Regression Models - Course Project

Pier Lorenzo Paracchini

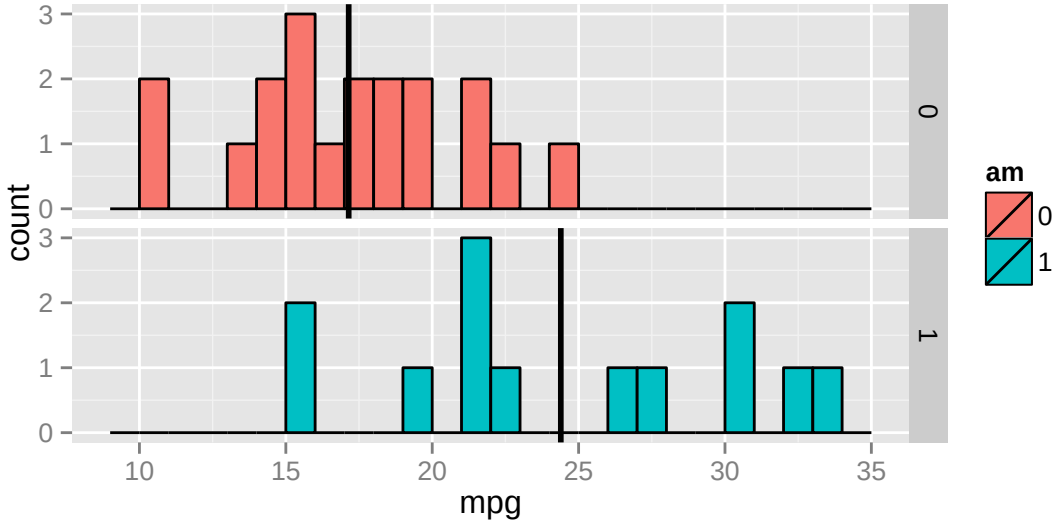
Executive Summary

Data Analysis

Exploratory Data Analysis

The dataset include 32 observations of 11 features. Each onservations comprises information for a specific automobiles (1973 - 1974 model). More information about the features can be found running `help(mtcars)`.

Focus is on two specific features **mpg** (miles per US gallon) and **am** (type of transmission - **0: automatic**, **1: manual**). The available datase contains 19 automatic car models and 13 manual car models. The observations of `mpg` by type of transmission `am` can be seen in the hystograms below (**note!** black line represents the sample mean for the group).

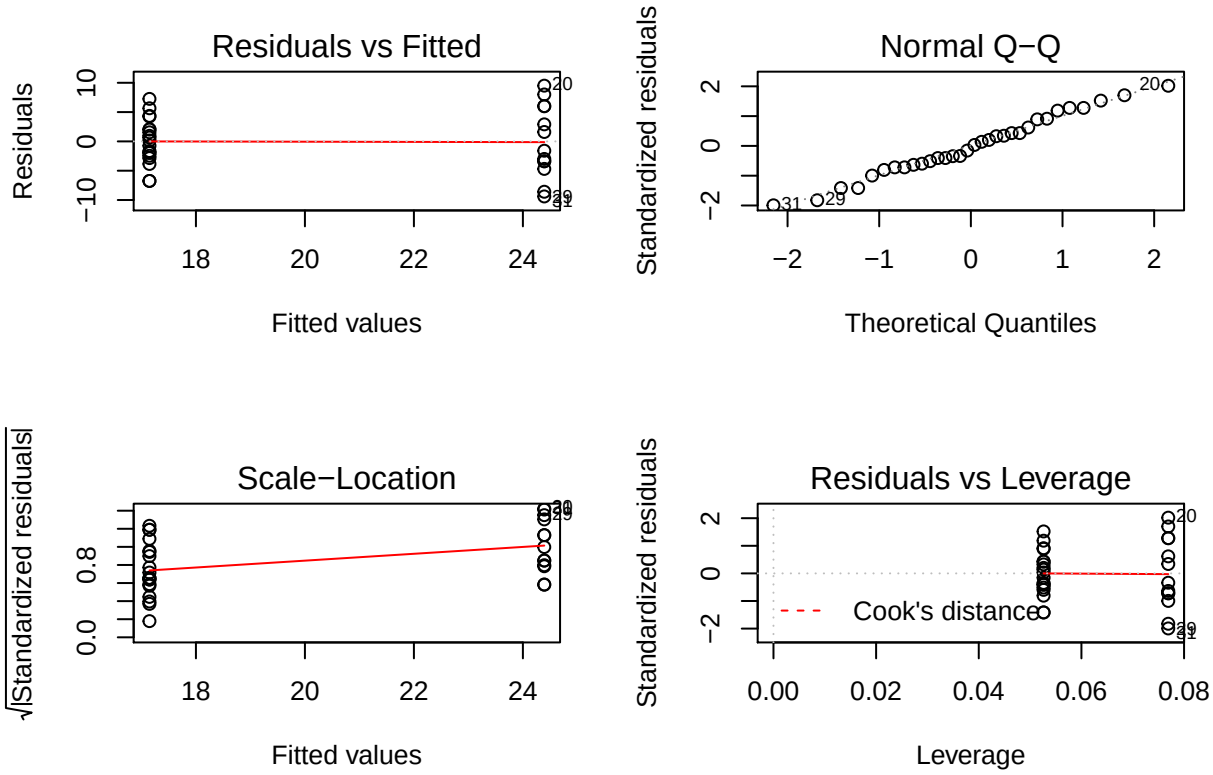


The `am` feature (predictor) is visibly related to the `mpg` feature (outcome) as we can see from the sample mean of each group. The "automatic" group has a lower sample mean (17.147 Miles/ Gallon) than the "manual" group (24.392 Miles/ Gallon).

The Regression Model

Let's fit a **linear model** using `mpg` (as outcome) with `am` as predictor with errors $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The coefficients of the models are

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```



From the **Residual vs Fitted** plot we can see that the model is just able to predict two possible values for the estimated `mpg` based on the predictors - **17.147 for automatic models** and **24.392 for manual models** with an estimated residuals variation of **4.902**.

Is an automatic or manual transmission better for MPG?

Looking at the coefficients

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

- **automatic** models (reference group) have an estimated mpg of 17.147 (miles per gallon) with a standard error of 1.125 (miles per gallon).
- **manual** models have an increased estimated mpg of 7.245 (miles per gallon) (over the reference group) with a standard error of 1.764 (miles per gallon). The P-value of 0 (statistically significant) confirms that the **null hypothersis** (having an increase/ decrease over the reference group null) can be rejected.

According to this very simple linear model, **automatic** transmission is better for **mpg** than **manual** transmission.

Quantify the MPG difference between automatic and manual transmissions

Based on the linear model previosly created we can state that

- **automatic** models use an estimated mpg of 17.147 (miles per gallon) with a 95% confidence interval included in [14.851 - 19.444] miles per gallon.
- **manual** models use an increased estimated mpg of 7.245 (miles per gallon) over the reference group with a 95% confidence interval included in [3.642 - 10.848] miles per gallon.

Description (ToBeDeleted)

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"

Constraints Report must be:

- Written as a PDF printout of a compiled (using knitr) R markdown document.
- Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures.
- Include a first paragraph executive summary.

Appendix

Code chunks and plots used in the previous sections.
##Dataset Structure

```
str(mtcars)
```

```
## '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 ...
```

Some more details about `mpg` and `am` features

Some `mpg` relevat statistics.

```
summary(mtcars$mpg)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

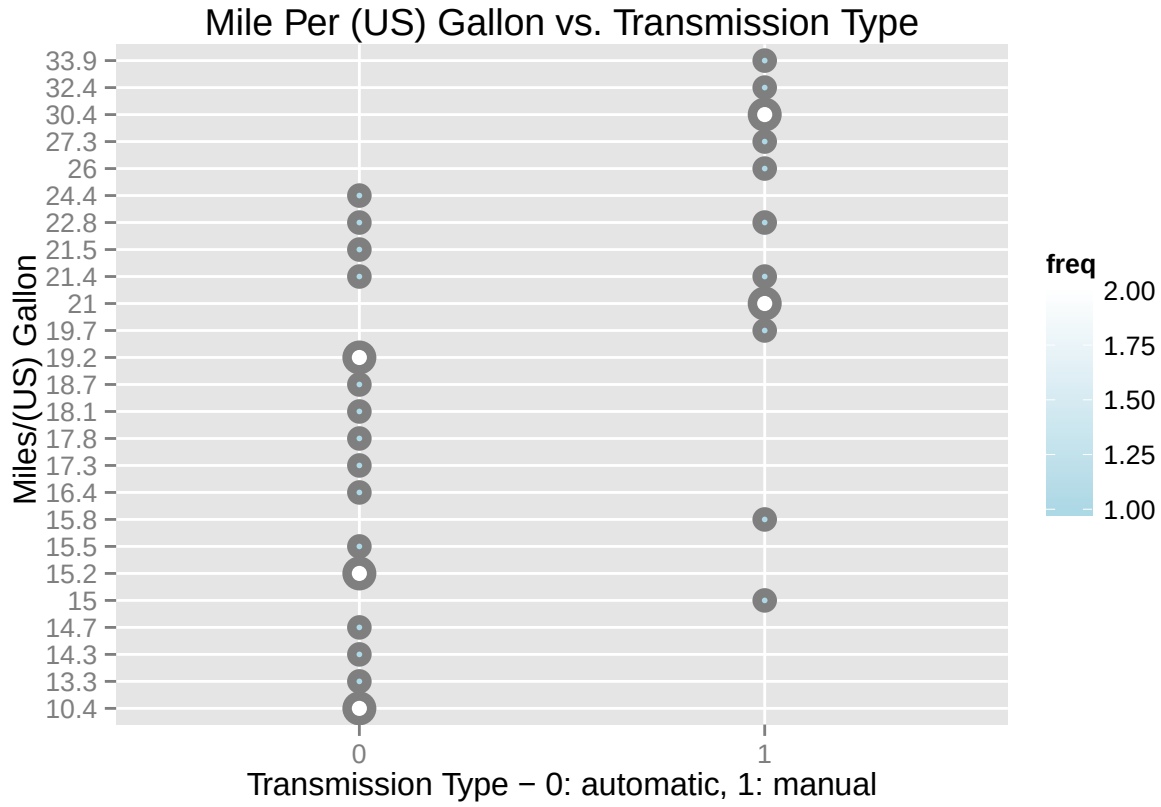
Number of car models by type of transmission.

```
table(mtcars$am)
##
##  0  1
## 19 13
```

Plot showing the number of car models by type of transmission and mpg.

```
require(dplyr)
require(ggplot2)
freqData <- as.data.frame(table(mtcars$mpg, mtcars$am))
names(freqData) <- c("mpg", "am", "freq")

g <- ggplot(filter(freqData, freq > 0), aes(x = am, y = mpg))
g <- g  + scale_size(guide = "none" )
g <- g + geom_point(colour="grey50", aes(size = freq+2, show_guide = TRUE))
g <- g + geom_point(aes(colour=freq, size = freq))
g <- g + scale_colour_gradient(low = "lightblue", high="white")
g <- g + ggtitle("Mile Per (US) Gallon vs. Transmission Type")
g <- g + xlab("Transmission Type - 0: automatic, 1: manual")
g <- g + ylab(" Miles/(US) Gallon")
g
```



Histograms for `mpg` by `am` (transmission type)

```
require(ggplot2)
require(dplyr)

mpg_avg_0 <- mean(filter(mtcars, am == 0)$mpg)
mpg_avg_1 <- mean(filter(mtcars, am == 1)$mpg)


#Other possibility
data_0 <- data.frame(mpg = mtcars$mpg, am = as.factor(mtcars$am))
data_1 <- data.frame(am = c("0", "1"), mean = c(mpg_avg_0, mpg_avg_1))

g <- ggplot(data_0, aes(x = mpg, fill = am))
g <- g + geom_histogram(colour = "black", binwidth=1)
g <- g + facet_grid(am ~ .)
g <- g + geom_vline(data = data_1, aes(xintercept = mean), lwd = 1)
g
```