Coursera Regression models peer assessment

Thomas Guenther, Germany

October 26, 2014

1 Executive summary

As many may know: cars with a manual transmission performing more efficient than cars with automatic transmission. One reason is that the automatic transmission was built with comfort first in head. Furthermore it has a higher friction loss than a manual transmission and the necessary technology used to make them work is much more complicated. The goal of this report is to answer the following particulary questions:

- Is an automatic or manual transition better for MPG?
- Quantify the MPG differene between automatic and manual transmission

Additionally to the questions above we will do a little inference to examine our car data set. Code This report is divided into several parts for better overview.

2 Automatic or manual transition? (Exploratory adventures)

First we will take a quick look at our data set *mtcars*. We have *11 variables with 32 observations* in our data set. For a description of the variables please see in the R-manual. Next we will do a box plot to see visually if manual transmission performs better than an automatic one or vise versa.

```
p <- ggplot(mtcars.with.factored.am, aes(am, mpg, fill=am)) +
  geom_boxplot() +
  coord_cartesian(ylim=c(10,35)) +
  xlab("Type of transmission") +
  ylab("Miles per gallon") +
  ggtitle("Miles per gallon by type of transmission") +
  theme(plot.title = element_text(color="blue", size=10, vjust=1.0))
print(p)
```

Figure 1 shows us where the middle 50% of the data can be found and the min/max values of mpg for each submission type lies. We will also see that the mpg-values of the manual transmission are right skewed. This means the values for lower mpg's are closer together.

2.1 Examining means via t-test

We will perform a 2-sided t-test to find out if we can find differences in the mean of mpg by transmission type:

- H_0: The means of mpg by transmission are equal
- H_1: The means of mpg by transmission are different

```
t.test(mpg ~ am, mtcars, var.equal = FALSE)
```

Please see **R-Code output 1** for the result of the code above. As we see the p-value is much lower than 0.05 ($0.001374 < 0.05$) and the confidence interval don't contains zero. So we should reject the Null-hypothesis H_0 that the means are equal. So we can conclude there is a difference between automatic and manual transmission as the box plot in **Figure 1** already turned out.

2.2 Regression analysis

In this section we want quantify the MPG differene between automatic and manual transmission. For this we will start with a simple linear regression and afterwards we will analyse a multivariate linear regression. A comparison of our ceated models including some diagnostics will round up the made effort.

2.2.1 Simple linear regression

The **R-Code output** 2 shows the simple regression model with mpg as dependent variable and am as independent variable. The R-squared error is very low. That tells us that the model performance is very poor. We will now try to make it more precise with multivariate regression. The model was generated by the following code:

```
mtcars.simple.lm <- lm(mpg ˜ am, data = mtcars)
```

2.2.2 Multivariate linear regression

Now let's see if we can find a better model. To do this an approach called ***"All subsets regression"*** will be applied. This algorithm inspects all possible predictor combinations and returns the ***nbest*** ones. We will use ***nbest=4***. That means that we get a result containing the 4 best combinations with one predictor, the 4 best combination with two predictors, and so on, until all variables are included. To accomplish our work we use the package *leaps* and the following command to find the best 4 predictor combinations:

```
best4.subsets <- regsubsets(mpg ˜ ., data=mtcars, nbest=4, force.in = "am", nvmax = NULL)
```

We drop in ***mpg*** and all predictors, we will force that ***am*** is always in the evaluation set and we want all predictor combinations tested, but only the 4 best reported back. Then we will do a plot of the regsubsets result (see **Figure** 2) and sort out the best predicor combinations. For simplicity we will use the ***adjusted R^2***. There are other methods like **Mallows Cp** to find the best combinations, but the R^2 is already well known.
The plot shows us the best predictor combinations with the best computed R^2 on top. The combinations with a R^2 of 0.84 in addition to our simple and full regression models are the ones we will should examine a bit deeper.

```
simple.fit <- lm(mpg ˜ am, data = mtcars)
full.fit <- lm(mpg ˜ ., data = mtcars)
subset.fit1 <- lm(mpg ˜ am + disp + hp + wt + qsec, data = mtcars)
subset.fit2 <- lm(mpg ˜ am + hp + wt + qsec, data = mtcars)
subset.fit3 <- lm(mpg ˜ am + wt + qsec+ carb, data = mtcars)
```

2.2.3 Diagnostics and final model selection

Now we have to select one model from our collected 5 models above. First we will generate Akaike's An Information Criterion, hoping to find a good model that suites our needs. We do this that way:

```
AIC(simple.fit, full.fit, subset.fit1, subset.fit2, subset.fit3)
```

The output can be shown in **R-Code output** 3. We look for models having the lowest AIC. As we can see he models ***subset.fit1 to subset.fit3*** looks promising.
My winner is model **subset.fit3** even if it hasn't the lowest AIC but the three subset models are close to each other. Because of the page limit i cannot plot more than one diagnostic plots to show my decision but i found the diagnostics better for that model. The dignostic plot goes to **Figure** 3.

- *Normality:* The Q-Q-Plot in the lower right (and also in the upper middle but without confidence envelop) in **Figure** 3 contains a 95% confidence envelope and shows that most of the points are close to the line within this envelop. There are some outliers we should examine but this is left to the reader. I suggest that normality is given.
- *Linearity:* The residuals vs. fitted shows no special pattern to me. All points are randomly cluttered.
- *Homoscedasticity:*The plot in the upper right (Scale-location) forms a random band around the line and so have a constant variance. I suggest that there is no homoscedasdicity.
- *Residual vs. Leverage:* The Residual versus Leverage shows some outliers (i.e. Chrysler Imperial). This plot is hard to interpret for me. Here we could try to remove these observations from our data set if we can find significance for our outliers and see if we get better results (This could be done with the *outlierTest() function of the car package*). But we will let it for future examinations.

If we look at the summary for model *subset.fit3* we will see that thhe predictors *am, wt and qseq* are significant. The predictor variables account for approx. 83.6% of the variance in miles per gallon. It could be say: on average cars with manual transmission gets **3.5114 miles per gallon** more than cars with automatic transmission (adjusted by wt, qsec and carb).

3 Conclusion and assumptions

We saw that cars with manual transmission are better than with automatic transmission what we showed visually via boxplot and via a 2-sided t-test. Afterwards we quantified the difference and found out that cars with manual transmission gets **3.5114 miles per gallon** more than cars with automatic transmission. But we saw also that we did not fit a perfect model. In future sessions another regression approaches (like glm) could be applied and analyzed. We saw that we had some outliers in our data this is a further point of investigation. Some more dignostics like ANOVA, BIC and so on could also lead to beter models. But this could be analyzed another day...

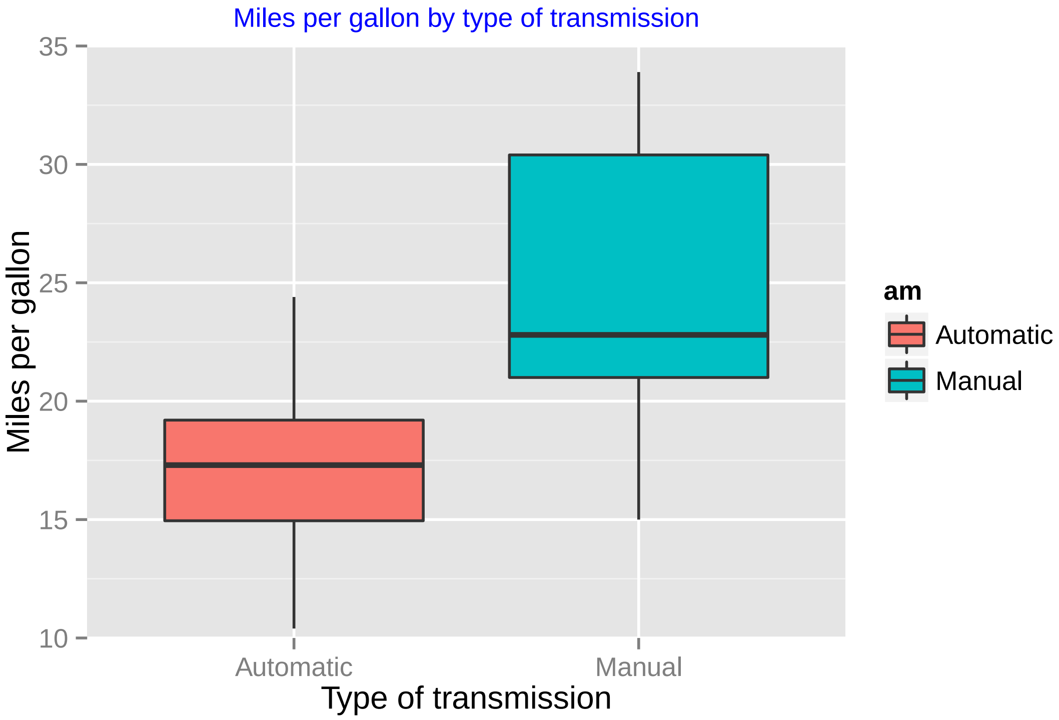

Figure 1: Box plot of MPG vs. transmission type

4 Appendix

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

R-Code output 1: Result of the 2-sided t-test for our simple regression model

```
##
## 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
```

R-Code output 2: Summary of the simple linear regression model

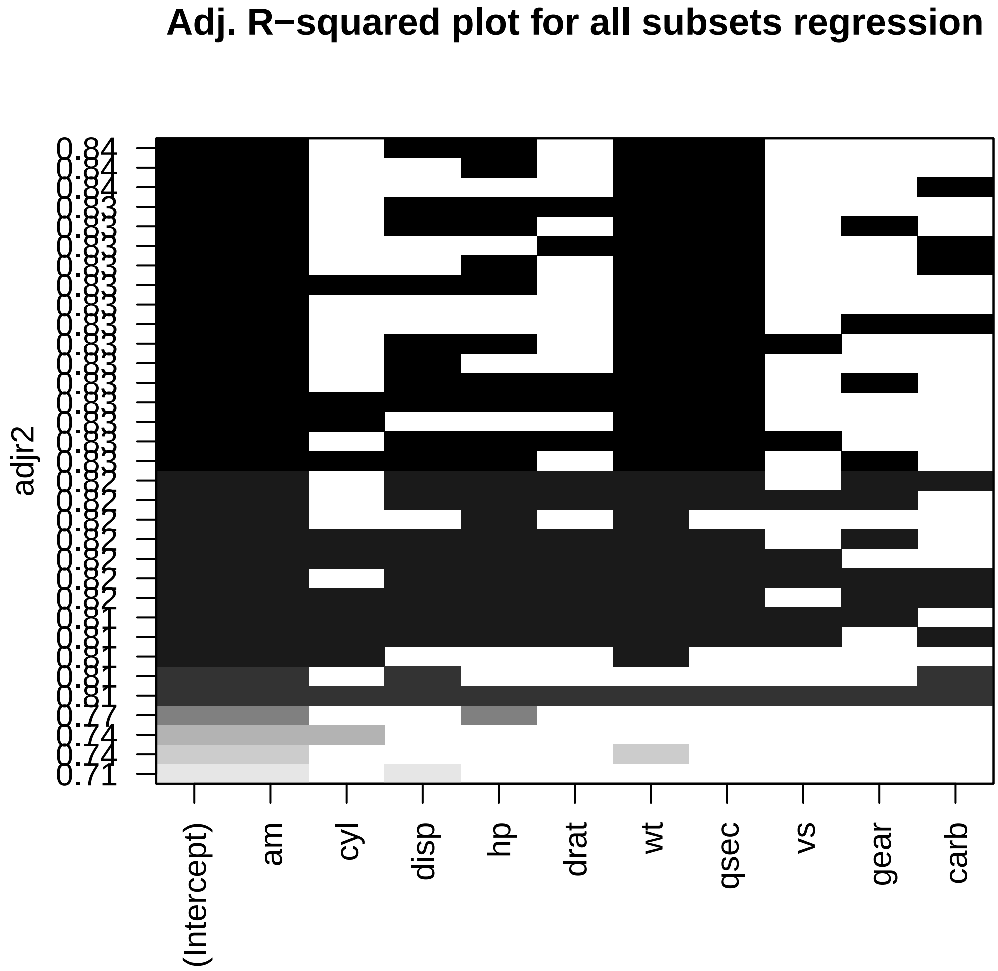

Figure 2: Subset regression plot

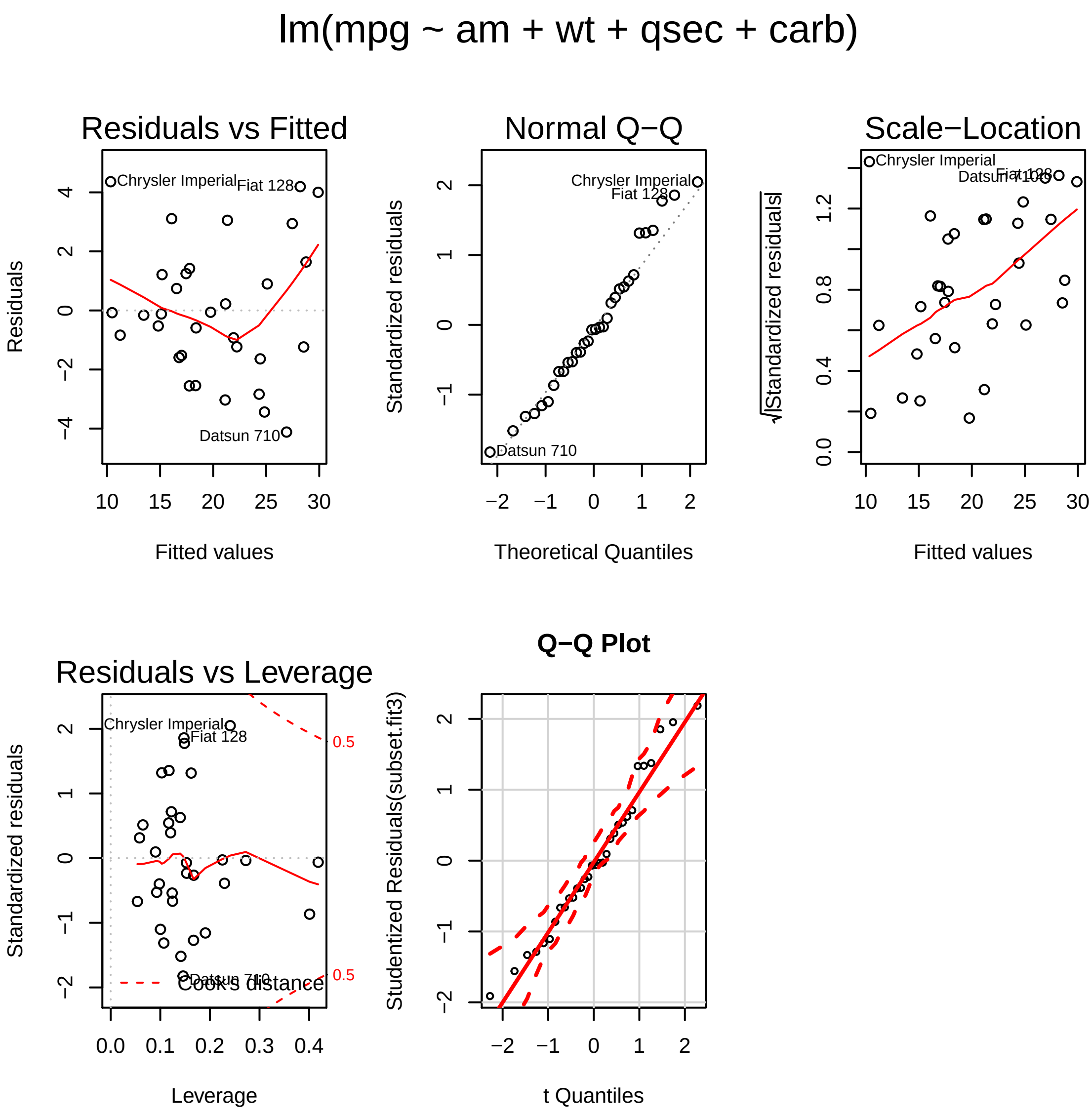

Figure 3: Diagnostic plot of our final model

```
##             df      AIC
## simple.fit   3 196.4844
## full.fit    12 163.7098
## subset.fit1  7 154.9740
## subset.fit2  6 154.3274
## subset.fit3  6 154.5631
##
## Call:
## lm(formula = mpg ˜ am + wt + qsec + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1184 -1.5414 -0.1392  1.2917  4.3604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.8972     7.4725   1.726 0.095784 .
## am            3.5114     1.4875   2.361 0.025721 *
## wt           -3.4343     0.8200  -4.188 0.000269 ***
## qsec          1.0191     0.3378   3.017 0.005507 **
## carb         -0.4886     0.4212  -1.160 0.256212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.444 on 27 degrees of freedom
## Multiple R-squared:  0.8568,Adjusted R-squared:  0.8356
## F-statistic: 40.39 on 4 and 27 DF,  p-value: 5.064e-11
```

R-Code output 3: AIC statistics and summary of our best multivariate regression model