Regression Analysis - Course Project

Mathieu Wauters

11 mei 2016

Executive Summary

We were contacted by Motor Trend, an automobile industry magazine. Looking at a dataset of a collection of cars, we were asked to analyze the relationship of a number of car characteristics with the Miles Per Gallon (MPG). We came to the following conclusions:

- A manual transmission leads to an increase of MPG of 1.81, holding all other variables constant.
- The other variables that were included in our statistical model were the number of cylinders, horsepower and weight.

Exploratory Analysis

The `mtcars` dataset counts 32 observations and 11 variables. A summary is provided below. In the next steps, we will have to determine which of those 11 variables play a significant role in predicting the MPG.

```
library(ggplot2)
```

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

```
data(mtcars)
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```r
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Regression analysis

Manual vs. automatic transmission

Prior to testing the relationship between transmission and MPG, it is a good idea to visually explore the relationship. The supporting graph for this relationship can be found in the `Appendix` section. Next, we test to see if the observed difference between transmission type and MPG is statistically significant. The below analysis suggests there is a significant difference (p-value of 0.001374<0.05).

```r
t.test(mpg~am,data=mtcars)
```

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

Model selection

In order to predict the MPG, we have examined an initial model containing only `am` as the explanatory variable, a complete model containing all explanatory variables and a final model that uses Akaike's Information Criterion to find out what the best model should be (this is achieved using the `step` function).

```r
model.initial<-lm(mpg ~ am,data=mtcars)
model.allvars<-lm(mpg ~ .,data=mtcars)
model.final<-step(model.allvars,direction="both",trace=0)
summary(model.final)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Note: summaries of the models can be found in the Appendix. We test whether there is a significant difference between the initial and final model. Phrased alternatively, do the variables added in `model.final` contribute to performance in a statistically significant manner? The below analysis points out that there is a significant difference (p-value < 0.05). Hence, we reject the null hypothesis and continue with `model.final`.

```
anova(model.initial,model.final)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model interpretation

The coefficients of `model.final` yield the following insights:

- Cars with a manual transmission have an MPG value that is 1.81 higher than cars with an automatic transmission, keeping all other variables constant.
- An increase of 1,000 lb in weight leads on average to an MPG decrease of 2.50, keeping all other variables constant.
- Horsepower only has a very small effect on MPG (0.32 increase per 10 hp increase)
- A car with more cylinders has a smaller MPG. The MPG is 3.03 lower when going from a 4 cylinder car to a 6 cylinder car and 2.16 lower when changing from a 6 cylinder car to an 8 cylinder car.

Diagnostics

Finally, we run some diagnostics to confirm we can use linear regression (i.e. check a number of the assumptions of linear regression) but also to identify influential data points. The graphs depicted in the Appendix lead to the following conclusions:

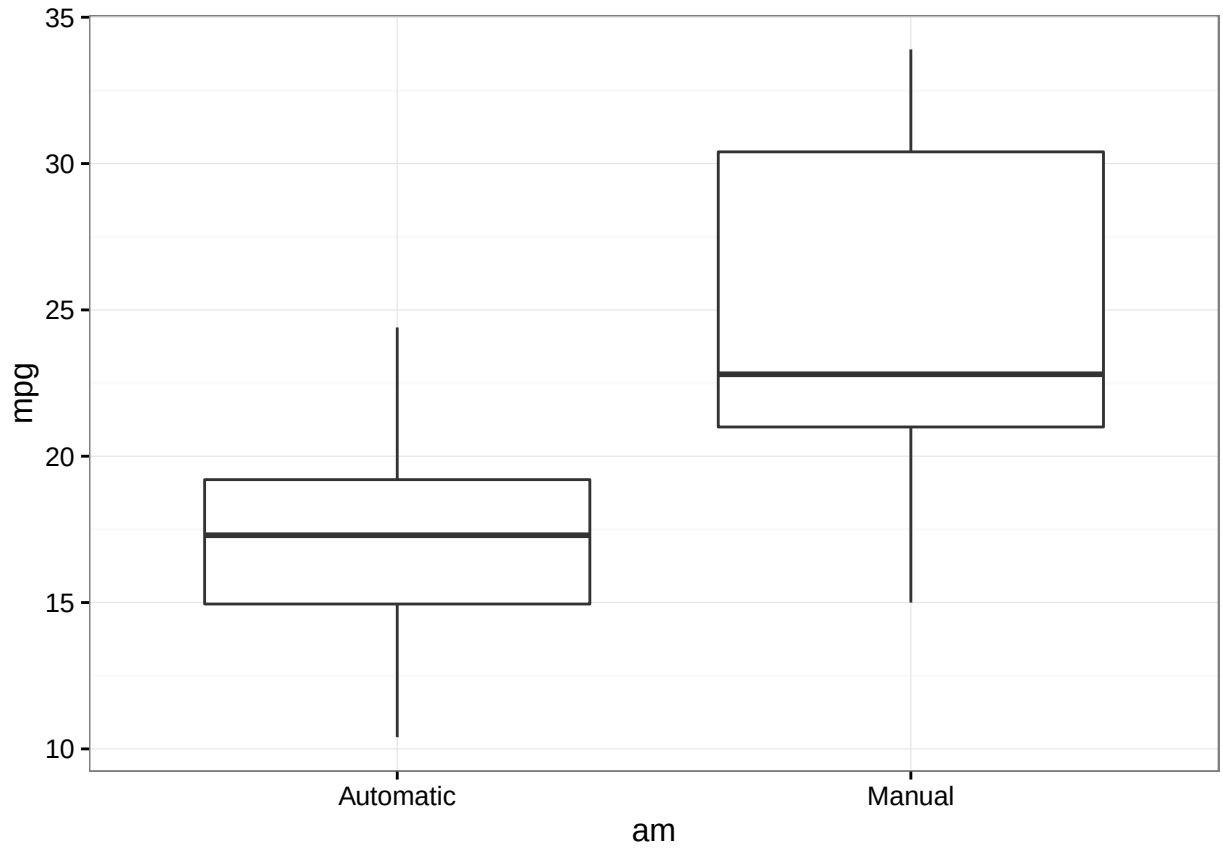
- A comparison of the residuals and fitted values shows there is no consistent trend
- The residuals are very closely normally distributed, as evidenced by the Q-Qplot.
- The scale-location plot suggests homoskedasticity.
- Some points have increased leverage (cf. labelled points in residuals vs. leverage plot)

The points with increased leverage can also be found in the Appendix.

Appendix

Exploratory graph:

```
ggplot(mtcars,aes(x=am,y=mpg))+geom_boxplot()+theme_bw()
```



Model summaries:

```
summary(model.initial)
```

```
##
## 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 ***
## 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
```

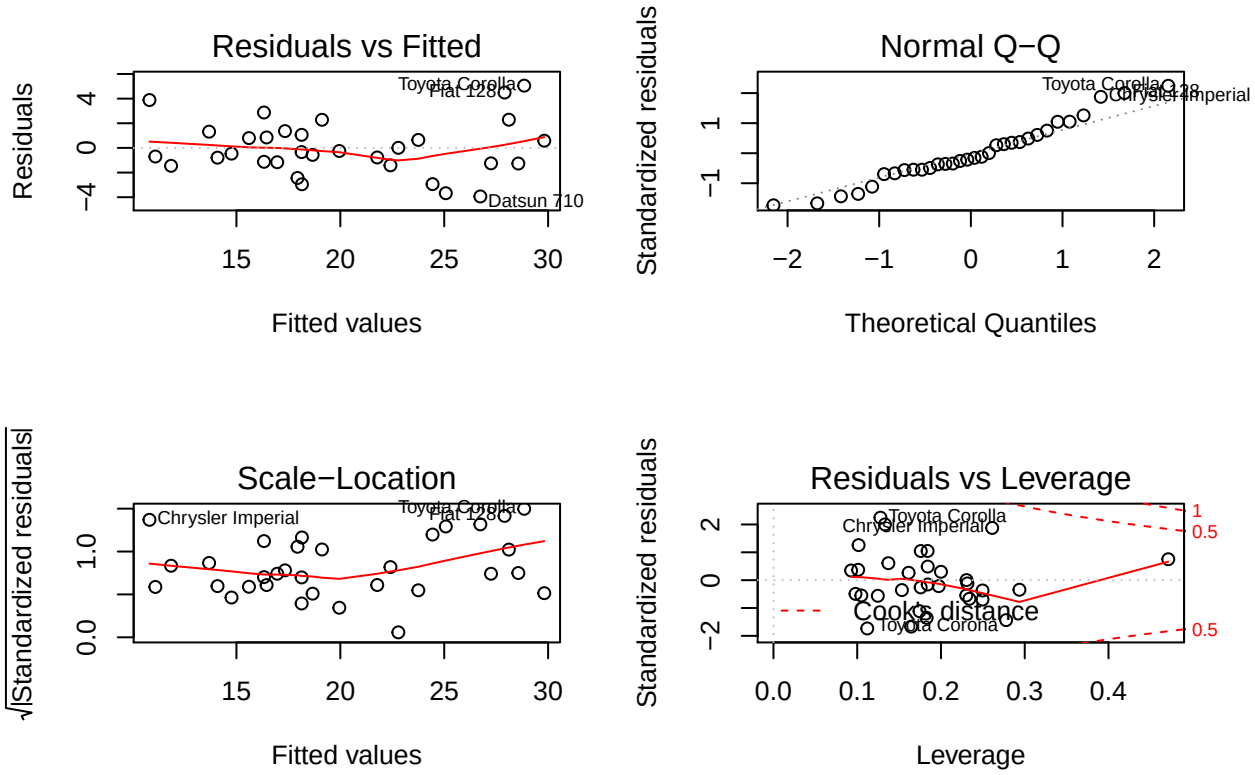
```
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
summary(model.allvars)
```

```
##
## 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
## amManual     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
```

Diagnostics:

```
par(mfrow=c(2,2))
plot(model.final)
```



Leverage points:

```
sort(hatvalues(model.final))[1:5]
```

```
##       Merc 450SE      Merc 450SLC       Merc 450SL Pontiac Firebird
##       0.09272718       0.09794356       0.10113822       0.10164174
##    Porsche 914-2
##       0.10489565
```

```
sort(dfbetas(model.final))[1:5]
```

```
## [1] -0.8709338 -0.5133810 -0.5073819 -0.4734464 -0.4334728
```