MPG Regression Analysis

Michael Garcia

4/13/2018

Exeutive Summary

Motor Trend provides information, opinions, and tips about cars to its readers. A topic of interest is energy efficiency of vehicles, specifically autmatic and manual transimission and miles per gallon. The analysis will provide insight into the methods used and the results for answering the question is miles per gallon for automatic vehicles greater than, less than, or equal to vehicles with manual transmission. * "Is an automatic or manual transmission better for MPG" * "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Analysis

```
data(cars)
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
mtcarsdf <- as.data.frame(mtcars)
```

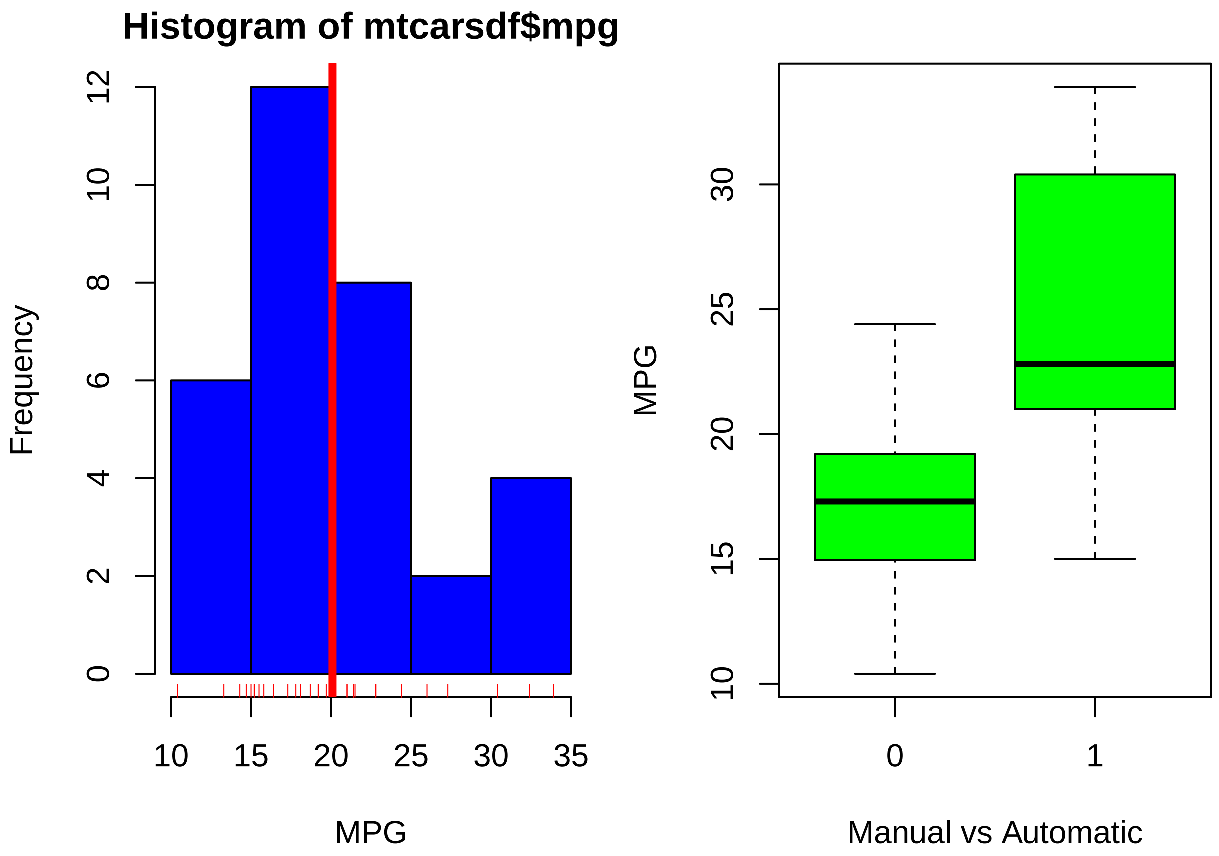
Exploratory Data Analysis- Distribution

The distributions for the mpg for the total dataset are reflected

```
mtcarsdf$mpg_rnd <- round(mtcarsdf$mpg,0)

par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
hist(mtcarsdf$mpg,col = "blue", freq = TRUE, xlab = "MPG")
rug(mtcarsdf$mpg, col = "red")
```

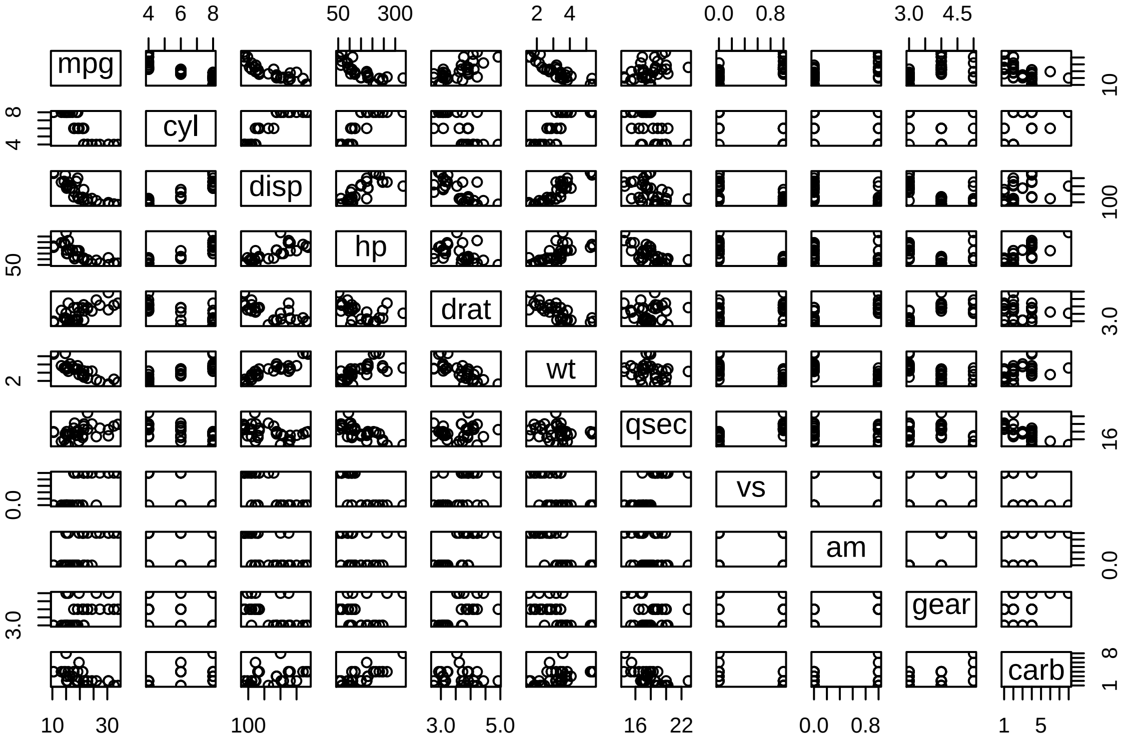
```
abline(v = mean(mtcarsdf$mpg), col = "red", lwd = 4)
boxplot(mpg ~ am, data = mtcarsdf, col = "green", xlab = "Manual vs Automatic", ylab = "MPG")
```



Scatterplot Matrix

The plot displays the points for pairs of variables.

```
pairs(mpg ~ ., data = mtcars, col = "black")
```



You can also embed plots, for example:

Nested Fitting - Generalized Linear Model

```
Model1 <- glm(am ~ mpg , data = mtcars, family = "binomial")
Model2 <- glm(am ~ mpg + wt, data = mtcars, family = "binomial")
Model3 <- glm(am ~ mpg + wt + hp , data = mtcars, family = "binomial")
Model4 <- glm(am ~ mpg + wt + hp+ disp, data = mtcars, family = "binomial")
#Model5 <- glm(am ~ ., data = mtcars, family = "binomial")
```

ANOVA GLM

```
anova(Model1, Model2, Model3, Model4)
```

```
## Analysis of Deviance Table
##
## Model 1: am ~ mpg
## Model 2: am ~ mpg + wt
## Model 3: am ~ mpg + wt + hp
## Model 4: am ~ mpg + wt + hp + disp
##   Resid. Df Resid. Dev Df Deviance
## 1        30    29.6752
## 2        29    17.1843  1  12.4909
## 3        28     8.7661  1   8.4181
## 4        27     8.1620  1   0.6041
```

Nested Fitting - Linear Model

```
LModel1 <- lm(mpg ~ am , data = mtcars)
LModel2 <- lm(mpg ~ am + wt, data = mtcars)
```

```
LModel3 <- lm(mpg ~ am + wt + hp , data = mtcars)
LModel4 <- lm(mpg ~ am + wt + hp+ disp, data = mtcars)
LModel5 <- lm(mpg ~ ., data = mtcars)
```

ANOVA LM

```
anova(LModel1, LModel2, LModel3, LModel4, LModel5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + disp
## Model 5: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 180.29  1     98.03 13.9571  0.001219 **
## 4     27 179.91  1      0.38  0.0546  0.817510
## 5     21 147.49  6     32.41  0.7692  0.602559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GLM

The model supports that the MPG increases for vehicles that are automatic or not automatic. We use binomial general linear model given that 1 of 2 outcomes is possible for mileage per gallon.

The model is given by: probautomatic = .307MPG - 6.6035. So for every increase in distance of .307MPG theres a higher probabilty that the vehicle is automatic.

```
logCars <- glm(mtcars$am~ mtcars$mpg, family = "binomial")
summary(logCars)
```

```
##
## Call:
## glm(formula = mtcars$am ~ mtcars$mpg, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5701  -0.7531  -0.4245   0.5866   2.0617
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6035     2.3514  -2.808  0.00498 **
## mtcars$mpg    0.3070     0.1148   2.673  0.00751 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 29.675  on 30  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 5
```

```
logCars$fitted
```

```
##          1          2          3          4          5          6
## 0.46109512 0.46109512 0.59789839 0.49171990 0.29690087 0.25993307
##          7          8          9         10         11         12
## 0.09858705 0.70846924 0.59789839 0.32991148 0.24260966 0.17246396
##         13         14         15         16         17         18
## 0.21552479 0.12601104 0.03197098 0.03197098 0.11005178 0.96591395
##         19         20         21         22         23         24
## 0.93878132 0.97821971 0.49939484 0.13650937 0.12601104 0.07446438
##         25         26         27         28         29         30
## 0.32991148 0.85549212 0.79886349 0.93878132 0.14773451 0.36468861
##         31         32
## 0.11940215 0.49171990
```

GLM Summary

The models for the GLM are summarized here. The Akaike Information Criterion (aic) is proper for the model as we are looking at the likelihood of the vehicle being either automatic or manual. The aic measures the dispersion of data points for models of likelihood. The AM = .307 - 6.60 has the largest AIC compared to the rest of the models

```
summary(Model1)
```

```
##
## Call:
## glm(formula = am ~ mpg, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5701  -0.7531  -0.4245   0.5866   2.0617
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6035     2.3514  -2.808  0.00498 **
## mpg           0.3070     0.1148   2.673  0.00751 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 29.675  on 30  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 5
```

```
summary(Model2)
```

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

```
## glm(formula = am ~ mpg + wt, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50806  -0.45191  -0.04684   0.24664   2.01168
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  25.8866    12.1935   2.123   0.0338 *
## mpg          -0.3242     0.2395  -1.354   0.1759
## wt           -6.4162     2.5466  -2.519   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 17.184  on 29  degrees of freedom
## AIC: 23.184
##
## Number of Fisher Scoring iterations: 7
```

```r
summary(Model3)
```

```
##
## Call:
## glm(formula = am ~ mpg + wt + hp, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93381  -0.09191  -0.00913   0.01139   1.47331
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.72137   40.00281  -0.393   0.6943
## mpg           1.22930    1.58109   0.778   0.4369
## wt           -6.95492    3.35297  -2.074   0.0381 *
## hp            0.08389    0.08228   1.020   0.3079
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.2297  on 31  degrees of freedom
## Residual deviance:  8.7661  on 28  degrees of freedom
## AIC: 16.766
##
## Number of Fisher Scoring iterations: 10
```

```r
summary(Model4)
```

```
##
## Call:
## glm(formula = am ~ mpg + wt + hp + disp, family = "binomial",
##     data = mtcars)
```

```
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.84992  -0.15966  -0.00615   0.01257   1.46081  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.48207   40.90451  -0.452    0.651
## mpg           1.13503    1.55720   0.729    0.466
## wt           -4.80560    3.97978  -1.208    0.227
## hp            0.10871    0.09837   1.105    0.269
## disp         -0.02588    0.04087  -0.633    0.527
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance:  8.162  on 27  degrees of freedom
## AIC: 18.162
## 
## Number of Fisher Scoring iterations: 9
```

Evaluating the AIC

Looking

```r
1-pchisq(Model1$aic,Model1$df.residual)
```

```
## [1] 0.2940046
```

```r
1-pchisq(Model2$aic,Model2$df.residual)
```

```
## [1] 0.768044
```

```r
1-pchisq(Model3$aic,Model3$df.residual)
```

```
## [1] 0.953067
```

```r
1-pchisq(Model4$aic,Model4$df.residual)
```

```
## [1] 0.8984913
```

```r
exp(logCars$coefficients)
```

```
## (Intercept)  mtcars$mpg 
## 0.001355579 1.359379288
```

```r
exp(confint(logCars))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %     97.5 %
## (Intercept) 4.425443e-06 0.06255158
## mtcars$mpg  1.129764e+00 1.79946863
```

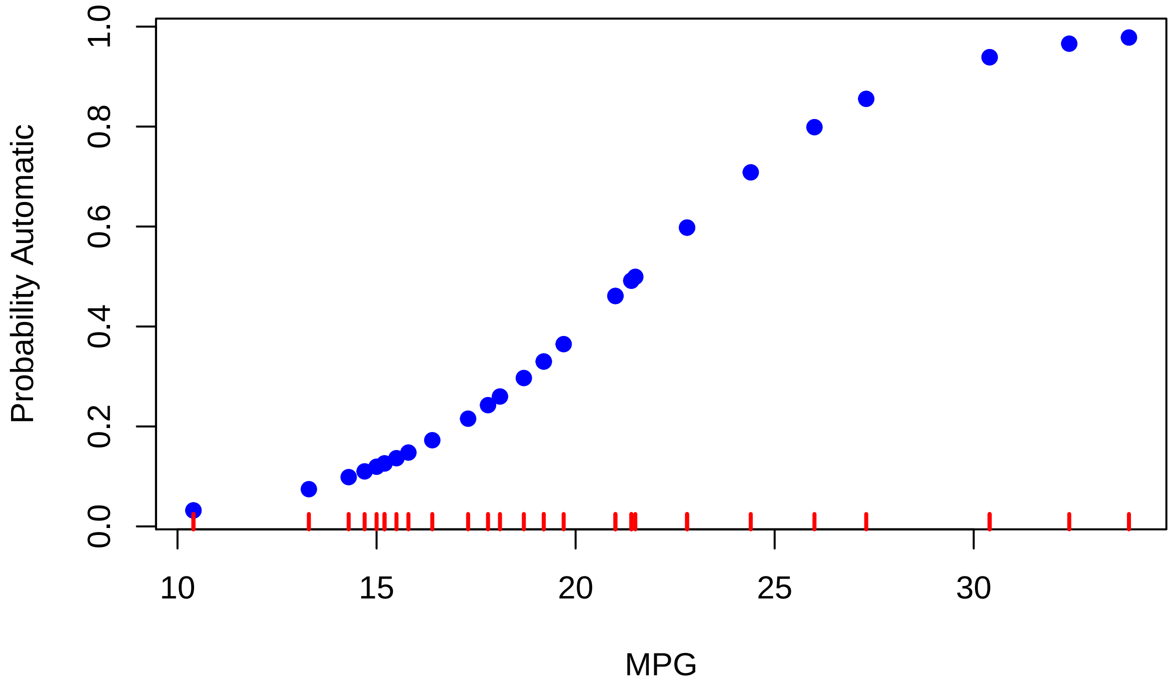
```r
anova(logCars, test = "Chisq")
```

```
## Analysis of Deviance Table
## 
## Model: binomial, link: logit
```

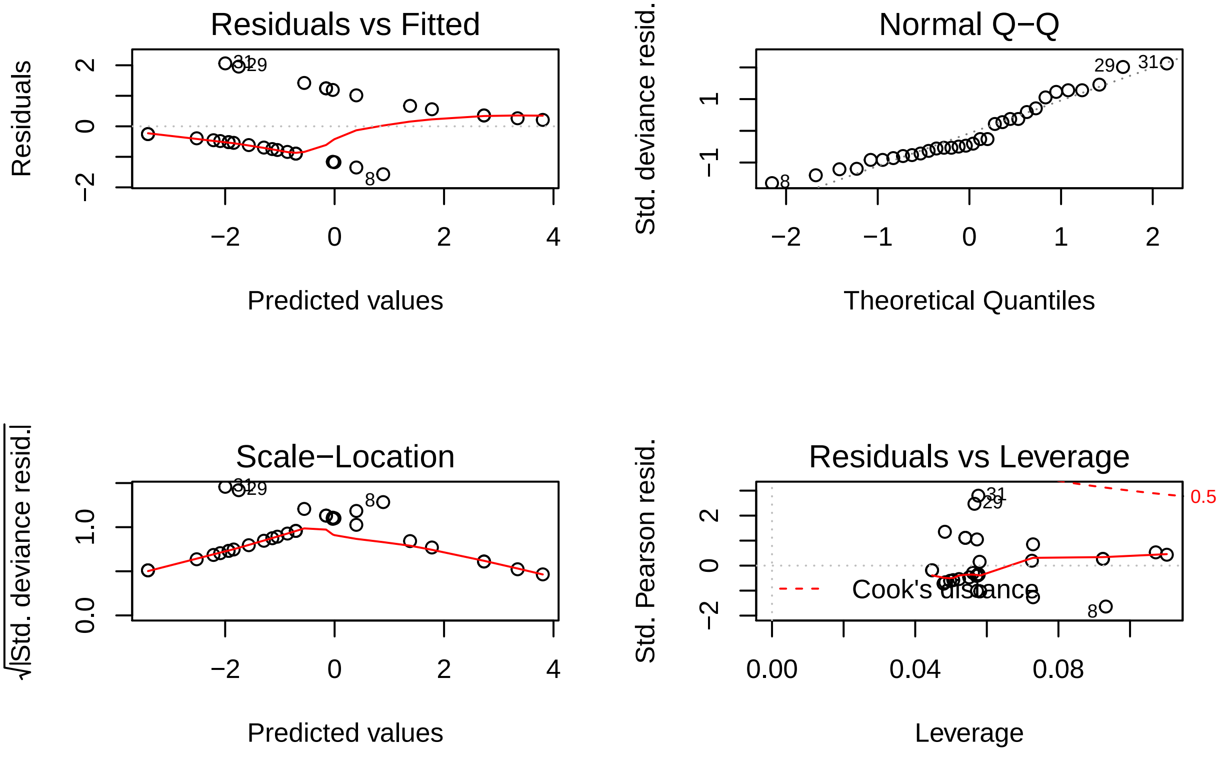
```
##
## Response: mtcars$am
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          31     43.230
## mtcars$mpg  1   13.555        30     29.675 0.0002317 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Probability Plot Automatic Transmission

```
plot(mtcars$mpg,logCars$fitted,pch=19,col="blue",xlab="MPG",ylab="Probability Automatic")
rug(mtcars$mpg, lwd = 2, col = "red")
```



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


Residuals vs Fitted
Residuals
2
0
−2
31
29
8
−2
0
2
4
Predicted values
Normal Q−Q
Std. deviance resid.
1
−1
29
31
8
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
1.0
0.0
31
29
8
−2
0
2
4
Predicted values
Residuals vs Leverage
Std. Pearson resid.
2
0
−2
31
29
8
0.5
Cook's distance
0.00
0.04
0.08
Leverage