Regression Models Project

Konstantinos Papastamos

25 Octomber 2015

Executive Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are 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"

Exploratory Data Analysis

So we have a dataframe of 32 observations and 11 variables. Let's do some necessary transformations and have a look at our data.

```
data(mtcars)
data = mtcars
data$am[data$am==0] <- "Automatic"
data$am[data$am==1] <- "Manual"
data$am <- as.factor(data$am)
str(data)
```

```
## '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  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Modelling and Testing

Let's do a t test in 95% confidence level to check if transmission type has an influence on Miles per Gallon.

```
t.test(data$mpg ~ data$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$mpg by data$am
```

```
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Here we can see that the p-value is lower than 0.05 so a first assumption is that transmission type actually affects mpg.

So let's create our first regression model with am as the predictor and mpg as the outcome.

```
rmodel1 = lm(mpg ~ am,data)
summary(rmodel1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = data)
##
## 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 ***
## amManual       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
```

The results suggest that manual cars have an average of 7.245 higher mpg than automatic ones.However, this model only explains 36% of the total variance. So let's create an improved model based on more variables.

First we create a correlation heatmap to see which variables are highly correlated and thus not include them in the improved model. See the appendix for the heatmap.

After we consult the heatmap we use the step(lm(mpg ~ .,data)) function in order to find the most important variables to include. Finally we construct our improved model.

```
rmodel2 = lm(mpg ~ am + wt + qsec,data)
summary(rmodel2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = data)
##
## 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
## amManual      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
##
## 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
```

This model incorporates the weight and qsec variables and explains 85% of the variance.

Let's compare the two models:

```
anova(rmodel1,rmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova table we can see that the Pr value for the second model is lower than 0.05 and so the addition of the qsec variable was necessary to improve the model. Thus it makes sense that we use the second one.

Conclusion

In order to have a complete conclusion, we needed to incorporate the qsec variable in our model. No more variables were used in order to avoid overfitting.

The final conlcusion is that manual transmission cars have on average 2.9358 more mpg than cars with automatic transmission.

Diagnostics

See appendix for the plots

Our model seems to work since:

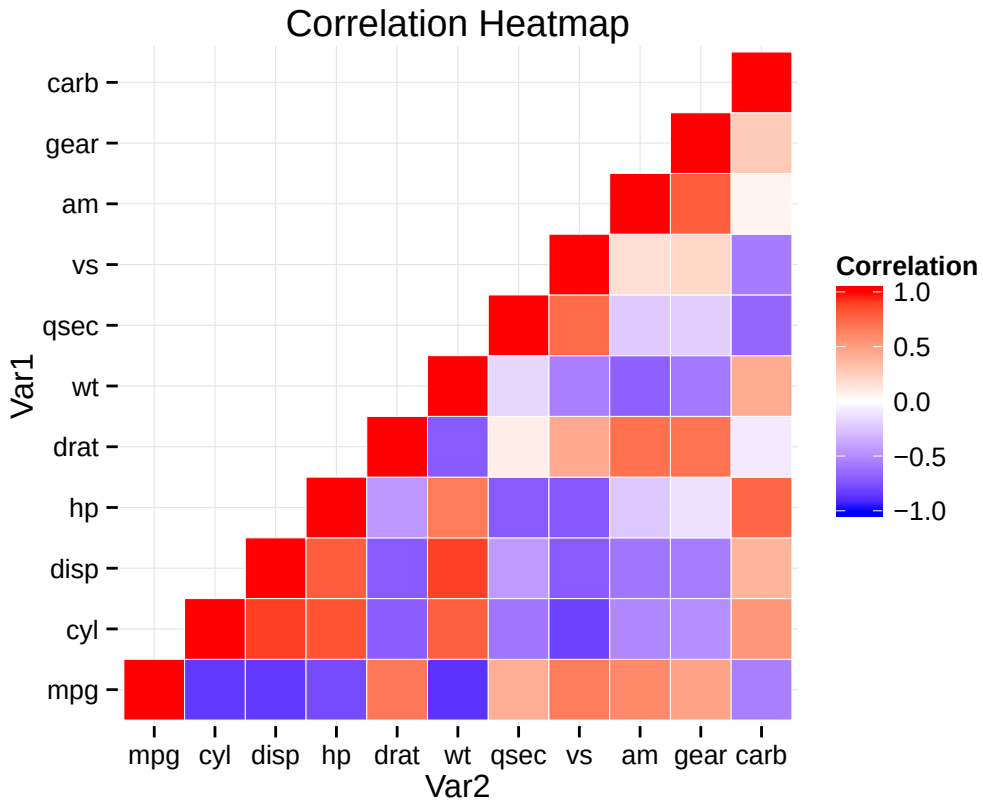
- There is no discernible pattern in the Residuals vs Fitted plot (it's more clear if we plot it without the other 3 plots)
- Residuals follow a Normal distribution
- There aren't any important outliers

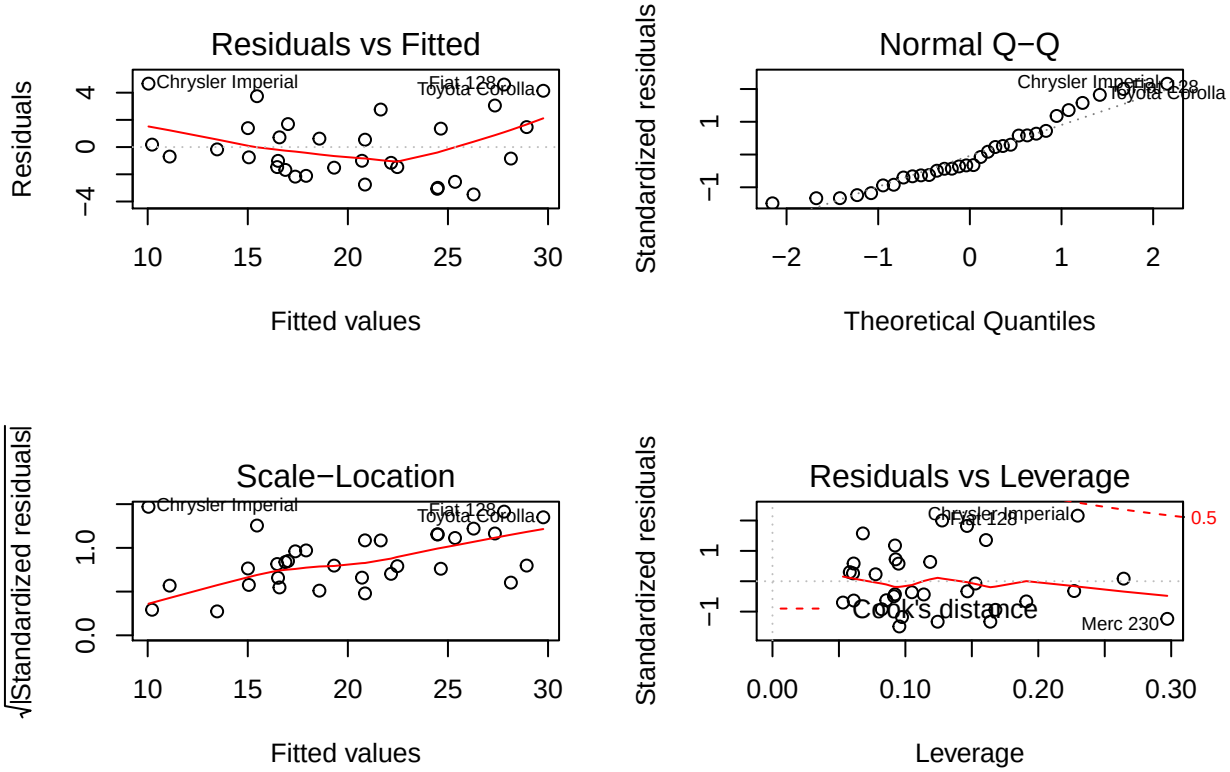
Appendix

```
library(ggplot2)
library(reshape2)
corheatmap = round(cor(mtcars),2)
corheatmap[lower.tri(corheatmap)]<- NA
melted <- melt(corheatmap)
melted <- na.omit(melted)



ggplot(data = melted, aes(Var2, Var1, fill = value))+
ggtitle("Correlation Heatmap")+
geom_tile(color = "white")+
scale_fill_gradient2(low = "blue",
high = "red", mid = "white",
midpoint = 0, limit = c(-1,1), name="Correlation")+
theme_minimal()+coord_fixed()
```




Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage