Regression_analysis_mtcars

Ana Laguna Pradas

Analysis of the MPG difference between automatic and manual transmissions:

- Executive summary
- Looking at the data set
- Exploratory data analyses
- Model fitting and hypothesis testing
- Two samples t-test
- Simple linear regression model

Executive summary

This analysis is performed for the Motor Trend, a magazine about the automobile industry. By looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) as outcome. We are particularly interested to explore:

- "Is an automatic or manual transmission better for MPG"

- "Quantify the MPG difference between automatic and manual transmissions"

In order to answer these questions we performed exploratory data analyses, and used hypothesis testing and linear regression as methodologies to make inference. We established both simple and multivariate linear regression analysis. However the result of the multivariable regression model is more promising as it includes the potential effect of other variables on MPG.

Using model selection strategy, we found out that among all variables weight and quarter mile time (acceleration) have significant impact in quantifying the difference of MPG between automatic and manual transmission cars.

Looking at the data set

For the purpose of this analysis we use mtcars dataset which is a dataset that 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). Below is a brief description of the variables in the data set:

```
# [, 1]   mpg     Miles/(US) gallon
# [, 2]   cyl     Number of cylinders
# [, 3]   disp    Displacement (cu.in.)
# [, 4]   hp      Gross horsepower
# [, 5]   drat    Rear axle ratio
# [, 6]   wt      Weight (lb/1000)
# [, 7]   qsec    1/4 mile time
```

```
# [, 8]   vs      V/S
# [, 9]   am      Transmission (0 = automatic, 1 = manual)
# [,10]   gear    Number of forward gears
# [,11]   carb    Number of carburetors
```

Also the first six records of the dataset are shown below:

```
library(knitr)
library(printr)

kable(head(mtcars),align = 'c')
```

	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
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Notice that each line of mtcars represents one model of car, which we can see in the row names. Each column is then one attribute of that car, such as the miles per gallon (or fuel efficiency), the number of cylinders, the displacement (or volume) of the car's engine in cubic inches, whether the car has an automatic or manual transmission, and so on.

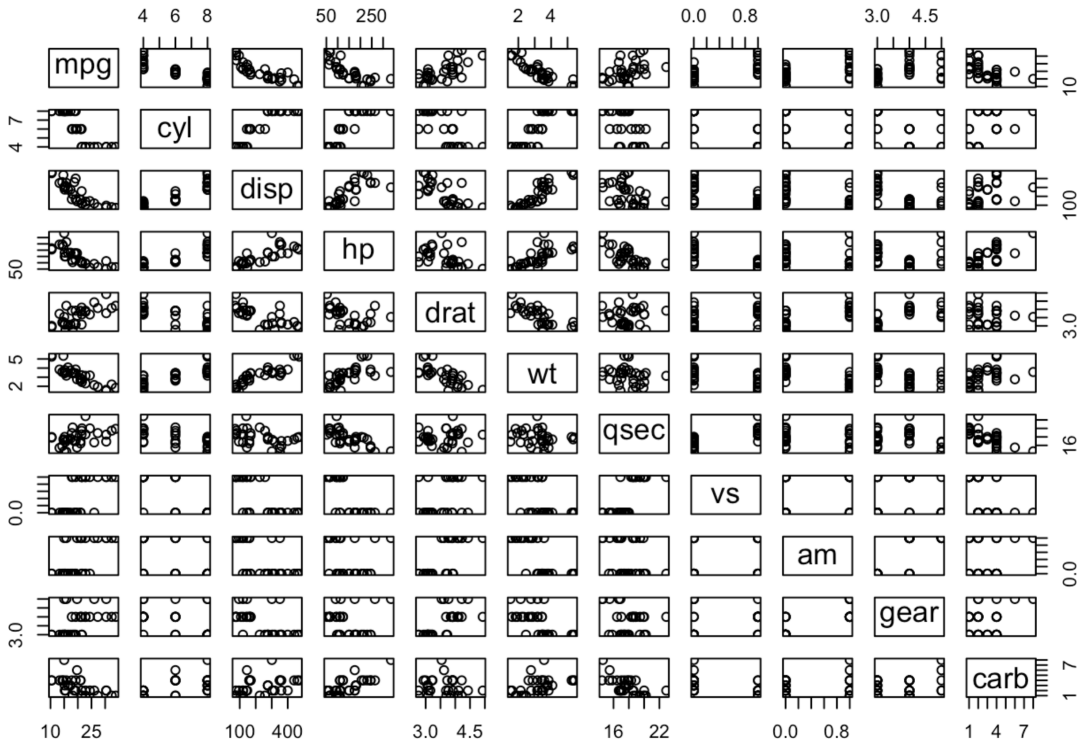
Exploratory data analyses

We begin the analysis by performing some initial explaratory data analysis to get a better idea of the existing patterns between variables in the data set. Normally in regression analysis scatter plot is a very effective tool. Below we create a nice pairwise scatter plots. This is a nice way to investigate the relationship between all the variables in this data set.

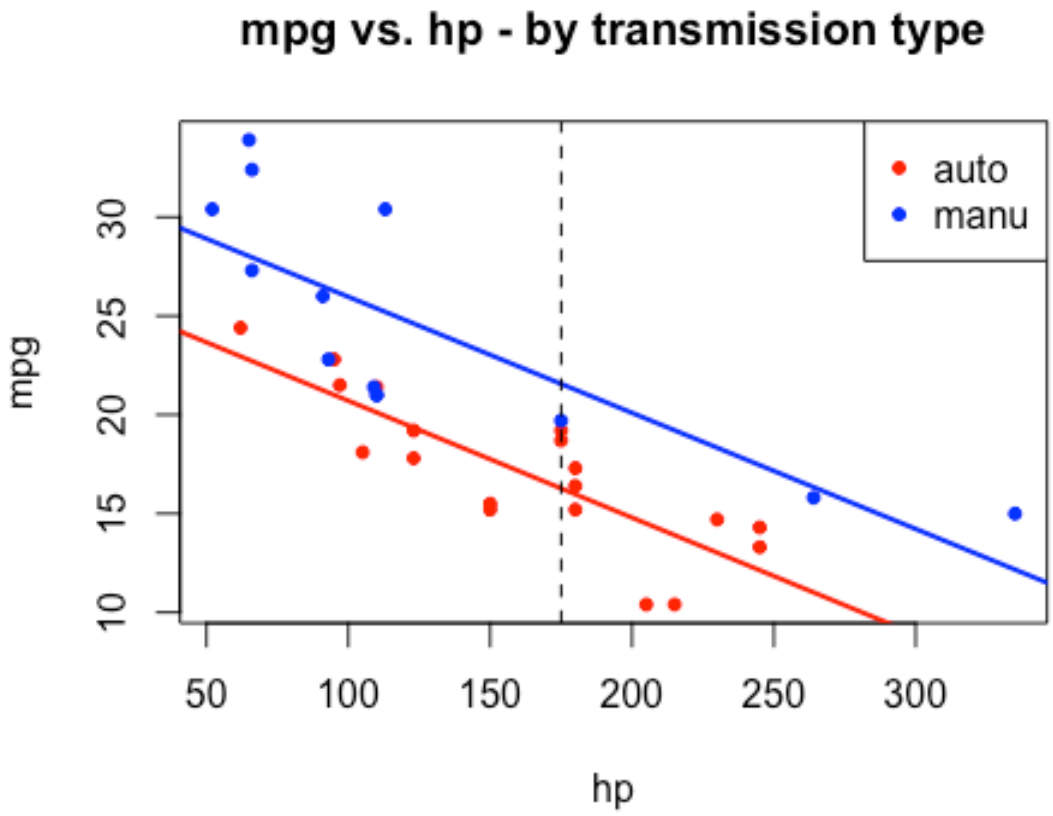
Bivariate analysis

In this chapter I analyse the behavior of target variable (mpg) conditional on a set of explanatory variables.

```
#### generate subset: automatic and manual cars ####
cars_auto = subset(mtcars, am == 0)
cars_manu = subset(mtcars, am == 1)

#### Visual inspection of all covariates ####
pairs(mtcars)
```

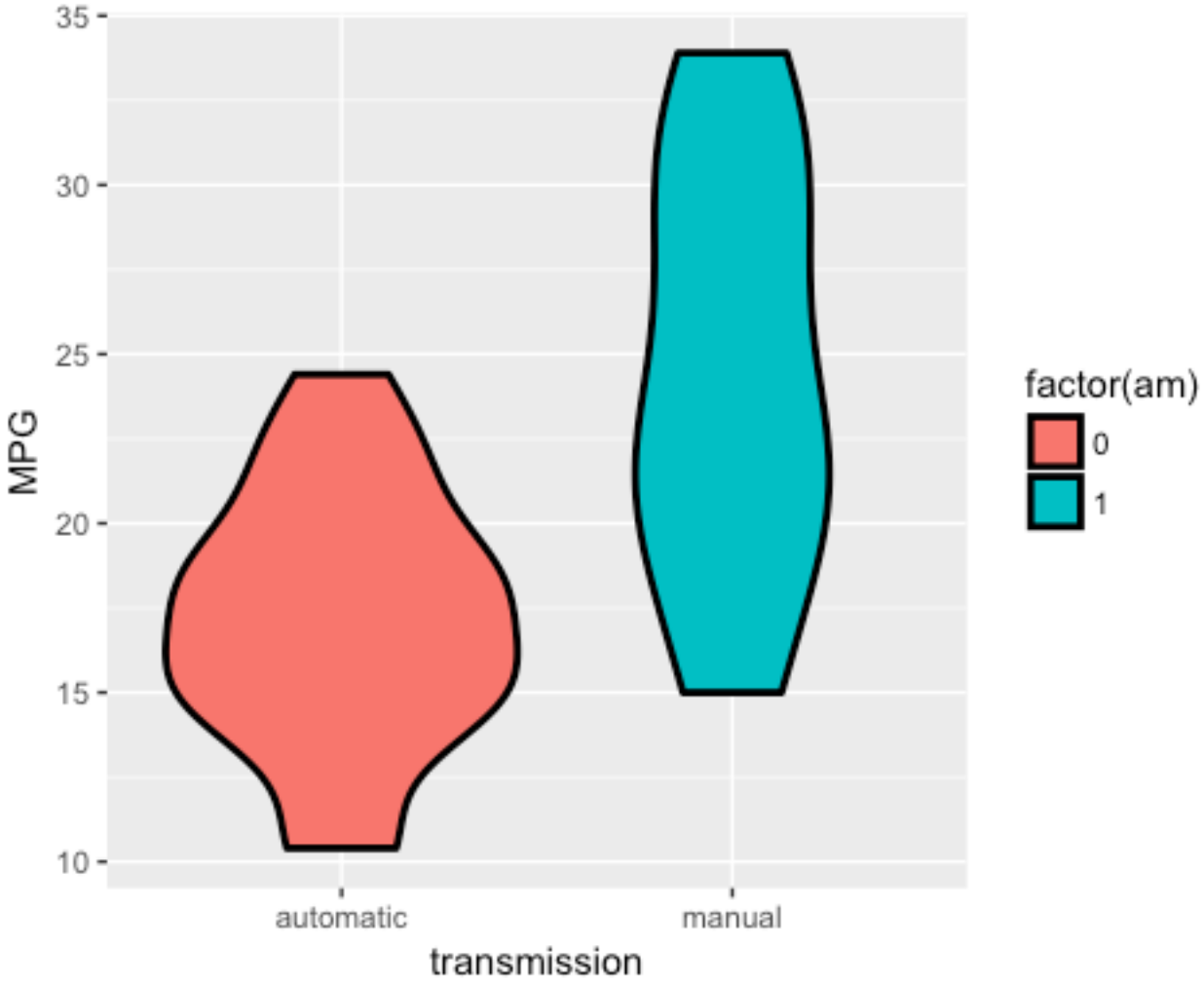



```
#### 4 bivariate analysis: hp / wt / drat / disp ####

# plot1
with(mtcars, plot(hp, mpg, type = "n", main = "mpg vs. hp - by transmi
ssion type")) # no data
with(cars_auto, points(hp, mpg, col = "red", pch = 20))
with(cars_manu, points(hp, mpg, col = "blue", pch = 20))
legend("topright", pch = 20, col = c("red", "blue"), legend = c("auto"
, "manu")) # add legend
model1_auto = lm(mpg ~ hp, data = cars_auto)
model1_manu = lm(mpg ~ hp, data = cars_manu)
abline(model1_auto, col = "red", lwd = 2)
abline(model1_manu, col = "blue", lwd = 2)
abline(v = 175, lty = 2)
```



mpg vs. hp: linear negative relation: as horse power of the engine (hp) increases, the mileage (mpg) reduces. Cars with manual transmission seems however to be more efficient: the group restricted regression (blue) has a higher intercept. It has to be highlighted however that the parameters of blue regression might be influenced by two extreme values with high hp - the regression should be re-estimated by removing the two datapoints.

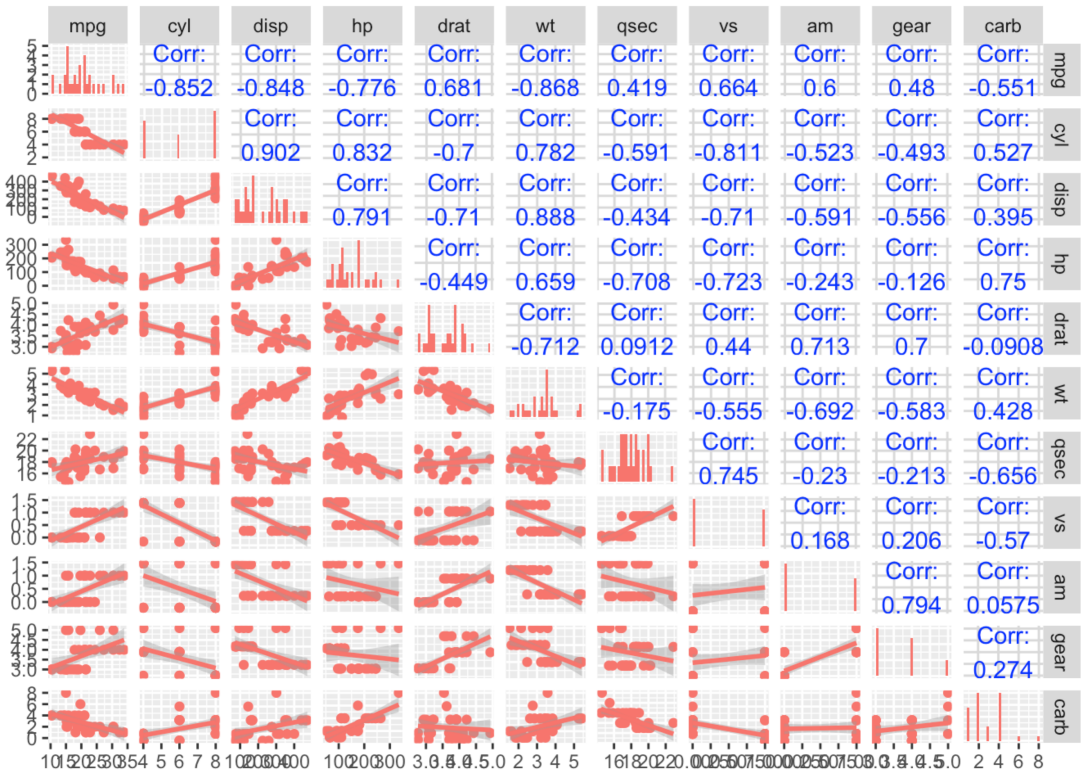
```
library(ggplot2)

library(stats)
ggplot(mtcars, aes(y=mpg, x=factor(am, labels = c("automatic", "manual
")), fill=factor(am)))+
        geom_violin(colour="black", size=1)+
        xlab("transmission") + ylab("MPG")
```



We can form a clear hypothesis from this visualization: it appears that automatic cars have a lower miles per gallon, and therefore a lower fuel efficiency, than manual cars do. But it is possible that this apparent pattern happened by random chance- that is, that we just happened to pick a group of automatic cars with low efficiency and a group of manual cars with higher efficiency. So to check whether that's the case, we have to use a statistical test.

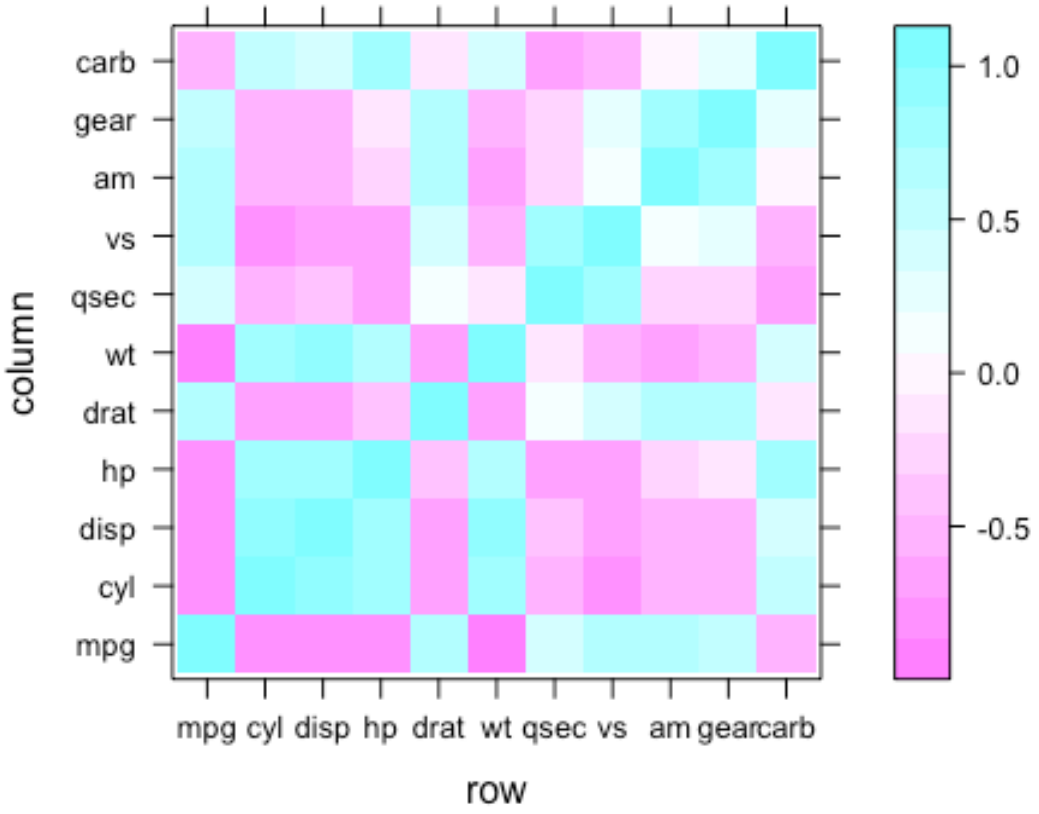
Multivariate analysis

```
library(GGally)
library(ggplot2)


ggpairs(mtcars,
        lower = list(continuous = "smooth"),
        diag=list(continuous="bar"),
        upper=list(continuous = "cor"),
        axisLabels='show',
        ggplot2::aes(colour="blue"))
```



Analysis of covariance matrix:

```
### analyse covariance matrix for regressor selection:
z <- cor(mtcars)
require(lattice)

## Loading required package: lattice

## Loading required package: lattice
levelplot(z)
```



It is also worthwhile check how MPG varies by automatic versus manual transmission. For that purpose we create a Violin plot of MPG by automatic and manual transmissions. In our dataset 0 represents an automatic transmission and 1 means a manual transmission.

Model fitting and hypothesis testing

Two samples t-test

We are interested to know if an automatic or manual transmission better for MPG. So first we test the hypothesis that cars with an automatic transmission use more fuel than cars with a manual transmission. To compare two samples to see if they have different means, we use two sample T-test.

```
test <- t.test(mpg ~ am, data= mtcars, var.equal = FALSE, paired=FALSE
,conf.level = .95)
result <- data.frame( "t-statistic"  = test$statistic,
                      "df" = test$parameter,
                       "p-value"  = test$p.value,
                       "lower CL" = test$conf.int[1],
                       "upper CL" = test$conf.int[2],
                       "automatic mean" = test$estimate[1],
                       "manual mean" = test$estimate[2],
                       row.names = "")
kable(x = round(result,3),align = 'c')
```

t.statistic	df	p.value	lower.CL	upper.CL	automatic.mean	manual.mean
-3.767	18.332	0.001	-11.28	-3.21	17.147	24.392

```
# t.statistic   df  p.value lower.CL    upper.CL    automatic.mean  manual.mean
# -3.767    18.332  0.001   -11.28  -3.21   17.147  24.392
```

p-value that shows the probability that this apparent difference between the two groups could appear by chance is very low. The confidence interval also describes how much lower the miles per gallon is in manual cars than it is in automatic cars. We can be confident that the true difference is between 3.2 and 11.3.

Simple linear regression model

We can also fit factor variables as regressors and come up with thing like analysis of variance as a special case of linear regression models. From the "dummy variables" point of view, there's nothing special about analysis of variance (ANOVA). It's just linear regression in the special case that all predictor variables are categorical. Our factor variable in this case is Transmission (am).

```
mtcars$amfactor <- factor(mtcars$am, labels = c("automatic", "manual")
)
summary(lm(mpg ~ factor(amfactor), data = mtcars))$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
factor(amfactor)manual	7.244939	1.764422	4.106127	0.000285

```
# Estimate  Std. Error  t value Pr(>|t|)
# (Intercept)   17.147368   1.124602    15.247492   0.000000
# factor(amfactor)manual    7.244939    1.764422    4.106127    0.0002
85
```

CONCLUSION

All the estimates provided here are in comparison with automatic transmission. The intercept of 17.14 is simply the mean MPG of automatic transmission. The slope of 7.24 is the change in the mean between manual transmission and automatic transmission. You can verify that from the plot as well. The p-value of 0.000285 for the mean MPG difference between manual and automatic transmission is significant. Therefore we conclude that according to this model manual transmission if more fuel efficient.

Source: https://rpubs.com/davoodastaraky/mtRegression