GAUTHAM MEENAKSHISUNDARAM

3rd January 2020

EXECUTIVE SUMMARY

Motor Trend, an automobile magazine is interested in analyzing the 'mtcars' dataset which contains information on a collection of 32 cars. It is interested in exploring the relationship between a set of 10 predictor variables and the response variable 'miles per gallon (mpg)'. The magazine particularly hopes to answer the following questions.

- Is automatic or manual transmission better for mpg?
- Quantify the mpg difference between automatic and manual transmissions

DATA

```
data1 <- data.frame(mtcars)
data1
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
```

```
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Response Variable (Continuous):

mpg : miles per gallon

Predictor Variables (Continuous):

disp : displacement measures the overall volume of the engine
hp : gross horsepower measures the theoretical power output of engine
drat : rear axle ratio is influenced by transmission configuration; a high ratio provides more torque
wt : overall weight of the vehicle
qsec : measure of acceleration
gear : number of gears - automatic (3 or 4) and manual (4 or 5)
carb : number of carburetor barrels; engines with high displacement have more barrels

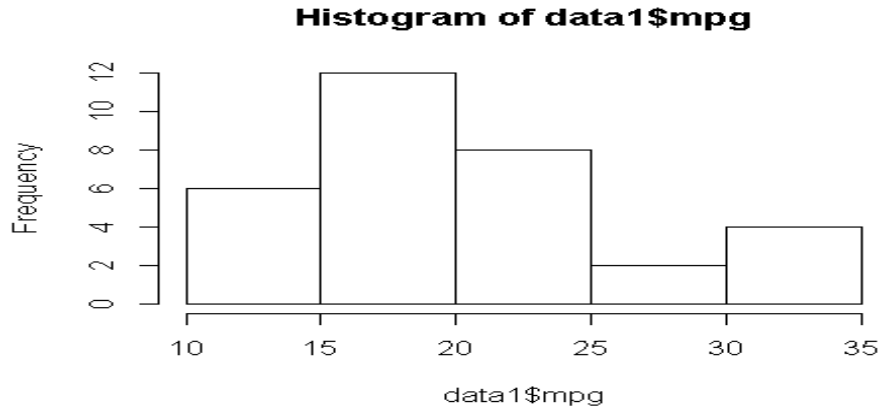
Predictor Variables (Categorical):

cyl : number of cylinders in the engine - 4, 6, 8
vs : engine cylinder configuration - V shape (0) or Straight Line (1); influences lots of specifications
am : transmission configuration - automatic (0) or manual (1)

EXPLORATORY DATA ANALYSIS

Response Variable (Continuous):

```
hist(data1$mpg)
```



The response variable is checked for normality and its histogram does not exactly show a normal distribution but it is not non-normal or skewed either. The distribution can be considered to be at least symmetric and does not require any transformation.

Predictor Variables (Categorical):

```
table(data1$cyl)

##
##  4  6  8
## 11  7 14

table(data1$vs)

##
##  0  1
## 18 14

table(data1$am)

##
##  0  1
## 19 13
```

The frequency counts of the categorical predictor variables is largely distributed evenly between their categories and there is no need to change the choice of reference category in any of these variables.

Predictor Variables (Continuous):

```
data2 <- data1[c(1,3:7,10:11)]
data2

##                     mpg  disp  hp drat    wt  qsec gear carb
## Mazda RX4          21.0 160.0 110 3.90 2.620 16.46    4    4
## Mazda RX4 Wag      21.0 160.0 110 3.90 2.875 17.02    4    4
## Datsun 710         22.8 108.0  93 3.85 2.320 18.61    4    1
## Hornet 4 Drive     21.4 258.0 110 3.08 3.215 19.44    3    1
## Hornet Sportabout  18.7 360.0 175 3.15 3.440 17.02    3    2
## Valiant            18.1 225.0 105 2.76 3.460 20.22    3    1
## Duster 360         14.3 360.0 245 3.21 3.570 15.84    3    4
## Merc 240D          24.4 146.7  62 3.69 3.190 20.00    4    2
## Merc 230           22.8 140.8  95 3.92 3.150 22.90    4    2
## Merc 280           19.2 167.6 123 3.92 3.440 18.30    4    4
## Merc 280C          17.8 167.6 123 3.92 3.440 18.90    4    4
## Merc 450SE         16.4 275.8 180 3.07 4.070 17.40    3    3
## Merc 450SL         17.3 275.8 180 3.07 3.730 17.60    3    3
## Merc 450SLC        15.2 275.8 180 3.07 3.780 18.00    3    3
## Cadillac Fleetwood 10.4 472.0 205 2.93 5.250 17.98    3    4
## Lincoln Continental 10.4 460.0 215 3.00 5.424 17.82    3    4
## Chrysler Imperial  14.7 440.0 230 3.23 5.345 17.42    3    4
## Fiat 128           32.4  78.7  66 4.08 2.200 19.47    4    1
```

```
## Honda Civic          30.4  75.7  52 4.93 1.615 18.52    4    2
## Toyota Corolla       33.9  71.1  65 4.22 1.835 19.90    4    1
## Toyota Corona        21.5 120.1  97 3.70 2.465 20.01    3    1
## Dodge Challenger     15.5 318.0 150 2.76 3.520 16.87    3    2
## AMC Javelin          15.2 304.0 150 3.15 3.435 17.30    3    2
## Camaro Z28           13.3 350.0 245 3.73 3.840 15.41    3    4
## Pontiac Firebird     19.2 400.0 175 3.08 3.845 17.05    3    2
## Fiat X1-9            27.3  79.0  66 4.08 1.935 18.90    4    1
## Porsche 914-2        26.0 120.3  91 4.43 2.140 16.70    5    2
## Lotus Europa         30.4  95.1 113 3.77 1.513 16.90    5    2
## Ford Pantera L       15.8 351.0 264 4.22 3.170 14.50    5    4
## Ferrari Dino         19.7 145.0 175 3.62 2.770 15.50    5    6
## Maserati Bora        15.0 301.0 335 3.54 3.570 14.60    5    8
## Volvo 142E           21.4 121.0 109 4.11 2.780 18.60    4    2
```

```
library(GGally)
```

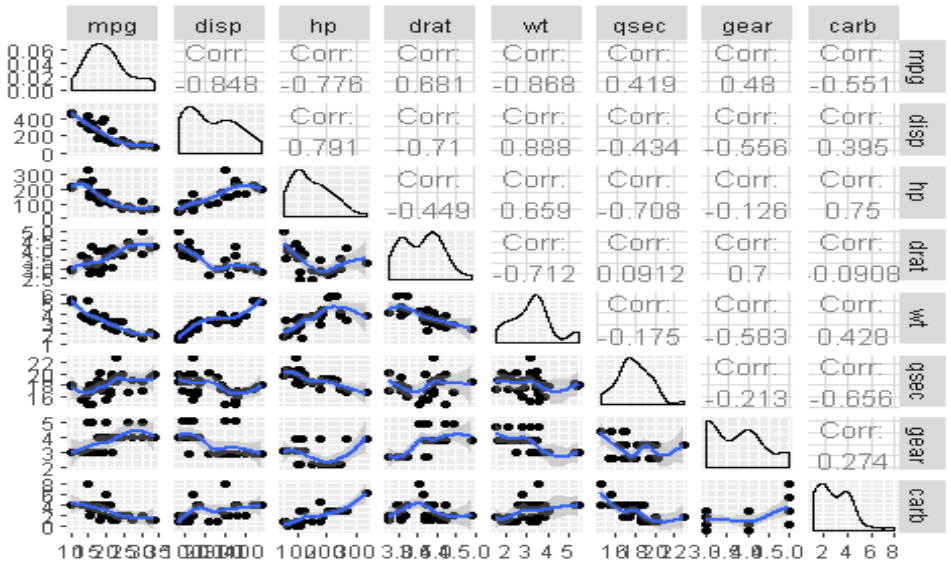
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
```

```
trendline <- function(data2, mapping, method="loess", ...){
      p <- ggplot(data = data2, mapping = mapping) +
      geom_point() +
      geom_smooth(method=method, ...)
      p
    }
```

```
ggpairs(data2, lower = list(continuous = trendline))
```



- drat and qsec shows a weak positive linear relationship with the response variable mpg
- disp, hp, wt and carb shows a weak negative linear relationship with the response variable mpg
- gear shows no relationship with the response variable mpg

None of the continuous predictor variables exhibit strong collinearity and hence, they all can be considered for model fitting.

Is automatic or manual transmission better for mpg? Quantify the mpg difference between automatic and manual transmissions

```
model_am <- glm(mpg~factor(am) - 1, family=gaussian, data=data1)
summary(model_am)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## factor(am)0 17.14737   1.124603 15.24749 1.133983e-15
## factor(am)1 24.39231   1.359578 17.94109 1.376283e-17
```

The manual transmission (am)1 provides 7.244494 mpg more than automatic transmission (am)0.

UNIVARIATE REGRESSION MODELS

```
library(car)
```

```
## Loading required package: carData
```

```
model_cyl <- glm(mpg~factor(cyl), family=gaussian, data=data1)
model_disp <- glm(mpg~disp, family=gaussian, data=data1)
model_hp <- glm(mpg~hp, family=gaussian, data=data1)
model_drat <- glm(mpg~drat, family=gaussian, data=data1)
model_wt <- glm(mpg~wt, family=gaussian, data=data1)
model_qsec <- glm(mpg~qsec, family=gaussian, data=data1)
model_vs <- glm(mpg~factor(vs), family=gaussian, data=data1)
model_am <- glm(mpg~factor(am), family=gaussian, data=data1)
model_gear <- glm(mpg~gear, family=gaussian, data=data1)
model_carb <- glm(mpg~carb, family=gaussian, data=data1)

Anova(model_cyl)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##             LR Chisq Df Pr(>Chisq)
## factor(cyl)   79.395  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_disp)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##      LR Chisq Df Pr(>Chisq)
## disp   76.513  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_hp)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##    LR Chisq Df Pr(>Chisq)
## hp    45.46  1  1.558e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_drat)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##      LR Chisq Df Pr(>Chisq)
## drat    25.97  1  3.468e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_wt)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##    LR Chisq Df Pr(>Chisq)
## wt   91.375  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_qsec)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##      LR Chisq Df Pr(>Chisq)
## qsec   6.3767  1    0.01156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(model_vs)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##            LR Chisq Df Pr(>Chisq)
## factor(vs)   23.662  1  1.148e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_am)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##            LR Chisq Df Pr(>Chisq)
## factor(am)    16.86  1  4.023e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_gear)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##      LR Chisq Df Pr(>Chisq)
## gear   8.9951  1   0.002707 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_carb)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: mpg
##      LR Chisq Df Pr(>Chisq)
## carb   13.074  1  0.0002995 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Each predictor variable is fitted against the response variable mpg individually and checked for the p value being less than 0.2 for inclusion in multivariate regression models. All predictor variables can be considered for multivariate regression.

MULTIVARIATE REGRESSION MODELS

```
model_full <-
step(glm(mpg~factor(am)+factor(cyl)+factor(vs)+disp+hp+drat+wt+qsec+gear+carb
, family=gaussian, data=data1))
```

```
## Start:  AIC=162.48
## mpg ~ factor(am) + factor(cyl) + factor(vs) + disp + hp + drat +
##     wt + qsec + gear + carb
```

```
##
##               Df Deviance    AIC
## - drat         1   133.32 160.48
## - gear         1   135.16 160.91
## - carb         1   135.27 160.94
## - factor(vs)   1   137.26 161.41
## - disp         1   137.58 161.48
## - factor(cyl)  2   147.57 161.73
## - qsec         1   138.68 161.74
## <none>             133.32 162.48
## - factor(am)   1   144.69 163.09
## - hp           1   152.61 164.80
## - wt           1   161.64 166.64
##
## Step:  AIC=160.48
## mpg ~ factor(am) + factor(cyl) + factor(vs) + disp + hp + wt +
##     qsec + gear + carb
##
##               Df Deviance    AIC
## - gear         1   135.18 158.92
## - carb         1   135.55 159.01
## - factor(vs)   1   137.40 159.44
## - disp         1   137.66 159.50
## - qsec         1   138.70 159.74
## - factor(cyl)  2   149.56 160.15
## <none>             133.32 160.48
## - factor(am)   1   145.03 161.17
## - hp           1   154.09 163.11
## - wt           1   162.57 164.82
##
## Step:  AIC=158.92
## mpg ~ factor(am) + factor(cyl) + factor(vs) + disp + hp + wt +
##     qsec + carb
##
##               Df Deviance    AIC
## - factor(vs)   1   139.43 157.91
## - carb         1   139.99 158.04
## - disp         1   140.08 158.06
## - qsec         1   140.10 158.06
## - factor(cyl)  2   152.28 158.73
## <none>             135.18 158.92
## - factor(am)   1   152.01 160.68
## - hp           1   155.07 161.31
## - wt           1   168.73 164.01
##
## Step:  AIC=157.91
## mpg ~ factor(am) + factor(cyl) + disp + hp + wt + qsec + carb
##
##               Df Deviance    AIC
## - carb         1   142.33 156.57
```

```
## - disp         1   143.65 156.86
## - factor(cyl)  2   153.43 156.97
## <none>             139.43 157.91
## - qsec         1   150.15 158.28
## - factor(am)   1   153.79 159.05
## - hp           1   155.08 159.31
## - wt           1   175.77 163.32
##
## Step:  AIC=156.57
## mpg ~ factor(am) + factor(cyl) + disp + hp + wt + qsec
##
##               Df Deviance    AIC
## - disp         1   143.98 154.94
## - factor(cyl)  2   153.44 154.97
## - qsec         1   150.41 156.34
## <none>             142.33 156.57
## - hp           1   157.73 157.86
## - factor(am)   1   159.75 158.26
## - wt           1   183.04 162.62
##
## Step:  AIC=154.94
## mpg ~ factor(am) + factor(cyl) + hp + wt + qsec
##
##               Df Deviance    AIC
## - factor(cyl)  2   160.07 154.33
## - qsec         1   151.03 154.47
## <none>             143.98 154.94
## - hp           1   159.42 156.20
## - factor(am)   1   160.55 156.42
## - wt           1   196.91 162.96
##
## Step:  AIC=154.33
## mpg ~ factor(am) + hp + wt + qsec
##
##              Df Deviance    AIC
## - hp          1   169.29 154.12
## <none>            160.07 154.33
## - qsec        1   180.29 156.13
## - factor(am)  1   186.06 157.14
## - wt          1   238.56 165.10
##
## Step:  AIC=154.12
## mpg ~ factor(am) + wt + qsec
##
##              Df Deviance    AIC
## <none>            169.29 154.12
## - factor(am)  1   195.46 156.72
## - qsec        1   278.32 168.03
## - wt          1   352.63 175.60
```

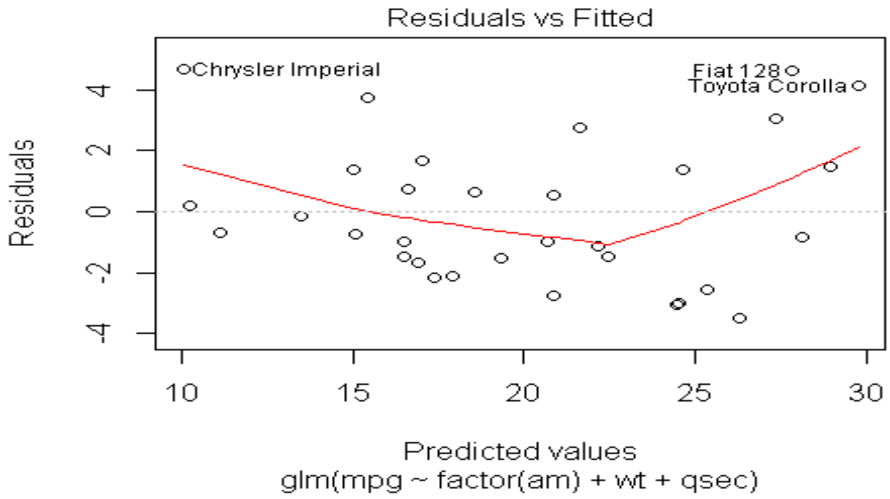
The model with the least Akaike Information Criterion (AIC) score is the best fit model. The model with predictor variables am, qsec and wt against response variable mpg has the least AIC score of 154.12

MODEL DIAGNOSTICS

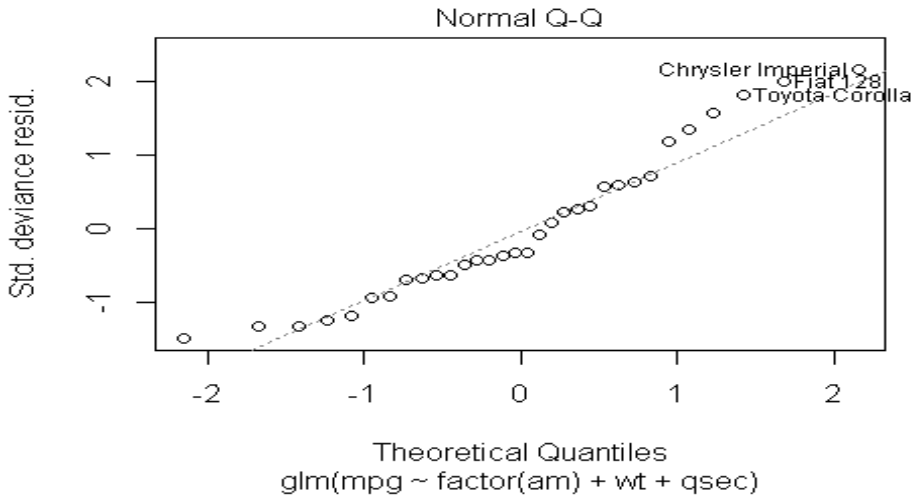
```
best_model <- glm(mpg~factor(am)+wt+qsec, family=gaussian, data=data1)

plot(best_model)
```

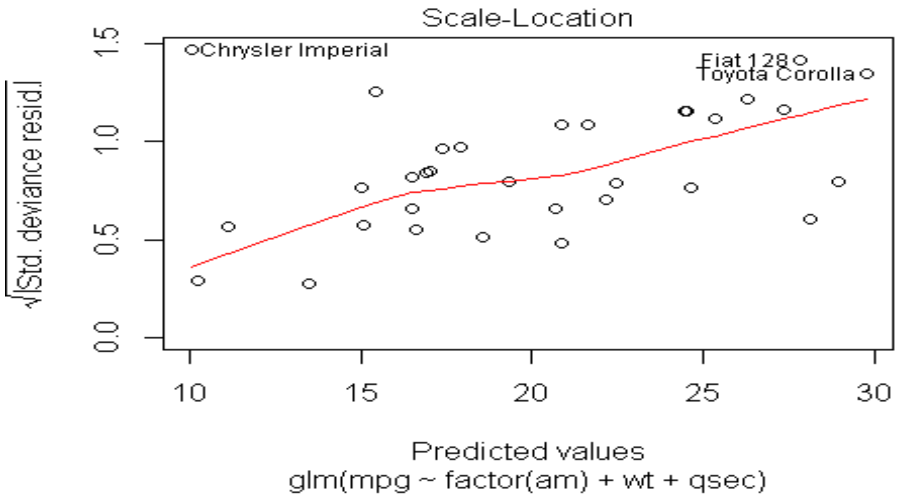
Residuals vs Fitted plot: Residuals are in a random scatter largely around the zero line



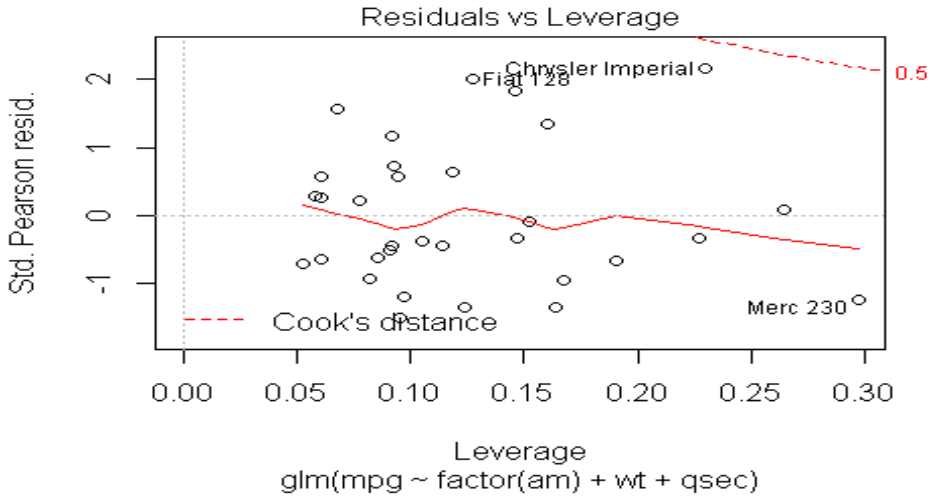
Normal Q-Q plot: Residuals are not exactly lined up on the dashed line but are largely distributed around it, indicating normality



Scale-Location plot: Residuals begin to spread wider and wider along the x axis on a steep angle which is not ideal

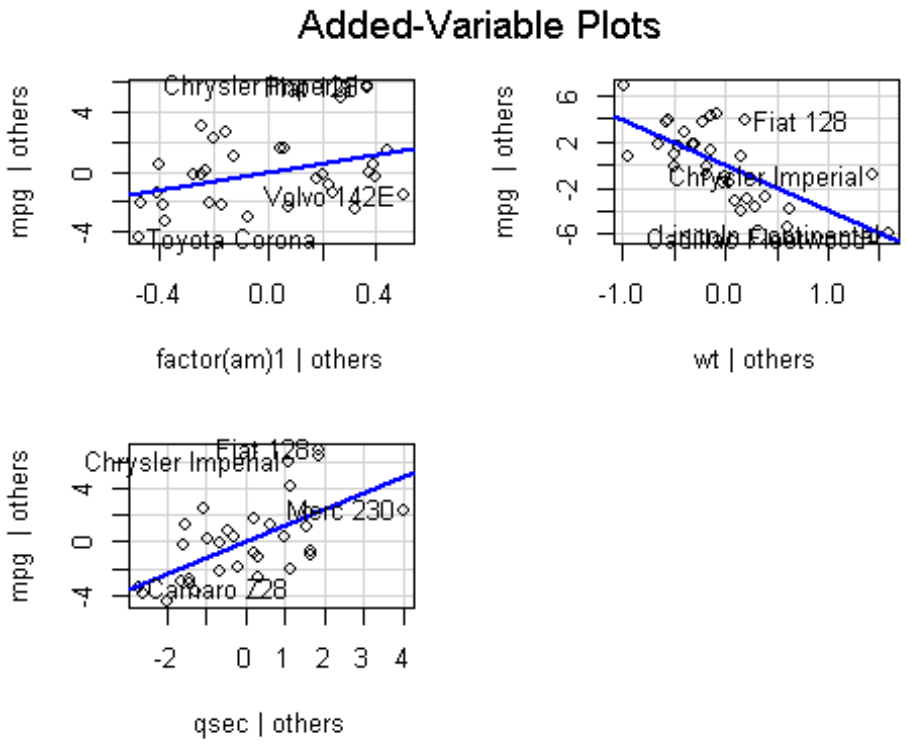


Residuals vs Leverage plot: No residuals outside the Cook's distance line influencing the regression



```
avPlots(best_model)
```

Added Variable plot: Linear relationship exhibited between all the variables added in the model and the response variable



MODEL INTERPRETATION

```
summary(best_model)
```

```
##
## Call:
## glm(formula = mpg ~ factor(am) + wt + qsec, family = gaussian,
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4811  -1.5555  -0.7257   1.4110   4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 6.045926)
##
##     Null deviance: 1126.05  on 31  degrees of freedom
## Residual deviance:  169.29  on 28  degrees of freedom
## AIC: 154.12
##
## Number of Fisher Scoring iterations: 2
```

Model Equation: y = 9.6178 + 2.9358 am1 - 3.9165 wt + 1.2259 qsec

- The manual transmission increases miles per gallon at a rate of exp(2.9358) which is 18.8366 times more than automatic transmission
- The weight decreases miles per gallon at a rate of exp(-3.9165) by 0.0200 times
- The acceleration increases miles per gallon at a rate of exp(1.2259) by 3.4072 times