Course Project - Regression Models

This document is created for "Course Project" assignment in the framework of Regression Models course (part of Data Science Specialization by Johns Hopkins University Bloomberg School of Public Health on Coursera).

The Challenge

Working 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. 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?

Executive Summary

To be detailed...

Data Overview

Data to be analysed for this exercise reside as a built-in object, called `mtcars`, comes along with any R distribution. Let's take a first look at it:

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

From R documentation (`?mtcars`) the following information can be obtained about the dataset:

Variable	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)

Variable	Description
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Exploratory Data Analyses

Let's take a look at the outcome variable (miles per gallon). Observed mean is **20.09** with **6.03** standard deviation; for automatic cars the mean is *17.15* (standard deviation = *3.83*), for manual it is *24.39* (standard deviation = *6.17*). For distribution graphs please refer to *Figure 1* in the appendix.

To explore relationship patterns between variables and Miles per Gallon measure I plotted all (normalized) variables against the outcome (Figure 2).

Model fit / model selection strategy

This part is about selecting the most relevant variables that explain miles per gallon measure most precisely. In order to find the best fitting linear regression model I have revised outcomes of some model selection algorythm.

Stepwise variable selection

The first, most obvious try for variable selection is the stepwise algorythm, provided by `step()` function. The following model has been selected in a *backward* stepwise process as the best fit, that incorporates the following variables: (Intercept), wt, qsec, am with the corresponding coefficients of 9.6177805, -3.9165037, 1.225886, 2.9358372.

Regression subset selection including exhaustive search

To verify the selected variables, I performed a different method for variable selection provided by `leaps` R library. (Figure 3)

To be detailed...

Appendix

Figure 1: Distribution of Miles Per Gallon

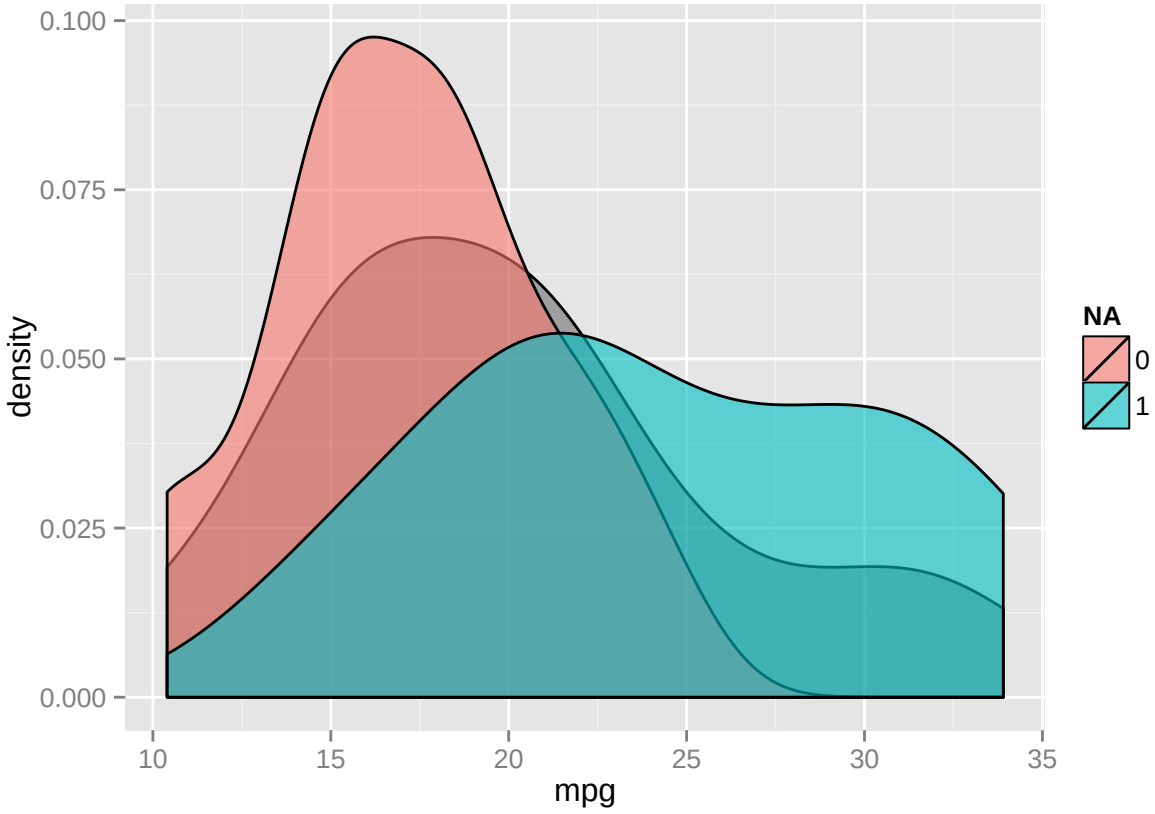


Figure 2: Outcome - Variable Regression Matrix

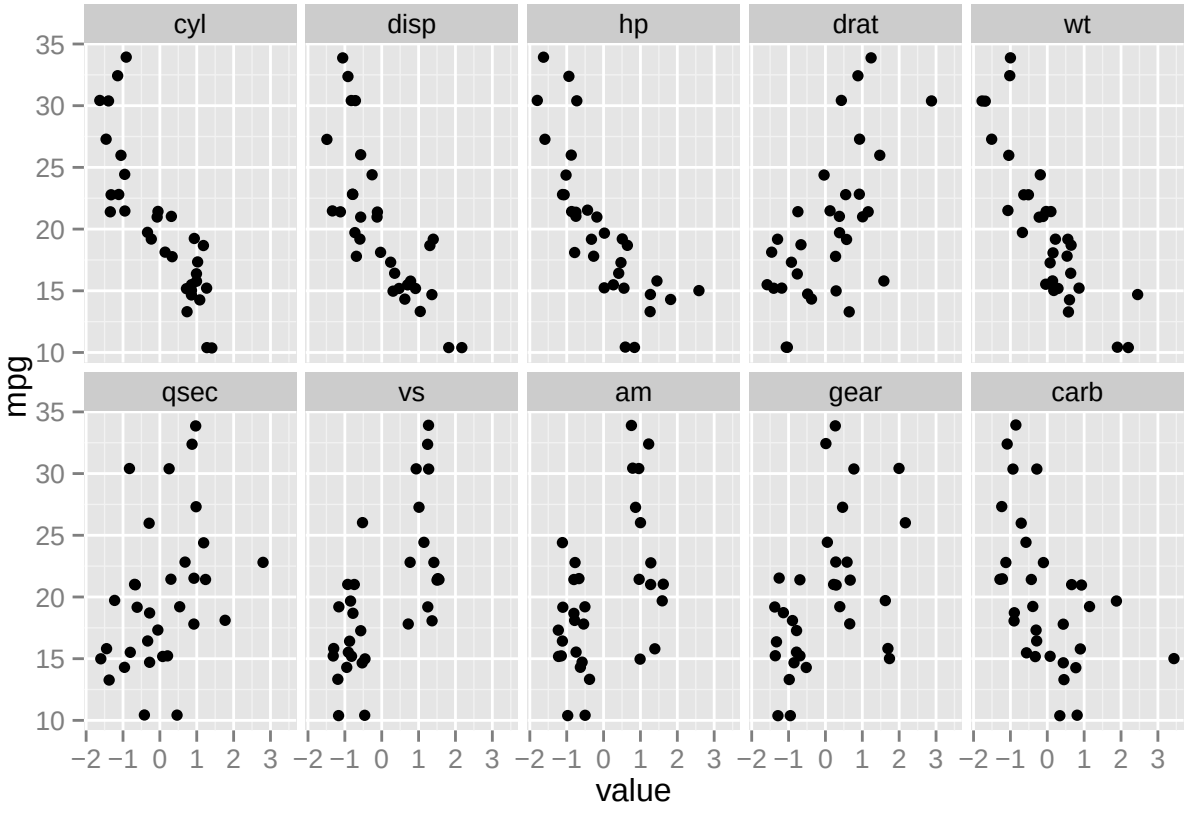
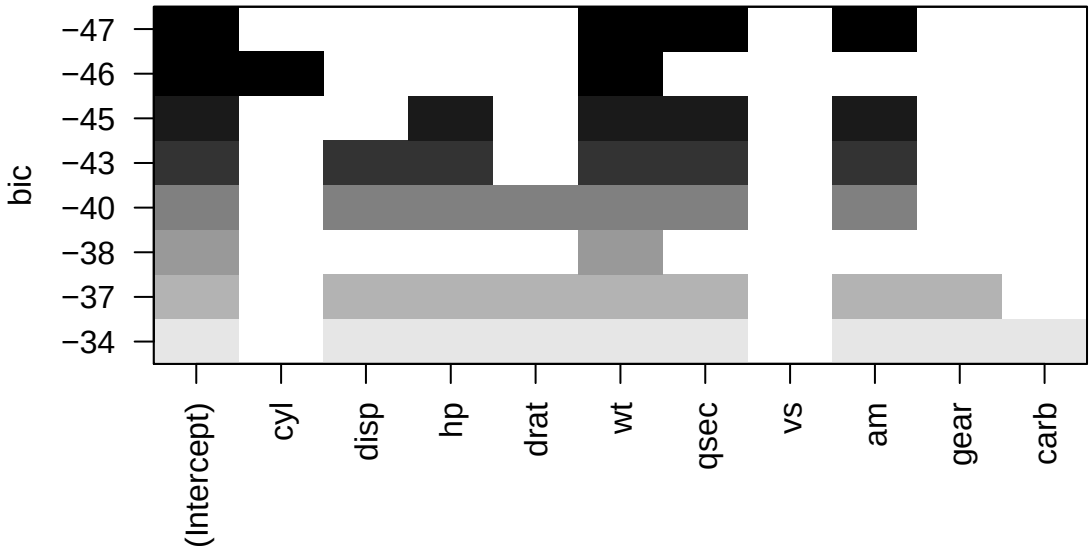


Figure 3: Regression subset selection including exhaustive search



Pointing criteria

Todo:

- interpret the coefficients
- do some exploratory data analyses
- fit multiple models and detail their strategy for model selection
- answer the questions of interest or detail why the question(s) is (are) not answerable
- do a residual plot and some diagnostics
- quantify the uncertainty in their conclusions and/or perform an inference correctly
- brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures

Done:

- include an executive summary
- done in Rmd (knitr)

Obsolete content

```
table(mtcars$am)
am_means <- data.frame(value=rbind(totalmean=mean(mtcars$mpg), automean=mean(mtcars[mtcars$am==1,]$mpg)
ggplot(mtcars, aes(y=mpg, x=factor(am))) + geom_boxplot()
mtcars_melted <- melt(mtcars, id="mpg")

null <- lm(mpg ~ 1, mtcars)

step(null, scope=list(lower=null, upper=full), direction="forward")
step(full, data=mtcars, direction="backward")
step(null, scope = list(upper=full), data=mtcars, direction="both")
```