Motor Trend Data Analysis

Jeffrey M. Hunter

02 June, 2019

Contents

Course Project

Regression Models Course Project

Peer-graded Assignment

- This course project is available on GitHub

  Motor Trend Data Analysis

Executive Summary

This analysis is being performed for Motor Trend, a popular American automobile magazine, to evaluate the relationship between transmission type (manual or automatic) and fuel consumption in miles per gallon (MPG) in automobiles. The analysis extends beyond transmission type to also include other possible variables that explain variance in fuel consumption (MPG).

As part of this analysis, Motor Trend is particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

The analysis will be conducted using exploratory and inferential data analyses and linear regression models using the `mtcars` dataset.

Data Description

The mtcars dataset is comprised of data that was was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

The mtcars dataset is a data frame with 32 observations on 11 (numeric) variables:

- `mpg` Miles/(US) gallon
- `cyl` Number of cylinders

- `disp` Displacement (cu.in.)
- `hp` Gross horsepower
- `drat` Rear axle ratio
- `wt` Weight (1000 lbs)
- `qsec` 1/4 mile time
- `vs` Engine (0 = V-shaped, 1 = straight)
- `am` Transmission (0 = automatic, 1 = manual)
- `gear` Number of forward gears
- `carb` Number of carburetors

Environment Setup

```
if (!require(knitr)) {
    install.packages("knitr")
    library(knitr)
}
if (!require(kableExtra)) {
    install.packages("kableExtra")
    library(kableExtra)
}
if (!require(ggplot2)) {
    install.packages("ggplot2")
    library(ggplot2)
}
if (!require(GGally)) {
    install.packages("GGally")
    library(GGally)
}
if (!require(MASS)) {
    install.packages("MASS")
    library(MASS)
}
```

Load Data

Load the mtcars dataset and display the internal structure of the variables.

```
library(datasets)
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
```

```
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Data Analysis

Perform some basic exploratory and inferential data analysis of the data to study the relationship between transmission type (manual or automatic) and automobile fuel consumption in miles per gallon (MPG).

Basic Data Summary

The mtcars dataset includes 1 target variable (mpg) and 10 independent control variables with 32 observations.

See A.1 Basic Data Summary in the Appendix section which shows the range and quartiles for each variable.

Relative Mean

Display the relative mean of automobile fuel consumption data grouped by transmission type.

```
by(data = mtcars$mpg,
   INDICES = list(factor(mtcars$am, labels = c("Automatic", "Manual"))), summary)
```

```
## : Automatic
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
## --------------------------------------------------------
## : Manual
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

Impact of Transmission Type on Fuel Consumption

Figure A.2.1 (Boxplot) and Figure A.2.2 (Histogram) in the Appendix section plot the relationship between transmission type and fuel consumption in automobiles.

Inferential Statistics

Hypothesis testing will be conducted to study the impact of transmission type on fuel consumption in automobiles. A t-test will be performed on the null hypothesis that transmission type has no effect on automobile fuel consumption.

```
t1 <- t.test(mpg ~ am, data = mtcars, conf.level = 0.95)
paste0("p-value = ", round(t1$p.value, 4))
paste0("confidence interval = (",
       round(t1$conf.int[1], 4),
       ", ",
       round(t1$conf.int[2], 4), ")")
t1$estimate
```

```
## [1] "p-value = 0.0014"
## [1] "confidence interval = (-11.2802, -3.2097)"
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The observed p-value 0.0014 is less than 0.05 and the 95% confidence interval does not contain zero. This indicates strong evidence against the null hypothesis so the null hypothesis can be rejected.

Observation

The calculated mean for both transmission types and the provided plots show a significant increase in better fuel consumption for automobiles with a manual transmission versus automatic.

The difference between mean fuel consumption of automatic and manual transmission is significantly different where the estimated difference favors a manual transmission by 7.24 MPG.

Linear Models

Linear regression analysis will be used to extend beyond our initial interest in the relationship between the transmission type variable only and fuel consumption. Other possible variables in the mtcars dataset may better explain variance in fuel consumption.

Simple Linear Regression Model

Based on our initial interest with only transmission type, start by building a simple linear regression model between the response variable (MPG) and the single predictor (transmission type).

```
singleModelFit <- lm(mpg ~ am, data = mtcars)
summary(singleModelFit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Here we see that the adjusted R^2 is only 0.3385 which suggests that this model can only explain 33.8% of the variance in fuel consumption (approximately one third) based on transmission type alone.

Looking to the correlation table below, it's possible that other variables in the dataset can better explain the outcome:

```
round(cor(mtcars, method = "pearson")[1,], 2)
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

Also see A.3 Correlation Matrix Plot in the Appendix section which provides a plot showing the correlation coefficients between all variables.

Multiple Linear Regression Model

Perform stepwise regression using the `stepAIC()` function from the `MASS` package to find the subset of variables which result in the best model fit (a model that lowers prediction error).

Start by building an initial model with all variables as predictors. Stepwise regression will select the significant predictors for the final model which is the best model. The AIC algorithm runs `lm` multiple times to build multiple regression models and selects the best variables from them using both forward selection and backward elimination methods.

```r
initialModel <- lm(mpg ~ ., data = mtcars)
stepReg <- stepAIC(initialModel, direction = "both")
```

See A.4 Stepwise Regression for stepwise regression output.

Show results of stepwise regression variable selection.

```r
print(stepReg$anova)
```

```
...
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
## Final Model:
## mpg ~ wt + qsec + am
...
```

As shown above, the best model obtained from the stepwise regression procedure consists of the predictors weight (wt) and 1/4 mile time (qsec) in addition to transmission type (am).

```r
bestModelFit <- lm(mpg ~ wt + qsec + am, data = mtcars)
summary(bestModelFit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## 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
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Here we observe that the adjusted R^2 is now 0.8336 which suggests that the new model (including the three predictors) can explain 84% of the variance in fuel consumption.

Analysis of Residuals

See A.6 Residuals Plot in the Appendix section to find the residuals plots for the multiple linear regression model (best model fit).

Overall, the fit of the multiple linear regression model and its residuals appear to support the prerequisites for a linear model and adequately explain the variance in fuel consumption.

The points in the *Residuals vs. Fitted* plot appear to be random which shows the data are independent. The plot also reveals potential outliers for the Chrysler Imperial, Fiat 128, and Toyota Corolla. The adjusted R^2 may be improved by removing those data values and studying them independently.

The points of the *Normal Q-Q* plot closely follow the line which show that the residuals are normally distributed.

The points on the *Scale-Location* plot appear to be spread equally along the horizontal line with equally (randomly) spread points allowing us to conclude equal variance (homoscedasticity).

The *Residuals vs. Leverage* plot doesn't show any influential cases as all of the cases are within the the dashed Cook's distance line. All points are within the 0.05 lines which conclude there are no outliers.

Conclusion

This analysis concludes the following:

1. **Is an automatic or manual transmission better for MPG?**

   Automobiles with a manual transmission yield better gas mileage than vehicles with an automatic transmission. However, determining fuel consumption based on transmission type alone showed that the relationship was not as statistically significant as first thought. Models were built with confounding variables such as weight (wt) and 1/4 mile time (qsec) in addition to transmission type (am) that better explain variance in fuel consumption.

2. **Quantify the MPG difference between automatic and manual transmissions.**

   Based on our simple linear regression model that only considered transmission type, the mean difference in fuel consumption increased to 7.24 MPG favoring a manual transmission.

   However, when the variables weight (wt) and 1/4 mile time (qsec) were added to the best fitted multiple regression model, the advantage of a manual transmission decreased to 2.94 MPG.

Appendix

A.1 Basic Data Summary

Provide a basic summary of the data.

```r
# target variable
kable(summary(mtcars[1]),
              row.names = FALSE,
              align = c("l"),
```

Table 1: Target Variable (MPG)

mpg
Min. :10.40
1st Qu.:15.43
Median :19.20
Mean :20.09
3rd Qu.:22.80
Max. :33.90

Table 2: Control Variables

cyl	disp	hp	drat	wt
Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513
1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581
Median :6.000	Median :196.3	Median :123.0	Median :3.695	Median :3.325
Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597	Mean :3.217
3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920	3rd Qu.:3.610
Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930	Max. :5.424

```
              caption = "Target Variable (MPG)") %>%
                kable_styling(position = "center")

# independent control variables
kable(summary(mtcars[2:6]),
              row.names = FALSE,
              align = c(rep("l", 5)),
              caption = "Control Variables") %>%
                kable_styling(position = "left")

kable(summary(mtcars[7:11]),
              row.names = FALSE,
              align = rep('l', 5),
              caption = "Control Variables (cont)") %>%
                kable_styling(position = "left")
```

A.2 Plot Impact of Transmission Type on Fuel Consumption

Plot the relationship of automobile fuel consumption as a function of transmission type.

Figure A.2.1 (Boxplot)

```
g <- ggplot(data = mtcars,
            aes(x = factor(am, labels = c("Automatic", "Manual")),
                y = mpg, fill = factor(am, labels = c("Automatic", "Manual"))))
g <- g + geom_boxplot()
g <- g + scale_colour_discrete(name = "Transmission Type")
g <- g + scale_fill_discrete(name = "Transmission Type")
g <- g + xlab("Transmission Type")
g <- g + ylab("Fuel Consumption (MPG)")
g <- g + theme(plot.title = element_text(size = 14,
                                         hjust = 0.5,
```

Table 3: Control Variables (cont)

qsec	vs	am	gear	carb
Min. :14.50	Min. :0.0000	Min. :0.0000	Min. :3.000	Min. :1.000
1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2.000
Median :17.71	Median :0.0000	Median :0.0000	Median :4.000	Median :2.000
Mean :17.85	Mean :0.4375	Mean :0.4062	Mean :3.688	Mean :2.812
3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:4.000
Max. :22.90	Max. :1.0000	Max. :1.0000	Max. :5.000	Max. :8.000

```
                                         vjust = 0.5,
                                         margin = margin(b = 12, unit = "pt")),
              axis.text.x = element_text(angle = 0,
                                         hjust = 0.5,
                                         vjust = 0.5,
                                         margin = margin(b = 10, unit = "pt")),
              axis.text.y = element_text(angle = 0,
                                         hjust = 0.5,
                                         vjust = 0.5,
                                         margin = margin(l = 10, unit = "pt")))
g <- g + ggtitle("Impact of Transmission Type on Fuel Consumption")
print(g)
```

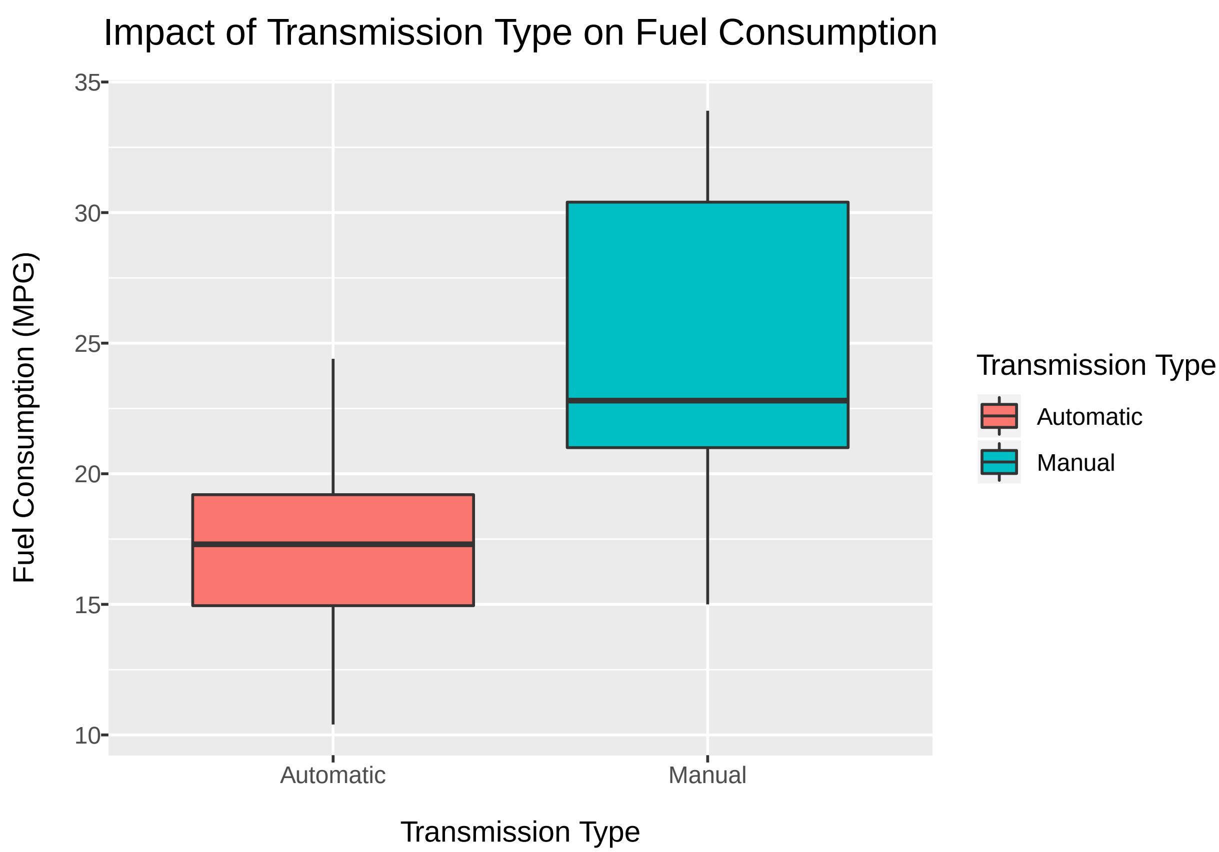
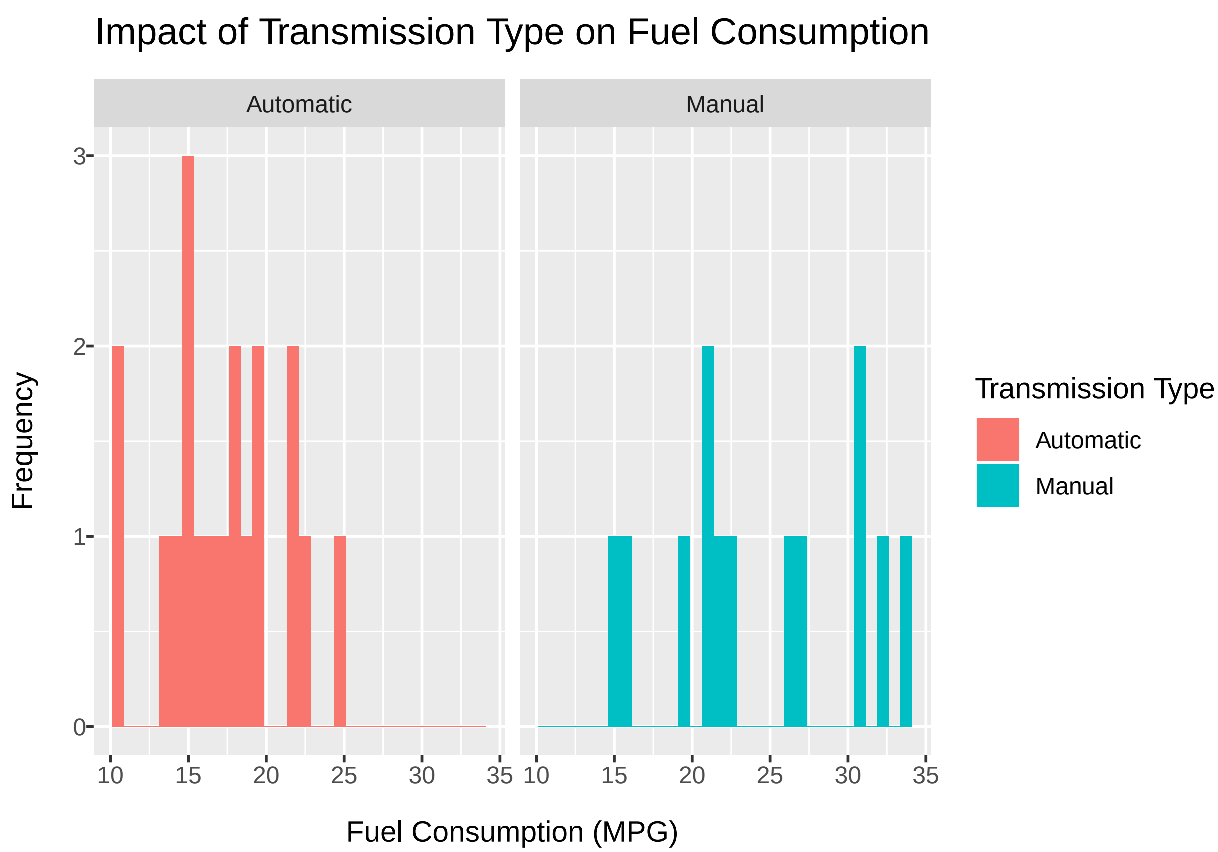


Figure A.2.2 (Histogram)

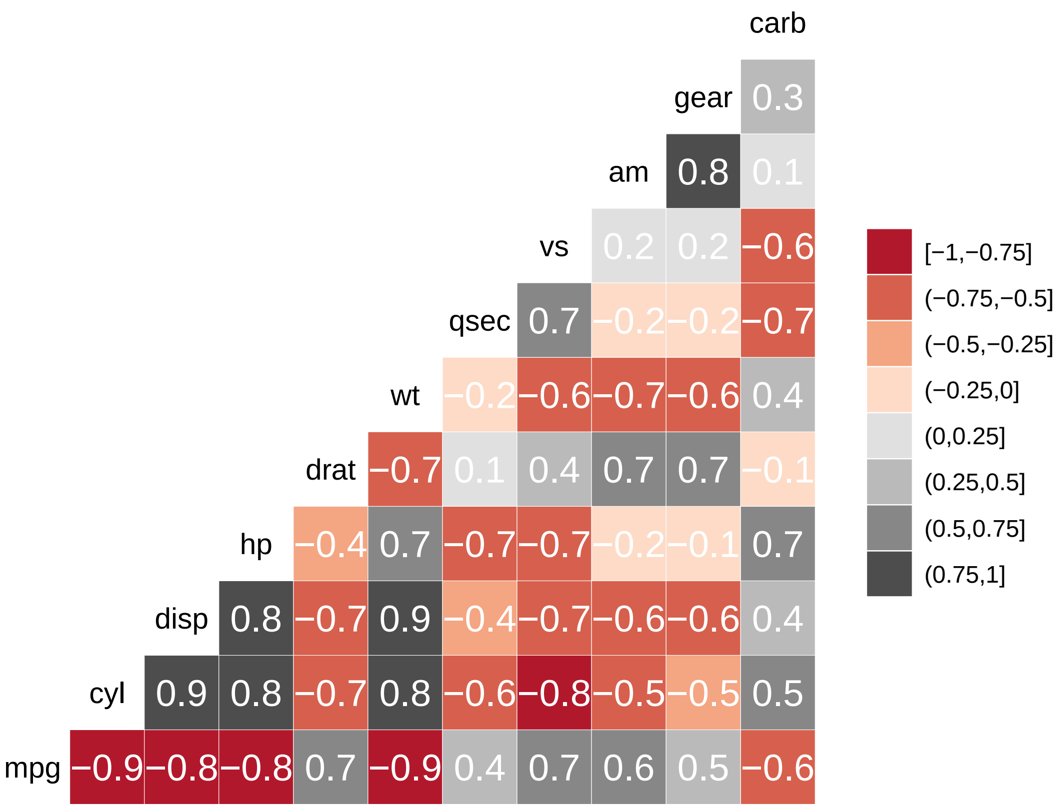
```r
g <- ggplot(data = mtcars, aes(x = mpg, y = ..count..))
g <- g + geom_histogram(binwidth = 0.75,
                        aes(fill = factor(am, labels = c("Automatic", "Manual"))))
g <- g + facet_grid(. ~ factor(am, labels = c("Automatic", "Manual")))
g <- g + scale_colour_discrete(name = "Transmission Type")
g <- g + scale_fill_discrete(name = "Transmission Type")
g <- g + xlab("Fuel Consumption (MPG)")
g <- g + ylab("Frequency")
g <- g + theme(plot.title = element_text(size = 14,
                                         hjust = 0.5,
                                         vjust = 0.5,
                                         margin = margin(b = 12, unit = "pt")),
           axis.text.x = element_text(angle = 0,
                                      hjust = 0.5,
                                      vjust = 0.5,
                                      margin = margin(b = 10, unit = "pt")),
           axis.text.y = element_text(angle = 0,
                                      hjust = 0.5,
                                      vjust = 0.5,
                                      margin = margin(l = 10, unit = "pt")))
g <- g + ggtitle("Impact of Transmission Type on Fuel Consumption")
print(g)
```



A.3 Correlation Matrix Plot

Show the correlation coefficients for all variables.

```r
ggcorr(data = mtcars,
       method = c("pairwise", "pearson"),
       nbreaks=8,
       palette='RdGy',
       label=TRUE,
       label_size=5,
       label_color='white')
```



A.4 Stepwise Regression

```r
initialModel <- lm(mpg ~ ., data = mtcars)
stepReg <- stepAIC(initialModel, direction="both")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##         Df Sum of Sq    RSS    AIC
## - cyl    1    0.0799 147.57 68.915
## - vs     1    0.1601 147.66 68.932
## - carb   1    0.4067 147.90 68.986
## - gear   1    1.3531 148.85 69.190
## - drat   1    1.6270 149.12 69.249
## - disp   1    3.9167 151.41 69.736
## - hp     1    6.8399 154.33 70.348
## - qsec   1    8.8641 156.36 70.765
## <none>               147.49 70.898
## - am     1   10.5467 158.04 71.108
```

```
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## + cyl   1    0.0799 147.49 70.898
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## + vs    1    0.2685 147.57 68.915
## + cyl   1    0.1883 147.66 68.932
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## + carb  1     0.685 147.84 66.973
## + vs    1     0.434 148.09 67.028
## + cyl   1     0.414 148.11 67.032
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
```

```
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## + gear  1     1.565 148.53 65.121
## + cyl   1     1.003 149.09 65.242
## + vs    1     0.645 149.45 65.319
## + carb  1     0.107 149.99 65.434
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## + drat  1     3.345 150.09 63.457
## + gear  1     2.977 150.46 63.535
## + cyl   1     2.447 150.99 63.648
## + vs    1     1.121 152.32 63.927
## + carb  1     0.011 153.43 64.160
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## + disp  1     6.629 153.44 62.162
## + carb  1     3.227 156.84 62.864
## + drat  1     1.428 158.64 63.229
## - qsec  1    20.225 180.29 63.323
## + cyl   1     0.249 159.82 63.465
## + vs    1     0.249 159.82 63.466
## + gear  1     0.171 159.90 63.481
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## + hp    1     9.219 160.07 61.515
## + carb  1     8.036 161.25 61.751
## + disp  1     3.276 166.01 62.682
## + cyl   1     1.501 167.78 63.022
## + drat  1     1.400 167.89 63.042
```

```
## + gear  1     0.123 169.16 63.284
## + vs    1     0.000 169.29 63.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

A.5 Model Coefficients

```r
coefficients(singleModelFit)
```

```
## (Intercept)          am
##   17.147368    7.244939
```

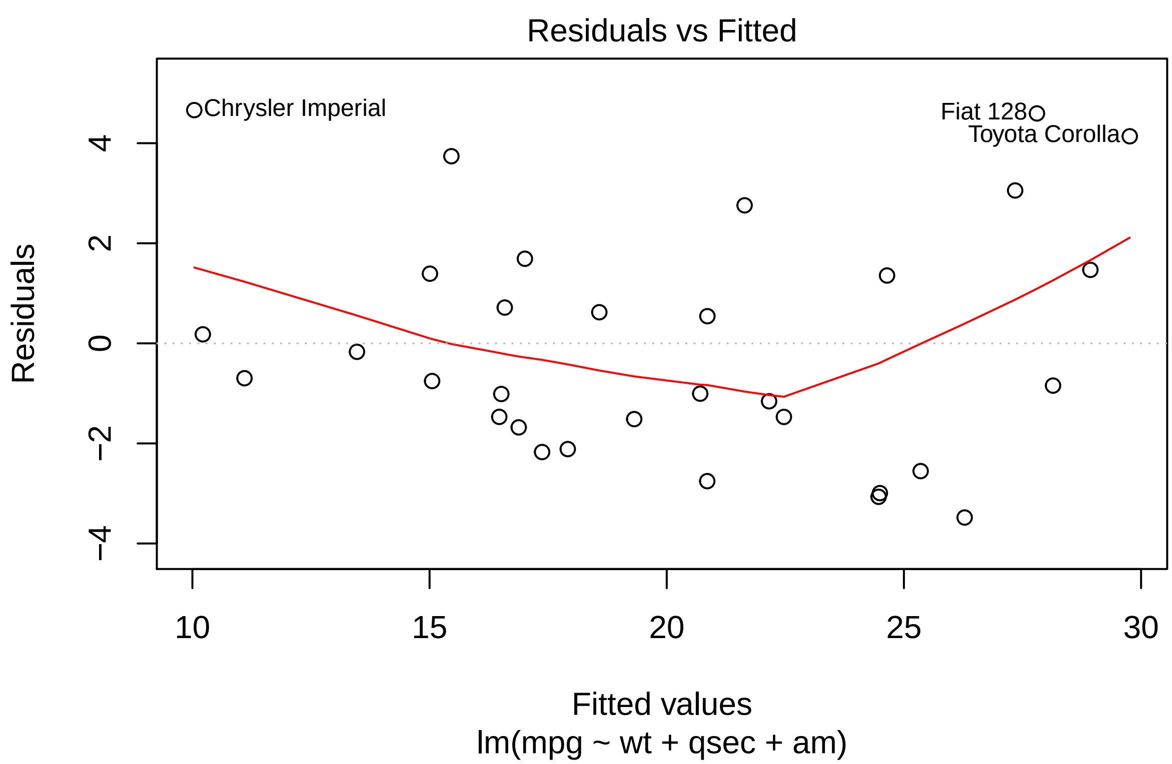
```r
coefficients(bestModelFit)
```

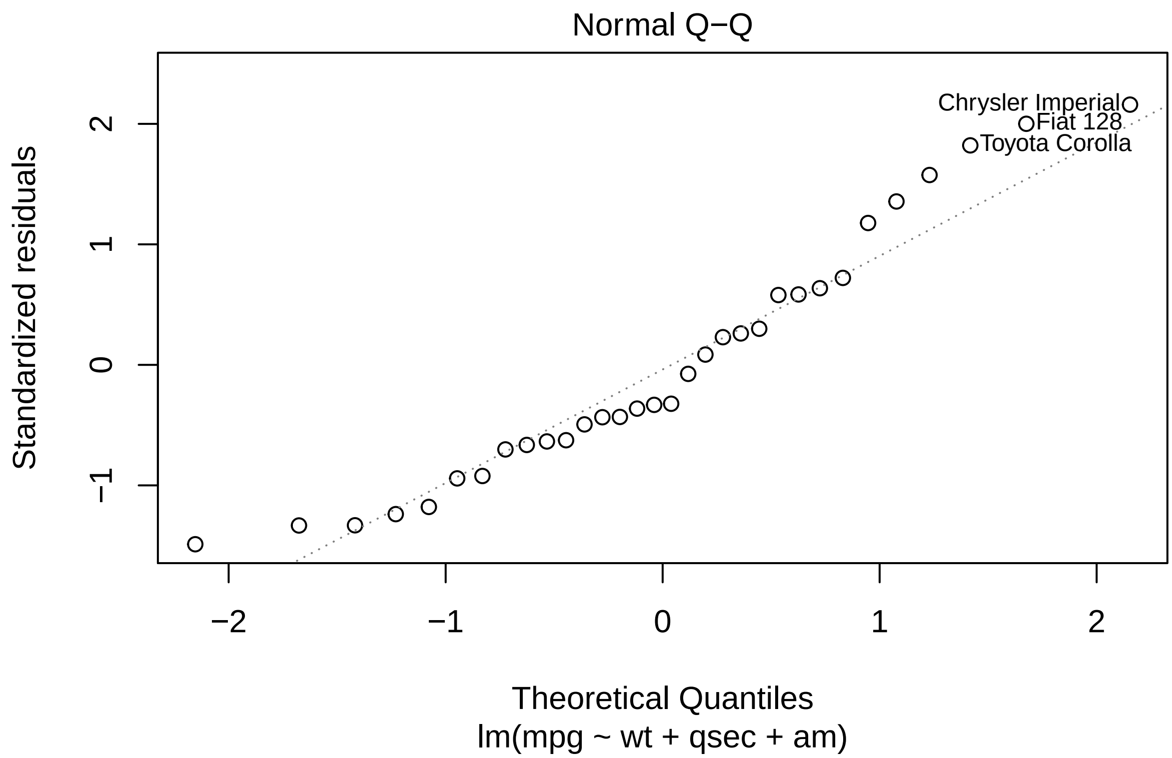
```
## (Intercept)          wt        qsec          am
##    9.617781   -3.916504    1.225886    2.935837
```

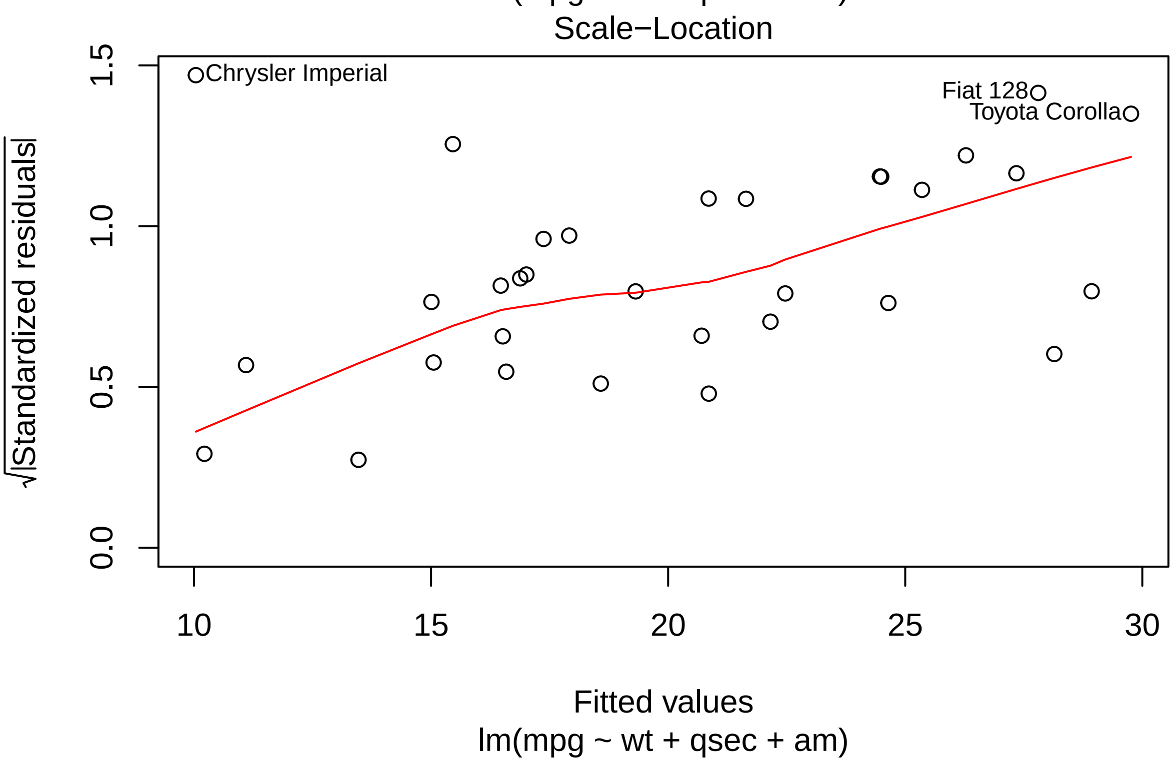
A.6 Residuals Plot

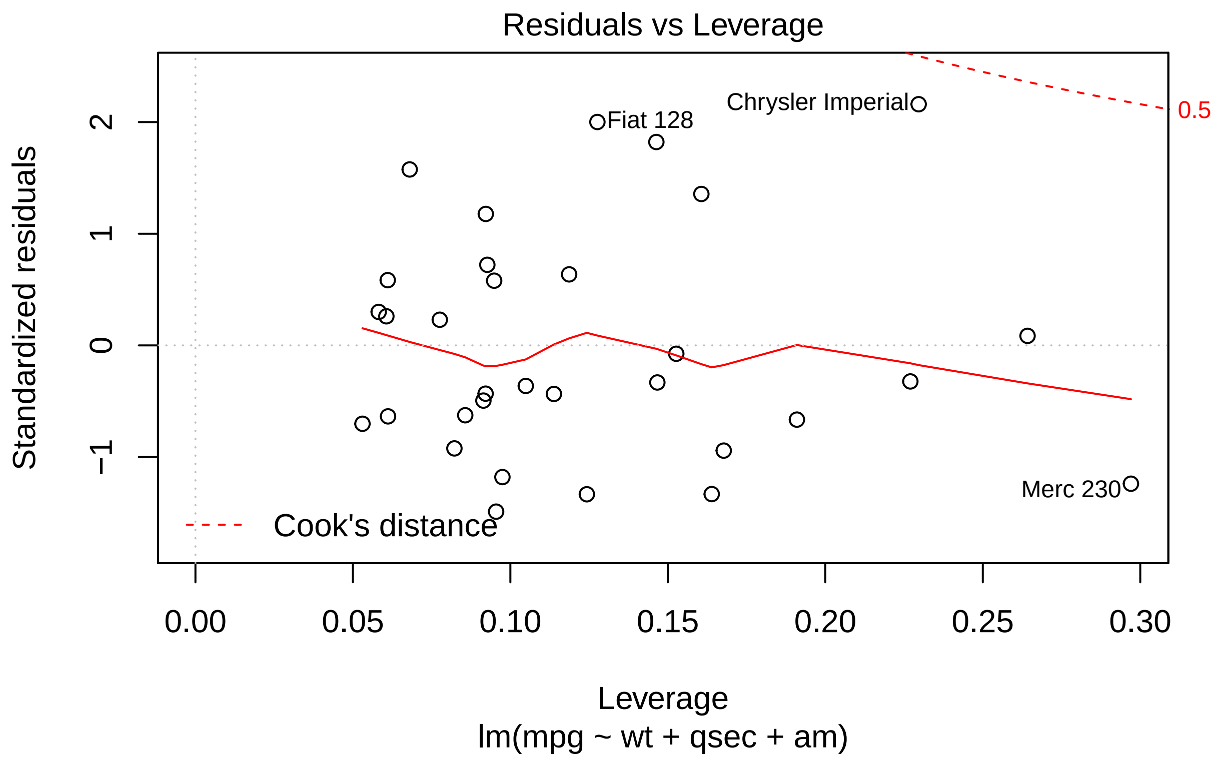
Residuals for the multiple linear regression model (best model fit).

```r
par(mfrow = c(1, 1))
plot(bestModelFit)
```




Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ wt + qsec + am)


Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
1.5
1.0
0.5
0.0
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)


Residuals vs Leverage
Standardized residuals
2
1
0
−1
Fiat 128
Chrysler Imperial
0.5
Merc 230
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(mpg ~ wt + qsec + am)