Course Project Regression Models: Motor Trend

Bert Lijnen

21 August 2017

1. Executive Summary

- Using simple linear regression, we have found a signficant difference between the mean mpg for automatic and manual transmission cars, with the latter having 7.245 more mpg on average.
- However, in order to adjust for other confounding variables such as the weight and horsepower of the car, we have run a multivariate regression in order to estimate the mpg more accurately.
- After validating the model using ANOVA, we can conclude that manual transmission cars get 1.809 mpg more than automatic transmission cars.

2. The *mtcars* dataset

We load the *mtcars* dataset and the packages we will use.

```
data(mtcars)
library(pander)
```

The dataset consists of 32 observations (car brands) and 11 variables. All variables are numeric, but some are essentially categorical variables so we have transformed them into factors (*cyl, vs, am, gear* and *carb*). The first 6 observations are shown in appendix 1.

3. Exploratory data analysis

In appendix 2 we have plotted a correlation matrix plot in order to better understand the associations between the numeric variables (in particular *mpg*). We can see a strong negative correlation between *mpg* and *displacement* (when you have more displacement, you must burn more fuel; a larger cubic inch engine will get less mileage), *mpg* and *horsepower*, *mpg* and *weight* (the heavier the car, the lower the gas mileage), but a strong positive correlation between *mpg* and the *rear axle ratio*.

In the boxplot (appendix 3) we have compared the mean value of mpg for both transmission types. We can see that the mpg in cars with manual transmission is clearly higher than in cars with automatic transmission.

4. Effect of transmission type on mpg: initial regression model

We want to know whether a manual or automatic transmission is better for mpg. The boxplot in appendix 3 shows that the mpg varies according to which transmission type is used. We fit a linear regression in order to quantify the difference.

```
m1<-lm(mpg~am, data=mtcars)
pander(summary(m1), add.significance.stars = TRUE, caption="")
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	17.15	1.125	15.25	1.134e-15	* * *
amManual	7.245	1.764	4.106	0.000285	* * *

Observations	Residual Std. Error	R^2	Adjusted R^2
32	4.902	0.3598	0.3385

We interpret the coefficient for *amManual* as the difference between *manual transmission* and the baseline (i.e. *automatic transmission*). The intercept estimate gives use the estimate of mpg for cars with automatic transmission (=17.15). When we add 7.245 (=coefficient for manual transmission) we find the estimate of mpg for cars with manual transmission (=24.395).

So, when driving a car with manual transmission, the mpg will be on average 7.245 higher than when a car with automatic transmission is used. The difference is highly significant ($p << 0.05$) and the standard error is low which means that we can be quite confident about the coefficient estimate. We verify this by calculating the 95% confidence interval:

```
pander(confint(m1))
```

	2.5 %	97.5 %
(Intercept)	14.85	19.44
amManual	3.642	10.85

The 95% confidence interval for both the estimated coefficient and the intercept include the respective estimates.

The regression output shows that 36% of the variation in mpg can be explained by the transmission type (R^2=0.3598). This means that we can improve the model by adding more predictors.

5. Improving the initial model

We first fit the full model and then apply stepwise regression in order to select the most relevant variables while avoiding overfitting.

```
full_model<-lm(mpg~., data=mtcars)
m_step<-step(full_model, trace=FALSE)
```

The output in appendix 4 shows that the best predictors for mpg are *6 cylinders* (baseline being 4 cylinders), *horsepower* and *weight*. The coefficient estimates of these variables are all highly significant ($p < 0.05$). We can also see that adding predictors (especially *weight* and *horsepower*) to the initial model confounds the relationship between *transmission type* and *mpg*. Now we can conclude that, on average, manual transmission cars have 1.809 mpg more than automatic transmission cars.

The adjusted R^2 is now 0.8401 which means that the final model explains 84% of the variation in *mpg*. This is an improvement of 48% compared to the initial model.

In appendix 6 we check the residuals using diagnostic plots and quantiles. The residuals range from -3.939 to 5.051 which means that the predictions will be quite accurate. The quantiles also suggest that the residuals are fairly symmetric around the mean.

This is corroborated by the diagnostic plot which do not show an obvious pattern between the fitted values and the residuals. The QQ-plot suggests that the residuals follow an approximate normal distribution.

6. Comparing nested models with ANOVA

Since the initial model is nested within the final model, we can use an ANOVA procedure to compare the two models in a more formal way.

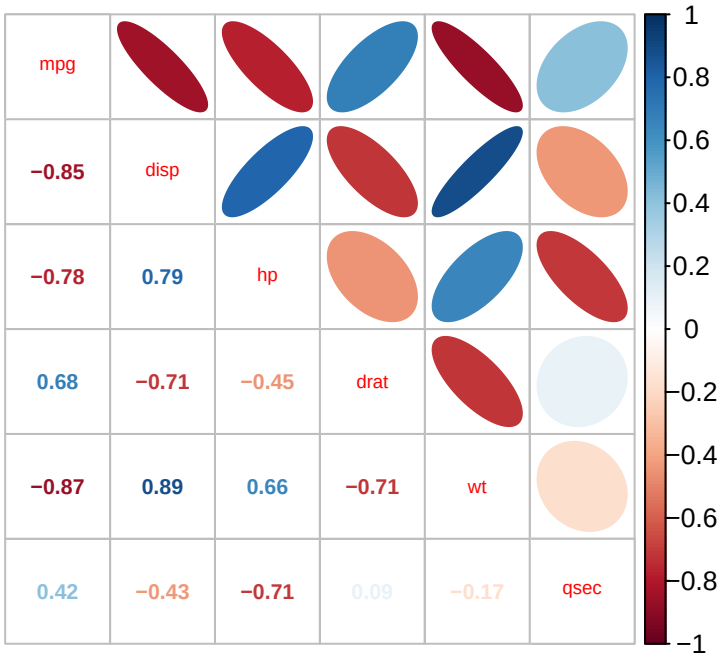
The ANOVA table in appendix 6 shows that the p-value is below 0.05. Therefore, we can reject the null hypothesis that the two models are statistically identical.

Appendices

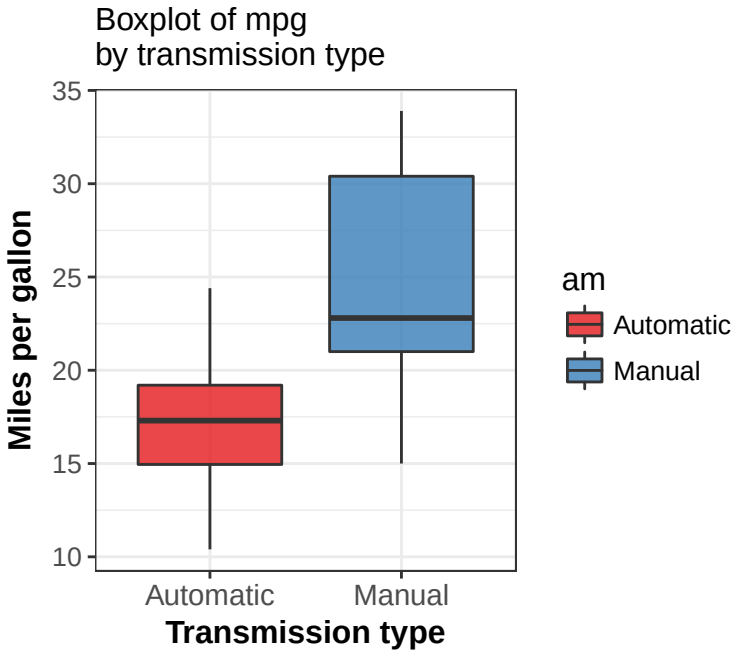
1. Data structure

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21	6	160	110	3.9	2.62	16.46	V	Manual	4	4
21	6	160	110	3.9	2.875	17.02	V	Manual	4	4
22.8	4	108	93	3.85	2.32	18.61	S	Manual	4	1
21.4	6	258	110	3.08	3.215	19.44	S	Automatic	3	1
18.7	8	360	175	3.15	3.44	17.02	V	Automatic	3	2
18.1	6	225	105	2.76	3.46	20.22	S	Automatic	3	1

2. Correlation matrix plot



3. Boxplot



4. Model summary of final model after applying stepwise regression

```
pander(summary(m_step), caption="")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.71	2.605	12.94	7.733e-13
cyl6	-3.031	1.407	-2.154	0.04068
cyl8	-2.164	2.284	-0.9472	0.3523
hp	-0.03211	0.01369	-2.345	0.02693
wt	-2.497	0.8856	-2.819	0.009081
amManual	1.809	1.396	1.296	0.2065

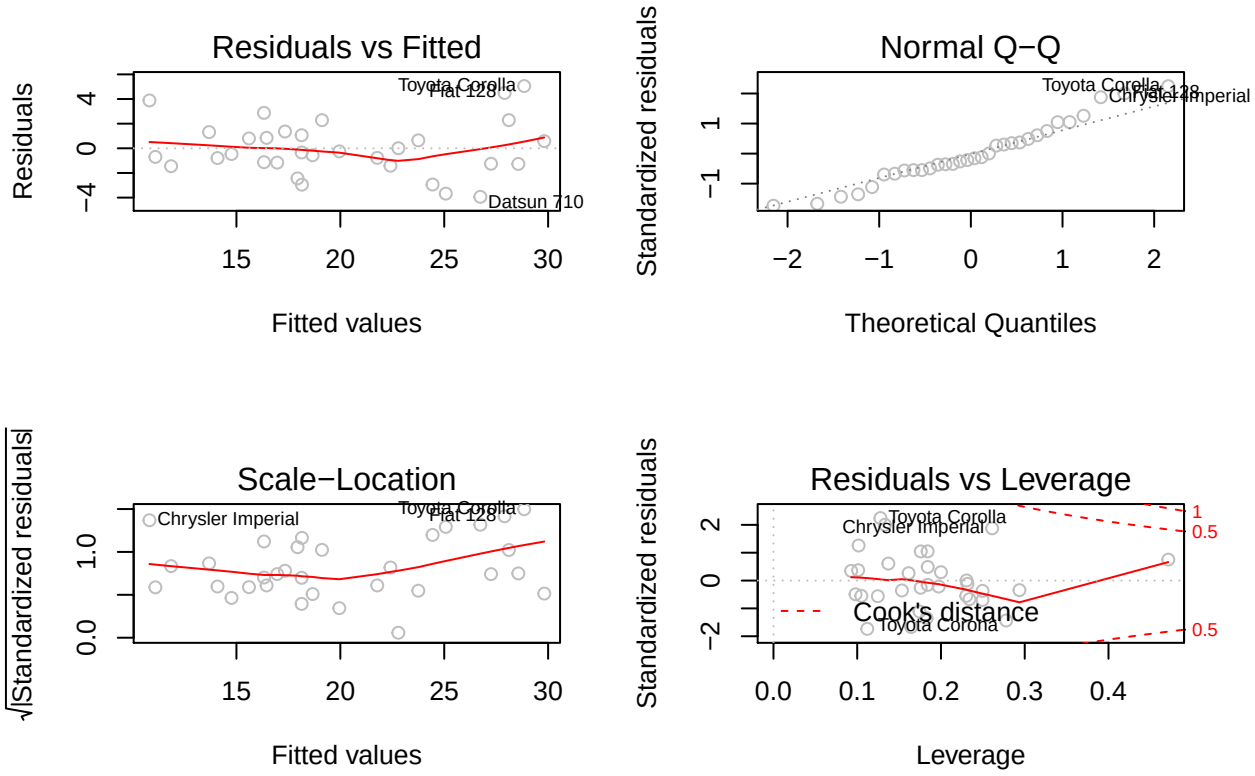
Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.41	0.8659	0.8401

5. Residuals

```
pander(summary(m_step$residuals))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.939	-1.256	-0.4013	1.301e-17	1.125	5.051

```
par(mfrow=c(2,2))
plot(m_step, col="grey")
```



6. ANOVA table

```
pander(anova(m1, m_step), caption="", missing="")
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9				
26	151	4	569.9	24.53	1.688e-08