Coursera - Regression Models Class Project

Max Gribov

March 3, 2018

Executive Summary

This report analyzes miles per gallon (MPG) performance of various cars based on mtcars data set, and addresses the specific question of whether automatic or manual transmission is better for MPG.

According to the findings, when only transmission type is considered, cars with manual transmission get approximately 7 additional miles per gallon. However, considering only transmission type results in a low Adjsuted R-squared value of 0.3385 for the model, which implies there are other significant predictors for MPG.

Additionally, if only car weight is considered, it plays a significant role as well, with approximately 5 MPG decrease per additional 1 ton of weight. Adjusted R-squared value of 0.7446 is much higher for this model, further indicating its significance, and making it a better predictor than transmission type.

Finally, the best model found uses Transmission Type, Weight, and 1/4 Mile Time ("am", "wt", "qsec" columns), resulting in a model with all coefficients having statistical significance (p-value < 0.05) and Adjusted R-squared of 0.8336, explaining 83% of the variance in MPG.

This final model suggests that MPG will increase by about 3 if the car is using manual transmission, and the weight and 1/4 mile time stay constant, but the p-value for this coeffcient (0.046716) is just barely under significance treshold. Weight seems to play biggest role, with 4 MPG decrease for every additional 1 ton of weight, with p-value of 6.95e-06.

Data

The data is the mtcars data set, which has MPG and other data for 32 different cars. The model examines effect of the transmission ("am" column) and weight ("wt" column) on MPG performance ("mpg" column).

```r
# load the dataset
data(mtcars)

# 32 observations of 11 variables
dim(mtcars)
```

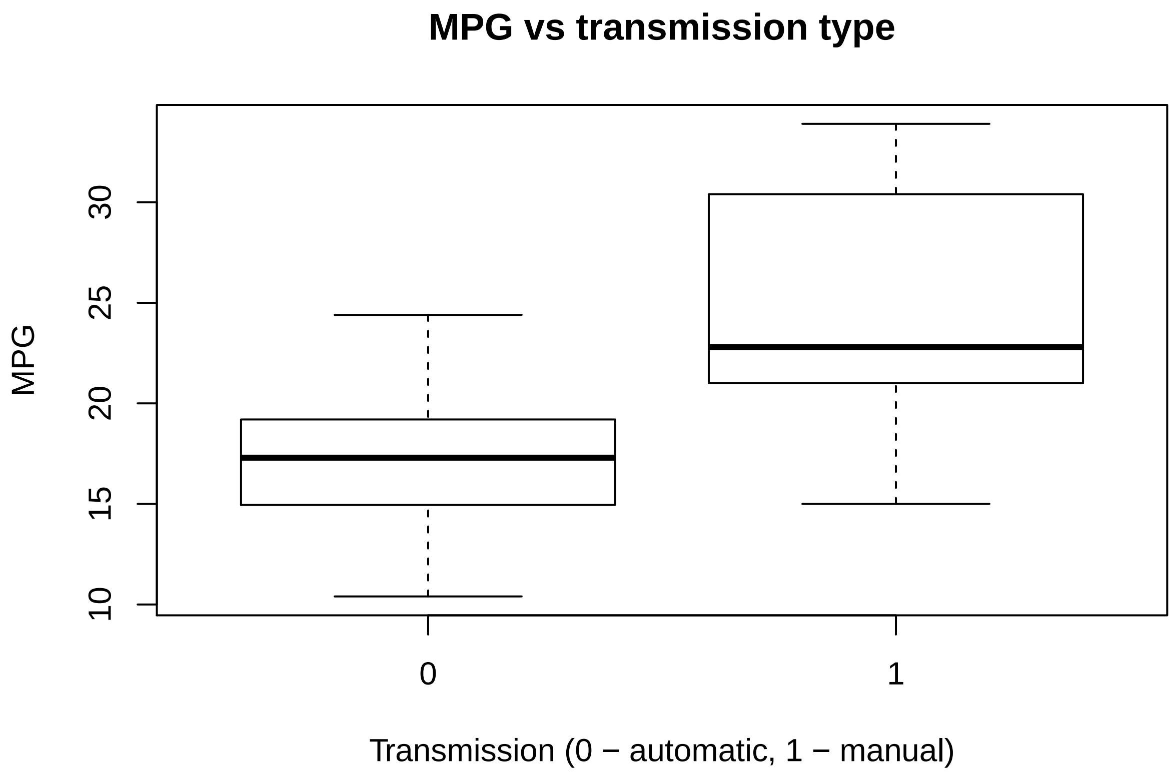
```
## [1] 32 11
```

```r
# sample of the data
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Data Exploration

Our primary goal is to analyze the impact of transmission type on MPG. Since there are 2 possible values for the "am" column, we can create a simple boxplot showing that mean MPG is different for these 2 different populations.




```
## [1] "mean mpg for automatic: 17.1473684210526 mean mpg for manual: 24.3923076923077"
```

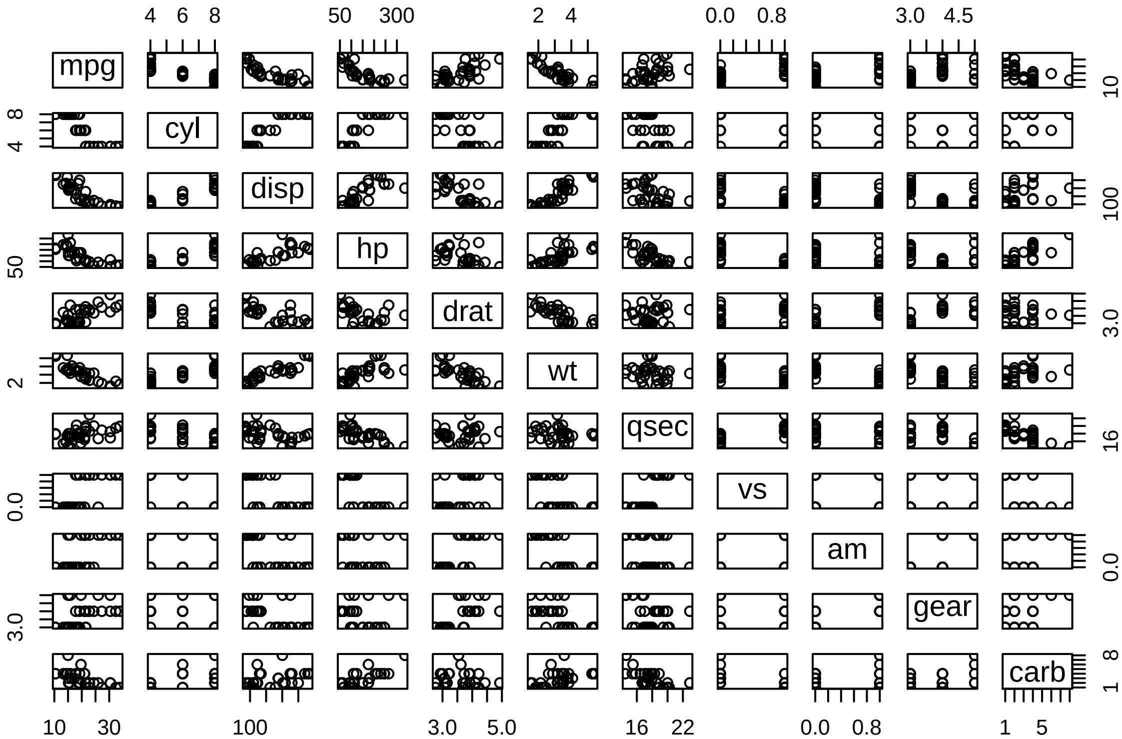
We can also use T-Test to compare the 2 populations.

```
t.test(mtcars[mtcars$am == 0, 'mpg'], mtcars[mtcars$am == 1, 'mpg'])
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 0, "mpg"] and mtcars[mtcars$am == 1, "mpg"]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Very low p-value of 0.001374 (below 0.05) indicates a significant differnce in the means, and that the data comes from 2 different populations.

We can also examine the pair graph for any other clearly visible linear relationships, and we can see one for MPG and car weight. There also appears to be linear relationships between MPG and Engine Displacement, as well as MPG and Horse Power, but these should be expected, as they are well known to directly affect the fuel consumption rate.



Model selection

MPG and transmission type

We can use a simple linear model to examine the relationship between MPG and transmission.

```r
# train the model
lm_am <- lm(mpg ~ am, data=mtcars)

# model summary
summary(lm_am)
```

```
##
## 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
```

This model shows that both intercept and "am" coefficient have very low p-values (below standard 0.05 treshold), which means we can reject the null hypotheses of both populations having same mean value. This proves that transmission type has a significant effect on resulting MPG performance. However, according to the Adjusted R-Squared value of 0.3385 this model can explain only 34% of the variance in MPG, which indicates there may be other significant predictors.

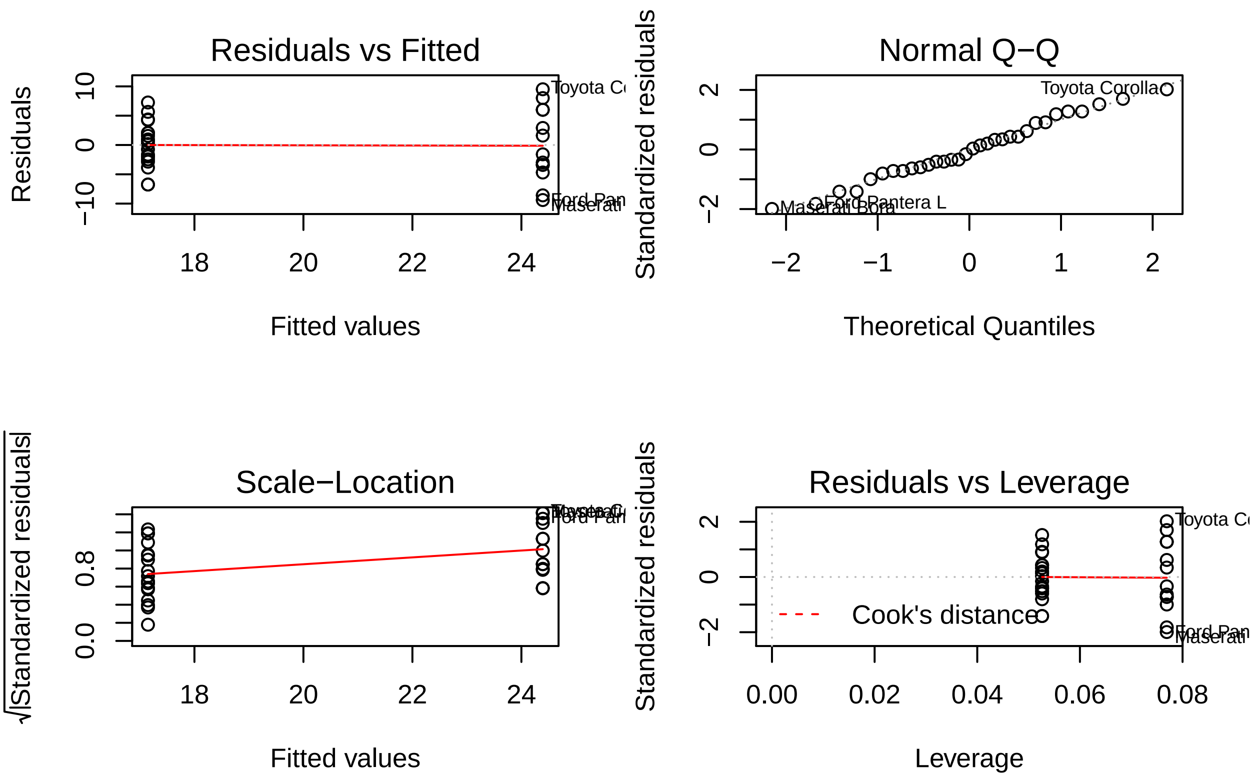
Coefficient value of 7.245 for "am" is the average increase in MPG value for 1 unit of increase in "am" variable. "am" variable has only two possible values, 0 and 1, and if we substitute 0 we will get the intercept term only, which is in turn the mean of the population with automatic transmission (17.147)

```
summary(lm_am)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

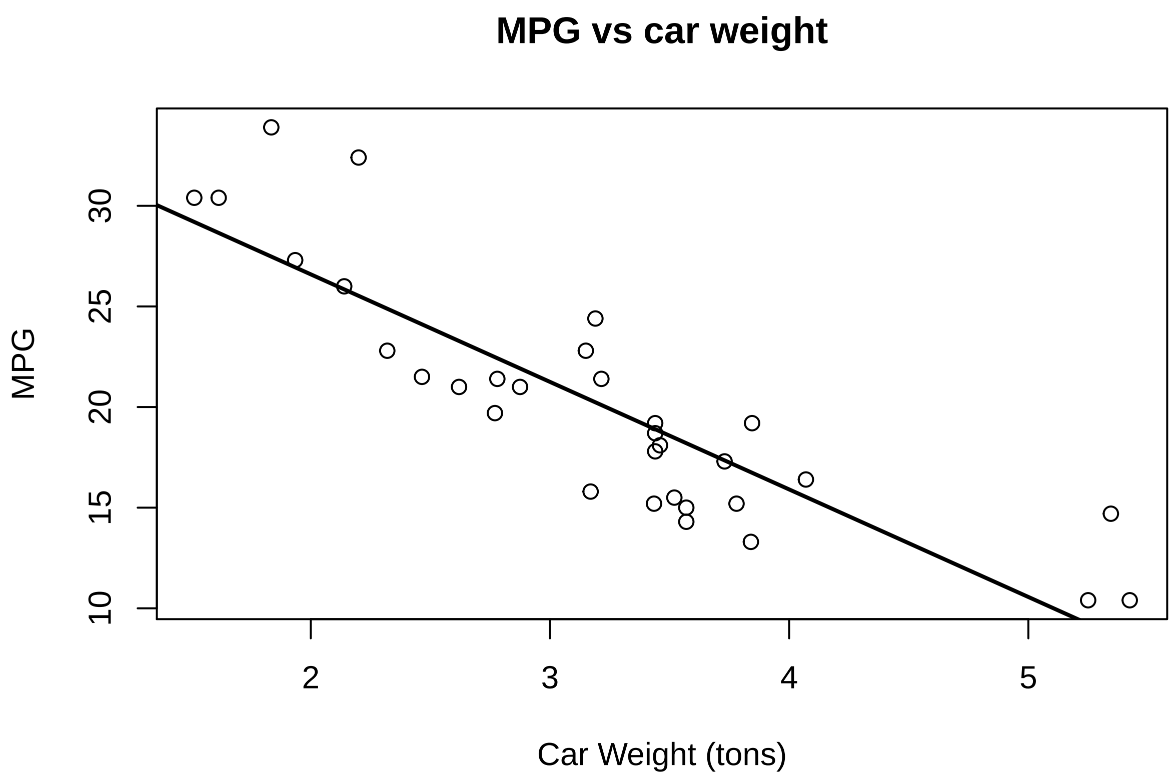
Analysis of residuals shows that they are very close to being normally distributed according to their normal Q-Q plot, which indicates that this model is valid.

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



MPG and car weight

Additionally, it appears that the car weight ("wt" column) have significant impact on MPG value. We can see some linearity in the following plot.



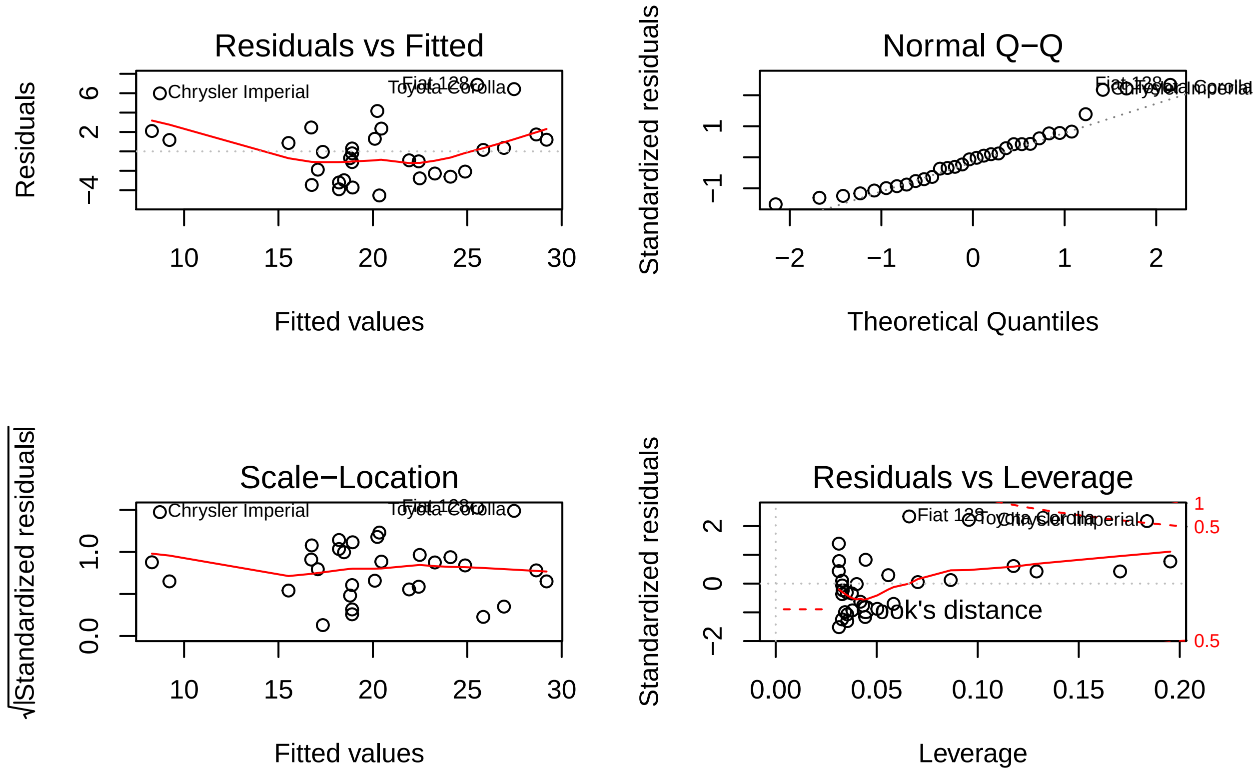

```r
# train the model
lm_wt <- lm(mpg ~ wt, data=mtcars)

# model summary
summary(lm_wt)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

According to a linear model predicting MPG based on weight, there is an average decrease of 5.34 MPG for every 1 ton increase in weight. The p-value of 1.29e-10 for the "wt" coefficient is very low, indicating its significance. This model also has a much higher Adjusted R-squared value of 0.7446, explaining 74% of the variation on MPG.

However, according to the normal Q-Q plot of the residuals, this model is not as good as the one based on transmission type, since the residuals deviate from the line.

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



MPG and car weight, transmission type, and 1/4 mile time

We will use step() to find the best model as compared to a full model.

First, we build model on the full data. However, this model does not have any significant coefficients.

```r
lm_full <- lm(mpg ~ ., data=mtcars)
summary(lm_full)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
```

```
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

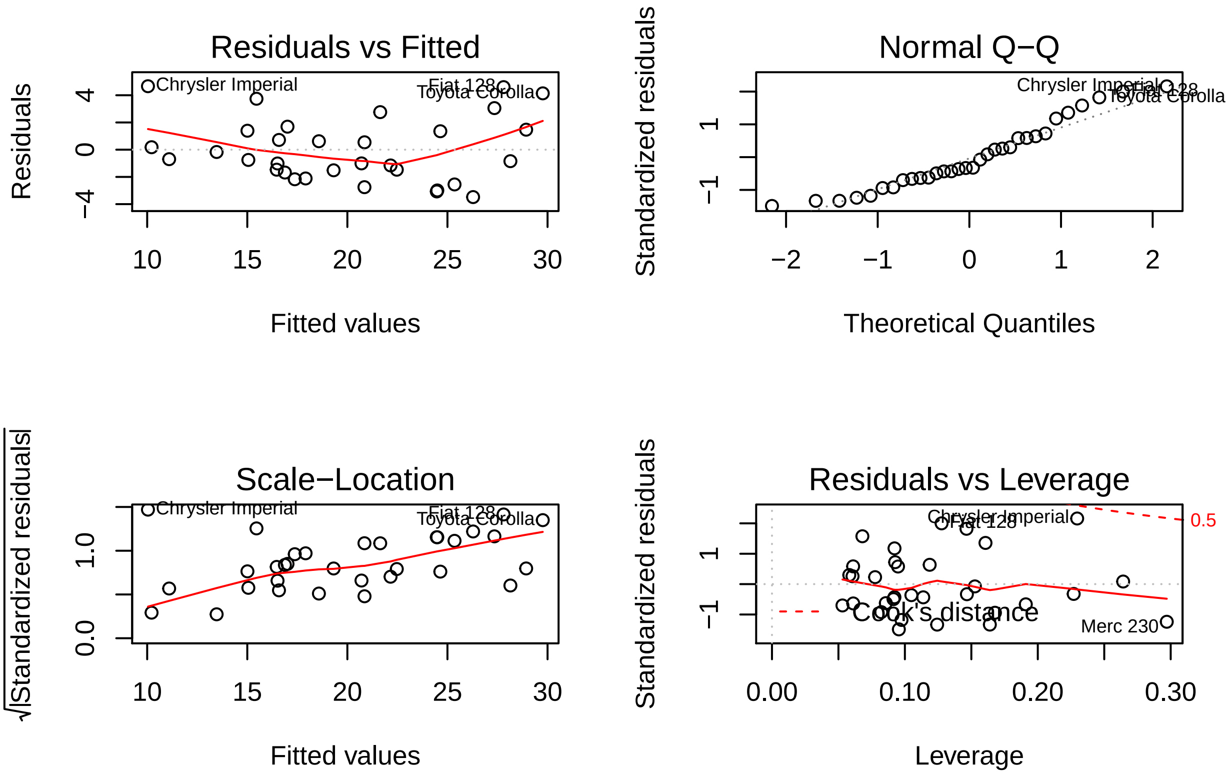
Next, we use step() to find best possible model.

```r
lm_step <- step(lm_full, trace=0)
summary(lm_step)
```

```
##
## 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
```

All coefficients in this model ("wt", "qsec", "am") are significant, with p-value < 0.05, the model residual standard error is low at 2.459 and adjusted r-squared value is high at 0.8336, explaining 83% variance in MPG. The residuals are close to normally distributed according to the normal Q-Q plot. This appears to be the best model.

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


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
−4
0
4
10
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
0.0
1.0
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5
0.00
0.10
0.20
0.30