Regression Models - Course Project

Nikolas Rohrmann

7/27/2021

Welcome

In the following steps I am going to introduce you to the mtcars data and evaluate the impact of automatic and manual transmission on mpg.

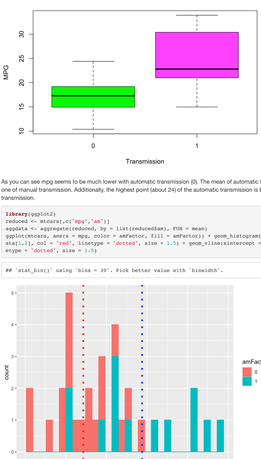
```
data("mtcars")
head(mtcars)
```

```
##                   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
```

The relevant variables for us are mpg, which is miles/(US) gallon and the variable am, which represents the transmission type. 0 is used for automatic and 1 for manual.

Exploratory Data Analysis

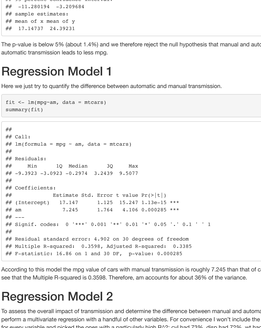
```
mtcars$amFactor <- as.factor(mtcars$am)
boxplot(mpg~ amFactor, mtcars, xlab = "Transmission", ylab = "MPG", col = c("green", "magenta"))
```



As you can see mpg seems to be much lower with automatic transmission (0). The mean of automatic transmission is significantly lower than the one of manual transmission. Additionally, the highest point (about 24) of the automatic transmission is belvo the 75th percentile of the manual transmission.

```
library(ggplot2)
reduced <- mtcars[,c('mpg','am')]
aggdata <- aggregate(reduced, by = list(reduced$am), FUN = mean)
ggplot(mtcars, aes(x = mpg, colour = amFactor, fill = amFactor)) + geom_histogram() + geom_vline(xintercept = aggd
ata[1,2], col = "red", linetype = "dotted", size = 1.5) + geom_vline(xintercept = aggdata[2,2], col = "blue", lin
etype = "dotted", size = 1.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Here is a histogram of data (Remeber 0 = automatic and 1 = manual) The dotted lines are the respective means of mpg by am. This supports the image of the data we got from the first plot.

Hypothesis Testing

We concluded from the boxplot and the histogram above that manual transmission likely results in more mpg than automatic transmission. Let's verify this using a t-test.

```
automatic <- reduced[reduced$am == 0,]
manual <- reduced[reduced$am == 1,]
t.test(automatic$mpg, manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic$mpg and manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The p-value is below 5% (about 1.4%) and we therefore reject the null hypothesis that manual and automatic transmission are equal. Thus, automatic transmission leads to less mpg.

Regression Model 1

Here we just try to quantify the difference between automatic and manual transmission.

```
fit <- lm(mpg~am, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

According to this model the mpg value of cars with manual transmission is roughly 7.245 than that of cars with automatic transmission. Also, we see that the Multiple R-squared is 0.3598. Therefore, am accounts for about 36% of the variance.

Regression Model 2

To assess the overall impact of transmission and determine the difference between manual and automatic transmission more accurately, I will perform a multivariate regression with a handful of other variables. For convenience I won't include the computation here, but I did the same thing for every variable and picked the ones with a particularly high R^2: cyl had 73%, disp had 72%, wt had 75% and hp had 60%. However, the number of cylinders (cyl) and displacement (disp) are correlated. Therefore, I decided to model with cyl, wt (weight) and hp (gross horsepower).

```
fit2 <- lm(mpg~ wt+ cyl+ hp + factor(am), data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp + factor(am), data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## wt          -2.60648    0.91984  -2.834  0.0084 **
## cyl         -0.74516    0.58279  -1.279  0.2119
## hp          -0.02495    0.01365  -1.828  0.0706 .
## factor(am)1  1.47805    1.44115   1.026  0.3142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

As you can see, this model accounts for roughly 85% of the variance. All factors negatively correlated with mpg, except for what is called factor(am)1. Due to the fact that 1 means manual, we can infer that having a manual transmission will lead to a 1.48 increase in mpg.

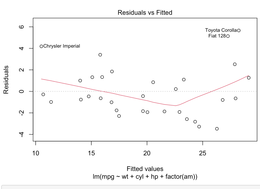
```
fit3 <- lm(mpg~ wt+ cyl+ hp, data = mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.9290 -1.5598 -0.5311  1.1850  5.8986
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
## wt          -3.16697    0.74058  -4.276 0.000199 ***
## cyl         -0.94162    0.55092  -1.709 0.098480 .
## hp          -0.01804    0.01188  -1.519 0.140015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.512 on 28 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263
## F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

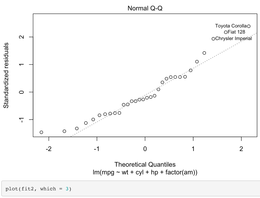
Still, note that without incorporating am, the model still accounts for 84% of the variance. This suggests that the impact of transmission is not that significant.

Appendix

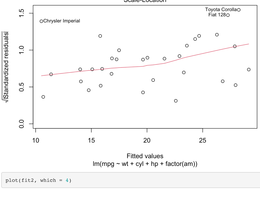
```
plot(fit2, which = 1)
```



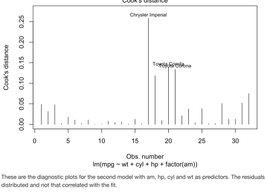

```
plot(fit2, which = 2)
```




```
plot(fit2, which = 3)
```




```
plot(fit2, which = 4)
```



These are the diagnostic plots for the second model with am, hp, cyl and wt as predictors. The residuals seem to be approximately identically distributed and not that correlated with the fit.