Regression Models - Course Project

Davide Rivola

25 September 2015

Executive summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

By exploratory analsys and statistical testing we found that the manual transmission cars have better MPG with a mean difference of 7 MPG.

We then considered other aspects such number of cylinders, horse power and weight and we found out that manual cars have additional 1.8 MPG than automatic cars (when keeping fixed cyl, hp and weight). The biggest impact on MPG is given by the weight, followed by horse power and cylinders number.

Exploratory analysis

Dataset description `help(mtcars)`:

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

```
head(mtcars,3)
```

```
##                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
```

Change variables to factors

```
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

By looking at density plot the automatic cars seems to have in general a lower mpg. (Appendix plot 1).

Pair graph shows a correlation between "mpg" and the "cyl", "disp", "hp", "drat", "wt", "vs" and "am" variables (Appendix plot 2).

Inference

We run a T-test

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

p-value = 0.001374, null hypothesys rejected. Therefore manual and automatic transmission cars are from different populations.This confirms what we saw visually in the density plot during exploratory analysis

Regression

```
fit.all.var <- lm(mpg ~ ., data = mtcars)
summary(fit.all.var)
```

By fitting considering all variables we see that no coefficients are significant (0.05).

We try then to do a backward selection

```
fit <- step(fit.all.var)
summary(fit)
```

The selected model is `mpg ~ cyl + hp + wt + am`

Residuals and diagnostics

Residuals

By looking at residuals plot in appendix plot 3 we can say:

- residuals vs fitted: OK, points randomly distributed
- normal Q-Q: OK, points close to the line, normally distrobuted
- scale-location: OK, points randomly distributed
- residuals vs leverage: OK, no outliers shown

dfbetas

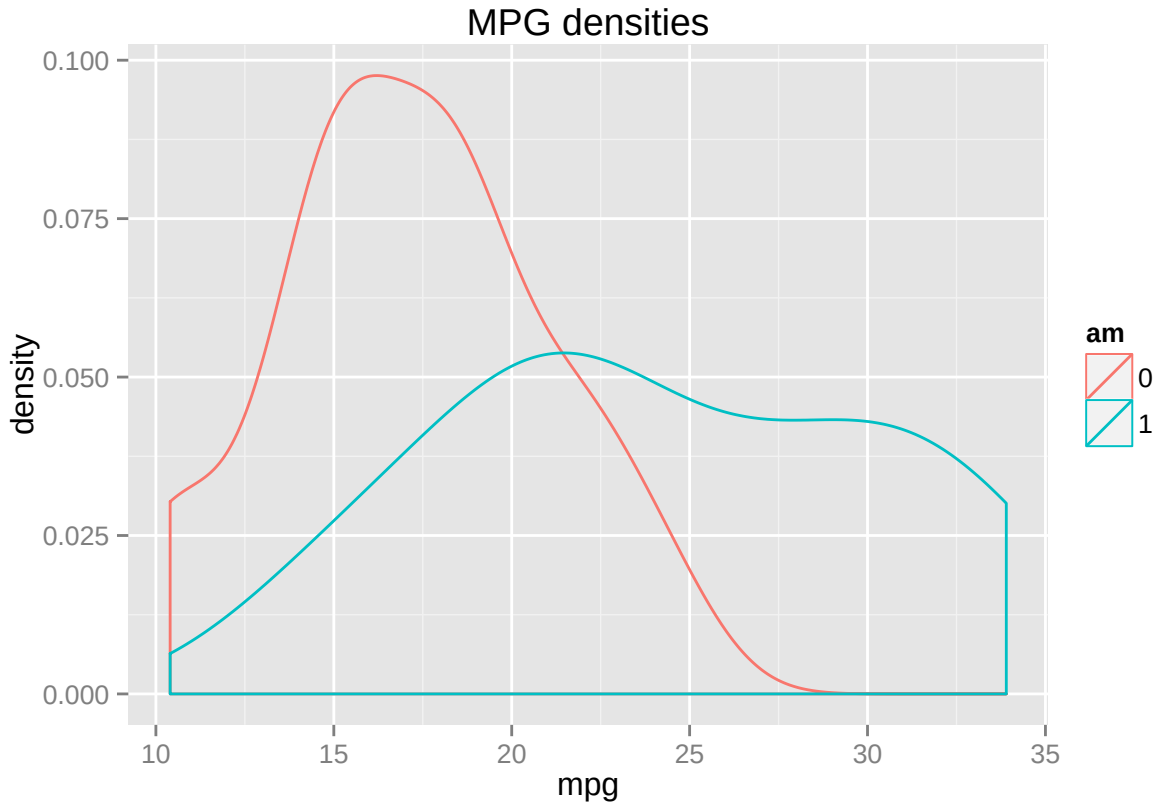
No influential values with dfbetas > 1 found

```
dfbetas <- abs(dfbetas(fit))
dfbetas[which(dfbetas > 1)]
```

Appendix

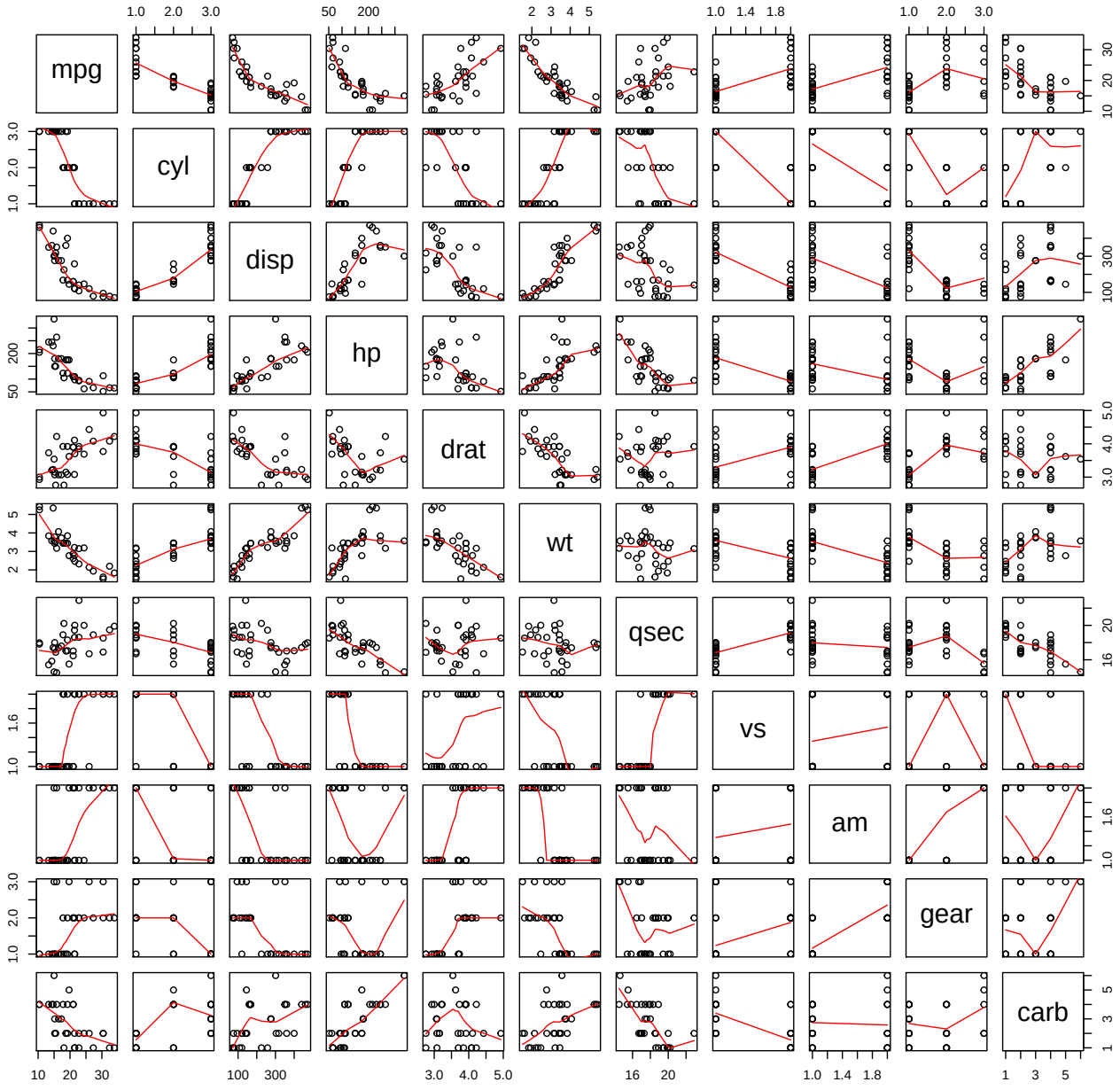
1. Density plot

```
library(ggplot2)
ggplot(mtcars, aes(x=mpg, colour=am)) + geom_density() + ggtitle("MPG densities")
```



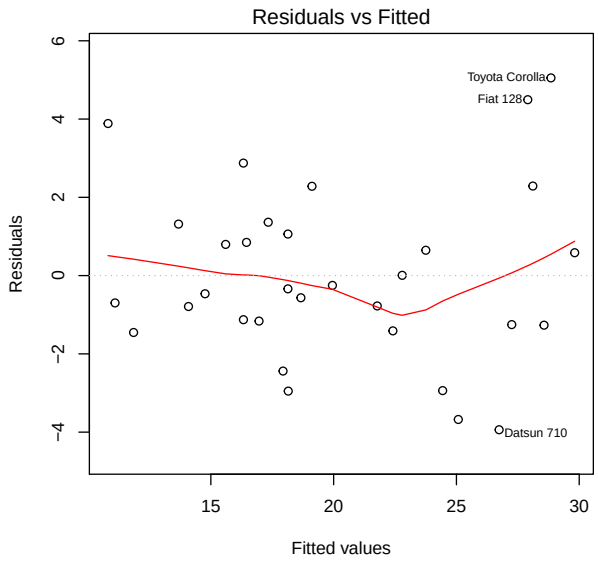
2. Pairs graph

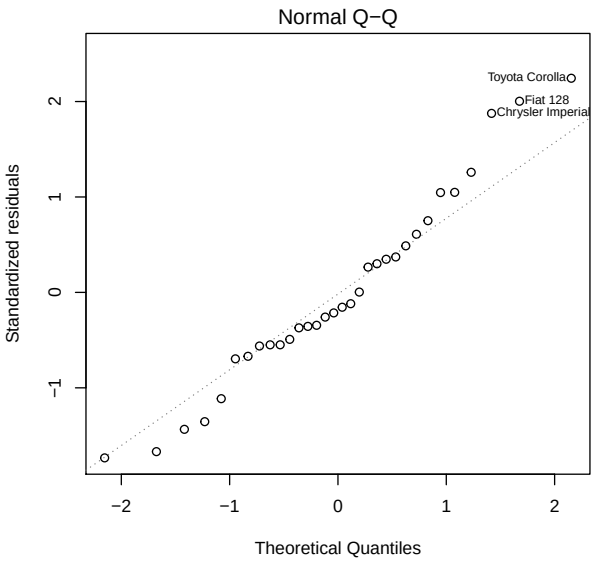
```
pairs(mtcars, panel = panel.smooth, main = "Pairs graph")
```

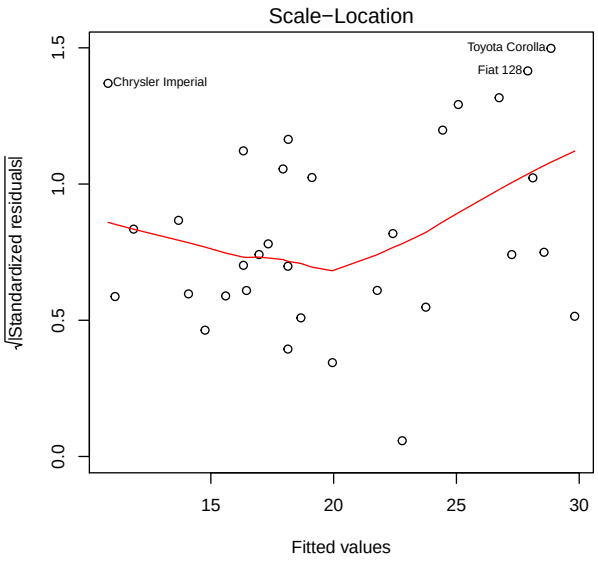


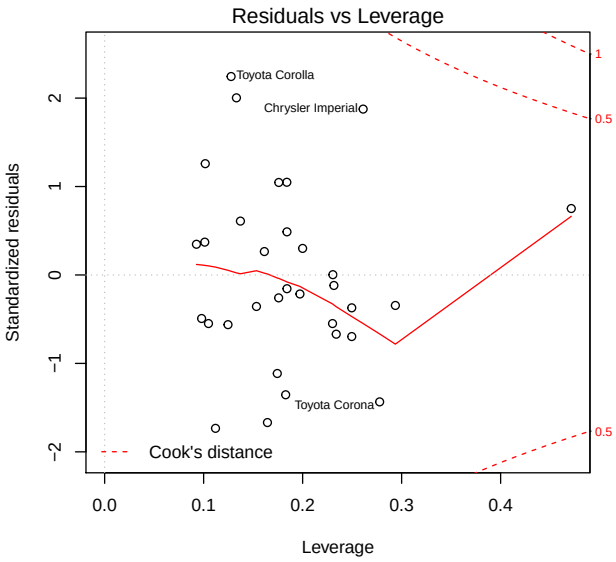
3. Residual plots

```
par(mfrow = c(2, 2))
plot(fit)
```


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Residuals
Fitted values


Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles


Scale−Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
Standardized residuals
Leverage