

Regression Models: Cars And MPG

DA

May 27, 2016

Executive Summary

Scenario: 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

This brief analysis considers cars fuel economy and a set of variables, in particular the car transmission (automatic versus manual). We use mtcars dataset made available by Motor Trend Magazine.

Technically, we evaluate this relationship using linear regression. ——— Based on the dataset we show that the manual transmission has an advantage over the automatic one when it comes to MPG (miles per gallon) - see Figure 1.

The Data - performing a bit of exploratory data analysis

Loading the mtcars dataset, data frame with 32 observations on 11 variables (source: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html>).

```
[, 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 (1000 lbs)
[, 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
```

Loading the libraries and supressing the feedback messages.

```
library(ggplot2)
library(UsingR)
data(mtcars)
```

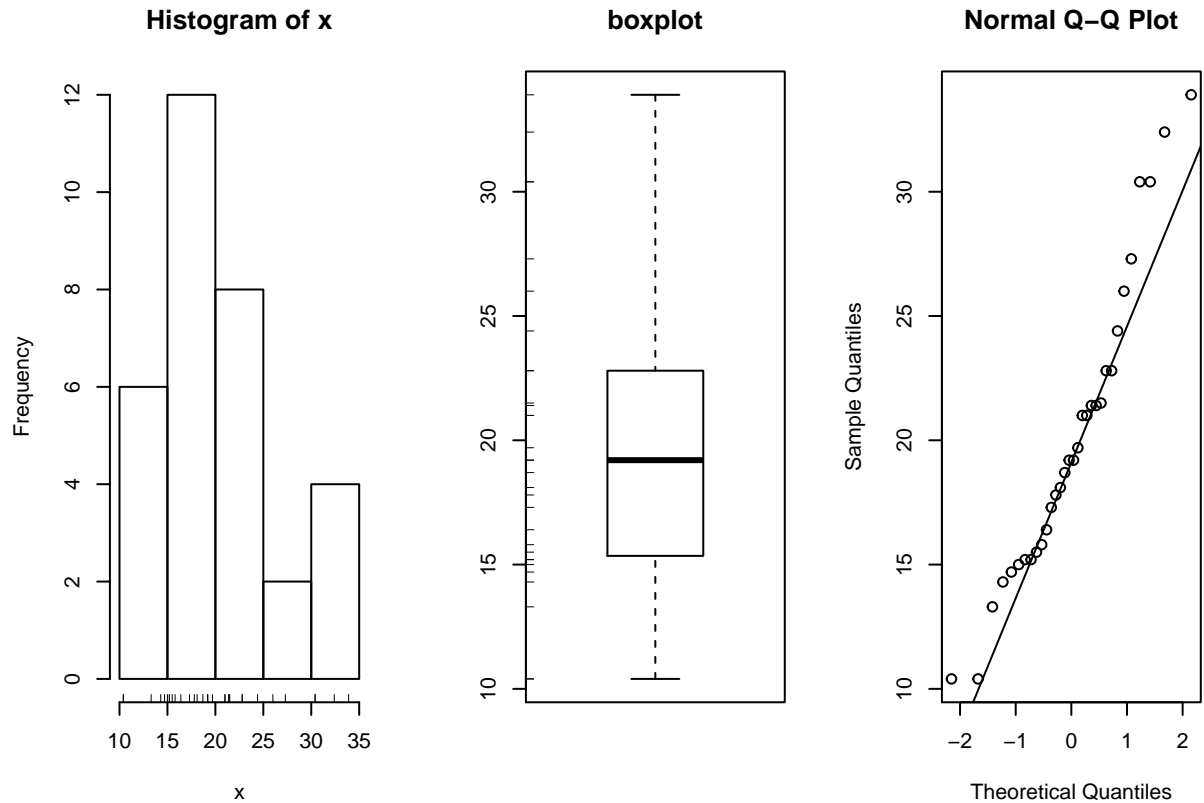
Looking a bit into how the mpg variable is distributed.

```
head(mtcars)
```

```
##           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
```

```
simple.eda(mtcars$mpg)
```



Shapiro-Wilk Test - checking on population

Using Shapiro-Wilk Normality Test and getting a p-value = 0.1229 which translates in the fact that we cannot reject the NULL hypothesis that the samples came from a normal distribution.

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.9476, p-value = 0.1229
```

T-test - comparing the means of MPG and type of transmission

We get a very small p-value which translates in the fact that the difference in means is not by chance but indeed there is clear difference in these two populations. So we reject the NULL hypothesis: “these two samples have the same means.”

```
t.test(mtcars$mpg, mtcars$am, var.equal=TRUE, paired=FALSE)
```

```
##
## Two Sample t-test
##
## data: mtcars$mpg and mtcars$am
## t = 18.4126, df = 62, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 17.54734 21.82141
## sample estimates:
## mean of x mean of y
## 20.09062 0.40625
```

Regression Models

Building a regression model using Backward Stepwise Regression (starting with all predictors and removes the ones that are not statistically significant).

```
initial.model <- lm(mpg ~., data= mtcars)
best.model <- step(initial.model, direction = "both")
```

Sumarizing the best regression model

The most significant variables are idenfied as being:

- the weight of the car (wt)
- the 1/4 mile time (qsec) The trasnmission seems to be of a lower importance as predictor of MPG.

We see a really small p-value (1.21e-11 much smaller than 0.05) meaning a good model. We also see an decent R-squared: 0.8497 which tranlates to outcome variance explained by this model.

```
summary(best.model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt          -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec         1.2259     0.2887   4.247 0.000216 ***
## am           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Appendix

Figure 1: Which transmission achieves better mileage (Miles per Gallon)?

As we can see from the boxplot below there is a distinct difference that favors the manual transmission over the automatic when it comes to get better millage.