mtcars analysis

Pascal P

24 August 2018

1 Summary

The aim of this project was to analyze the mtcars dataset to provide answers to the two following questions:

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

1. "Quantify the MPG difference between automatic and manual transmissions".

After a short data preparation and exploratory data analysis (with plots), we use a t-test which shows a difference in fuel consumption between manual and automatic transmission cars. We then select a multi-variable linear regression model (among several) using adjusted R^2 statistic and p-values for significance and investigate how the assumptions for linear regression are met, which suggest out model is not too inaccurate. We then state our conclusions, namely that manual transmission cars (mtc) have a better 'mpg' value (lower consumption) than automatic transmission ones. Our model quantified this difference by a factor of 2.94 for mtc.

2 Analysis

2.1 Exploratory data analysis

According to the help `?mtcars`, the dataset was extracted from the 1974 *Motor Trend* US magazine. It contains 32 samples (cars) measured by 11 features, the outcome is the fuel consumption (denoted mpg). Here is a brief overview of the dataset:

	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

The features `cyl`, `vs`, `gear`, `carb` and `am` are numerical (discrete) and should be treated as factor, so let's transform them.

According to appendix *Pairwise combination scatterplots*, we can see some links between `wt` and `mpg` (to be expected, as weight increases, mpg will decrease), as well as `hp`, `cyl`, `qsec` and `disp` with `mpg`. Also appendix *Box plots MPG vs transmission and weight vs transmission* shows that manual transmission cars seem to have lower `mpg` (which means lower consumption) than automatic ones while the later are heavier than the former (suggesting a possible interaction).

2.2 Fitting models and assessing their statistical relevance

2.2.1 Hypothesis testing

We want to check if the mean MPG difference between manual transmission cars and automatic ones is significant (the null hypothesis states that it is 0, i.e no difference). To do this, we will perform a two sample t-test (unpaired sample with non equal variance, given the sample):

t_stat	df	auto_mean	manual_mean	low_CI	upp_CI	p_value
-3.767123	18.33225	17.14737	24.39231	-11.28019	-3.209684	0.0013736

The results from the t-test show that the p-value is very small and the 95% confidence interval does not contain 0. We can therefore reject the null hypothesis and retain the alternative one which states that the true mean difference of`mpg` for cars with manual and automatic transmissions is different from zero.

2.2.2 Multivariables linear regression models

We want to determine the "best" combination of predictors (comprising *am*) to make accurate predictions about `mpg`. For this we are going to use the *Best Subset Selection* method and the `regsubsets()` function from the `leaps` package. There are multiple ways to quantified "best": adjusted R^2, C_p, `BIC`. We choose the adjusted R^2, cf. *Best subset regression* in appendix for the details.

The best model according to adjusted R^2 has five-predictors. Around this maximum we also have a four-predictor and a three-predictors (left) and above five-predictors (right). This is suggested by the plots in the appendix *Best subset regression* (on the right one, the top line on y-axis). It turns out that looking at the `p-values` for these models, some predictors are not significantly different from 0 (null hypothesis), suggesting that `mpg` and these predictors are not linearly related (cf. summary below, where `mxp` stands for model x predictors). *Please note that I am not showing the models above 5 predictors.*

	Estimate	Std. Error	t value	Pr(>|t|)
m5p_Intercept	31.1846139	3.4200237	9.1182449	0.0000000
m5p_cyl6	-2.0901087	1.6286796	-1.2833148	0.2111508
m5p_cyl8	0.2909754	3.1426983	0.0925878	0.9269690
m5p_hp	-0.0347503	0.0138188	-2.5147163	0.0187137
m5p_wt	-2.3733671	0.8876312	-2.6738212	0.0130226
m5p_vs1	1.9900040	1.7601846	1.1305655	0.2689680
m5p_ammanual	2.7038444	1.5985012	1.6914873	0.1031742
m4p_Intercept	33.7083239	2.6048862	12.9404210	0.0000000
m4p_cyl6	-3.0313445	1.4072835	-2.1540397	0.0406827
m4p_cyl8	-2.1636753	2.2842517	-0.9472140	0.3522509
m4p_hp	-0.0321094	0.0136926	-2.3450251	0.0269346
m4p_wt	-2.4968294	0.8855878	-2.8194036	0.0090814
m4p_ammanual	1.8092114	1.3963045	1.2957141	0.2064597
m3p_Intercept	9.6177805	6.9595930	1.3819458	0.1779152
m3p_wt	-3.9165037	0.7112016	-5.5068823	0.0000070
m3p_qsec	1.2258860	0.2886696	4.2466757	0.0002162
m3p_ammanual	2.9358372	1.4109045	2.0808192	0.0467155

The best three-predictors model (related to adjusted R^2) shows that the three predictors (`wt`, `qsec` and `am`) are significantly different from zero, **this is our model of choice**.

3 Conclusion

- The three-predictors model tells us that on average manual transmission cars have **2.94** more `mpg` than automatic transmission ones.
- It also explains 85% of the variance, as given by R^2 value below:

```
round(summary(best.model)$r.squared, 3)
```

```
[1] 0.85
```

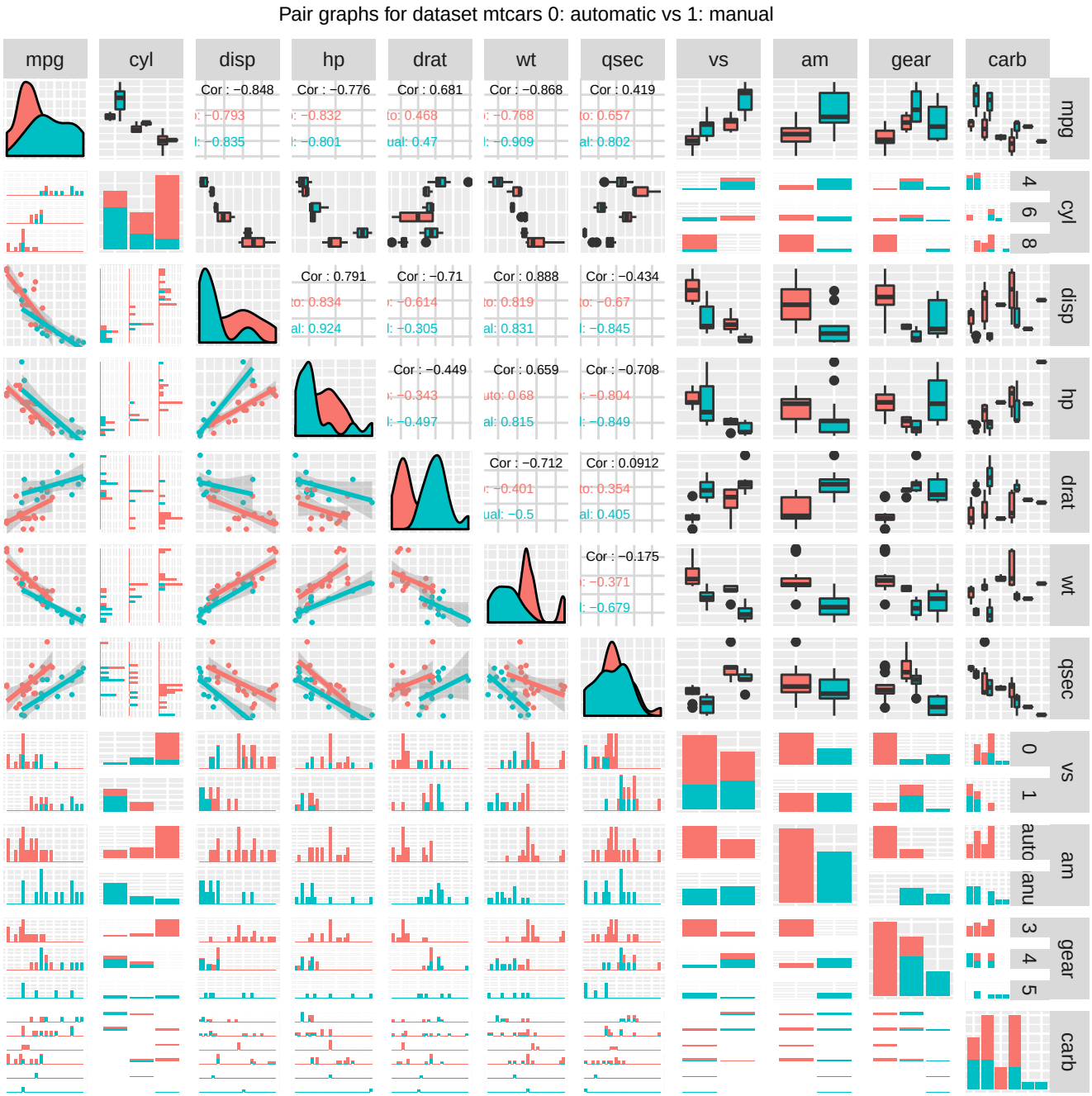
- The assumption of normality (cf. Q-Q plot in appendix) is limit acceptable, while the one about linearity seems to hold.
- The constant variance assumption (homoscedasticity), `ncvTest` (cf. appendix) show a p-value just above the 0.05 cutoff (again limit acceptable), while the spread level plot looks fine.
- With our three-predictor model, multicollinearity is not a problem (cf. appendix for details).
- The last plot appendix "unusual observations", shows that for our model, `Chrysler Imperial` and `Fiat 128` are outliers, `Merc 230` has high leverage, while `Chrysler Imperial` may have disproportionate influence on the parameter estimates.
- Last but not least, the `regsubset()` function does not seem to take into account potential interaction between predictors, which we noticed earlier (automatic transmission cars are heavier, at least on the given sample). We show an alternative (and admittedly better) model in the appendix.

4 Appendix

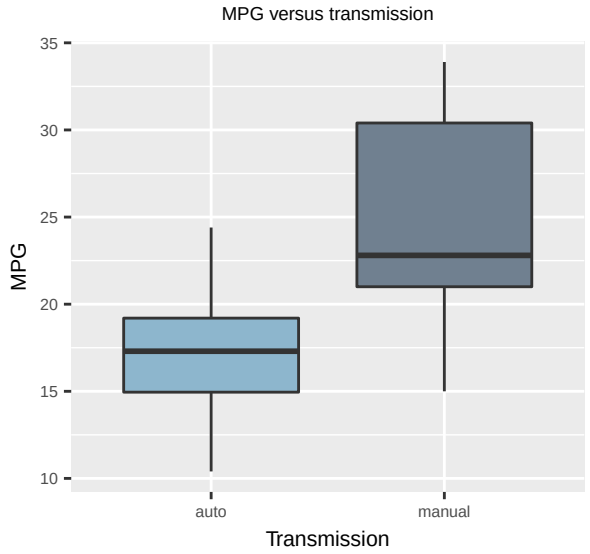
4.1 Data Preparation

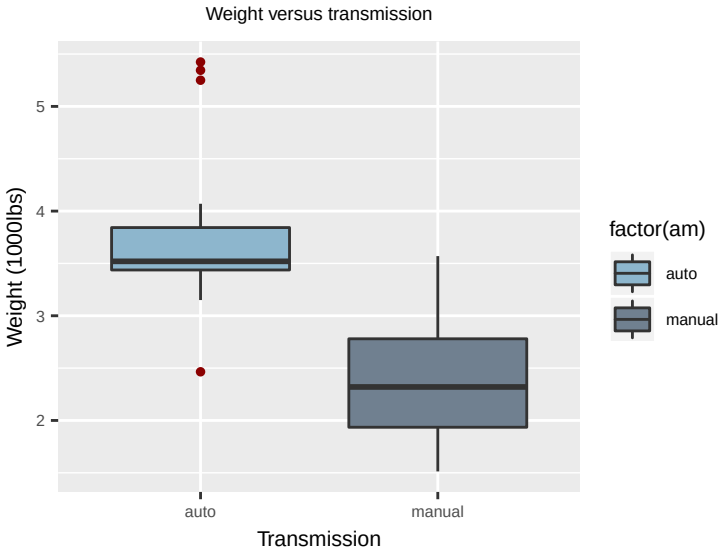
```
# side effect mtcars
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <-  factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am, labels=c("auto", "manual"))
```

4.2 Pairwise combination scatterplots



4.3 Box plots MPG vs transmission and weight vs transmission

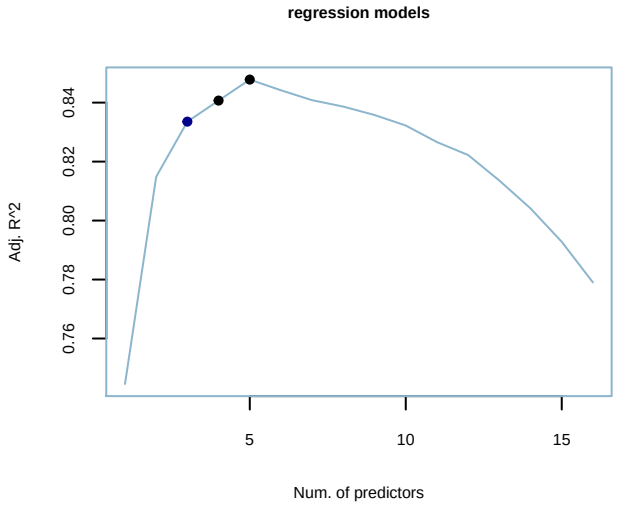


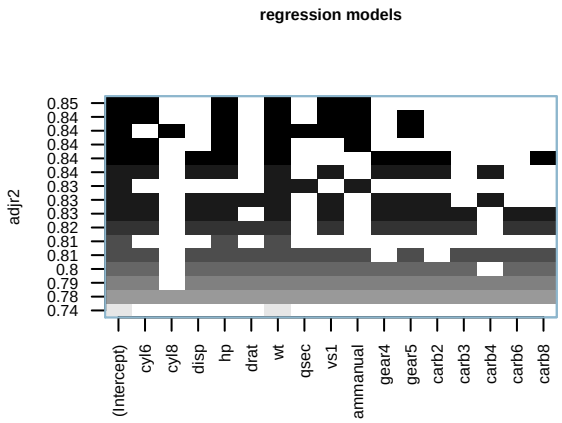


4.4 Best subset regression

```
library(leaps)
fit_mv <- regsubsets(mpg ~ ., data=mtcars, nvmax=18, nbest=1,
                     method="exhaustive")
ix <- which.max(summary(fit_mv)$adjr2) # index of 'best' model (rel. to adjr2)
coef(fit_mv, ix) # coefficients for 'best' model (5-predictors)
```

```
(Intercept)        cyl6          hp          wt         vs1    ammanual
31.28240981 -2.20519611 -0.03393442 -2.36781111  1.87741318  2.62111773
```





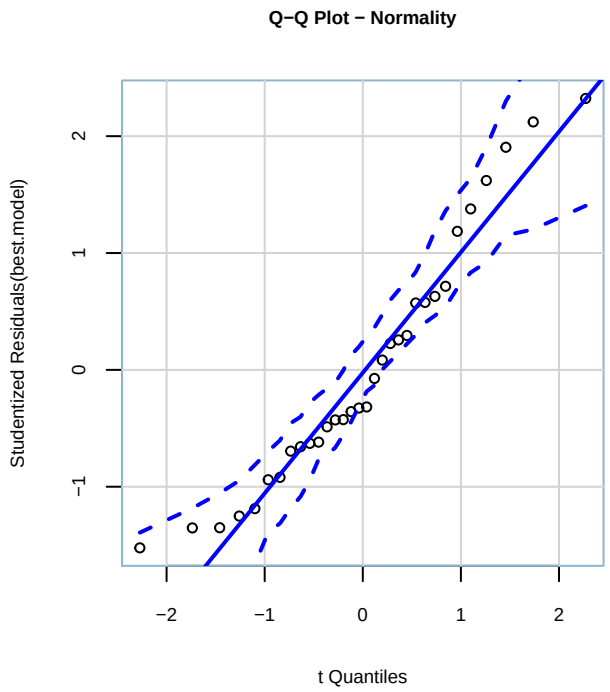
4.5 Regression diagnostics

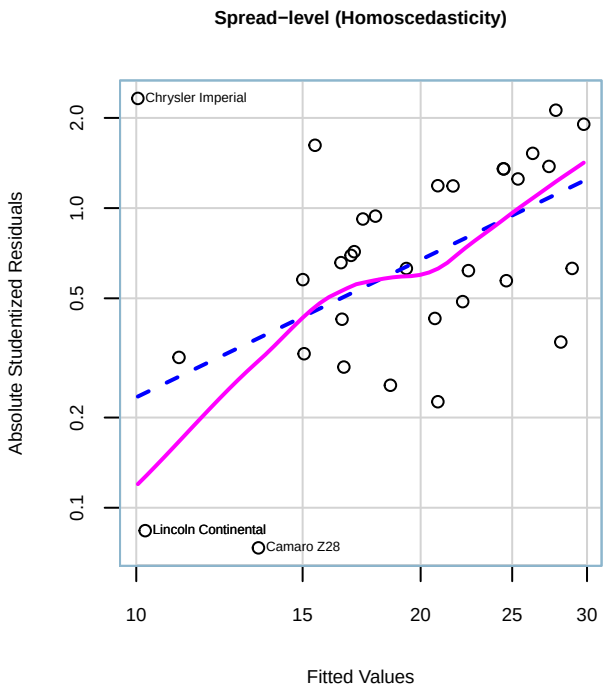
- Homoscedasticity

```
ncvTest(best.model)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.55815    Df = 1     p = 0.2119363
```

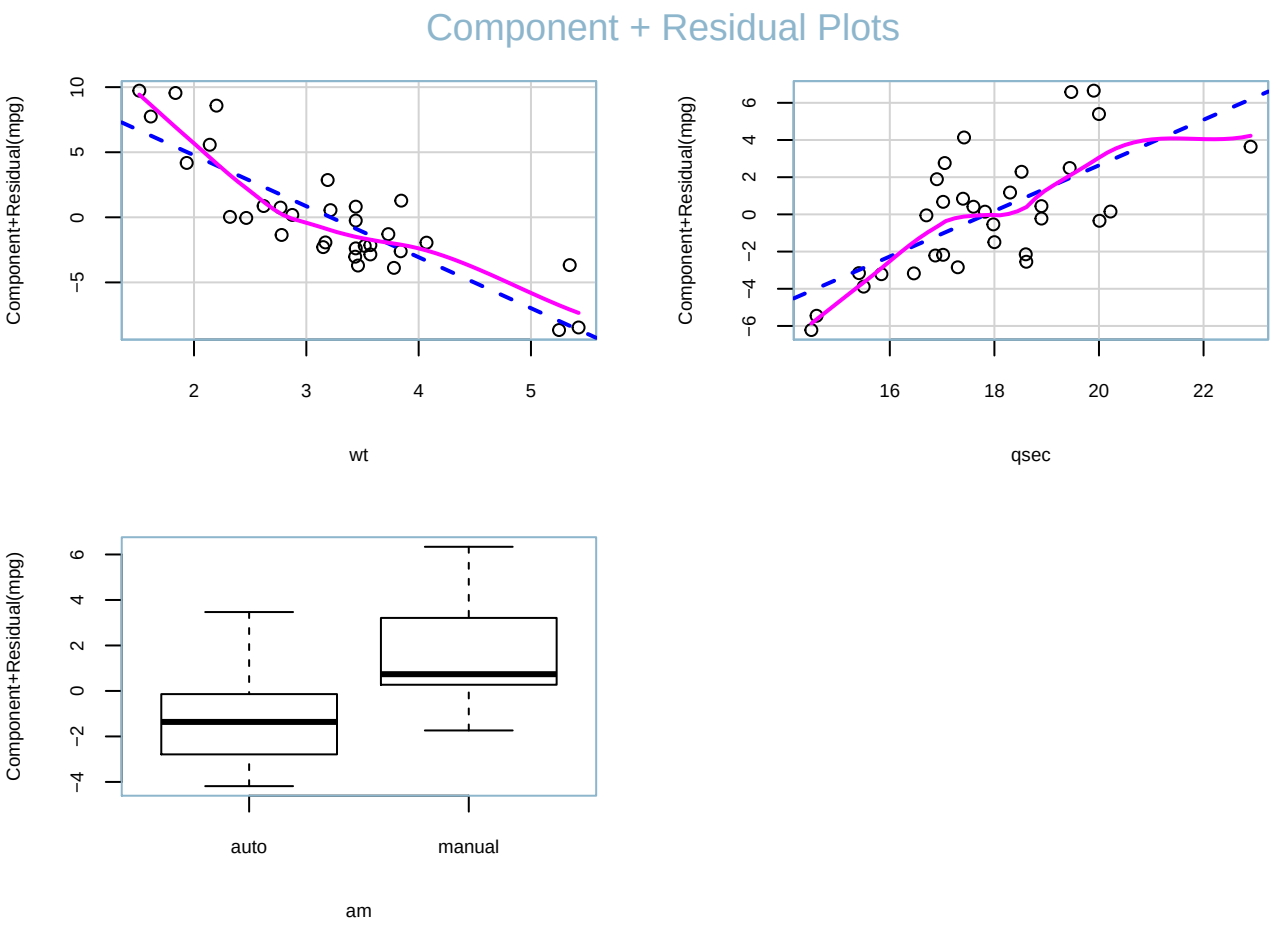
- Normality and Homoscedasticity






```
Suggested power transformation:  -0.5249755
```

- Linearity

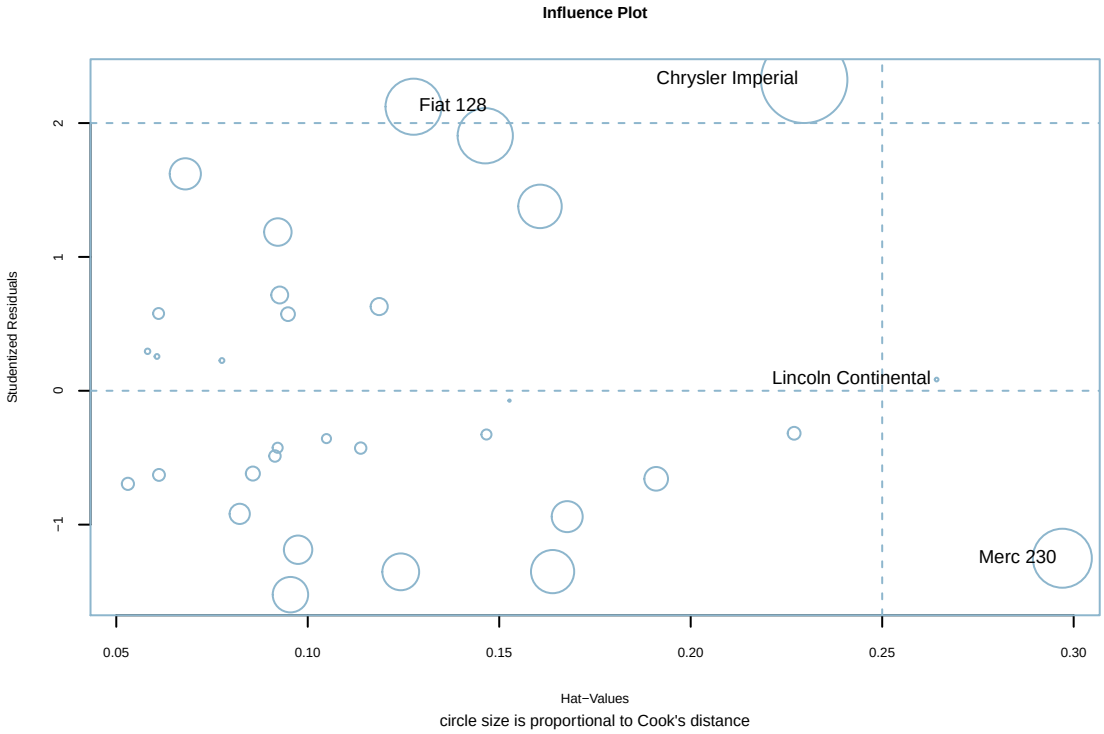


- Multicollinearity (problem if $\sqrt{(vif)} > 2$)

```
      wt     qsec       am
2.482952 1.364339 2.541437
```

```
   wt  qsec    am
FALSE FALSE FALSE
```

- Unusual observation



	StudRes	Hat	CookD
Merc 230	-1.2511056	0.2970422	0.1620827
Lincoln Continental	0.0838378	0.2642151	0.0006542
Chrysler Imperial	2.3231194	0.2296338	0.3475974
Fiat 128	2.1221458	0.1276313	0.1464019

4.6 Alternative model (with interaction)

Let us consider, the possible interaction between `wt` and `am` (as noted on our first plots for exploratory data analysis).

```
best.model_with_interaction <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(best.model_with_interaction)$coefficient;summary(best.model_with_interaction)$r.squared
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  9.723053  5.8990407  1.648243 0.1108925394
wt          -2.936531  0.6660253 -4.409038 0.0001488947
qsec         1.016974  0.2520152  4.035366 0.0004030165
ammanual    14.079428  3.4352512  4.098515 0.0003408693
wt:ammanual -4.141376  1.1968119 -3.460340 0.0018085763
```

```
[1] 0.8958514
```

```
anova(best.model, best.model_with_interaction)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
28	169.2859	NA	NA	NA	NA
27	117.2763	1	52.00966	11.97396	0.0018086

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)

This model looks better indeed, but to prove it, I would need to provide the same regression diagnostics as above, which I did (to some extend) in an optional document (*TODO* github repo here).

5 References

Beyond the coursera course itself [] (and all its resources), I also used (and keep using) the following books:

- *An Introduction to Statistical Learning: with Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani,
- *R in Action: Data Analysis and Graphics with R Second Edition* by Robert Kabacoff,
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* by Trevor Hastie and Robert Tibshirani