALE+AGE, data=healthData1)
S(my.mod)
```

```
## Call: lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC + EDUC +
##          NEWHSAT + FEMALE:AGE + FEMALE + AGE, data = healthData1)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.654234   0.278295  27.504  < 2e-16 ***
## HANDDUM      0.822075   0.079322  10.364  < 2e-16 ***
## HOSPVIS      0.620569   0.035884  17.294  < 2e-16 ***
## WORKING     -0.342599   0.077562  -4.417 1.00e-05 ***
## HHNINC      -0.840006   0.189640  -4.429 9.48e-06 ***
## EDUC        -0.007864   0.014578  -0.539 0.589607
## NEWHSAT     -0.814790   0.014508 -56.161  < 2e-16 ***
## FEMALE       0.964607   0.252755   3.816 0.000136 ***
## AGE          0.024197   0.004010   6.034 1.62e-09 ***
## FEMALE:AGE  -0.004794   0.005631  -0.851 0.394561
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.213 on 27287 degrees of freedom
## Multiple R-squared: 0.1612
## F-statistic: 582.6 on 9 and 27287 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 167621.4 167711.8
```

Since education and the interaction term with female and age are insignificant we will take those out.

```
my.mod = lm(DOCVIS~HANDDUM+HOSPVIS+WORKING+HHNINC+NEWHSAT+FEMALE+AGE,
            data=healthData1)
S(my.mod)
```

```
## Call: lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC + NEWHSAT
##          + FEMALE + AGE, data = healthData1)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.658290   0.203767  37.584  < 2e-16 ***
## HANDDUM      0.826015   0.079194  10.430  < 2e-16 ***
## HOSPVIS      0.621099   0.035878  17.312  < 2e-16 ***
## WORKING     -0.339599   0.077382  -4.389 1.15e-05 ***
## HHNINC      -0.857668   0.183676  -4.669 3.03e-06 ***
## NEWHSAT     -0.815451   0.014471 -56.352  < 2e-16 ***
## FEMALE       0.762582   0.068787  11.086  < 2e-16 ***
## AGE          0.022114   0.002961   7.468 8.39e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard deviation: 5.213 on 27289 degrees of freedom
## Multiple R-squared: 0.1611
## F-statistic: 748.9 on 7 and 27289 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 167618.4 167692.3
```

Going one step further, we know that age likely doesn't have a linear effect on doctor visits. Therefore, economically it makes sense to include an AGE^2 term.

```
my.mod = lm(DOCVIS~HANDDUM+HOSPVIS+WORKING+HHNINC+NEWHSAT+FEMALE+AGE+
              I(AGE^2), data=healthData1)
S(my.mod)
```

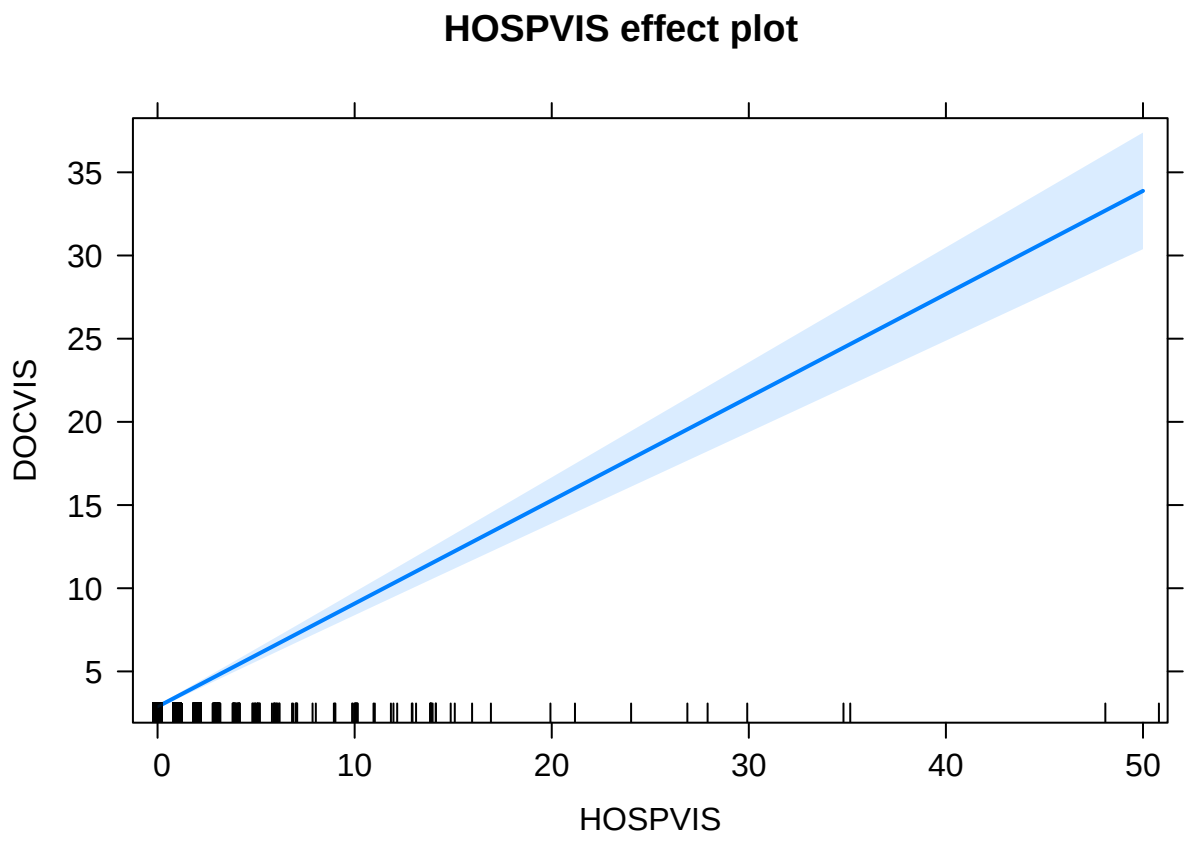
```
## Call: lm(formula = DOCVIS ~ HANDDUM + HOSPVIS + WORKING + HHNINC + NEWHSAT
##          + FEMALE + AGE + I(AGE^2), data = healthData1)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.6039407  0.5205402  20.371  < 2e-16 ***
## HANDDUM      0.8056877  0.0792098  10.172  < 2e-16 ***
## HOSPVIS      0.6201442  0.0358539  17.296  < 2e-16 ***
## WORKING     -0.2351652  0.0791733  -2.970 0.002978 **
## HHNINC      -0.6945191  0.1854603  -3.745 0.000181 ***
## NEWHSAT     -0.8178825  0.0144663 -56.537  < 2e-16 ***
## FEMALE       0.8006394  0.0690183  11.600  < 2e-16 ***
## AGE         -0.1284117  0.0246586  -5.208 1.93e-07 ***
## I(AGE^2)     0.0017207  0.0002798   6.149 7.91e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.209 on 27288 degrees of freedom
## Multiple R-squared: 0.1623
## F-statistic: 660.9 on 8 and 27288 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 167582.6 167664.8
```

Looking at the multiple R-squared and the AIC and BIC , this is the best model and I will be using this.

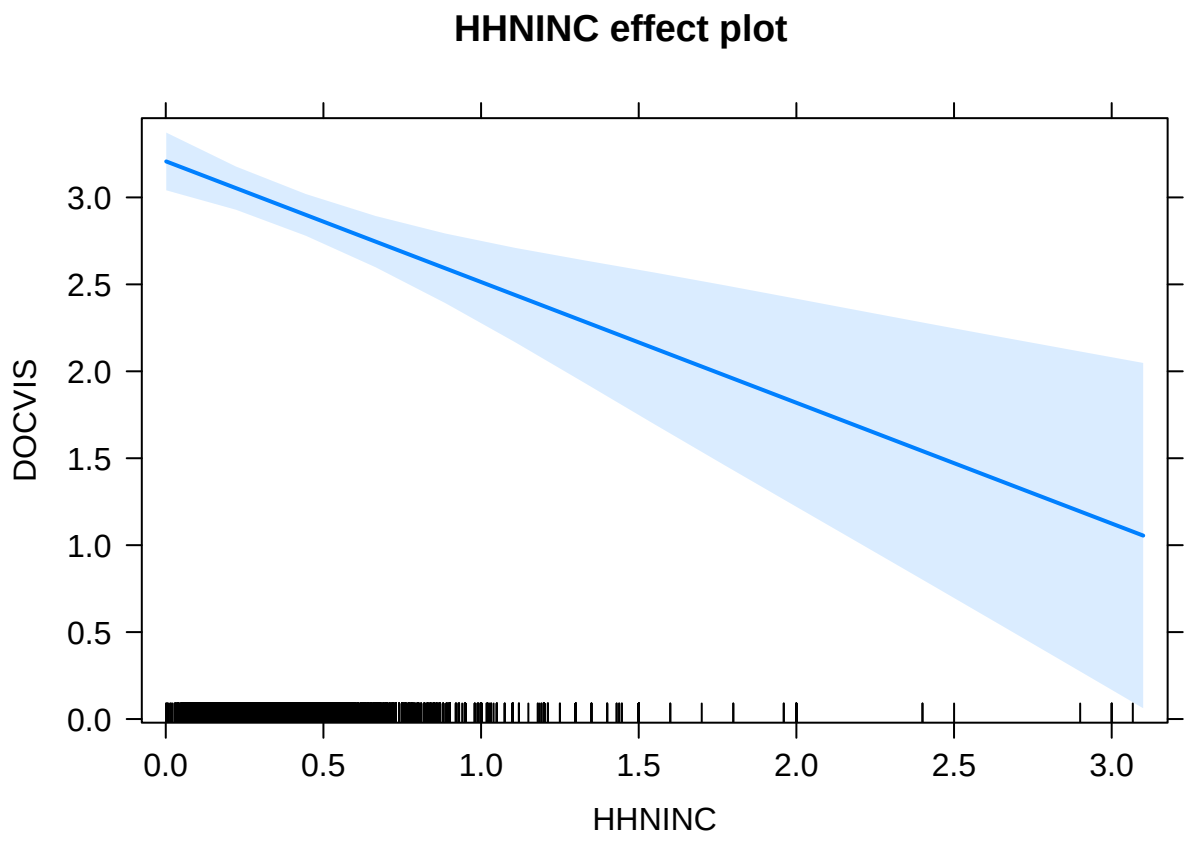
The output tells us that individuals that have a handicap, are female, or are older all tend to visit the doctor more within a three-month period.

On the other hand, working individuals with higher income and have higher levels of health satisfaction will visit the doctor less often. We can look at the marginal effects plots to understand the relationships

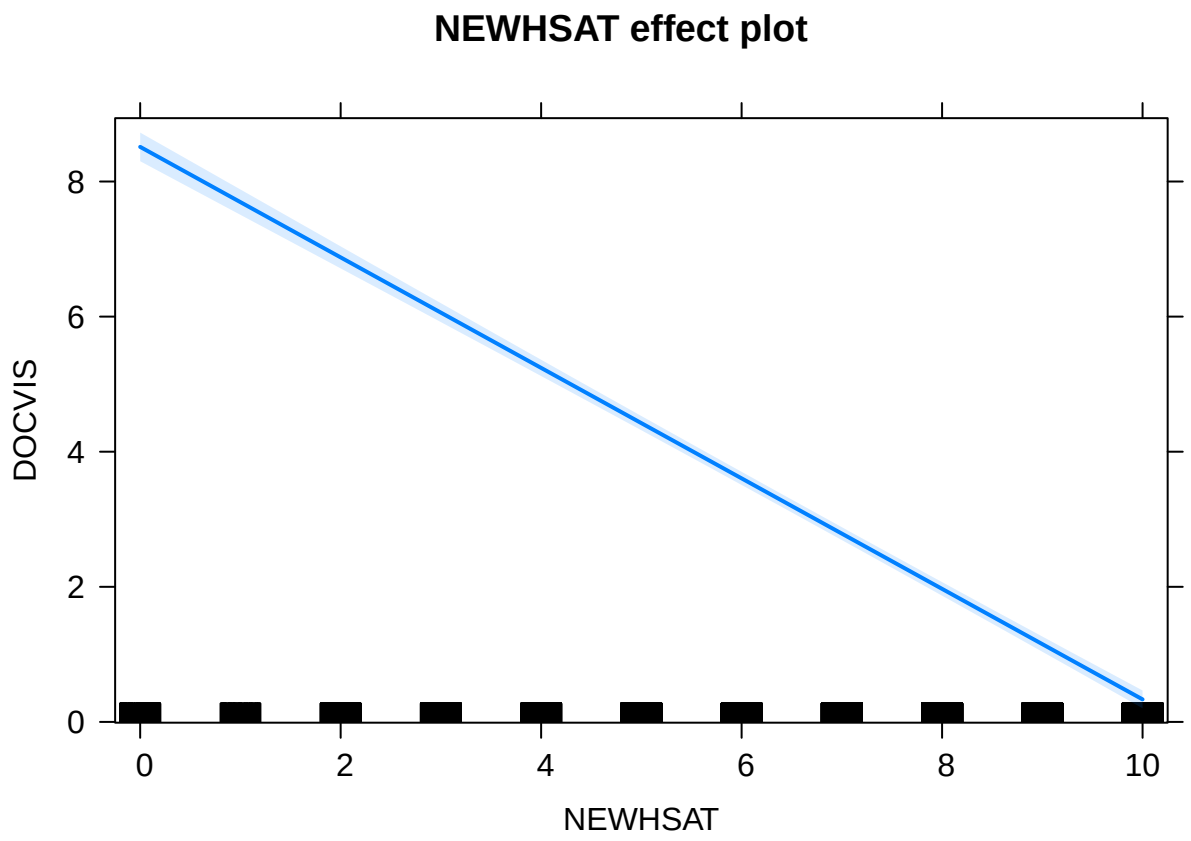
```
plot(effect(mod=my.mod,"HOSPVIS"))
```



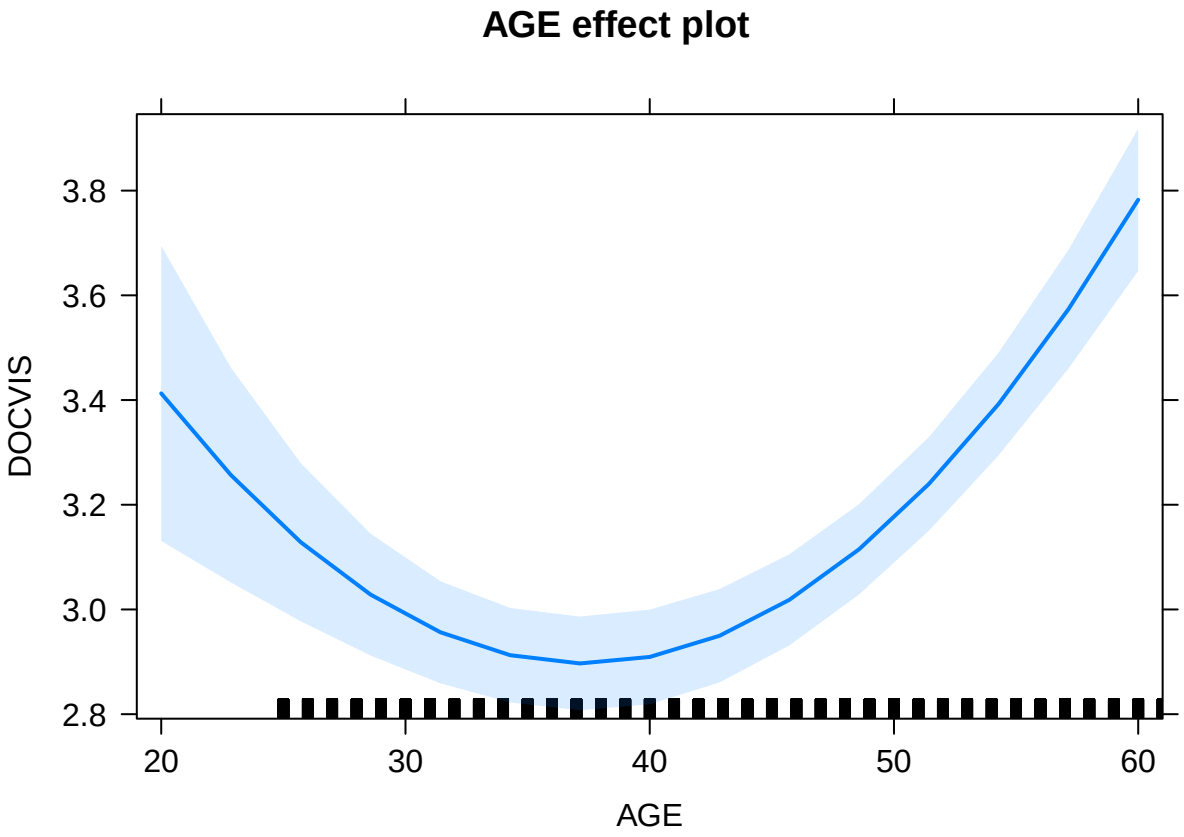

```
plot(effect(mod=my.mod,"HHNINC"))
```




```
plot(effect(mod=my.mod,"NEWHSAT"))
```




```
plot(effect(mod=my.mod,"AGE"))
```



The Marginal effects plots show us that an increase in hospital visits tends to increase with the number or doctor visits.

We can also see that doctor visits tend to decrease for those with higher incomes. This could be because they are better off and are therefore living in sanitary environments, and eating healthy food..etc. There is however, high variability amongs those with high incomes. This might be because those with higher incomes can afford to visit the doctor more. Therefore they might be less likely to ignore smaller issues.

When health satisfaction is reported at higher levels than the number of doctor visits are less frequent. This logically makes sense as individuals are less likely to go to the doctor if they are feeling healthy and satisfied with their health.

For age we can see that at younger ages and older ages, the number of doctor visits are typically higher. For older individuals, this is because age can have a lot of impacts on health. For younger individuals, there are lots of checkups that need to be done. One examply might be young women who are pregnant may visit the doctor more frequently.

To find the optimal level of age we can use the delta method.

```
deltaMethod(my.mod,"-b8/(2*b9)",parameterNames = paste("b",1:9,sep = ""))
```

```
##              Estimate      SE    2.5 %   97.5 %
## -b8/(2 * b9) 37.31352 1.35334 34.66102 39.96602
```

Therefore, the lowest level of doctorvisits will occur around the age of 37 years.

(b) Differences in Differences: In 1987 the German Government passed a series of legislations to improve healthcare access for unemployed people and women.

i. Determine whether or not the policy worked for women.

```
healthData1$TIME= ifelse(healthData1$YEAR>=1987,1,0)
healthData1$DID = healthData1$TIME * healthData1$FEMALE
attach(healthData1)
```

```
## The following objects are masked from healthData:
##
##     ABITUR, ADDON, AGE, ALC, BEAMT, BLUEC, DOCTOR, DOCVIS, EDUC,
##     FACHHS, FAMHIST, FEMALE, HANDDUM, HANDPER, HAUPTS, HEALTHY,
##     HHKIDS, HHNINC, HOSPITAL, HOSPVIS, HSAT, ID, LOGINC, MARRIED,
##     NEWHSAT, NUMOBS, PRESCRIP, PUBLIC, REALS, SELF, TI, UNEMPLOY,
##     UNIV, WHITEC, WORKING, YEAR, YEAR1984, YEAR1985, YEAR1986,
##     YEAR1987, YEAR1988, YEAR1991, YEAR1994
```

```
didreg = lm(DOCVIS~FEMALE+TIME+DID,data=healthData1)
S(didreg)
```

```
## Call: lm(formula = DOCVIS ~ FEMALE + TIME + DID, data = healthData1)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.64255    0.07335  36.026   <2e-16 ***
## FEMALE       1.27462    0.10590  12.036   <2e-16 ***
## TIME        -0.02944    0.09620  -0.306    0.760
## DID         -0.18560    0.13899  -1.335    0.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.661 on 27293 degrees of freedom
## Multiple R-squared: 0.01067
## F-statistic: 98.08 on 3 and 27293 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 172114.4 172155.5
```

Since the Difference-In-Difference estimator (DID) is insignificant, the policy did not work for women.

ii. Determine whether or not the policy worked for unemployed.

```
healthData1$DID = healthData1$TIME * healthData1$UNEMPLOY
didreg = lm(DOCVIS~UNEMPLOY+TIME+DID,data=healthData1)
S(didreg)
```

```
## Call: lm(formula = DOCVIS ~ UNEMPLOY + TIME + DID, data = healthData1)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.67609    0.06569  40.741   <2e-16 ***
## UNEMPLOY     1.63193    0.11037  14.786   <2e-16 ***
## TIME         0.04550    0.08480   0.537   0.5916
## DID         -0.25961    0.14752  -1.760   0.0784 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.648 on 27293 degrees of freedom
## Multiple R-squared: 0.01509
## F-statistic: 139.4 on 3 and 27293 DF,  p-value: < 2.2e-16
##    AIC    BIC
```

```
## 171992 172033
```

Here the DID estimator is significant to the 10% level so it somewhat worked for the unemployed.

(c) Test the hypothesis that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

```
fvis=0
mvis=0
dif.mod = lm(DOCVIS~FEMALE,data=healthData1)
S(dif.mod)
```

```
## Call: lm(formula = DOCVIS ~ FEMALE, data = healthData1)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.62543    0.04746   55.32   <2e-16 ***
## FEMALE       1.16707    0.06859   17.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.661 on 27295 degrees of freedom
## Multiple R-squared: 0.0105
## F-statistic: 289.5 on 1 and 27295 DF,  p-value: < 2.2e-16
##      AIC      BIC
## 172115.1 172139.7
```

```
for(i in 1:27297){
  if(FEMALE[i]==1){
    fvis[i]=DOCVIS[i]
  }else{
    mvis[i]=DOCVIS[i]
  }
}

fvis=na.omit(fvis)
mvis=na.omit(mvis)

t.test(fvis,mvis,alternative="greater", mu=0,paired = FALSE,var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  fvis and mvis
## t = 16.898, df = 25787, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.0532    Inf
## sample estimates:
## mean of x mean of y
##  3.792212  2.625431
```

This output tells us that we fail to accept the null that the true difference in means is not greater than zero. And therefore we conclude that the number of doctorvisits is greater for females than for males.

(d) Based on your findings propose and test your own hypothesis of interest using the linear functional form: $\lambda = c_1\beta_1 + c_2\beta_2 + ...$

The hypothesis I am interested in testing is whether an increase in age by one year and a decrease in level of health satisfaction by one unit will increase the number of doctor visits.

We start by creating a model.

```
mod.test <- lm(DOCVIS ~  AGE + NEWHSAT, data = healthData1)

summary(glht(mod.test, linfct = c("1*AGE - 1*NEWHSAT = 0")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ AGE + NEWHSAT, data = healthData1)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## 1 * AGE - 1 * NEWHSAT == 0  0.91302    0.01393   65.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The current model results in approximately 1 doctor visit every 3 months

Now we will change this slightly to incorporate a decrease in health satisfaction by two units and an inrease in age by two years.

```
mod.test <- lm(DOCVIS ~  AGE + NEWHSAT, data = healthData1)

summary(glht(mod.test, linfct = c("2*AGE - 2*NEWHSAT = 0")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = DOCVIS ~ AGE + NEWHSAT, data = healthData1)
##
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)
## 2 * AGE - 2 * NEWHSAT == 0  1.82604    0.02787   65.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

As we can see, the number of doctor visits increased to about 2 visits. We can therefore accept the hypothesis that an increase in age and decrease in health satisfaction by one unit will increase the number of doctor visits.