Regression Models Course Project

Azmi

April 21, 2016

Executive Summary

This article studies the relationship between the mileage (miles per gallon - MPG) and a set of variables. In particular, we are interested whether manual or automatic transmission will be better for mileage.

The data used for this study is the **mtcars** datasets available from within R. It contains 11 different features for 32 models of cars. The relationship between mileage and the other variables is studied using linear regression methods.

The result of the study suggest, given the current evidence, that the choice of transmission has no significant effect on the mileage. The rest of this article will provide arguments towards this conclusion.

Data Preparation & Exploratory Data Analysis

The study is initiated by loading the **mtcars** dataset. By looking at the features, it is decided that several features are best converted to factors before the dataset can be used.

```
data("mtcars")

# Convert some features to factor types
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"))
```

For brevity, the code and graphs in this section has been included in the **Appendix**. **Figure 1** shows a pairs plot for all the features in the dataset. By looking at either the first column or row, we can explore if there exists any direct relationship between MPG and any of the features. For example, we can see a linear relationship between *MPG* and *am* (transmission type), a potentially non-linear relationship between *MPG* and *disp* (displacement) and no obvious relationship between *MPG* and *qsec* (quarter mile time).

As the study focusses on the link between *MPG* and *am*, we look at *am* variable with a bit more detail. **Figure 2** shows a boxplot to compare the *MPG* for the 2 transmission types. The graph seems to suggest that there is a difference in the mileage with the automatic transmission on average giving better mileage.

Simple Linear Regression

A first model attempted in the analysis is a simple linear relationship between mileage and the transmission type. This model uses *mpg* as outcome and *am* as the predictor. The following code shows summary of the model.

```
model1 <- lm(mpg ~ am, data=mtcars)
summary(model1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

The results show that that intercept value (B0 coefficient), representing the expected mileage for the automatic transmission is **17.1473684** mpg and the slope value (B1 coefficient), representing the increase in mileage when the transmission is manual is **7.2449393**.

We look at the strength of this outcome using the t-test results. The p-value for B1 is **0.0002850207**.

The 95% confidence intervals for both transmission types are:

```
confint(model1)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## amManual     3.64151 10.84837
```

This provides significant evidence for us to reject the null hypothesis (at alpha=0.05) that there is no difference in the mileage between the two transmission types i.e. that the manual transmission gives better mileage when compared to the automatic transmission.

However to definitively assert this claim we need to look deeper at this regression result. **Figure** 3 shows the residual plot for this model. The fitted residuals (top-left graph) is as expected given that *am* is a factor variable and the Q-Q plot seems to show that the residuals following the normality assumption. However, if we look at the adjusted R2 value, this simple linear model only accounts for **33.85**% of the relationship which suggests that we need to include more variables in our regression.

Multivariate Regression

The next step is to discover which variables should be included in our linear model. Since we already have an initial, 'minimal' model, we next look at the 'maximal' model which includes all the variables in the dataset. Using *mpg* as the outcome and all other features as predictors, a linear model is built.

```
model.all <- lm(mpg~., data=mtcars)
```

Compared with the initial model, the adjusted R2 value is very high, explaining **77.9**% of the data. However if we look at the p-values (not shown here), all of the covariates are larger than alpha=0.05, we would lead us to reject the alternative hypothesis for each.

The fitted residuals plot (not shown here) exhibits a non-linear relationship existing in the residuals while the Q-Q plot shows that there are values deviating from the diagonal line. This leads to the conclusion that too many covariates have been included in the model.

The challenge then is to discover and select which covariates will lead us to an optimal, parsimonious description of the data. The content for the Regression Models course does not describe a method for doing so other than a simple nesting method. However, a quick search online highlights one method which is to use the Akaike Information Criterion (AIC). More information can be found on the wikipedia page (link provided in the References). To use the AIC method, the **step** function is utilised (Note: The outcome of this command is suppressed).

```
# Step
model.best <- step(model.all, direction="both")
```

```
summary(model.best)$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
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

The resulting model has 4 predictors: *cyl, hp, wt, am*. This new, 'optimal' model explains for **84.01**% of our data, which is even better than our 'maximal' model. It is good to see that our variable of interest *am* is automatically included in this model. Looking at the residuals in **Figure 4**, the plots are better compared with the 'maximal' model but there is still a slight non-linear relation ship in the fitted residuals graphs. This indicates that the model may include additional interactions between the covariates and not just the simple additive model that we have here. Or the data many need to be investigated further by looking at the more influential points. However due to the lack of time, these aspects are not explored further.

If we look at the coefficient for the manual transmission *amManual*, it shows that, keeping all other covariates fixed, an increase of 1.81 miles per gallon. The p-value for this covariate is 0.206 which is larger than alpha=0.05 and the 95% confidence interval is:

```
confint(model.best)[6,]
```

```
##     2.5 %    97.5 %
## -1.060934  4.679356
```

The interval includes the 0 value. This leads to the final conclusion that there is no significant evidence to reject the null hypothesis and thus there is no signicant difference in mileage between the automatic and manual transmission.

Conclusion

Using the best multivariate linear regression model selected for this data, there is no significant evidence that there is a difference in mileage between automatic and manual transmissions.

However the caveat for this conclusion is that more work needs to be done to improve the model by looking at alternative covariate selection methods. Perhaps a more confident result can be obtained by then averaging across a few 'best' models. Finally, due to the restrictions on length on this assignment, several code chunks and their output mentioned in the text have been omitted.

Reference

1. (https://en.wikipedia.org/wiki/Akaike_information_criterion)

Appendix.

Figure 1: Pairs plot of "mtcars" features

```
# Pairs plot comparing all the features
pairs(mpg~., data=mtcars)
```

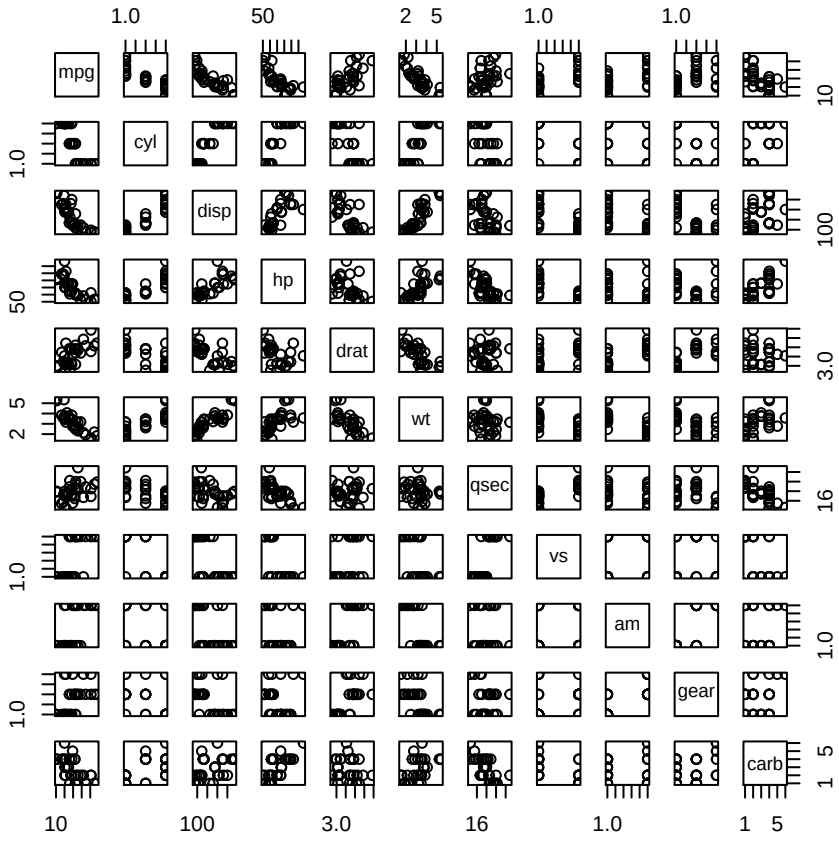

Figure 2: Boxplot of Miles per Gallon (mpg) against Transmission type

```
# Boxplot for the transmision type
boxplot(mpg ~ am, data=mtcars, col=(c("red","blue")),
        xlab="Transmission Type",ylab="Miles per Gallon",
        main="Mileage comparison for Automatic/Manual transmission")
```

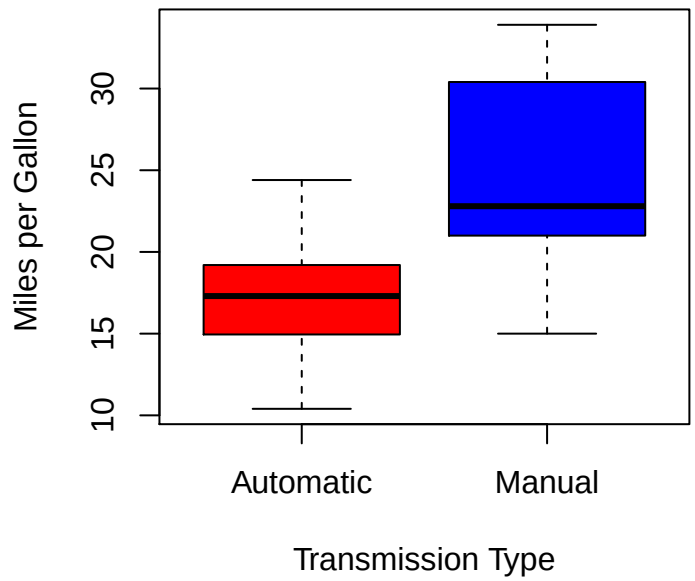

Figure 3: Residuals plot for minimal regression model

```
# plot residuals
par(mfrow=c(2,2), mar=c(2,2,2,2))
plot(model1)
```

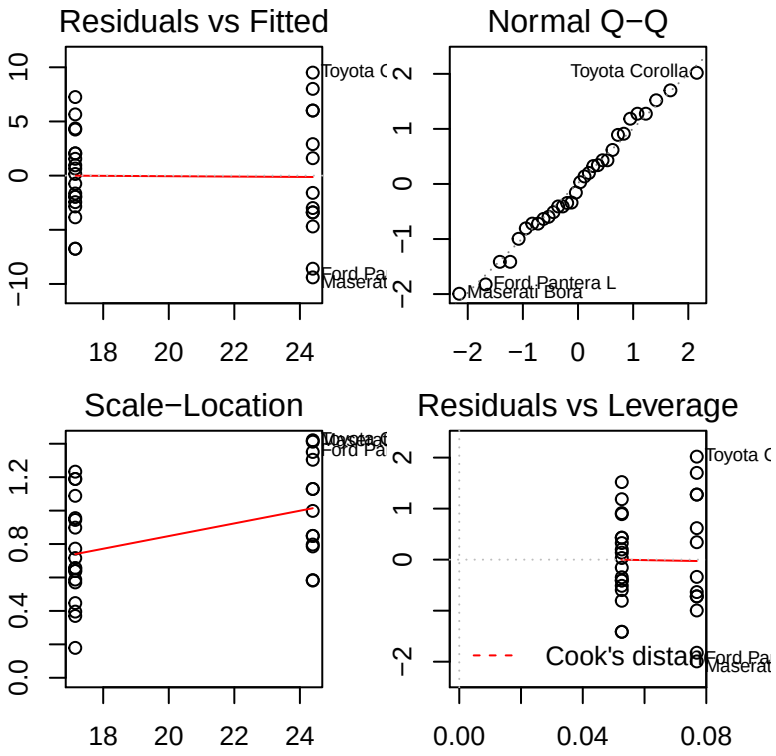

Figure 4: Residuals plot for 'best' regression model

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