MPG Regression Analysis

Michael Garcia

5/17/2018

Executive 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"

The analysis includes to sets of stepped models. Those with "Model#" are Generalized Linear Model using binomial distribution and logiscit regression; "LModel#" are the models using Linear Models.

The question involves a response of binary data: is it MPG automatic or manual. The models tested involve linear models and generalized linear model with binomial family function. The coefficient of determination is highest for LModel5. However, the predictors have p-values indiciating they are not significant to the model. The question is whether there probability of a vehicle being automatic with higher miles per gallon. So the a logistic GLM is the better approach for binomial outcome.

The logistic generalized model with the best fit is Model1 or AM = 0.307 MPG - 6.604.

Exploratory Analysis

s

```
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
```

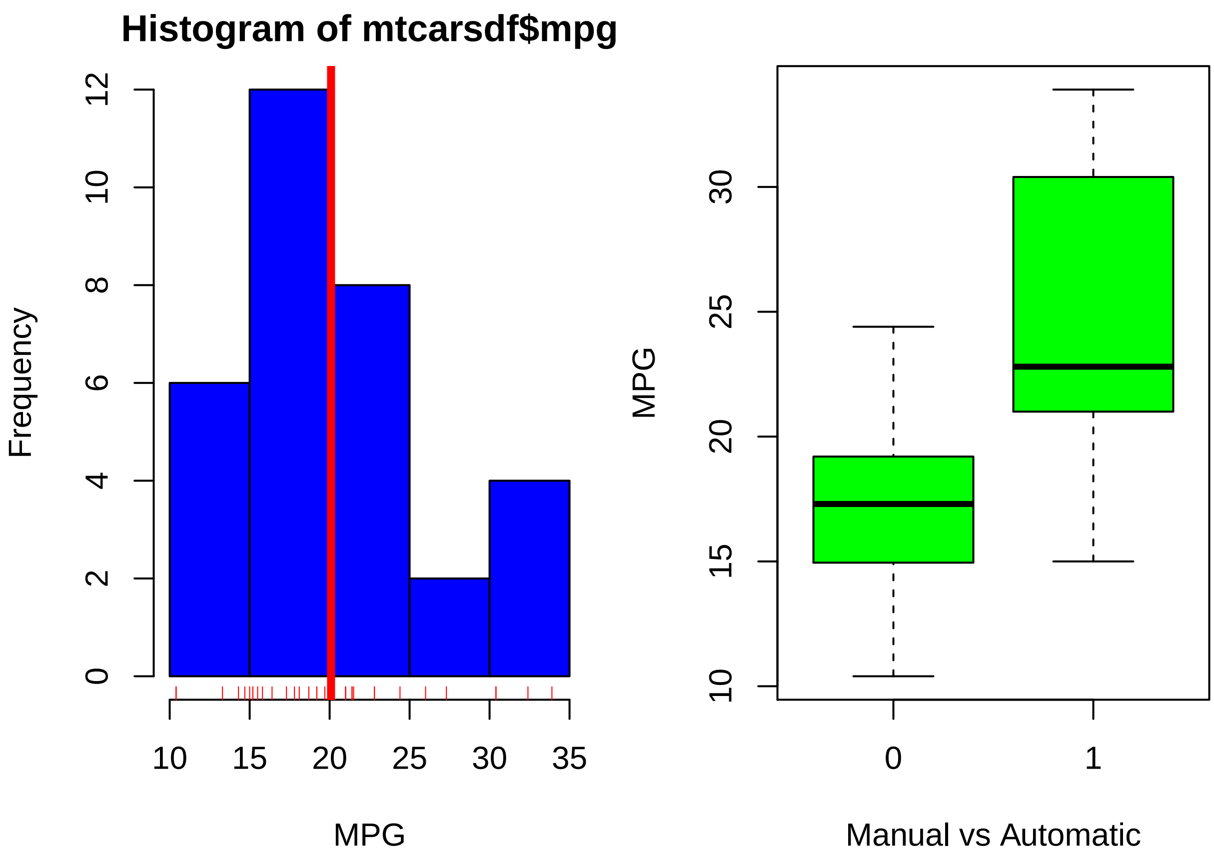
```r
mtcarsdf <- as.data.frame(mtcars)
```

Exploratory Data Analysis- Distribution

The distributions for the mpg for the total dataset are reflected

```r
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")
```



You can also embed plots, for example:

Nested Fitting - Generalized Linear Model

```r
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

```r
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

```r
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

```r
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
```

```r
LModelSum1 <- summary(LModel1)
LModelSum2 <- summary(LModel2)
LModelSum3 <- summary(LModel3)
LModelSum4 <- summary(LModel4)
LModelSum5 <- summary(LModel5)
LModelSum1$r.squared
```

```
## [1] 0.3597989
```

```r
LModelSum2$r.squared
```

```
## [1] 0.7528348
```

```r
LModelSum3$r.squared
```

```
## [1] 0.8398903
```

```r
LModelSum4$r.squared
```

```
## [1] 0.8402309
```

```r
LModelSum5$r.squared
```

```
## [1] 0.8690158
```

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.

```r
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
```

```r
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
```

```
logCars$coefficients
```

```
## (Intercept)  mtcars$mpg
##  -6.6035267   0.3070282
```

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
```

GLM Step and Best Model Selection

```r
base_model <- glm(am ~ ., data = mtcars)
optimal_model <- step(base_model, direction = "both")
```

```
## Start:  AIC=17.93
## am ~ mpg + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
##
##        Df Deviance    AIC
## - wt    1   1.5502 15.936
## - carb  1   1.5530 15.995
## - disp  1   1.5567 16.071
## - hp    1   1.5651 16.243
## - drat  1   1.5736 16.416
## - vs    1   1.6219 17.384
## - cyl   1   1.6494 17.922
## <none>      1.5497 17.926
## - mpg   1   1.6605 18.136
## - gear  1   1.6785 18.482
## - qsec  1   1.7247 19.350
##
## Step:  AIC=15.94
## am ~ mpg + cyl + disp + hp + drat + qsec + vs + gear + carb
##
##        Df Deviance    AIC
## - carb  1   1.5533 14.001
## - disp  1   1.5609 14.156
## - hp    1   1.5653 14.248
## - drat  1   1.5736 14.417
## - vs    1   1.6236 15.417
## <none>      1.5502 15.936
## - cyl   1   1.6531 15.993
## - mpg   1   1.6754 16.422
## - gear  1   1.6791 16.493
## + wt    1   1.5497 17.926
## - qsec  1   1.7941 18.613
##
## Step:  AIC=14
## am ~ mpg + cyl + disp + hp + drat + qsec + vs + gear
##
##        Df Deviance    AIC
## - disp  1   1.5613 12.164
## - hp    1   1.5654 12.248
## - drat  1   1.5745 12.434
```

```
## - vs    1   1.6238 13.420
## <none>      1.5533 14.001
## - cyl   1   1.6698 14.316
## - gear  1   1.6933 14.762
## - mpg   1   1.7304 15.456
## + carb  1   1.5502 15.936
## + wt    1   1.5530 15.995
## - qsec  1   1.8134 16.954
##
## Step:  AIC=12.16
## am ~ mpg + cyl + hp + drat + qsec + vs + gear
##
##        Df Deviance    AIC
## - hp    1   1.5677 10.296
## - drat  1   1.5842 10.630
## - vs    1   1.6255 11.454
## <none>      1.5613 12.164
## - cyl   1   1.7031 12.947
## - gear  1   1.7371 13.580
## - mpg   1   1.7512 13.839
## + disp  1   1.5533 14.001
## + wt    1   1.5567 14.071
## + carb  1   1.5609 14.156
## - qsec  1   1.8506 15.605
##
## Step:  AIC=10.3
## am ~ mpg + cyl + drat + qsec + vs + gear
##
##        Df Deviance     AIC
## - drat  1   1.5916  8.7807
## - vs    1   1.6264  9.4724
## <none>      1.5677 10.2955
## - cyl   1   1.7185 11.2350
## - mpg   1   1.7643 12.0771
## + hp    1   1.5613 12.1642
## + disp  1   1.5654 12.2479
## + wt    1   1.5659 12.2584
## + carb  1   1.5676 12.2950
## - gear  1   1.8101 12.8961
## - qsec  1   1.8965 14.3893
##
## Step:  AIC=8.78
## am ~ mpg + cyl + qsec + vs + gear
##
##        Df Deviance     AIC
## - vs    1   1.6505  7.9429
## <none>      1.5916  8.7807
## + drat  1   1.5677 10.2955
## + hp    1   1.5842 10.6305
## + disp  1   1.5887 10.7212
## + wt    1   1.5897 10.7418
## + carb  1   1.5904 10.7570
## - cyl   1   1.8074 10.8484
## - mpg   1   1.8148 10.9804
```

```
## - gear  1   1.9097 12.6110
## - qsec  1   1.9780 13.7359
##
## Step:  AIC=7.94
## am ~ mpg + cyl + qsec + gear
##
##        Df Deviance     AIC
## <none>      1.6505  7.9429
## + vs    1   1.5916  8.7807
## - cyl   1   1.8138  8.9619
## + drat  1   1.6264  9.4724
## - mpg   1   1.8666  9.8803
## + carb  1   1.6489  9.9113
## + hp    1   1.6492  9.9170
## + disp  1   1.6494  9.9219
## + wt    1   1.6505  9.9429
## - gear  1   1.9579 11.4076
## - qsec  1   2.3181 16.8127
```

```r
summary(optimal_model)
```

```
##
## Call:
## glm(formula = am ~ mpg + cyl + qsec + gear, data = mtcars)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51557  -0.15860  -0.00793   0.19350   0.35820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.36836    1.52965   1.548  0.13319
## mpg          0.02703    0.01438   1.880  0.07091 .
## cyl         -0.11052    0.06762  -1.634  0.11378
## qsec        -0.14810    0.04481  -3.305  0.00269 **
## gear         0.22288    0.09940   2.242  0.03335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.06112933)
##
##     Null deviance: 7.7188  on 31  degrees of freedom
## Residual deviance: 1.6505  on 27  degrees of freedom
## AIC: 7.9429
##
## Number of Fisher Scoring iterations: 2
```

Evaluating the AIC

```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
```

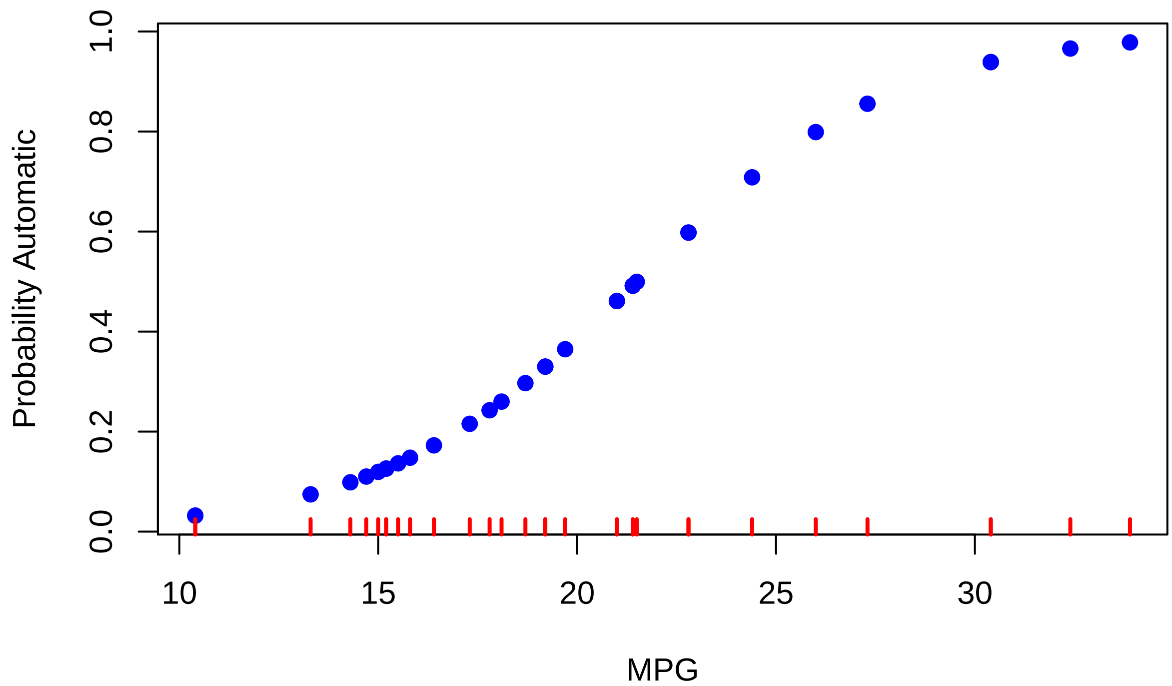
```r
anova(logCars, test = "Chisq")
```

```
## 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
```

Appendix

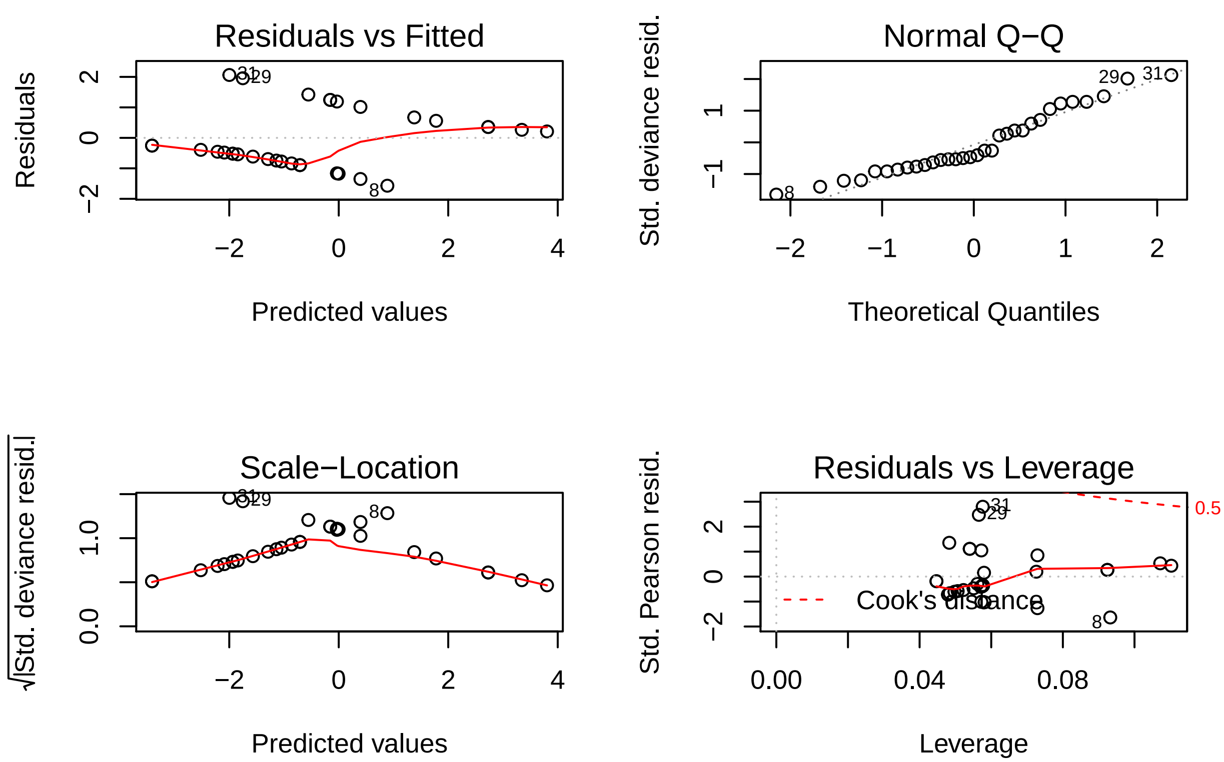
Probability Plot Automatic Transmission

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



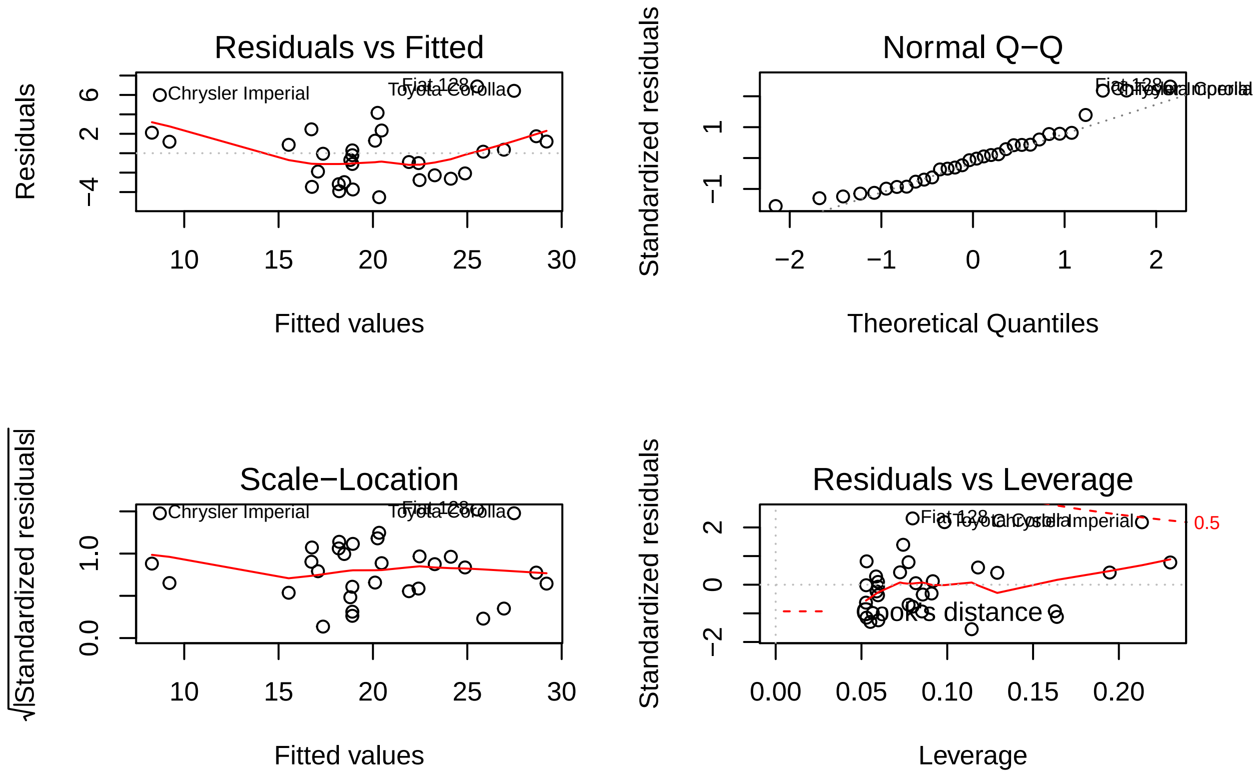
Residuals and Fit Plots - Logistic GLM Model

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



Residuals and Fit Plots - Linear Model- Best Model based on Significance and Coefficient of Determination

```r
par(mfrow = c(2,2))
plot(LModel2)
```



Scatterplot Matrix The plot displays the points for pairs of variables.

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