Project 1

Andreas Hochrein *ID: 4855928* *hochr007@umn.edu*

Due: Feb 12, 2016

1. Preparation

Setting up the environment:

Load required packages and import dataset into the global environment:

```
require(car)
```

```
## Loading required package: car

## Warning: package 'car' was built under R version 3.2.2
```

```
require(alr4)
```

```
## Loading required package: alr4
## Loading required package: effects
##
## Attaching package: 'effects'
##
## The following object is masked from 'package:car':
##
##     Prestige
```

```
fuel2001 <- read.csv("~/Desktop/STAT 3022/Projects/P1/fuel2001.csv")
```

Get an overview of the data:

```
summary(fuel2001)
```

```
##        X          Drivers             FuelC              Income
##  AK     : 1   Min.   :  328094   Min.   :  148769   Min.   :20993
##  AL     : 1   1st Qu.: 1087128   1st Qu.:  737361   1st Qu.:25323
##  AR     : 1   Median : 2718209   Median : 2048664   Median :27871
##  AZ     : 1   Mean   : 3750504   Mean   : 2542786   Mean   :28404
##  CA     : 1   3rd Qu.: 4424256   3rd Qu.: 3039932   3rd Qu.:31208
##  CO     : 1   Max.   :21623793   Max.   :14691753   Max.   :40640
##  (Other):45
##      Miles             MPC             Pop                Tax
##  Min.   :  1534   Min.   : 6556   Min.   :  381882   Min.   : 7.50
##  1st Qu.: 36586   1st Qu.: 9391   1st Qu.: 1162624   1st Qu.:18.00
##  Median : 78914   Median :10458   Median : 3115130   Median :20.00
##  Mean   : 77419   Mean   :10448   Mean   : 4257046   Mean   :20.15
##  3rd Qu.:112828   3rd Qu.:11311   3rd Qu.: 4845200   3rd Qu.:23.25
##  Max.   :300767   Max.   :17495   Max.   :25599275   Max.   :29.00
##
```

Data pre-processing

```
lapply(fuel2001, class)
```

```
## $X
## [1] "factor"
##
## $Drivers
## [1] "integer"
##
## $FuelC
## [1] "integer"
##
## $Income
## [1] "integer"
##
## $Miles
## [1] "integer"
##
## $MPC
## [1] "numeric"
##
## $Pop
## [1] "integer"
##
## $Tax
## [1] "numeric"
```

As we can see, the different variables already have the correct type. X, which gives the state from which the data is collected, is the only categorical variable in this analysis. All other variables are quantitative. State is more an identifier for the collected data here than an actual variable. Thus, it will not ne included in the model.

Prepare Data for Analysis

As described in the problem statement, we need to make the totals given by Drivers and FuelC comparable between states. This can be accomplised by making them variables that are relative to the state population. In particular, I will transform Drivers to Drivers per 100 residents and FuelC to FuelC per capita. I chose FuelC per capita instead of FuelC per Driver since we o/w would not consider the effect of fuel taxe rate on the people that chose not to drive at all. I also will replace Miles by log(Miles)

```
DriversP100 <- (fuel2001$Drivers/fuel2001$Pop)*100
FuelCPC <- (fuel2001$FuelC/fuel2001$Pop)
lnMiles <- log(fuel2001$Miles)

fuel2001b <- data.frame(fuel2001$X, DriversP100, FuelCPC, fuel2001$Income, lnMiles, fuel2001$MPC, fuel20
colnames(fuel2001b) <- c("State","DriversP100","FuelCPC","Income","lnMiles","MPC","Pop","Tax")

summary(fuel2001b)
```

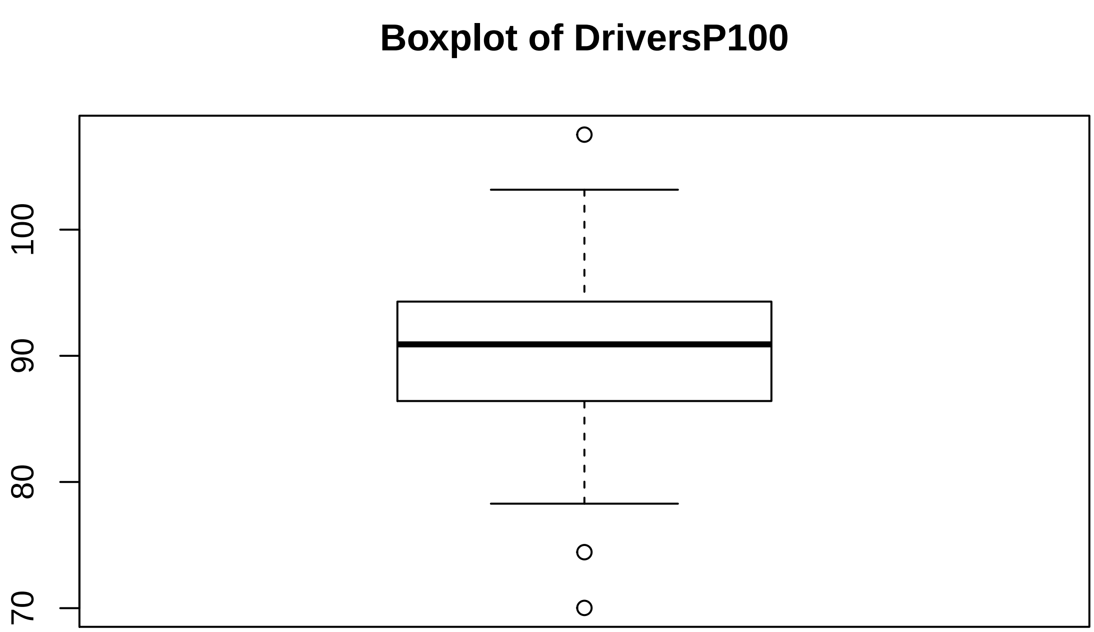
```
##      State     DriversP100        FuelCPC            Income
```

```
##  AK      : 1   Min.   : 70.02   Min.   :0.3175   Min.   :20993
##  AL      : 1   1st Qu.: 86.41   1st Qu.:0.5750   1st Qu.:25323
##  AR      : 1   Median : 90.91   Median :0.6260   Median :27871
##  AZ      : 1   Mean   : 90.37   Mean   :0.6131   Mean   :28404
##  CA      : 1   3rd Qu.: 94.30   3rd Qu.:0.6666   3rd Qu.:31208
##  CO      : 1   Max.   :107.53   Max.   :0.8428   Max.   :40640
##  (Other):45
##     lnMiles            MPC              Pop                Tax
##  Min.   : 7.336   Min.   : 6556   Min.   :  381882   Min.   : 7.50
##  1st Qu.:10.507   1st Qu.: 9391   1st Qu.: 1162624   1st Qu.:18.00
##  Median :11.276   Median :10458   Median : 3115130   Median :20.00
##  Mean   :10.914   Mean   :10448   Mean   : 4257046   Mean   :20.15
##  3rd Qu.:11.634   3rd Qu.:11311   3rd Qu.: 4845200   3rd Qu.:23.25
##  Max.   :12.614   Max.   :17495   Max.   :25599275   Max.   :29.00
##
```

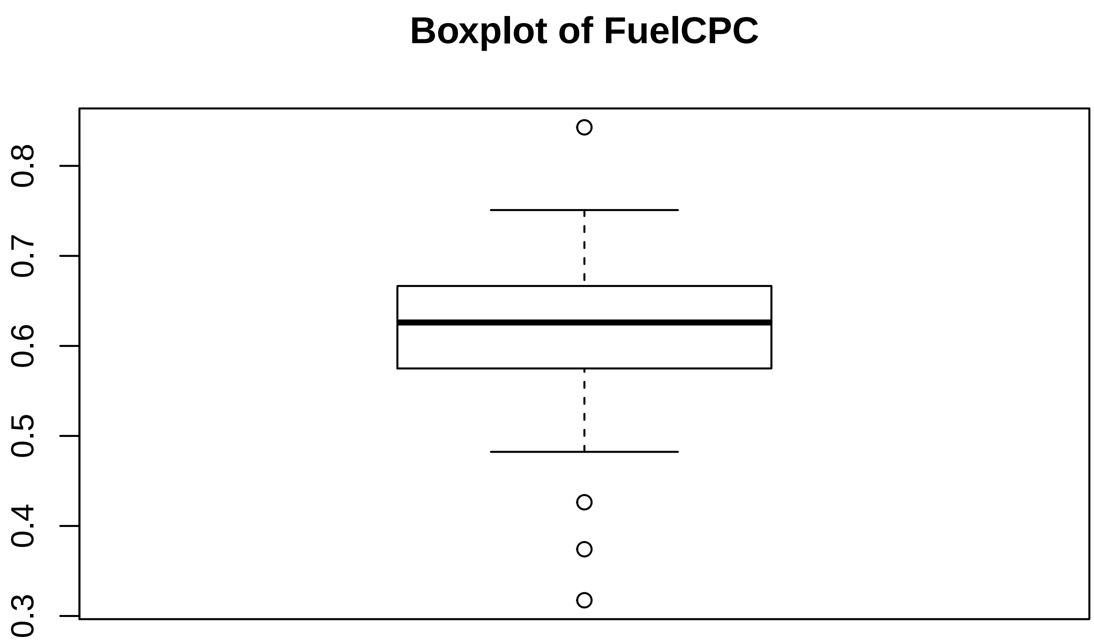
2.Graphical exploration of the data:

Boxplots

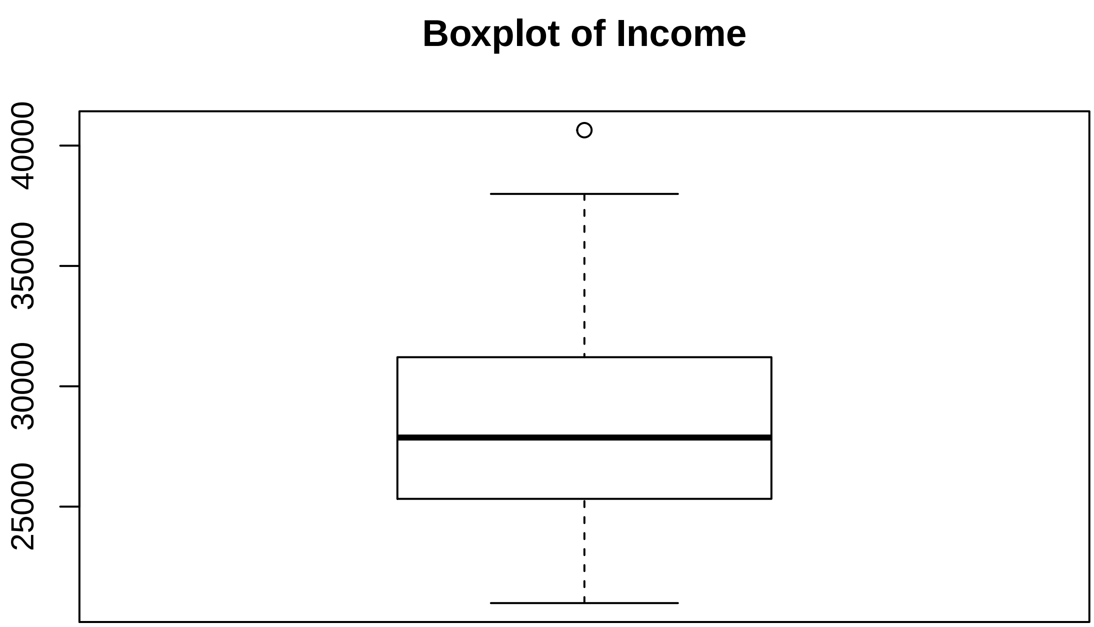
```
# We will start the analysis with a Boxplot of the quantitative variables
par(mfrow=c(1,1))
boxplot(fuel2001b$DriversP100, main="Boxplot of DriversP100")
```



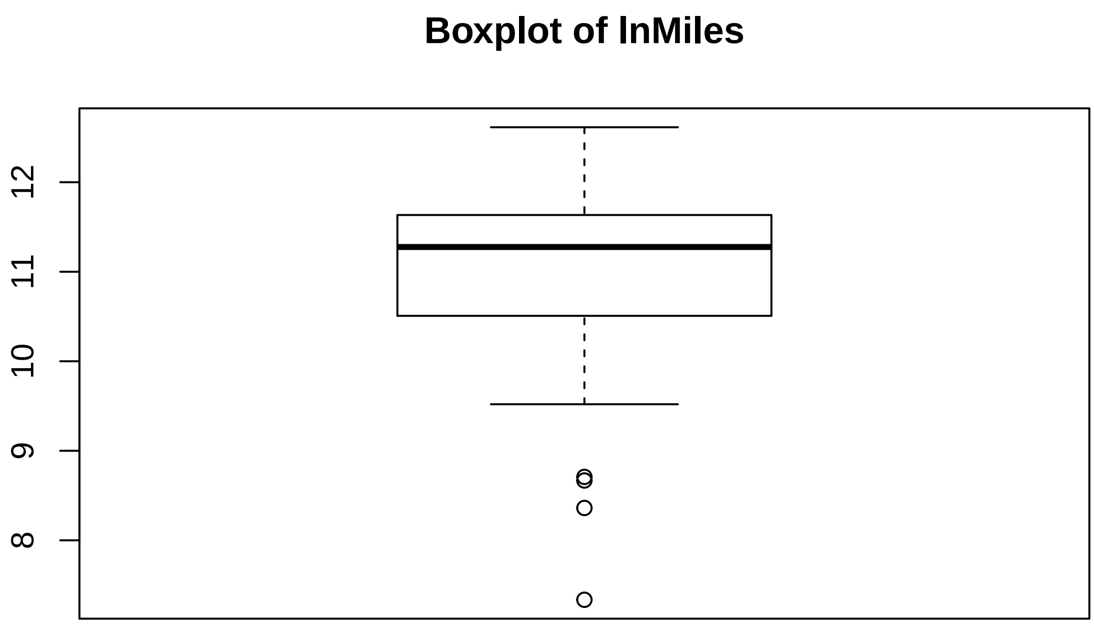

```
boxplot(fuel2001b$FuelCPC, main="Boxplot of FuelCPC")
```



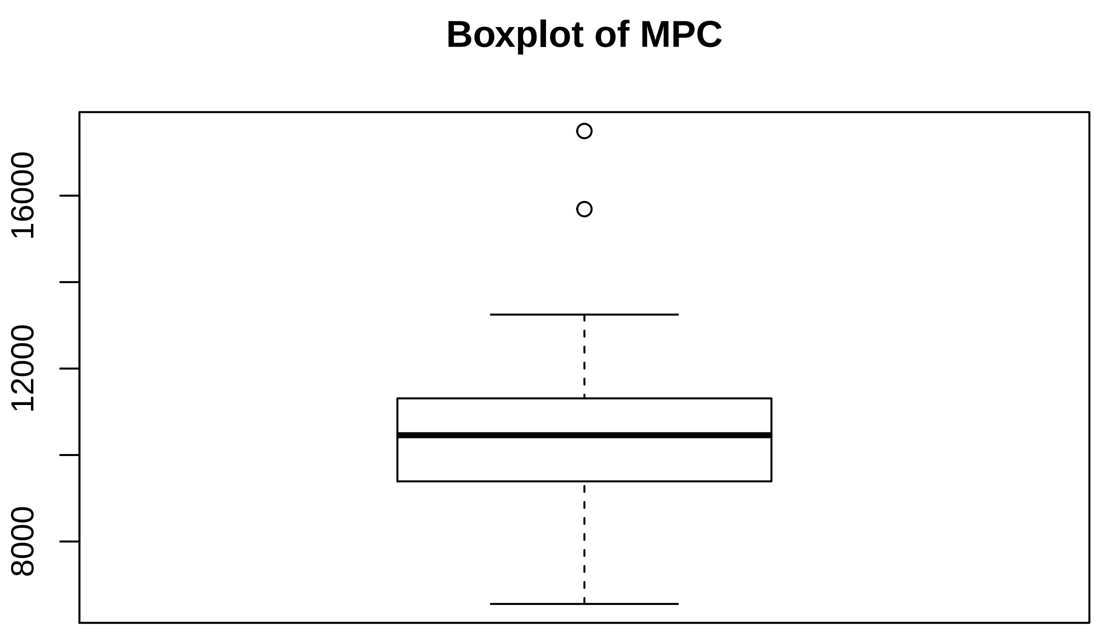

```
boxplot(fuel2001b$Income, main="Boxplot of Income")
```



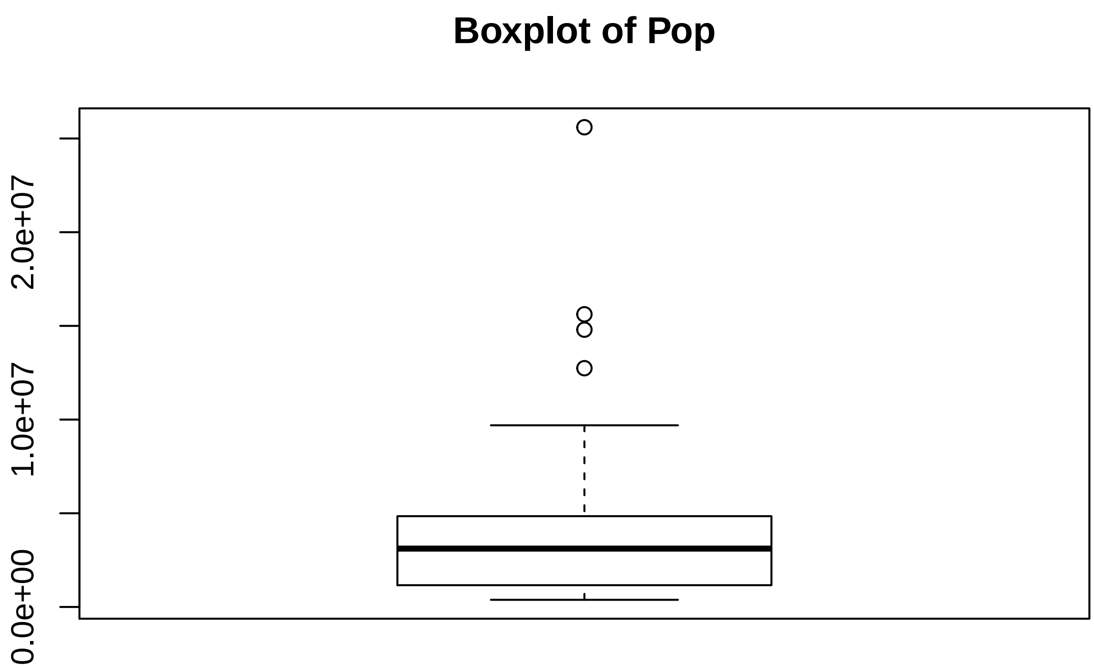

```
boxplot(fuel2001b$lnMiles, main="Boxplot of lnMiles")
```



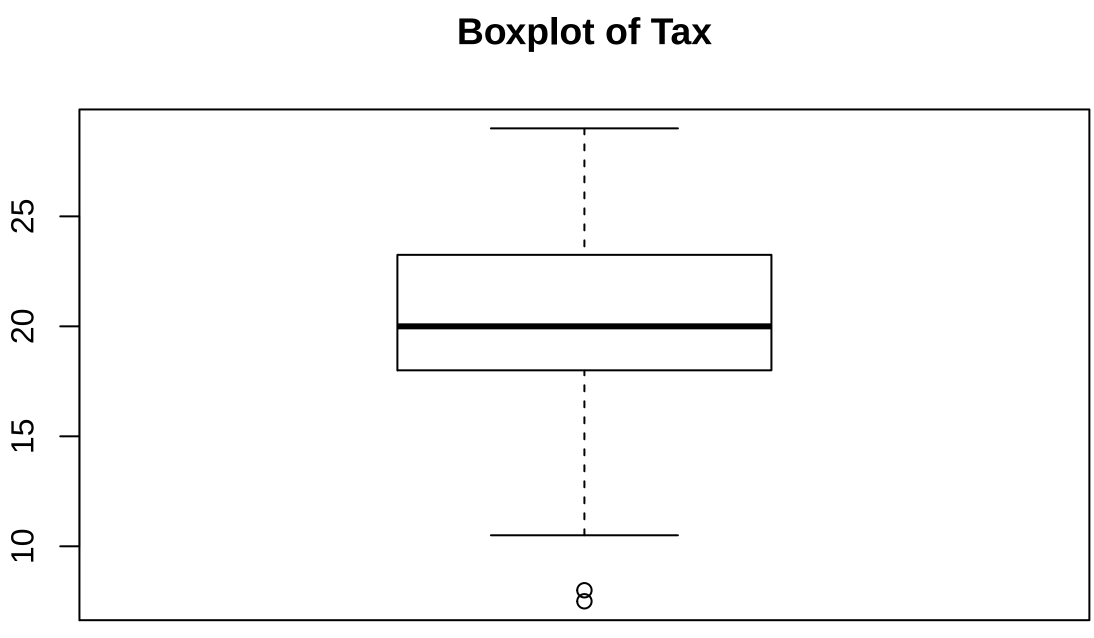

```r
boxplot(fuel2001b$MPC, main="Boxplot of MPC")
```




```r
boxplot(fuel2001b$Pop, main="Boxplot of Pop")
```




```
boxplot(fuel2001b$Tax, main="Boxplot of Tax")
```

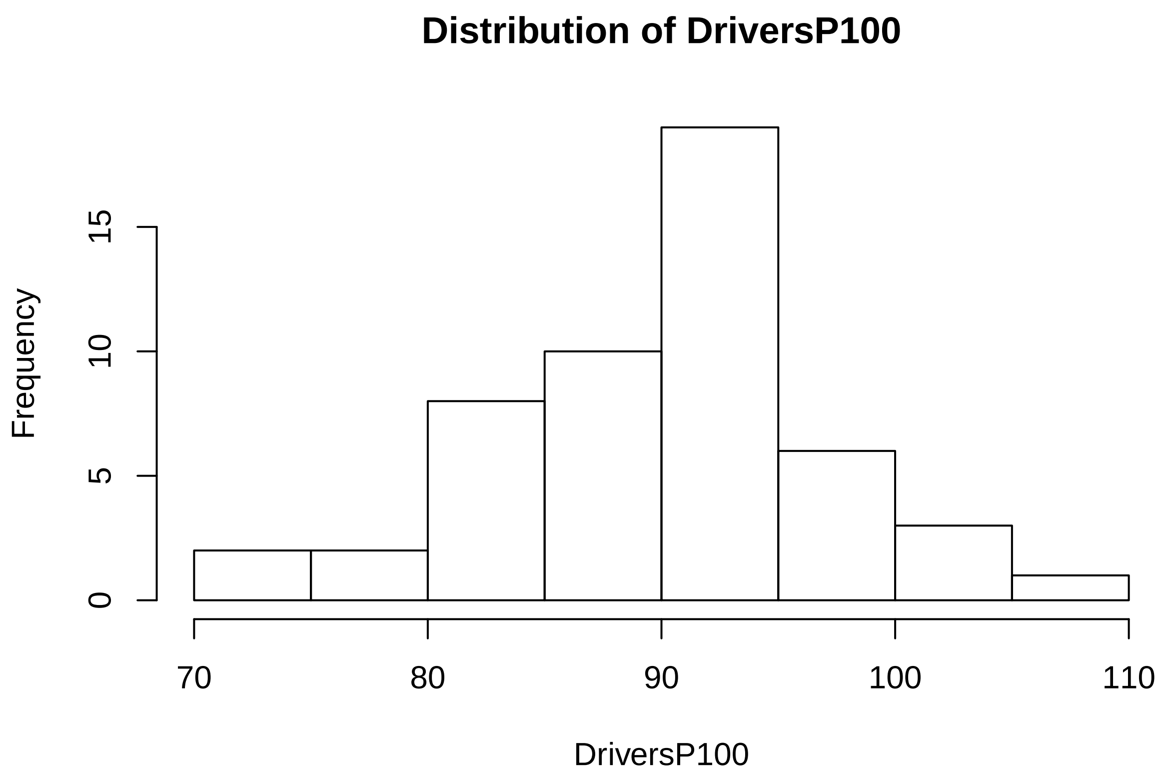


As we can see, the data for each of the variables contains outliers. This will require additional analysis.

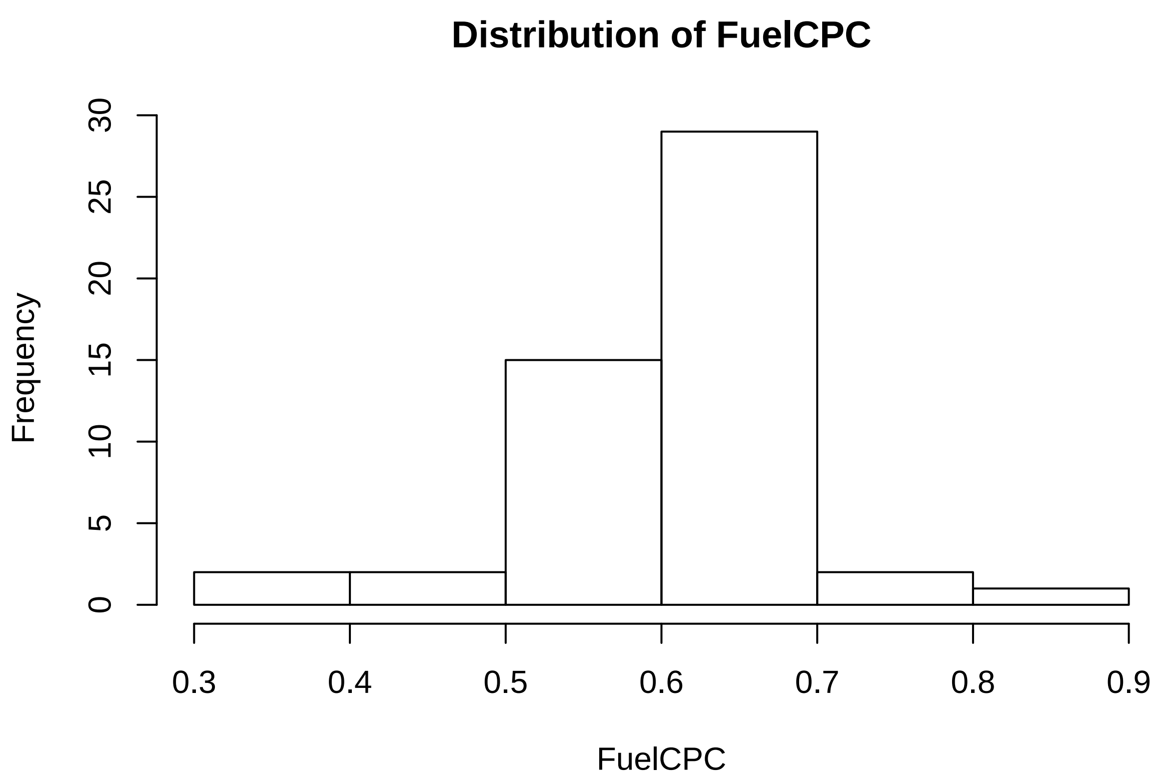
Histogram

First, though, I will create histograms for each quantitative variable to get a better understanding of the data distributions.

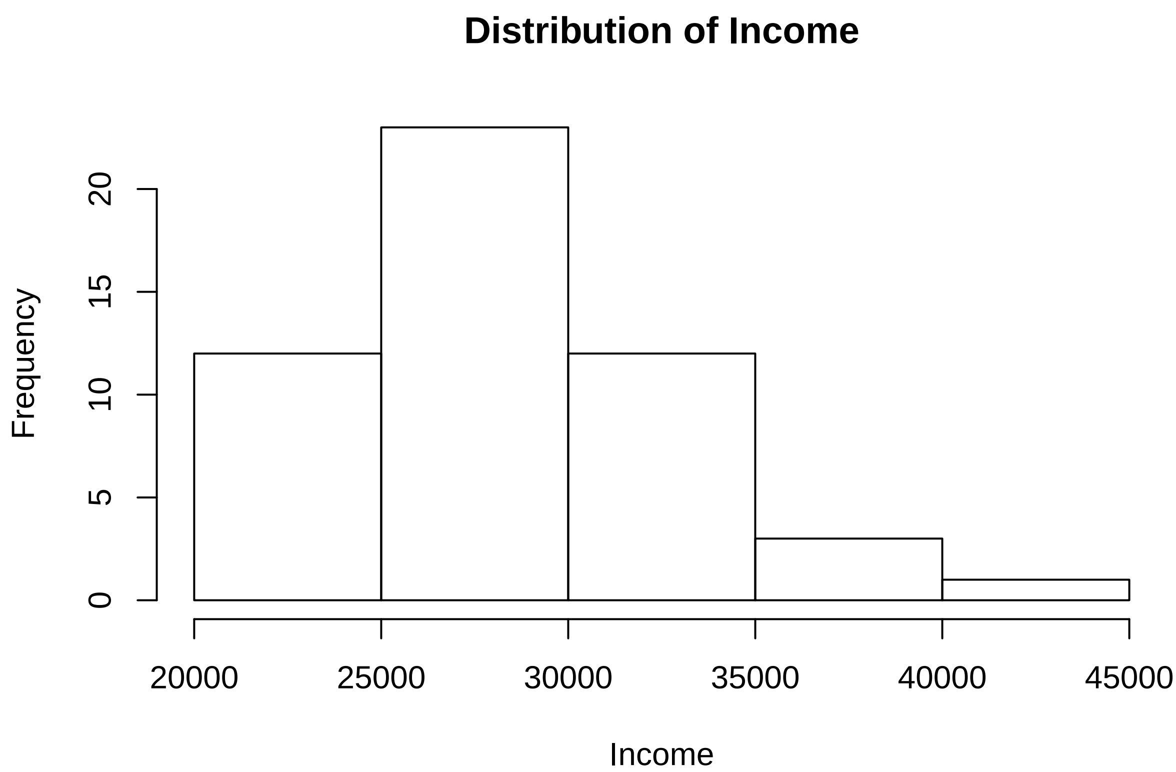
```
par(mfrow=c(1,1))
hist(fuel2001b$DriversP100,main ="Distribution of DriversP100", xlab = "DriversP100")
```



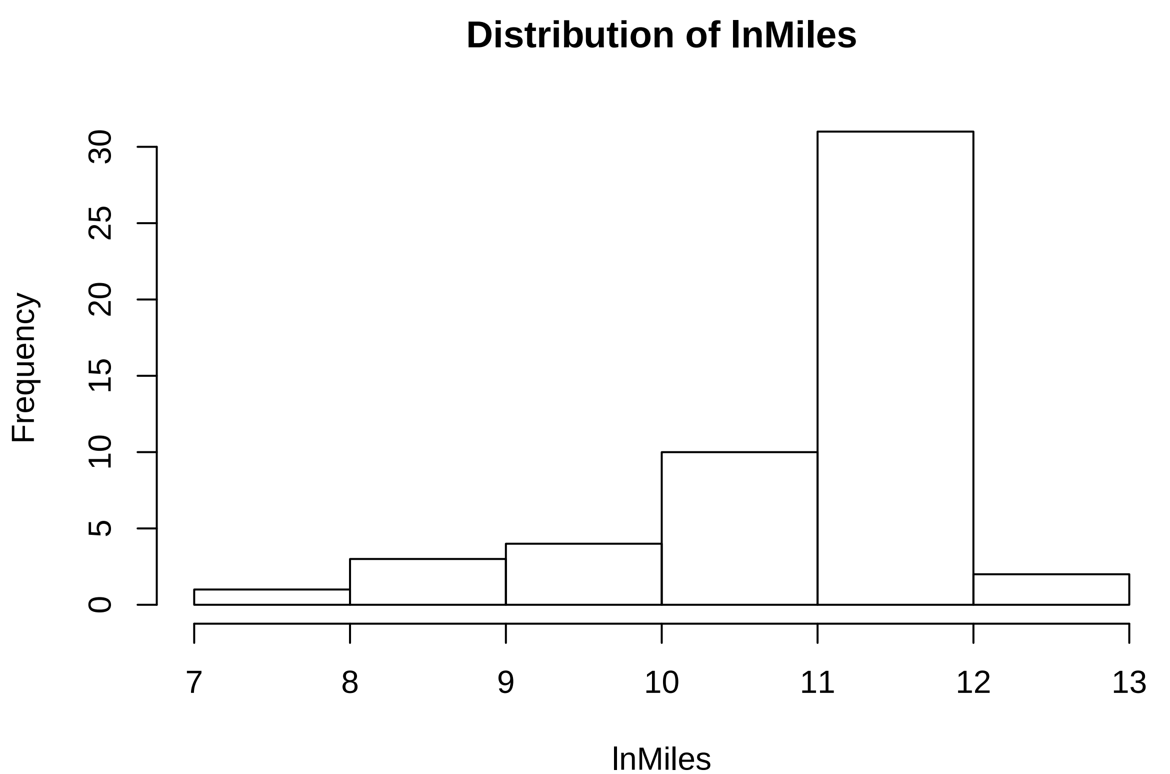

```r
hist(fuel2001b$FuelCPC,main ="Distribution of FuelCPC", xlab = "FuelCPC")
```



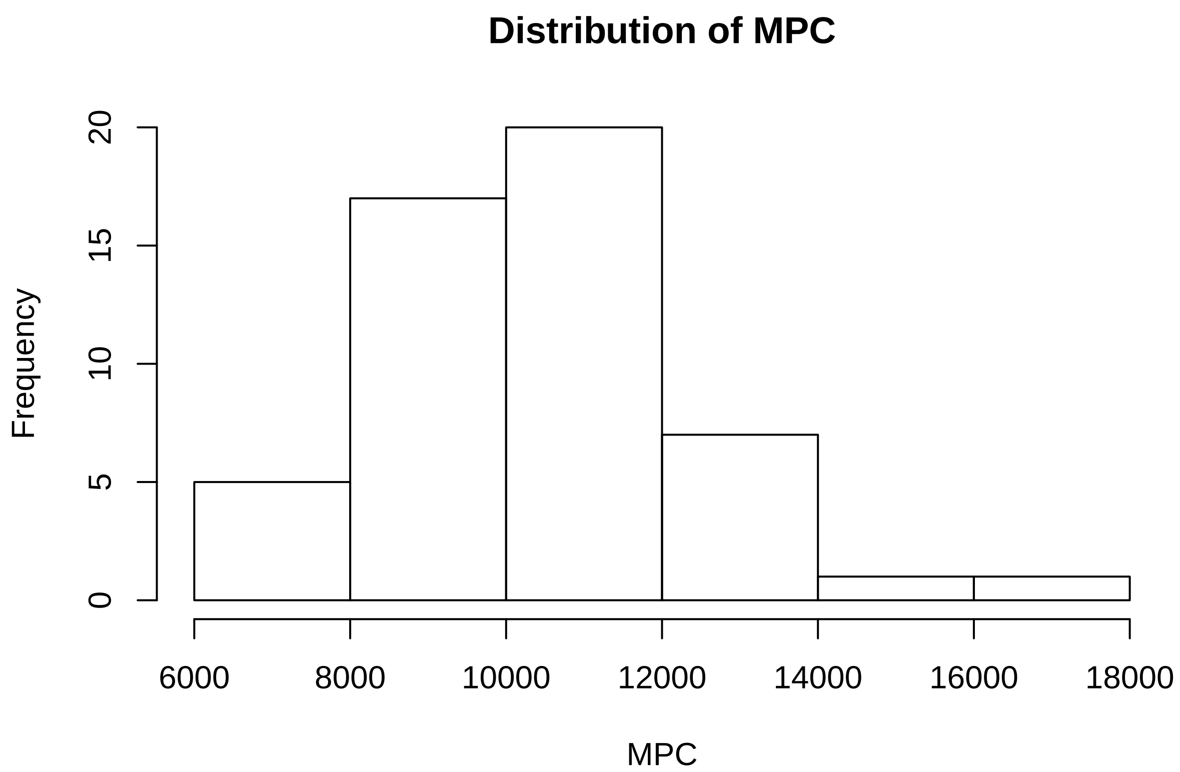

```r
hist(fuel2001b$Income,main ="Distribution of Income", xlab = "Income")
```



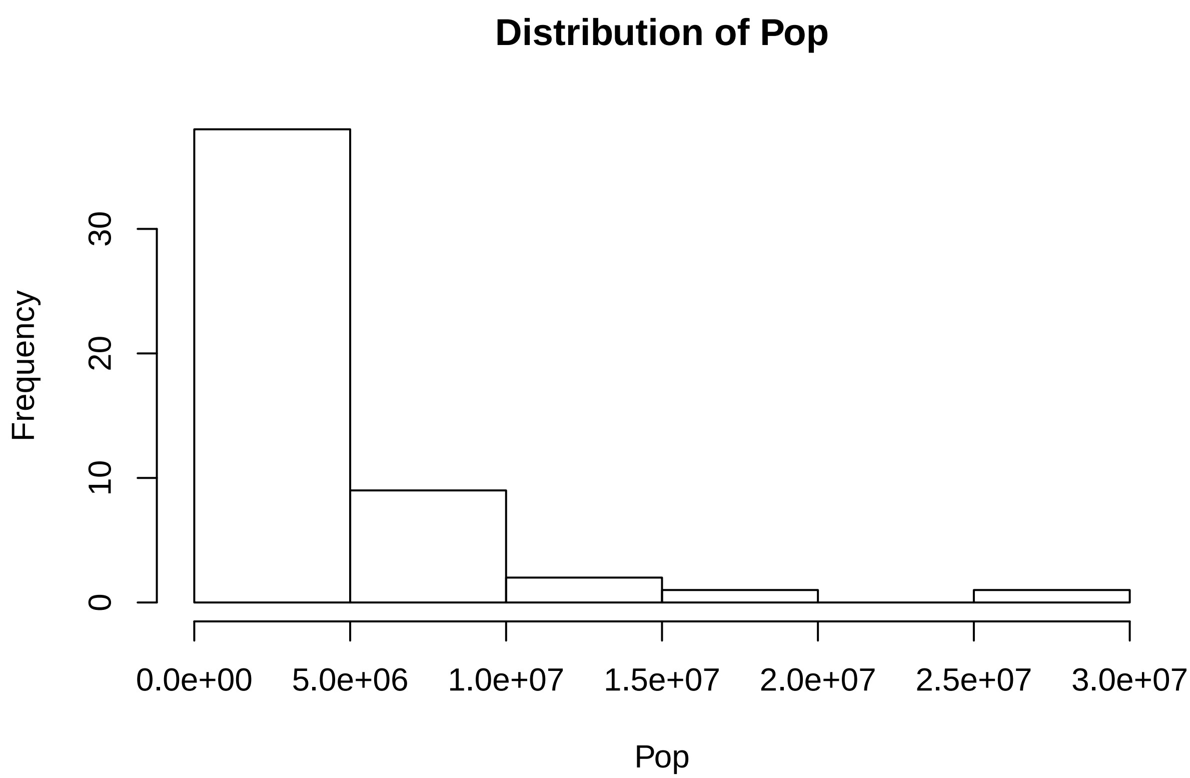

```
hist(fuel2001b$lnMiles,main ="Distribution of lnMiles", xlab = "lnMiles")
```



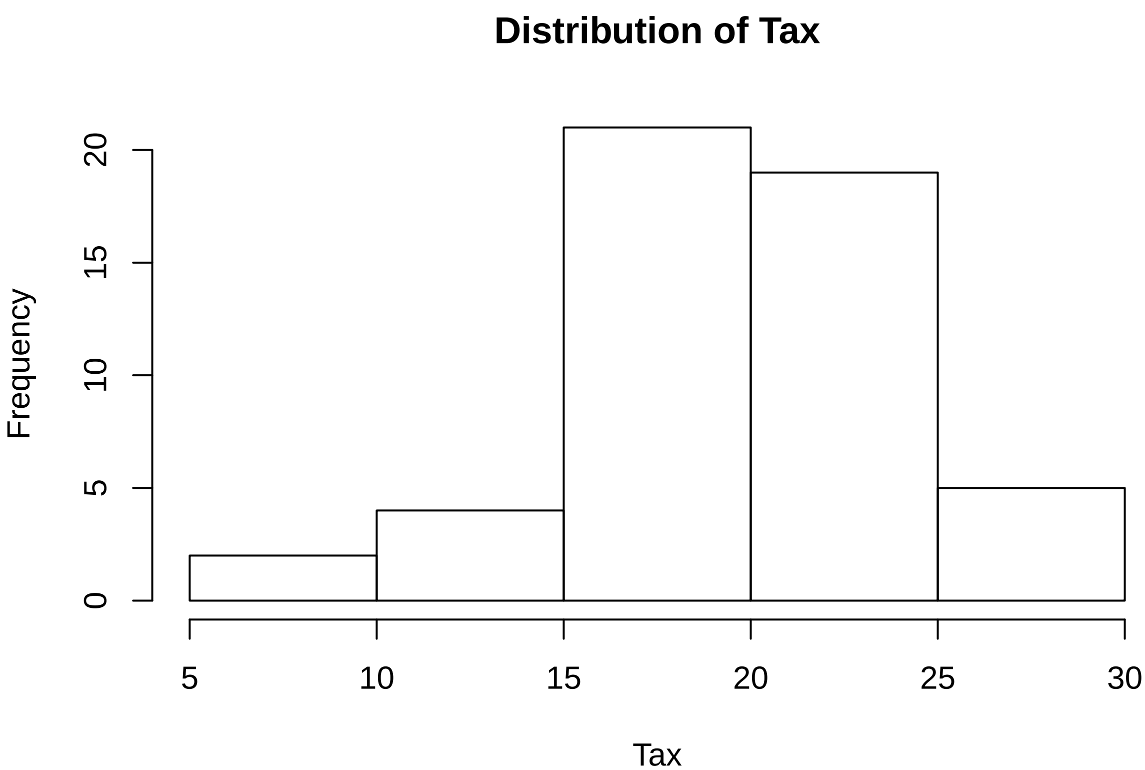

```
hist(fuel2001b$MPC,main ="Distribution of MPC", xlab = "MPC")
```




```r
hist(fuel2001b$Pop,main ="Distribution of Pop", xlab = "Pop")
```




```r
hist(fuel2001b$Tax,main ="Distribution of Tax", xlab = "Tax")
```

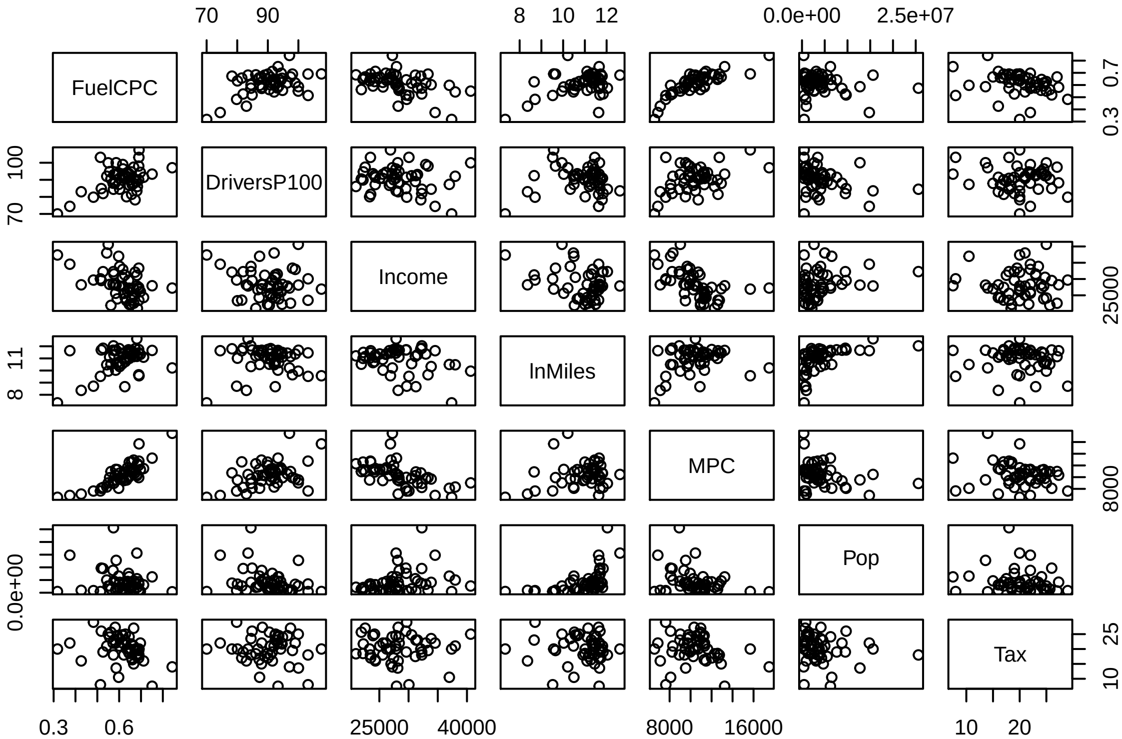


The histograms indicate that the distributions of data for DriversP100, FuelCPC, MPC, Tax are more or less symetric. The data distributions of Income and Pop appear right skewed. The distributions of lnMiles might be left skewed. Thus, we might have to transform some of the variables for the analysis.

Scatterplot

At last, I will create a scatterplot of all variables to get a better understanding of their relationships among each other.

```
with(fuel2001b, pairs(FuelCPC ~ DriversP100+Income+lnMiles+MPC+Pop+Tax))
```



There appears to be a negative correaltion between the two main variables of interest, FuelCPC and Tax. Additionally, Tax appears to be negatively correlated with DriversP100 and MPC.

---

3. Building the model

Now, I will build a first model without making any further corrections on the data.

```
m1 <- lm(FuelCPC ~ DriversP100+Income+lnMiles+MPC+Pop+Tax, data=fuel2001b)
summary(m1)
```

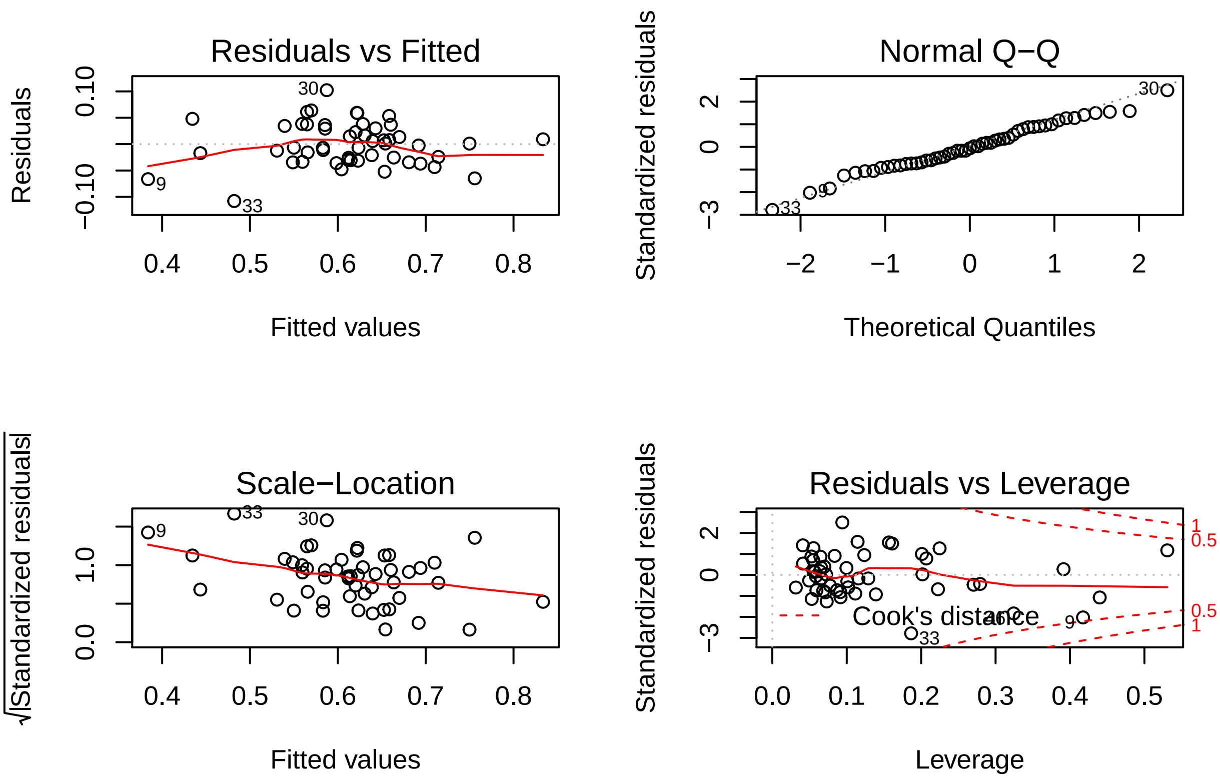
```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles + MPC +
##     Pop + Tax, data = fuel2001b)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.107890 -0.030227 -0.002622  0.032199  0.102253
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.494e-01  1.374e-01  -1.087  0.28281
## DriversP100  1.286e-03  9.649e-04   1.332  0.18958
## Income       1.741e-06  1.813e-06   0.960  0.34230
## lnMiles      3.266e-02  8.340e-03   3.916  0.00031 ***
## MPC          2.947e-05  4.186e-06   7.040 9.99e-09 ***
```

```
## Pop         -3.291e-09  1.902e-09  -1.730  0.09059 .
## Tax         -2.648e-03  1.413e-03  -1.874  0.06755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04296 on 44 degrees of freedom
## Multiple R-squared:  0.7948, Adjusted R-squared:  0.7668
## F-statistic:  28.4 on 6 and 44 DF,  p-value: 1.322e-13
```

From this first model, it appears that FuelCPC is only significantly related to Miles and MPC. This would mean that there is no significant correlation between our main variables of interest, FuealCPC and Tax.

Model Diagnostics

```r
par(mfrow=c(2,2))
plot(m1)
```



The Residual vs Fitted plot indicates non-constant variance. The Normal Q-Q plot is not perfectly straight, but it looks like the response variable is approx. normally distributed. The Residual vs Leverage plot does not show any outliers, but we still should test more formally for those.

Numerical analysis of the Model

a) Transformation

Next I will examine further if transforming the model might be a solution for the non-constant varaince problem.

```r
summary(powerTransform(cbind(FuelCPC,DriversP100,Income,lnMiles,MPC,Pop,Tax) ~1,fuel2001b))
```

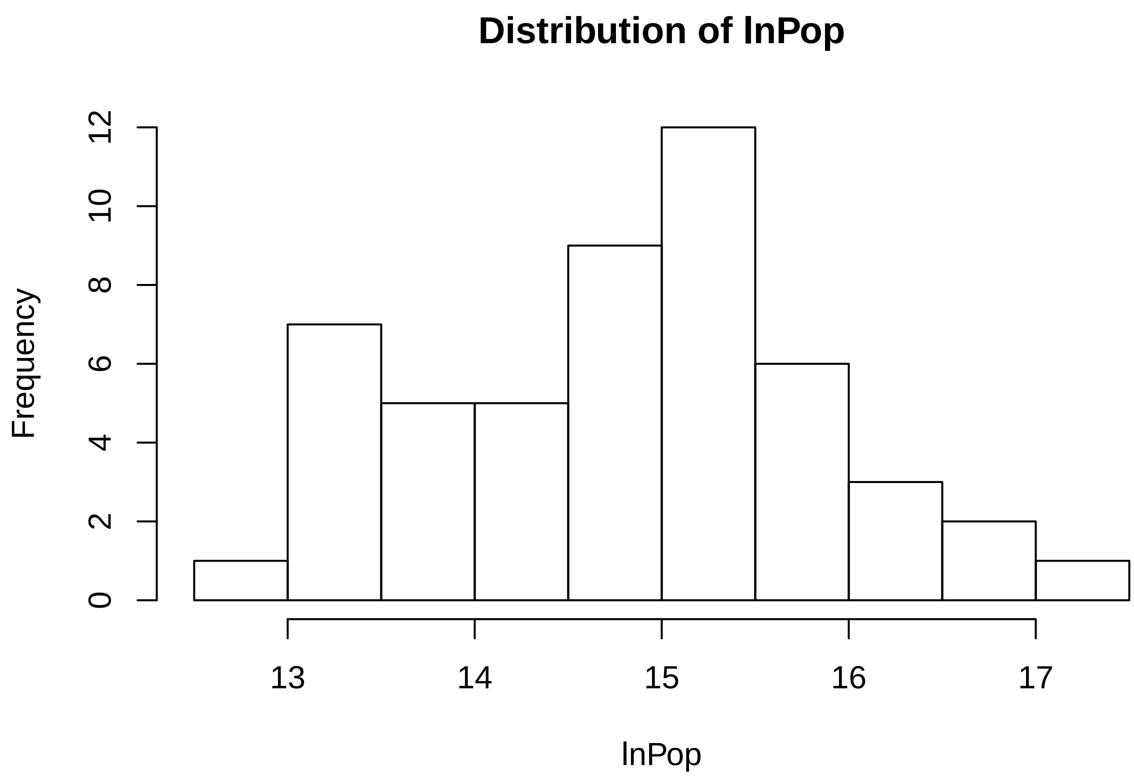
```
## bcPower Transformations to Multinormality
##
##             Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
## FuelCPC        1.6378   0.4691           0.7185           2.5572
## DriversP100    1.7297   1.3471          -0.9107           4.3700
## Income        -0.2349   0.7198          -1.6458           1.1760
## lnMiles        4.6385   1.2224           2.2427           7.0343
## MPC           -1.1692   0.4978          -2.1449          -0.1934
## Pop            0.1434   0.1003          -0.0532           0.3399
## Tax            1.7400   0.4447           0.8685           2.6115
##
## Likelihood ratio tests about transformation parameters
##                                          LRT df         pval
## LR test, lambda = (0 0 0 0 0 0 0)     70.50901  7 1.165734e-12
## LR test, lambda = (1 1 1 1 1 1 1)    131.78846  7 0.000000e+00
## LR test, lambda = (1 1 1 4.64 -1 0 1)  12.11305  7 9.690199e-02
```

The last row of the powerTransform LRT test suggests to transform the following variables:

- take the log of Pop
- take MPC to the power of -1
- take lnMiles to the power of 4.64. To keep it simple, I will do 5

I will do these transformations step by step and see if the output of powerTransform changes inbetween.

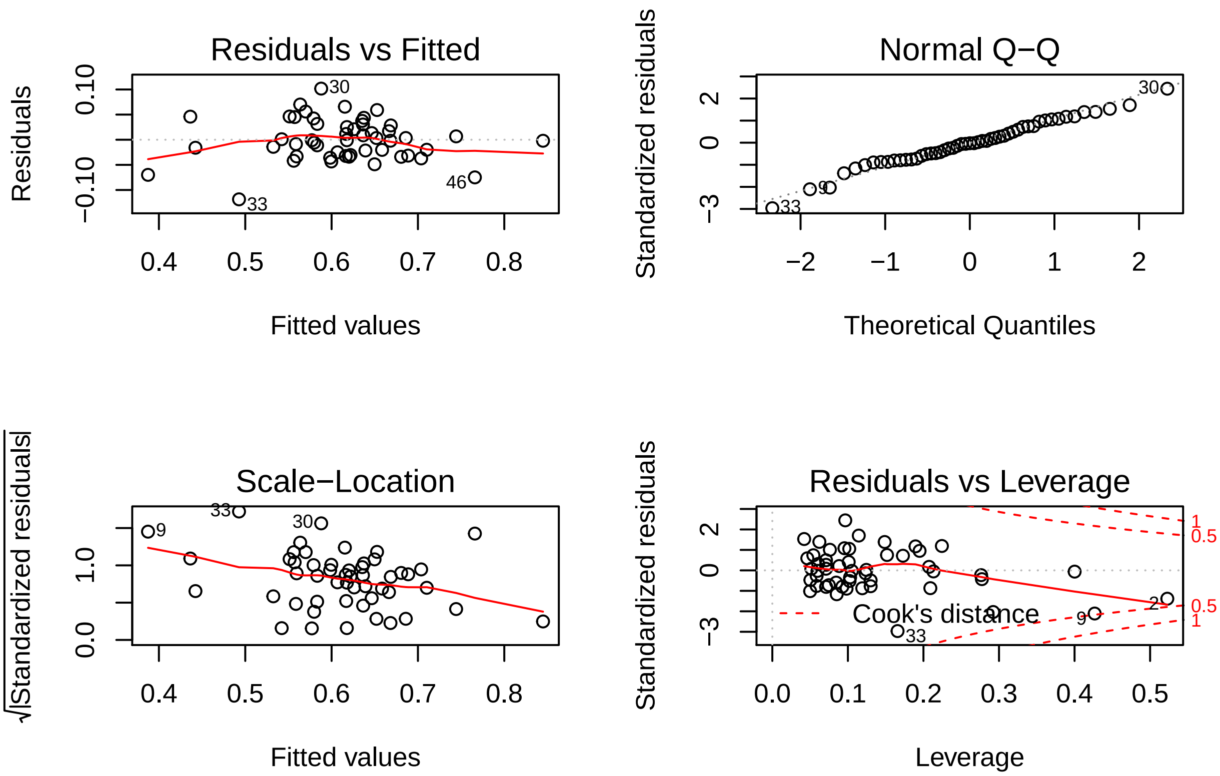
```r
fuel2001b.T1 <- with(fuel2001b, data.frame(DriversP100=DriversP100, FuelCPC=FuelCPC, Income=Income, lnMi

hist(fuel2001b.T1$lnPop,main ="Distribution of lnPop", xlab = "lnPop")
```




```r
m2 <- lm(FuelCPC ~ DriversP100+Income+lnMiles+MPC+lnPop+Tax, data=fuel2001b.T1)
summary(m2)
```

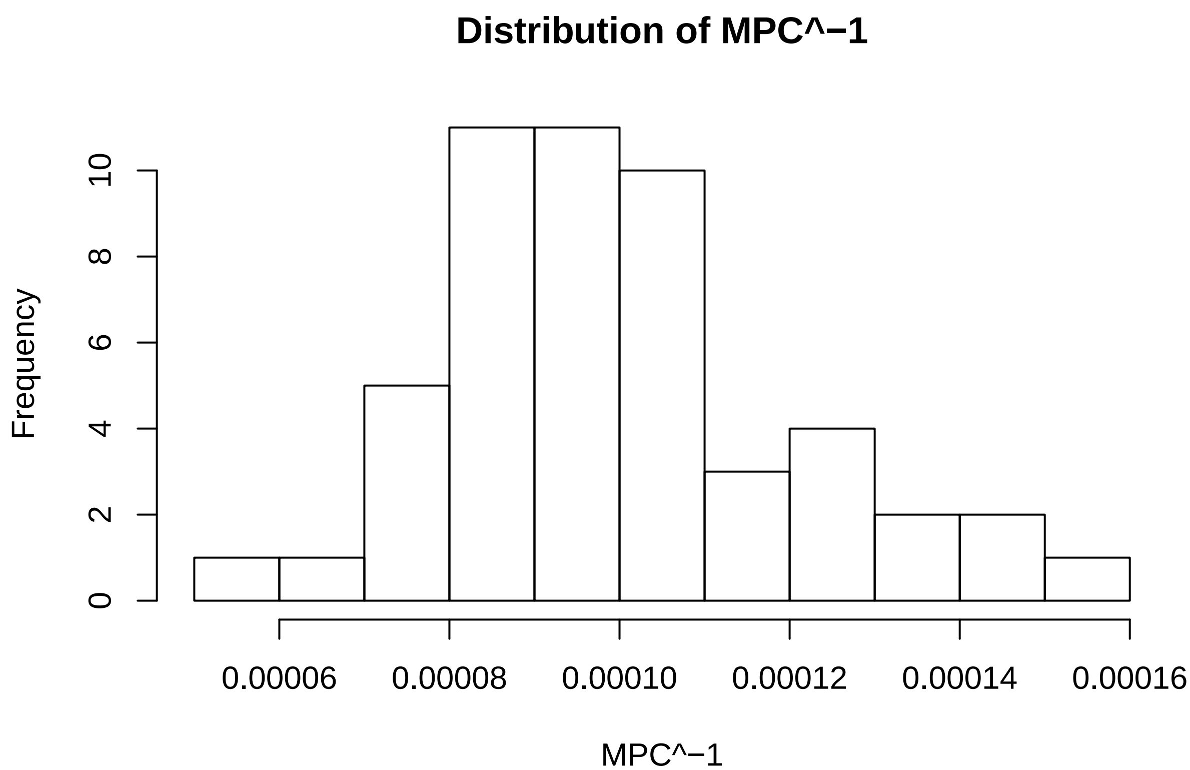
```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles + MPC +
##     lnPop + Tax, data = fuel2001b.T1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118602 -0.031030 -0.001041  0.029381  0.101706
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.125e-02  1.627e-01   0.069  0.94521
## DriversP100  1.367e-03  9.928e-04   1.377  0.17561
## Income       1.560e-06  1.910e-06   0.817  0.41854
## lnMiles      3.329e-02  1.118e-02   2.978  0.00471 **
## MPC          2.952e-05  4.386e-06   6.731 2.84e-08 ***
## lnPop       -1.270e-02  1.139e-02  -1.115  0.27094
## Tax         -2.491e-03  1.449e-03  -1.720  0.09250 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04379 on 44 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.7577
## F-statistic: 27.06 on 6 and 44 DF,  p-value: 2.99e-13
```

```r
par(mfrow=c(2,2))
plot(m2)
```



Non-constant varaince is still a problem. However, the transformation made the distribution of Pop data more symetric. Only MPC and Miles are significant in this model.

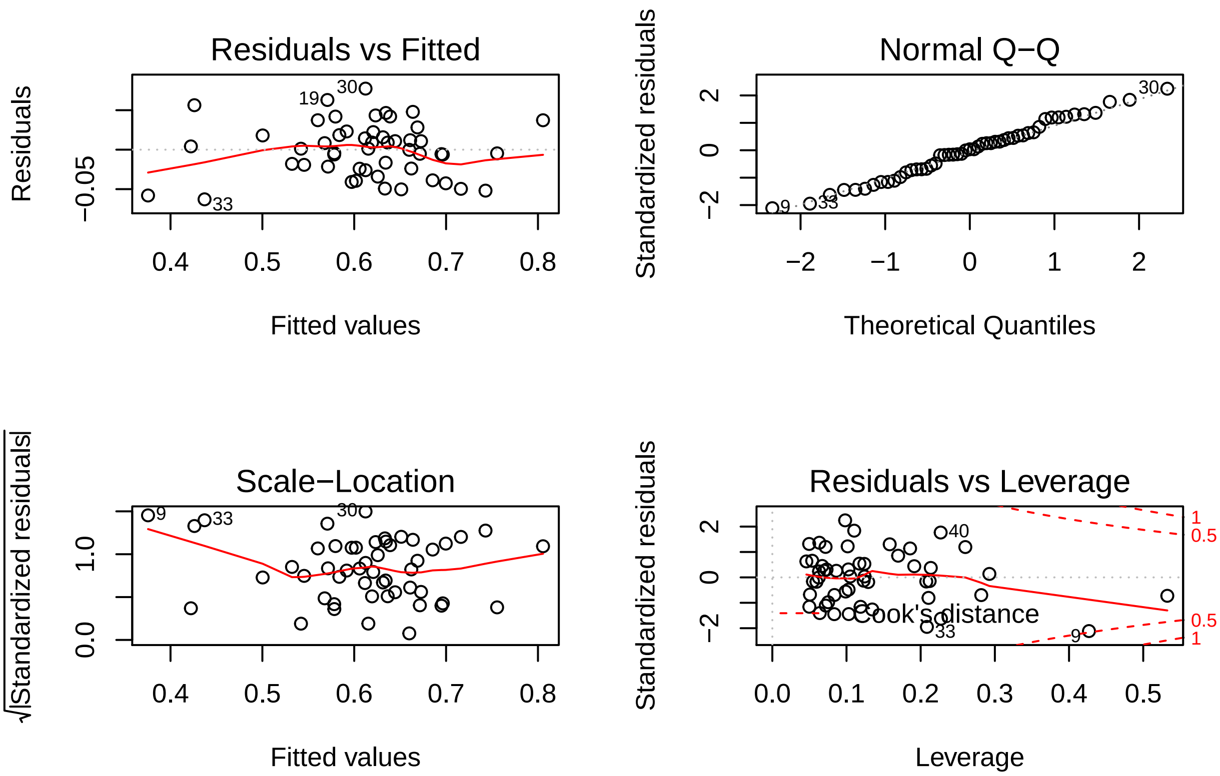
```r
fuel2001b.T2 <- with(fuel2001b.T1, data.frame(DriversP100=DriversP100, FuelCPC=FuelCPC, Income=Income, I

hist(fuel2001b.T2$MPC.i,main ="Distribution of MPC^-1", xlab = "MPC^-1")
```




```r
m3 <- lm(FuelCPC ~ DriversP100+Income+lnMiles+MPC.i+lnPop+Tax, data=fuel2001b.T2)
summary(m3)
```

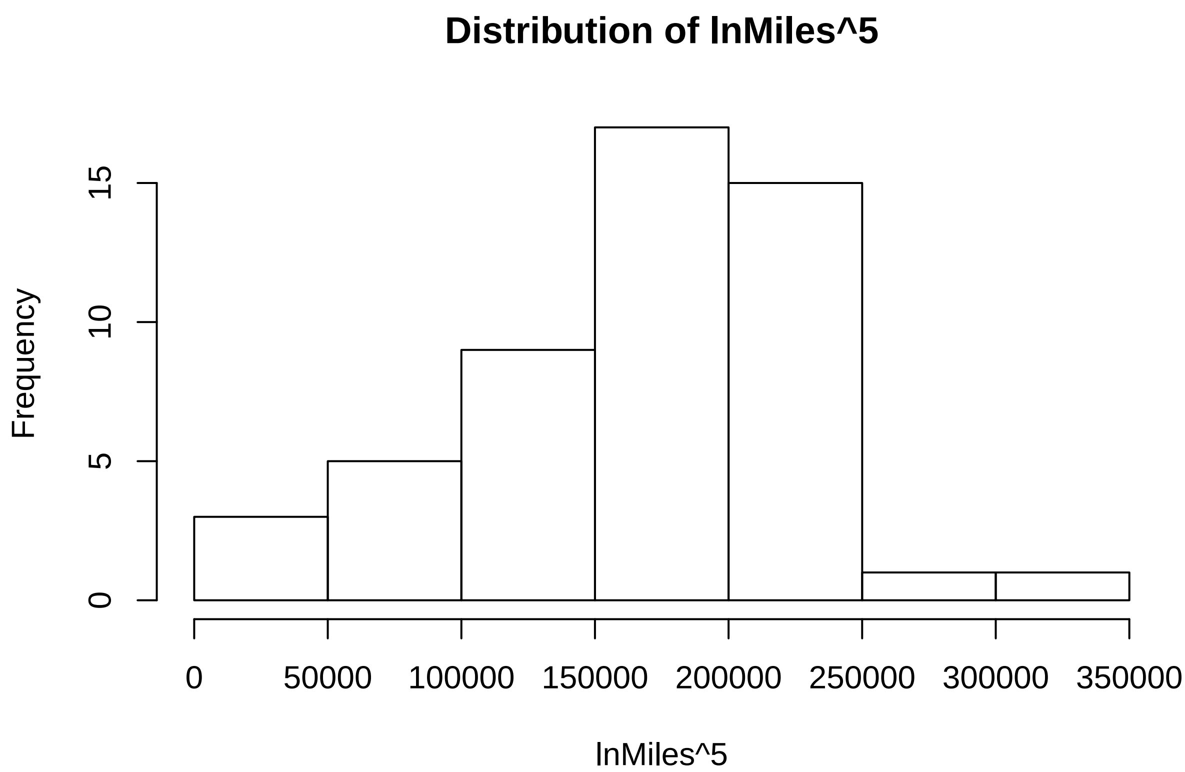
```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles + MPC.i +
##     lnPop + Tax, data = fuel2001b.T2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062901 -0.023993  0.001221  0.020364  0.077340
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.103e-01  1.365e-01   6.667 3.52e-08 ***
## DriversP100  1.987e-04  8.566e-04   0.232   0.8176
## Income       3.250e-06  1.625e-06   2.000   0.0517 .
## lnMiles      2.112e-02  9.598e-03   2.201   0.0331 *
## MPC.i       -3.910e+03  4.214e+02  -9.278 6.40e-12 ***
## lnPop       -1.204e-02  9.344e-03  -1.289   0.2043
## Tax         -3.580e-03  1.168e-03  -3.066   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03628 on 44 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8337
## F-statistic: 42.77 on 6 and 44 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m3)
```



Non-constant varaince remains an issue. The transformation also did not improve the distribution of MPC. Tax, MPC, and Miles are significant in this model.

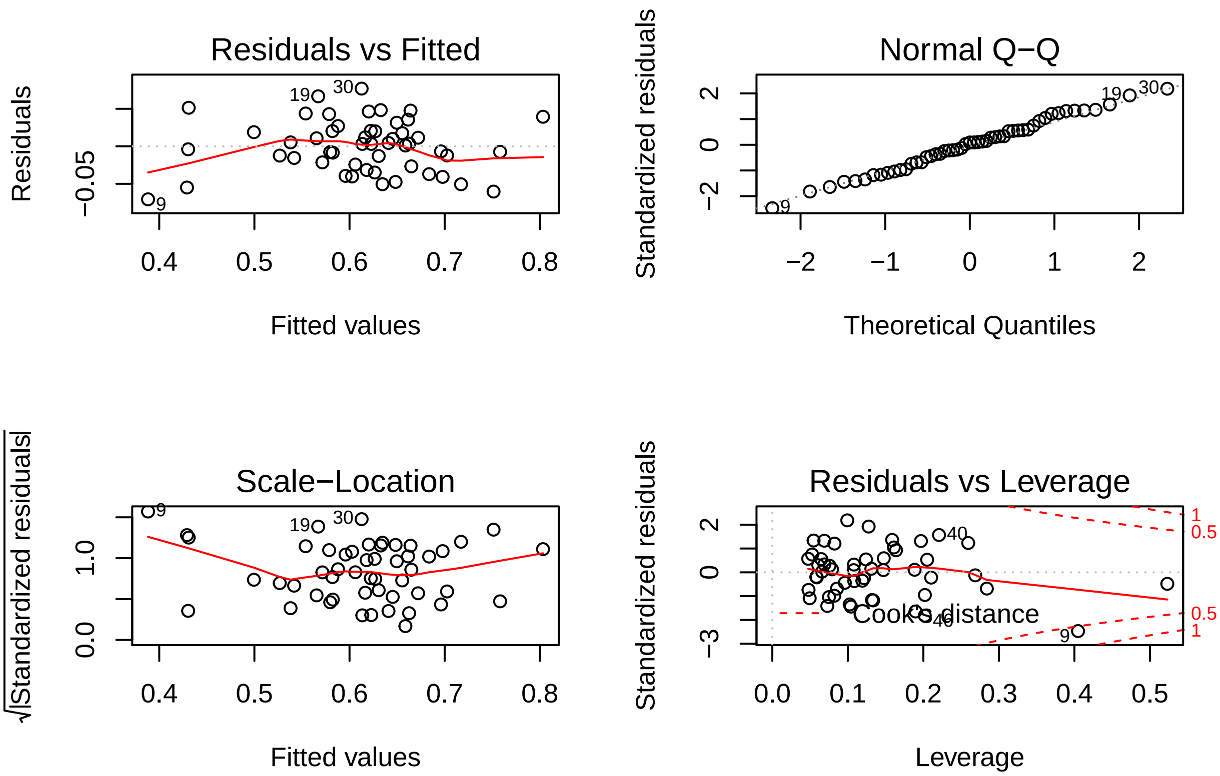
```
fuel2001b.T3 <- with(fuel2001b.T2, data.frame(DriversP100=DriversP100, FuelCPC=FuelCPC, Income=Income, I

hist(fuel2001b.T3$lnMiles.p5, main ="Distribution of lnMiles^5", xlab = "lnMiles^5")
```




```r
m4 <- lm(FuelCPC ~ DriversP100+Income+lnMiles.p5+MPC.i+lnPop+Tax, data=fuel2001b.T3)
summary(m4)
```

```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles.p5 + MPC.i +
##     lnPop + Tax, data = fuel2001b.T3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.070706 -0.025538  0.003172  0.020449  0.076922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.018e+00  1.474e-01   6.904 1.58e-08 ***
## DriversP100  5.004e-04  8.766e-04   0.571  0.57100
## Income       2.821e-06  1.642e-06   1.719  0.09272 .
## lnMiles.p5   2.539e-07  1.577e-07   1.610  0.11456
## MPC.i       -4.046e+03  4.214e+02  -9.601 2.31e-12 ***
## lnPop       -6.963e-03  9.148e-03  -0.761  0.45058
## Tax         -3.367e-03  1.186e-03  -2.838  0.00684 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03715 on 44 degrees of freedom
## Multiple R-squared:  0.8466, Adjusted R-squared:  0.8256
## F-statistic: 40.46 on 6 and 44 DF,  p-value: 2.475e-16
```

```r
par(mfrow=c(2,2))
plot(m4)
```



There is still non-constant variance. However, lnMiles.p5 has a more symetric distribution than lnMiles. Tax and MPC are significant in this model.

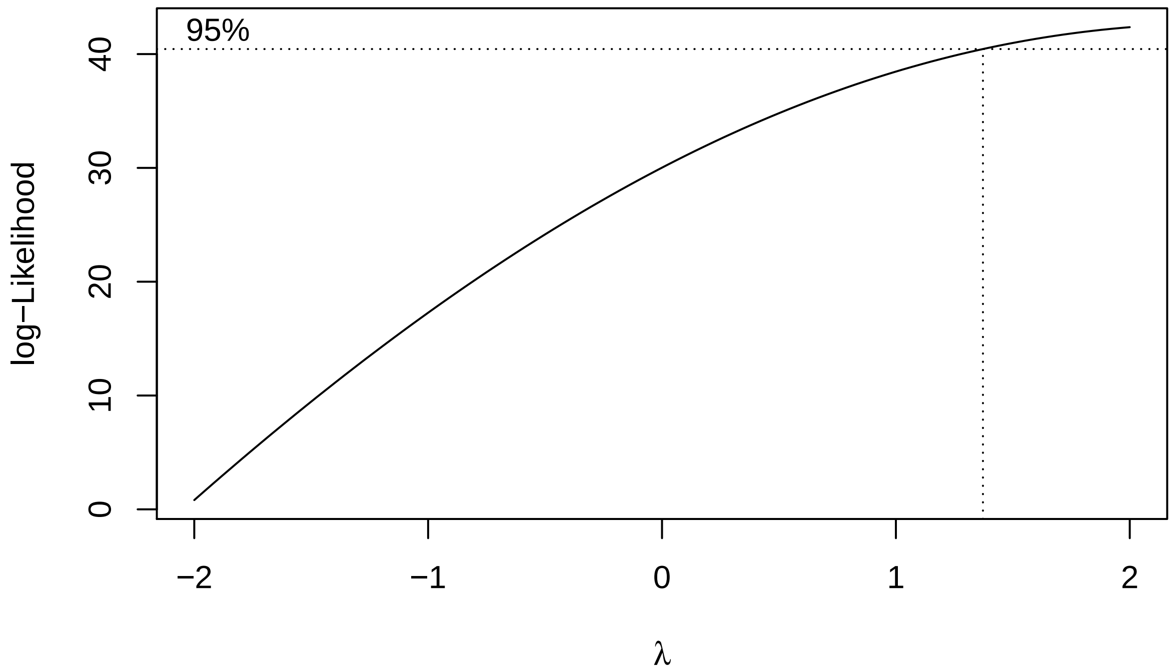
None of the transformations significantly improved the model in terms of nonconstant variance.

Does a transformation of the response variable maybe make sense?

```r
require(MASS)
```

```
## Loading required package: MASS
```

```r
b <- boxcox(m1)
```




```
with(b, x[which.max(y)])
```

```
## [1] 2
```

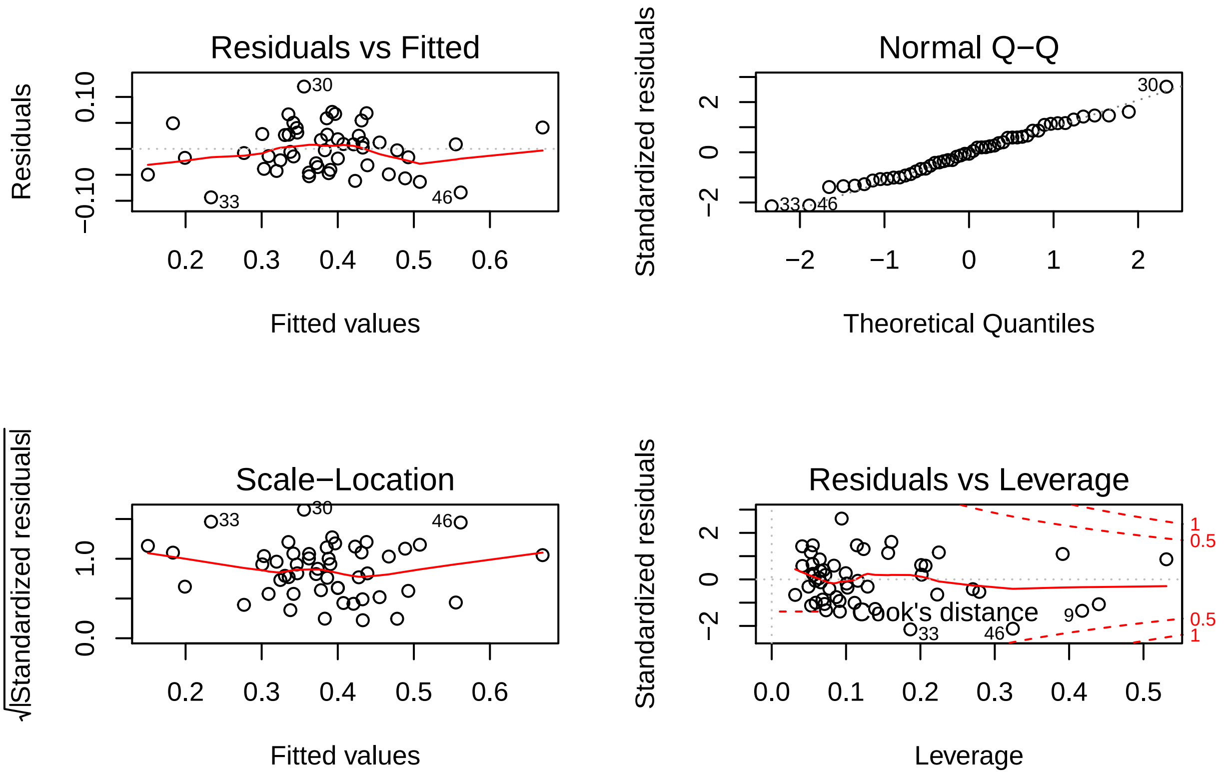
The boxcox method indicates that we should square FuelCPC, the response var.

```
fuel2001b.T4 <- with(fuel2001b, data.frame(DriversP100=DriversP100, FuelCPC.p2=(FuelCPC)^2, Income=Inco

m5 <- lm(FuelCPC.p2 ~ DriversP100+Income+lnMiles+MPC+Pop+Tax,data=fuel2001b.T4)
summary(m5)
```

```
##
## Call:
## lm(formula = FuelCPC.p2 ~ DriversP100 + Income + lnMiles + MPC +
##     Pop + Tax, data = fuel2001b.T4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.093448 -0.036856 -0.002747  0.027935  0.119970
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.903e-01  1.542e-01  -2.531   0.0150 *
## DriversP100  5.321e-04  1.083e-03   0.491   0.6256
## Income       2.702e-06  2.035e-06   1.328   0.1911
## lnMiles      3.111e-02  9.361e-03   3.324   0.0018 **
## MPC          3.833e-05  4.699e-06   8.157  2.4e-10 ***
## Pop         -3.506e-09  2.135e-09  -1.642   0.1077
## Tax         -3.767e-03  1.586e-03  -2.376   0.0219 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.04822 on 44 degrees of freedom
## Multiple R-squared:  0.8071, Adjusted R-squared:  0.7808
## F-statistic: 30.68 on 6 and 44 DF,  p-value: 3.476e-14
```

```
par(mfrow=c(2,2))
plot(m5)
```



Again, this transformation does not help the nonconstant varaince problem and it seems to harm the normality assumption for the response var. MPC, Tax and Miles are signifiant in the model.

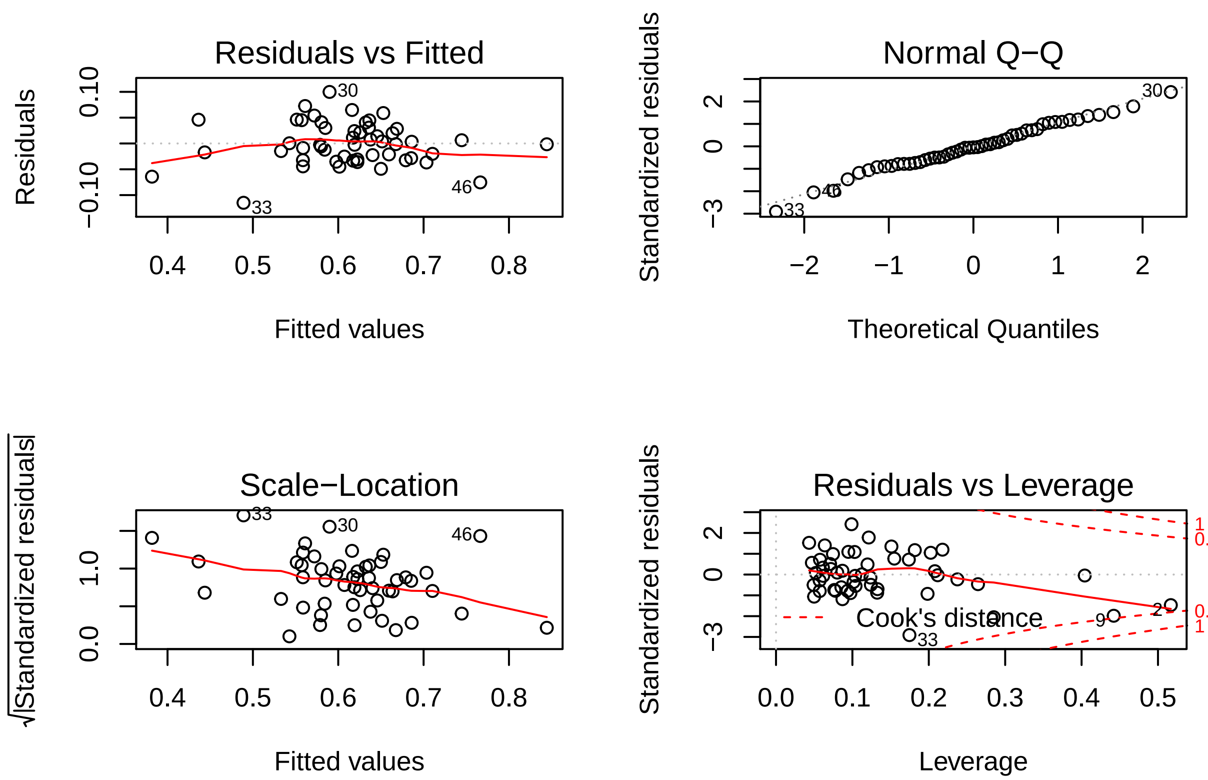
Next we will transform only the regressor variables for which log could make sense, i.e. where 0 is in the Wald C.I. as indicated by the powerTransform test earlier: DriversP100, Income, Pop

```
fuel2001b.T5 <- with(fuel2001b, data.frame(lnDriversP100=log(DriversP100), FuelCPC=FuelCPC, lnIncome=log

m6 <- lm(FuelCPC ~ lnDriversP100+lnIncome+lnMiles+MPC+lnPop+Tax,data=fuel2001b.T5)
summary(m6)
```

```
##
## Call:
## lm(formula = FuelCPC ~ lnDriversP100 + lnIncome + lnMiles + MPC +
##     lnPop + Tax, data = fuel2001b.T5)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.114889 -0.029653 -0.001548  0.029167  0.099802
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -9.073e-01  6.402e-01  -1.417  0.16343
## lnDriversP100   1.390e-01  8.709e-02   1.596  0.11759
## lnIncome        4.525e-02  5.450e-02   0.830  0.41084
## lnMiles         3.258e-02  1.102e-02   2.956  0.00499 **
## MPC             2.929e-05  4.350e-06   6.733 2.82e-08 ***
## lnPop          -1.222e-02  1.125e-02  -1.085  0.28362
## Tax            -2.474e-03  1.437e-03  -1.721  0.09224 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04345 on 44 degrees of freedom
## Multiple R-squared:  0.7901, Adjusted R-squared:  0.7615
## F-statistic:  27.6 on 6 and 44 DF,  p-value: 2.146e-13
```

```
par(mfrow=c(2,2))
plot(m6)
```



Again, the transformation did not change the non-constant variance issue.

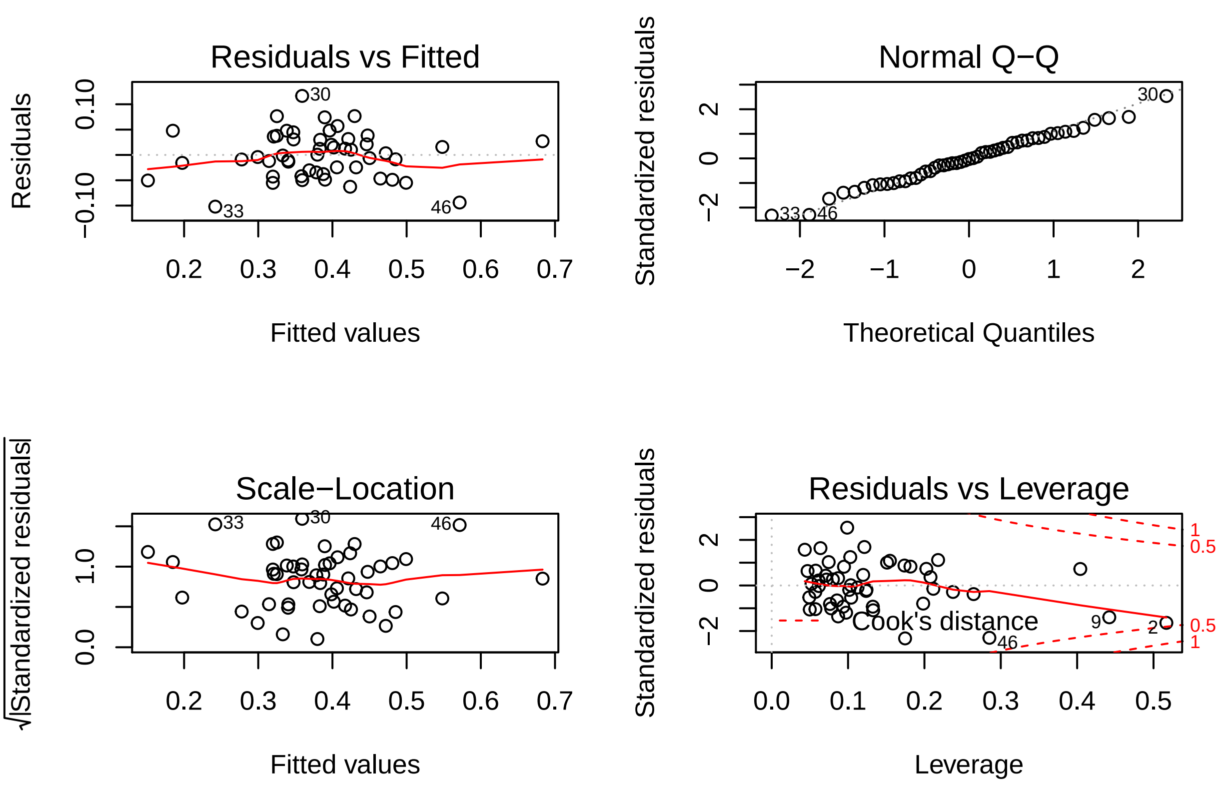
Last, I will fit a model with power transformation of the response and the log transformations of the regressors combined.

```
fuel2001b.T6 <- with(fuel2001b.T5, data.frame(lnDriversP100=lnDriversP100, FuelCPC.p2=(FuelCPC)^2, lnIn

m7 <- lm(FuelCPC.p2 ~ lnDriversP100+lnIncome+lnMiles+MPC+lnPop+Tax,data=fuel2001b.T6)
summary(m7)
```

```
##
```

```
## Call:
## lm(formula = FuelCPC.p2 ~ lnDriversP100 + lnIncome + lnMiles +
##     MPC + lnPop + Tax, data = fuel2001b.T6)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.102050 -0.036136 -0.001204  0.031017  0.116466
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.171e+00  7.125e-01  -1.644  0.10731
## lnDriversP100  6.121e-02  9.693e-02   0.631  0.53100
## lnIncome       8.244e-02  6.065e-02   1.359  0.18098
## lnMiles        3.516e-02  1.227e-02   2.866  0.00635 **
## MPC            3.756e-05  4.841e-06   7.759 8.99e-10 ***
## lnPop         -1.801e-02  1.253e-02  -1.438  0.15756
## Tax           -3.728e-03  1.600e-03  -2.330  0.02443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04836 on 44 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.7796
## F-statistic: 30.47 on 6 and 44 DF,  p-value: 3.925e-14
```

```r
par(mfrow=c(2,2))
plot(m7)
```



As the other transformations, no change in the non-constant variance issue.

Since none of the transformations could correct the nonconstant variance, I will choose the model with the highest R-squared value for now:

```
compR <-rbind(summary(m1)$r.squared,summary(m2)$r.squared,summary(m3)$r.squared,summary(m4)$r.squared,su
rownames(compR)<-c("m1","m2","m3","m4","m5","m6","m7")
colnames(compR)<-c("R^2")
compR
```

```
##          R^2
## m1 0.7947530
## m2 0.7868101
## m3 0.8536340
## m4 0.8465631
## m5 0.8070991
## m6 0.7900779
## m7 0.8060099
```

The model of choice is m3.

b) Correlation

It is time to examine the correlation between the different variables numerically:

```
with(fuel2001b.T2, cor(cbind(FuelCPC,DriversP100,Income,lnMiles,MPC.i,lnPop,Tax)))
```

```
##                 FuelCPC DriversP100      Income     lnMiles       MPC.i
## FuelCPC      1.00000000  0.46850627 -0.46440498  0.42203233 -0.89192646
## DriversP100  0.46850627  1.00000000 -0.17596063  0.03059068 -0.49065861
## Income      -0.46440498 -0.17596063  1.00000000 -0.29585136  0.61297986
## lnMiles      0.42203233  0.03059068 -0.29585136  1.00000000 -0.36410709
## MPC.i       -0.89192646 -0.49065861  0.61297986 -0.36410709  1.00000000
## lnPop       -0.07585706 -0.28074477  0.22272880  0.66345386  0.17824582
## Tax         -0.25944711 -0.08584424 -0.01068494 -0.04373696  0.09410773
##                   lnPop         Tax
## FuelCPC     -0.07585706 -0.25944711
## DriversP100 -0.28074477 -0.08584424
## Income       0.22272880 -0.01068494
## lnMiles      0.66345386 -0.04373696
## MPC.i        0.17824582  0.09410773
## lnPop        1.00000000 -0.12438570
## Tax         -0.12438570  1.00000000
```

None of the covariants is very highly correlated with another one. The response is highly correlated with MPC^-1. Nevertheless, there are some covariants that don't seem to explain much of the variance in the response right now.

```
summary(m3)
```

```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles + MPC.i +
```

```
##     lnPop + Tax, data = fuel2001b.T2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062901 -0.023993  0.001221  0.020364  0.077340
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.103e-01  1.365e-01   6.667 3.52e-08 ***
## DriversP100  1.987e-04  8.566e-04   0.232   0.8176
## Income       3.250e-06  1.625e-06   2.000   0.0517 .
## lnMiles      2.112e-02  9.598e-03   2.201   0.0331 *
## MPC.i       -3.910e+03  4.214e+02  -9.278 6.40e-12 ***
## lnPop       -1.204e-02  9.344e-03  -1.289   0.2043
## Tax         -3.580e-03  1.168e-03  -3.066   0.0037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03628 on 44 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8337
## F-statistic: 42.77 on 6 and 44 DF,  p-value: < 2.2e-16
```

Particularly, DriversP100 and lnPop. Thus, I will fit 2 more models without these variables and one more model without both of them to see how this will change the significance of the other variables as well as the overall model fit.

```r
m3a <- lm(FuelCPC ~ Income+lnMiles+MPC.i+lnPop+Tax, data=fuel2001b.T2)
m3b <- lm(FuelCPC ~ DriversP100+Income+lnMiles+MPC.i+Tax, data=fuel2001b.T2)
m3c <- lm(FuelCPC ~ Income+lnMiles+MPC.i+Tax, data=fuel2001b.T2)
summary(m3a);summary(m3b);summary(m3c);
```

```
##
## Call:
## lm(formula = FuelCPC ~ Income + lnMiles + MPC.i + lnPop + Tax,
##     data = fuel2001b.T2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064020 -0.023337  0.001982  0.020586  0.078248
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.352e-01  8.384e-02  11.154 1.52e-14 ***
## Income       3.339e-06  1.562e-06   2.137  0.03804 *
## lnMiles      2.133e-02  9.455e-03   2.256  0.02898 *
## MPC.i       -3.949e+03  3.806e+02 -10.378 1.61e-13 ***
## lnPop       -1.253e-02  9.004e-03  -1.392  0.17077
## Tax         -3.603e-03  1.151e-03  -3.129  0.00308 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0359 on 45 degrees of freedom
## Multiple R-squared:  0.8535, Adjusted R-squared:  0.8372
## F-statistic: 52.41 on 5 and 45 DF,  p-value: < 2.2e-16
```

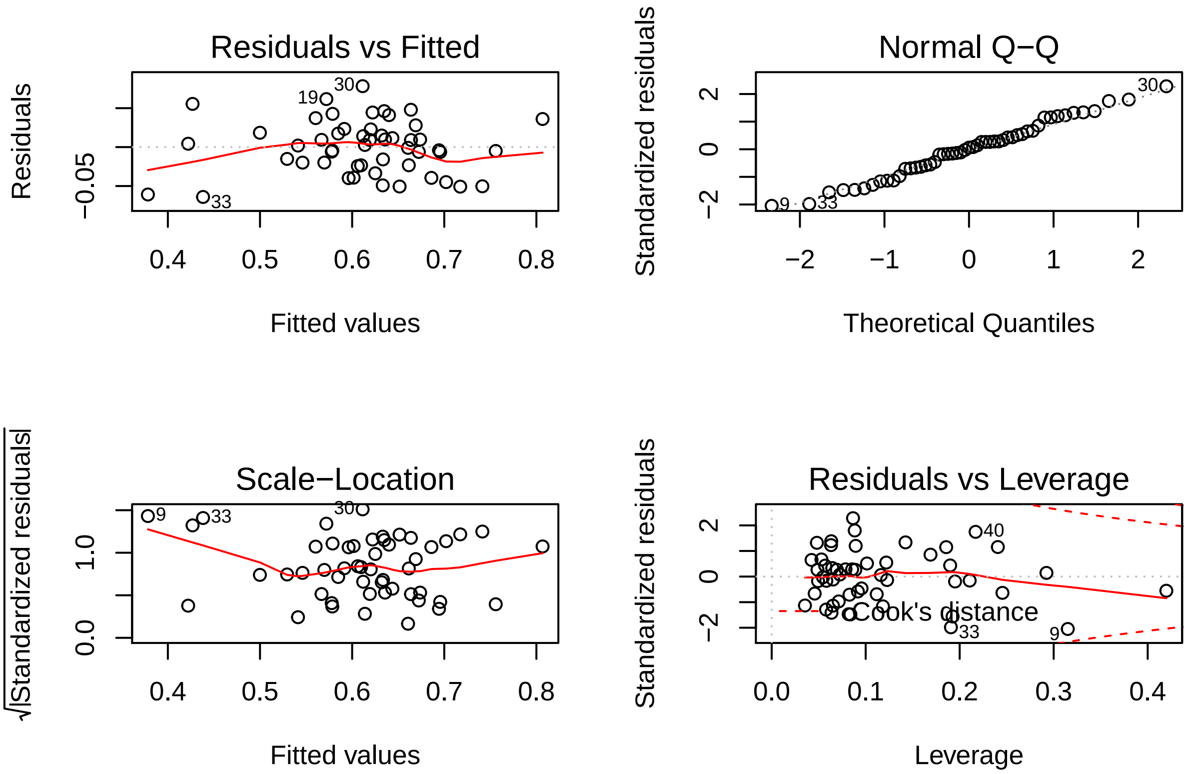
```
##
## Call:
## lm(formula = FuelCPC ~ DriversP100 + Income + lnMiles + MPC.i +
##     Tax, data = fuel2001b.T2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060847 -0.027607 -0.000199  0.022758  0.079032
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.534e-01  1.301e-01   6.558 4.63e-08 ***
## DriversP100  4.494e-04  8.403e-04   0.535  0.59543
## Income       2.420e-06  1.503e-06   1.610  0.11431
## lnMiles      1.094e-02  5.491e-03   1.993  0.05238 .
## MPC.i       -4.066e+03  4.065e+02 -10.001 5.18e-13 ***
## Tax         -3.250e-03  1.148e-03  -2.832  0.00689 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03655 on 45 degrees of freedom
## Multiple R-squared:  0.8481, Adjusted R-squared:  0.8312
## F-statistic: 50.25 on 5 and 45 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = FuelCPC ~ Income + lnMiles + MPC.i + Tax, data = fuel2001b.T2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.063316 -0.027567  0.000994  0.021082  0.081362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.071e-01  8.220e-02  11.035 1.62e-14 ***
## Income       2.551e-06  1.471e-06   1.734   0.0895 .
## lnMiles      1.045e-02  5.370e-03   1.945   0.0579 .
## MPC.i       -4.176e+03  3.474e+02 -12.020 8.54e-16 ***
## Tax         -3.271e-03  1.138e-03  -2.875   0.0061 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03626 on 46 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.8339
## F-statistic: 63.73 on 4 and 46 DF,  p-value: < 2.2e-16
```

Model m3a has DriversP100 removed. Interestingly, R-squared almost does not decrease and Income is now a significant variable.

How do the diagnostics look for this model?

```r
par(mfrow=c(2,2))
plot(m3a)
```



The non-constant variance is still an issue and the normal distribution of the response is still more or less given, even though the distribution has several small derivations form a straight line. It seems the model change did not further harm the assumptions. Thus, I will continue my analysis with model m3a.

c) Outliers

I will test numerically for outliers in the m3:

```
outlierTest(m3a)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 30 2.398089           0.020789           NA
```

According to the test with Bonferonni p-value, there is no outlier, thus I will not remove any of the data points given.

---

4. Conclusions

After my analysis, I conclude that m3a is the best model with the given data.

m3a: FuelCPC ~ Income + lnMiles + MPC.i + lnPop + Tax

Since the non-constant varaince assumption is not fulfilled, the estimates of the model are likely not very accurate for all possible values of FuelCPC.

Nevertheless, the coefficients of the model can be interpreted as follows: With all other variables kept at a constant level…

- a increase of one cent/gallon in the gas state tax decreases the gasoline sold per capita and year by 3.603 gallons.
- a one percent increase in miles of highway in a state will increase gasoline sold per capita by approx. .2133 gallons.
- a one dollar increase in average income per capita increases the the gasoline sold per capita by .000003 gallons.

Thus, Tax and Fuel Consumption per capita are negatively correlated while Fuel Consumption and total miles of federal highway in a state and the average income per capita are positively correlated. These findings make intuitively sense.

Since Pop is not significant, we cannot really interpret it's estimate sensically. The inverse of Miles per capita cannot be interpreted in a linear fashion.

The best aspect about this model is that the value of multiple R-squared and adjusted R-squared are both high and thus suggest a good model fit.