Regression Models - Course Project

Ioannis Moros

July 7, 2017

Executive summary

This paper explores the relationship between miles per US gallon and type of transmission, using the mtcars dataset in R.

Our analysis showed that manual transmission is better than automatic in regards to MPG. While accounting for number of cylinders, horsepower and weight, cars with automatic tranmission have 1.8 higher MPG than those with manual.

Data pre-processing

```
my_data <- mtcars
head(my_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
```

```
# Transform variables to factors where appropriate
my_data$cyl <- factor(my_data$cyl)
my_data$vs <- factor(my_data$vs)
my_data$am <- factor(my_data$am, labels = c("auto", "man"))
my_data$gear <- factor(my_data$gear, labels = c("3", "4", "5"))
my_data$carb <- factor(my_data$carb)
```

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

Let's get a first idea about the difference in average MPG between automatic and manual transmissions:

```
aggregate(my_data[, 1], list(my_data$am), mean)
```

```
##   Group.1        x
## 1    auto 17.14737
## 2     man 24.39231
```

There seems to be a clear difference (7.25 MPG) in the average mpg between automatic and manual transmission (see appendix for plot). Let's confirm it with a hypothesis test.

H_0: There is no significant difference in mpg between auto and man trans.

H_1: Automatic transmission is associated with lower values of mpg

```r
t.test(mpg ~ am, data = my_data, paired = FALSE, alt = "less")$p.value
```

```
## [1] 0.0006868192
```

```r
t.test(mpg ~ am, data = my_data, paired = FALSE, alt = "less")$estimate
```

```
## mean in group auto  mean in group man
##           17.14737           24.39231
```

The p-value is 0.0007 which means that we reject the null at any reasonable significance level, i.e. Manual transmission is better for MPG.

2. Quantify the MPG difference between types of transmission.

We've fit several linear regression models to quantify the difference in MPG between automatic and manual type of transmission.

The final model is shown below. The full model selection strategy and intermediate models can be found in the Appendix.

```r
mdl3 <- lm(mpg ~ cyl + hp + wt + am, data = my_data)

summary(mdl3)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amman        1.80921138 1.39630450  1.295714 2.064597e-01
```

```r
summary(mdl3)[8:9]
```

```
## $r.squared
## [1] 0.8658799
##
## $adj.r.squared
## [1] 0.8400875
```

In addition to the type of transmission, this model takes into account the number of cylinders, horsepower and weight. The model can explain 87% of total variance in MPG. While keeping all other variables constant, MPG increases by 1.8 miles/gallon from automatic to manual transmission.

Diagnostics

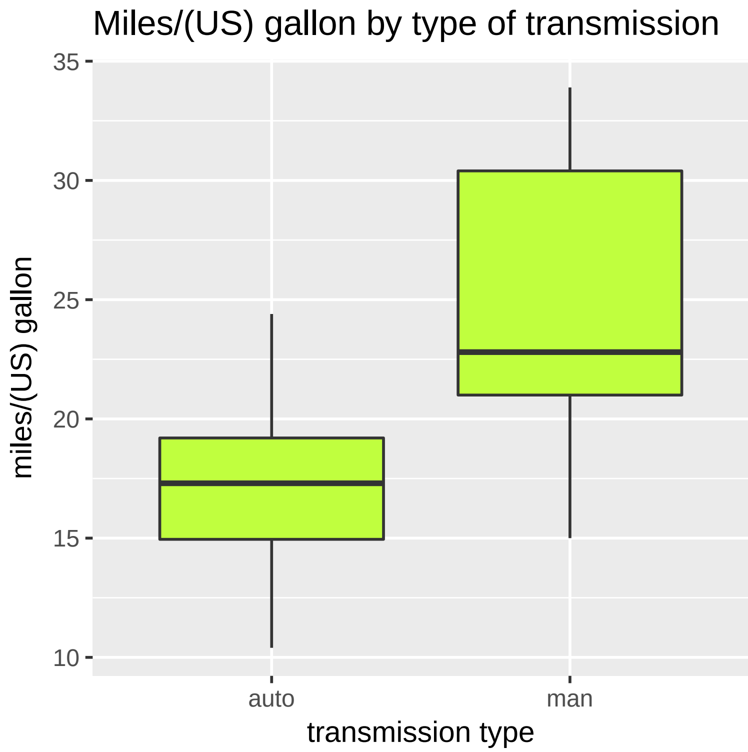
The results of the diagnostic tests are shown below. See Appendix for plots.

1. The residual vs fitted plot does not reveal any non-linear or other patterns
2. Testing the normality assumption - the qqplot is not a perfect straight line but does not appear to be concerning
3. There is no evidence of heteroscedasticity from the scale-location plot
4. All the residuals are within Cook's distance, so there's no reason to suspect influential data points in the dataset.

Appendix

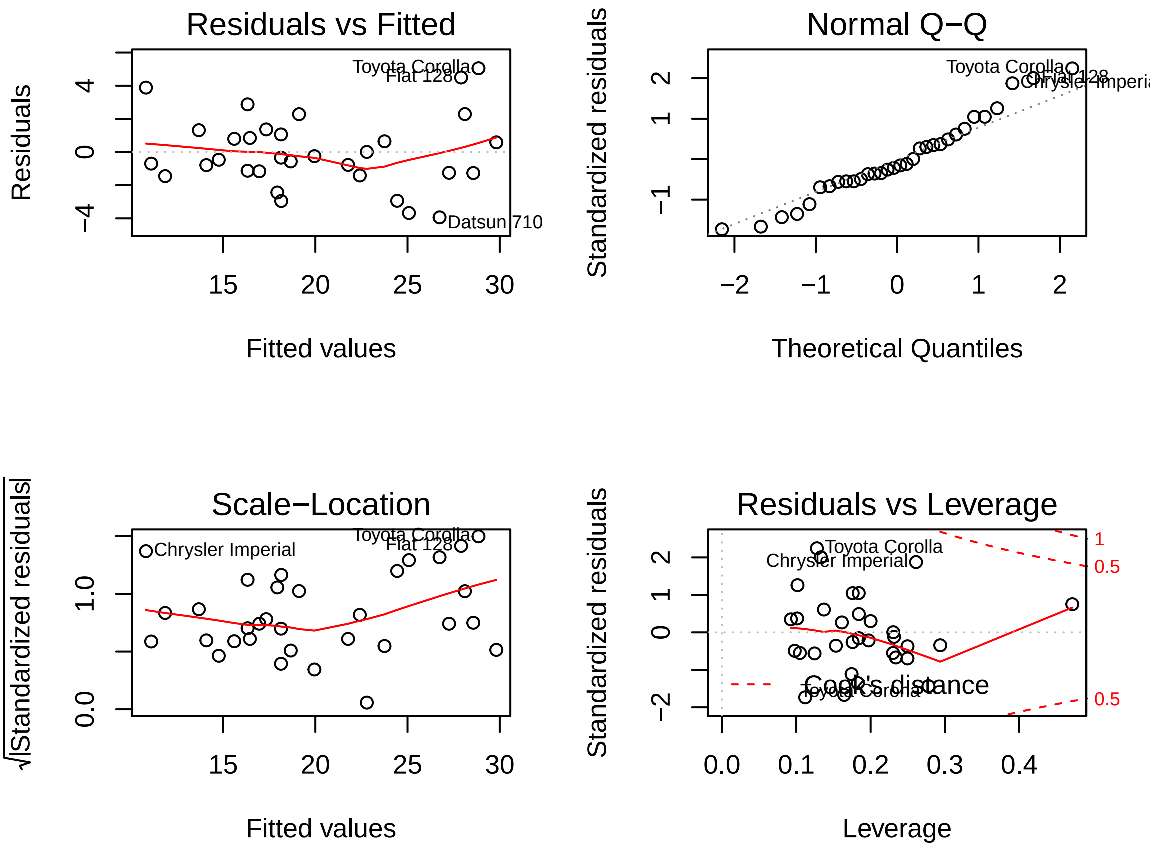
1. Boxplot of MPG by type of transmission

```
require(ggplot2)
g <- ggplot(my_data, aes(x = am, y = mpg))
g + geom_boxplot(aes(group = am), fill = "olivedrab1") +
        xlab("transmission type") +
        ylab("miles/(US) gallon") +
        ggtitle("Miles/(US) gallon by type of transmission")
```



2. Diagnostic plots for the final regression model

```
par(mfrow = c(2,2))
plot(mdl3)
```



3. Model selection strategy and intermediate models

Model1

First try a simple linear regression model with MPG as the dependent variable and transmission type (am) as the independent.

```
mdl1 <- lm(mpg ~ am, data = my_data)
summary(mdl1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = my_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 ***
## amman          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
```

From the summary we can see that am seems to have a significant effect on mpg (p-value < 0.05), and the difference in average MPG between the two levels of am (manual - auto) is 7.245, which matches the difference we've already observed. However, am alone does not appear to be enough to explain the variation in mpg, the R-squared is 0.3598, which means that this model can only explain 36% of the total variation in mpg. We will try to add some more independent variables to the model from the dataset to try and get a better fit.

Let's have a look at the correlations between mpg and the other variables in the dataset. We'll try a model that includes variables that appear to be highly correlated with mpg. Let's pick arbitratily the variables that have an absolute correlation higher than 0.7 plus the am variable.

Model2

```r
cor(mtcars, method = "pearson")[, "mpg"]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

```r
mdl2 <- lm(mpg ~ cyl + disp + hp + wt + am, data = my_data)
summary(mdl2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + wt + am, data = my_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## amman        1.806099   1.421079   1.271   0.2155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

Model 2 explains 87% of the variance and the adjusted R-squared is 83%. It is a much better fit than model 1, but includes independent variables that don't seem to have a significant effect (variables with p-value higher than 0.05). Transmission type doesn't seem to be significant in this model either. Let's try to remove variable "disp".