Regression Models - Project

Krishna V Iyer

April 17, 2016

Data

The data 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 Task

Examine the `mtcars` data set for relationships between a set of variables and miles per gallon (mpg). Employ Regression models to analyze how the **Transmission** factor variable [**automatic** (am = 0) and **manual** (am = 1)] affects the **mpg** variable.

Exploratory Data Analysis

Load the data and libraries. Identify and covert factor variables.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.2.3
```

```
data(mtcars)
head(mtcars) # Sample Data
```

```
##                   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
```

```
dim(mtcars)
```

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

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Null Hypothesis

The **mpg** does not depend on the type of transmission of the automobile. The assumption is the distribution is normal and the sample is representative of the population

```
result <- t.test(mpg ~ am, mtcars)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

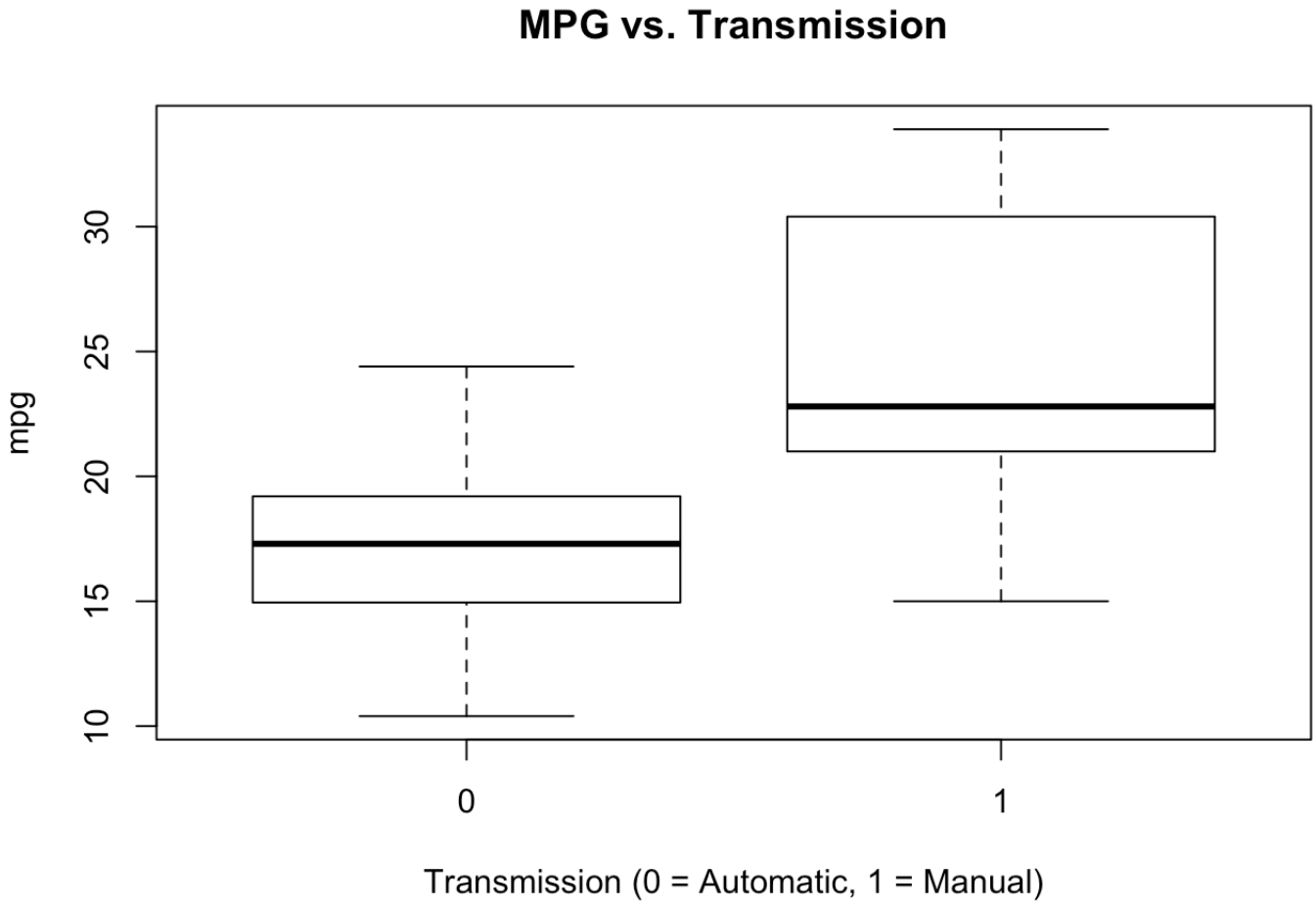
```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The p-value is about 0.001, and hence the null hypothesis is rejected. The means of the two samples differ by 7 mpg, the **automatic** transmission fairing better than the **manual** transmission.

Exploratory plotting of Data

Boxplot of MPG vs. Transmission

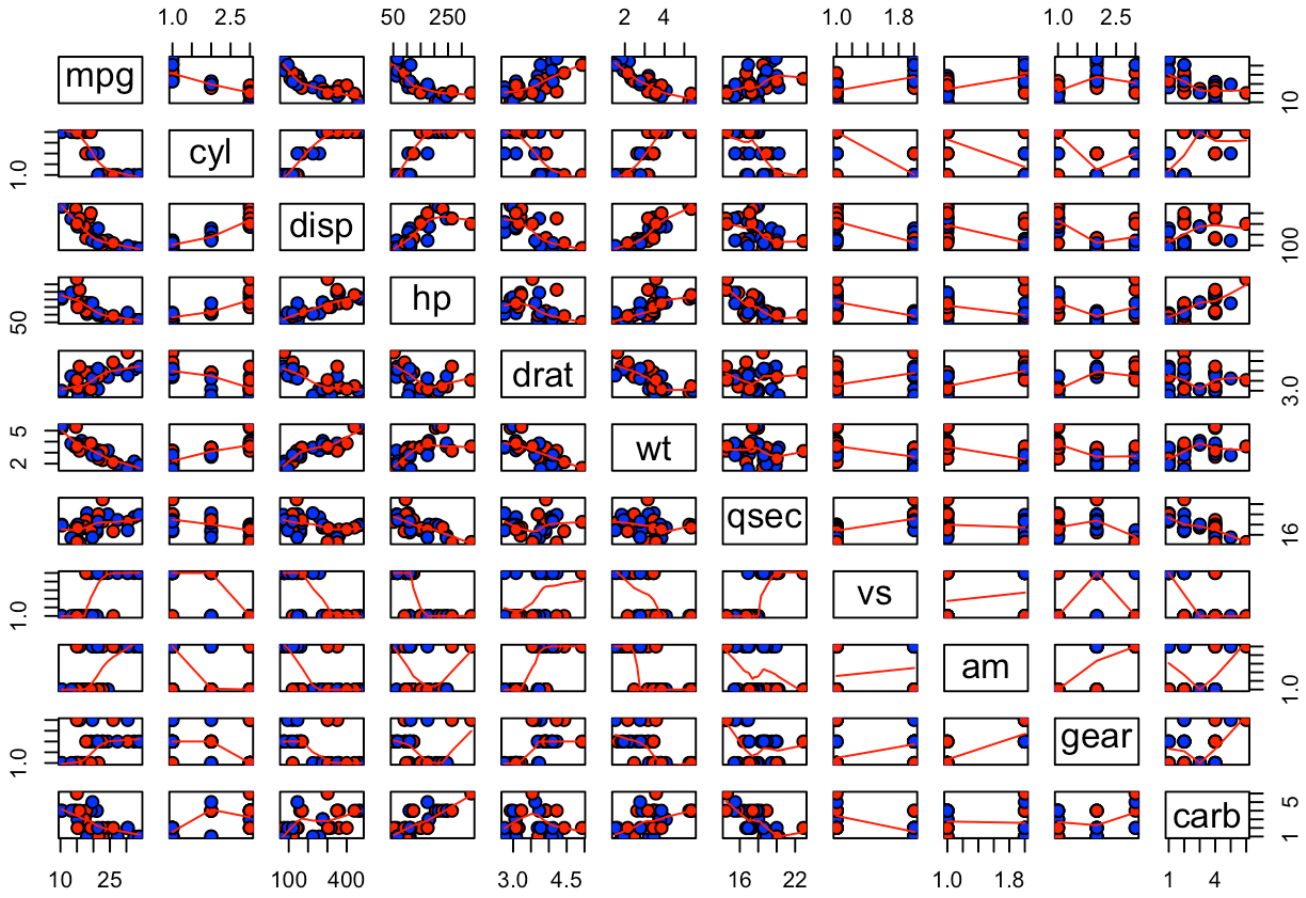
```
boxplot(mpg ~ am, mtcars, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="mpg"
,
        main="MPG vs. Transmission")
```



It is now clear that the differnce is statistically significant. Lets examine the data further to see if there are any other relationships between variables before proceeding with modeling them. The most informative plot is that can plot pairwise relationships between variables.

Pairplots of all variables in mpg dataset

```
pairs(mtcars, panel=panel.smooth, pch = 21, cex=1.2, bg = c('red', 'blue') )[unclass(
mtcars$am)]
```

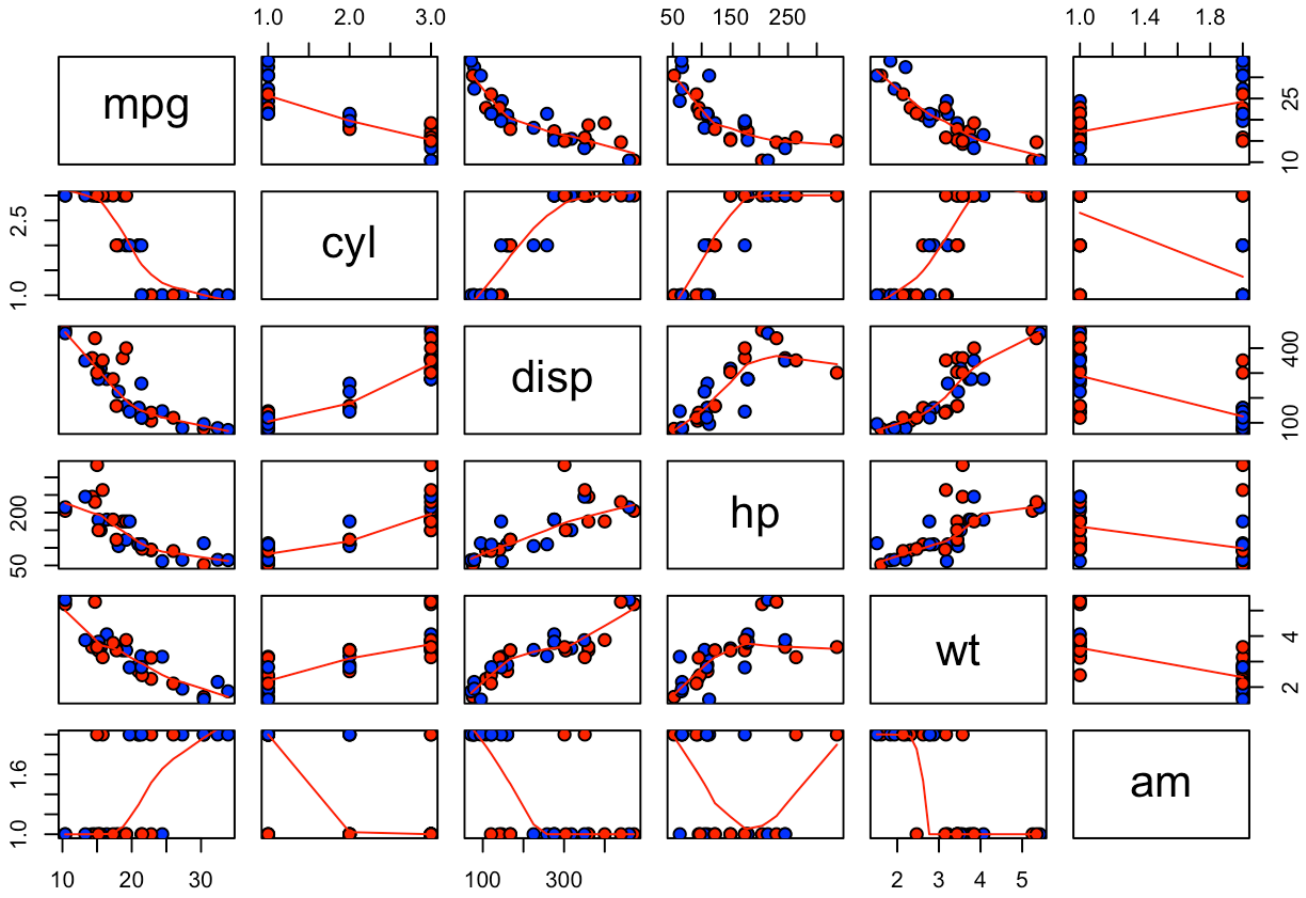



```
## NULL
```

We can see some relationships between mpg and cyl, disp, hp and am. Lets examine this further.

Pairplots of select variables in mpg dataset

```
pairs(mtcars[, c(1, 2, 3, 4, 6, 9)], panel=panel.smooth, pch = 21, cex=1.2, bg = c('r
ed', 'blue') )[unclass(mtcars$am)]
```




```
## NULL
```

At this point, we can clearly see a definite relationship. However, there are interactions between these variables that needs to be considered. The model should be able to address this. Lets now proceed with the model.

Regression Analysis

Lets start by fitting a full model…

```
model0 <- lm(mpg ~ ., data=mtcars)
summary(model0)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

The full model has an Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. It is interesting to note that none of the variables are coming out as significant estimator of the the label. This means that all the variables together does not seem to predict the outcome. Colinearity is one concern and some variables might not add to additional information. This can be addressed by stepwise regression analyis.

A stepwise regression of the full model is done as follows:

```
stepModel <- step(model0, k=log(nrow(mtcars)))
```

```
## Start:  AIC=101.32
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS     AIC
## - carb  5   13.5989 134.00  87.417
```

```
## - gear  2    3.9729 124.38  95.428
## - cyl   2   10.9314 131.33  97.170
## - am    1    1.1420 121.55  98.157
## - qsec  1    1.2413 121.64  98.183
## - drat  1    1.8208 122.22  98.335
## - vs    1    3.6299 124.03  98.806
## - disp  1    9.9672 130.37 100.400
## <none>              120.40 101.321
## - wt    1   25.5541 145.96 104.014
## - hp    1   25.6715 146.07 104.040
##
## Step:  AIC=87.42
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 81.662
## - cyl   2   12.5642 146.57 83.353
## - disp  1    0.9934 135.00 84.187
## - drat  1    1.1854 135.19 84.233
## - vs    1    3.6763 137.68 84.817
## - qsec  1    5.2634 139.26 85.184
## - am    1   11.9255 145.93 86.679
## <none>              134.00 87.417
## - wt    1   19.7963 153.80 88.360
## - hp    1   22.7935 156.79 88.978
##
## Step:  AIC=81.66
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - cyl   2   10.4247 149.45 77.045
## - drat  1    0.9672 139.99 78.418
## - disp  1    1.5483 140.57 78.551
## - vs    1    2.1829 141.21 78.695
## - qsec  1    3.6324 142.66 79.022
## <none>              139.02 81.662
## - am    1   16.5665 155.59 81.799
## - hp    1   18.1768 157.20 82.129
## - wt    1   31.1896 170.21 84.674
##
## Step:  AIC=77.04
## mpg ~ disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     0.645 150.09 73.717
## - drat  1     2.869 152.32 74.187
## - disp  1     9.111 158.56 75.473
## - qsec  1    12.573 162.02 76.164
## - hp    1    13.929 163.38 76.431
## <none>              149.45 77.045
```

```
## - am    1    20.457 169.91 77.684
## - wt    1    60.936 210.38 84.523
##
## Step:  AIC=73.72
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 70.956
## - disp  1     8.545 158.64 72.023
## - hp    1    13.285 163.38 72.965
## <none>              150.09 73.717
## - am    1    20.036 170.13 74.261
## - qsec  1    25.574 175.67 75.286
## - wt    1    67.572 217.66 82.146
##
## Step:  AIC=70.96
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 68.844
## - hp    1    12.572 166.01 70.011
## <none>              153.44 70.956
## - qsec  1    26.470 179.91 72.583
## - am    1    32.198 185.63 73.586
## - wt    1    69.043 222.48 79.380
##
## Step:  AIC=68.84
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 67.170
## <none>              160.07 68.844
## - qsec  1    20.225 180.29 69.186
## - am    1    25.993 186.06 70.193
## - wt    1    78.494 238.56 78.147
##
## Step:  AIC=67.17
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 67.170
## - am    1    26.178 195.46 68.306
## - qsec  1   109.034 278.32 79.614
## - wt    1   183.347 352.63 87.187
```

```
summary(stepModel)
```

```
##
## 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 ***
## am1           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
```

AIC is a goodness of fit measure that favours smaller residual error in the model, but penalises for including further predictors and helps avoiding overfitting. The model “mpg ~ wt + qsec + am” has the lowest AIC, and this model can explain about 83% of the variance of the mpg variable. All of the coefficients are significant with p-value less than 0.05.

According to the pair plot, it appears that there is an interaction term between “wt” variable and “am” variables. This can be due to the fact that cars with automatic transmission tend to weigh heavier than cars with manual transmission. The new model is done as follows:

```
model1 <-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am1           14.079      3.435   4.099 0.000341 ***
## wt:am1        -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

This model's Adjusted R-squared value is 0.8804, which means it can explain about 88% of the variance in the value of the label.

For understanding how the identified model performs in comparison to the baseline we which is fitting the label to 'am' variable only.

```
model2 <-lm(mpg ~ am, data=mtcars)
summary(model2)
```

```
##
## 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 ***
## am1            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
```

Though the model can separate the two types of transmission, the Adjusted R-squared value drops to 0.3385. This suggests the previous model might perform better for the given dataset.

Finally, we select the final model.

```
anova(model0, stepModel, model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ wt + qsec + am + wt:am
## Model 4: mpg ~ am
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     15 120.40
## 2     28 169.29 -13    -48.88  0.4685    0.9114
## 3     27 117.28   1     52.01  6.4795    0.0224 *
## 4     30 720.90  -3   -603.62 25.0667 4.295e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(model1)
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## wt          -4.3031019 -1.569960
## qsec         0.4998811  1.534066
## am1          7.0308746 21.127981
## wt:am1      -6.5970316 -1.685721
```

We end up selecting the model with the highest Adjusted R-squared value, “mpg ~ wt + qsec + am + wt:am”.

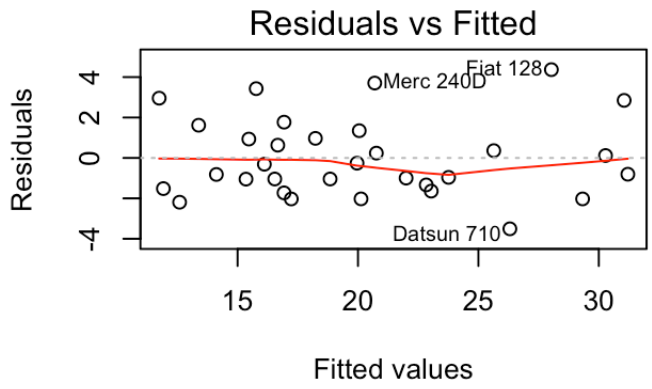
```
summary(model1)$coef
```

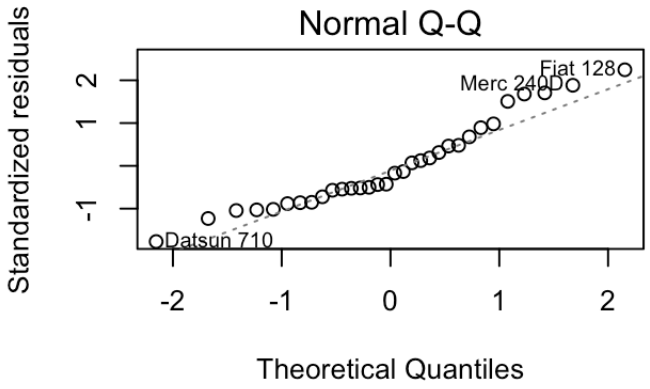
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

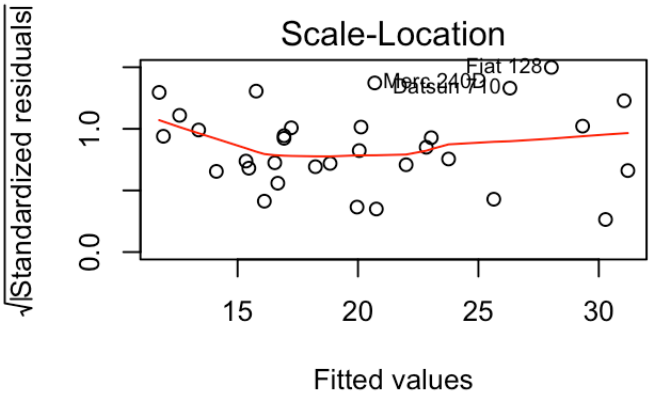
Thus, the result shows that when “wt” (weight lb/1000) and “qsec” (1/4 mile time) remain constant, cars with manual transmission add about 7 mpg (miles per gallon) more on average than cars with automatic transmission.

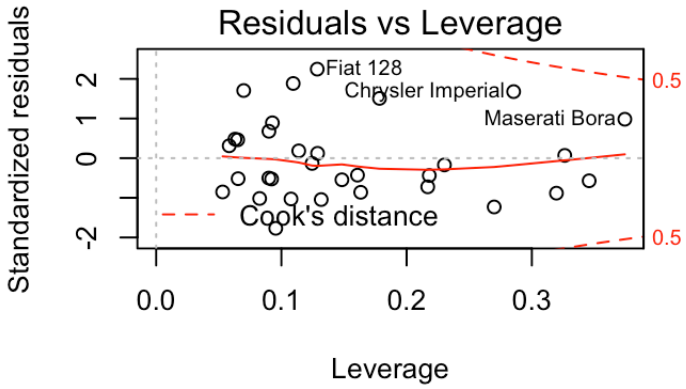
Residual Analysis and Diagnostics

```
par(mfrow = c(2, 2))
plot(model1)
```









1. The Residuals vs. Fitted plot suggests no underlying pattern in the data
2. The Normal Q-Q plot indicates that the residuals are normally distributed
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage suggests that no outliers are present