Regression Models Project

Antoine Mertz

2018-02-07

Synopsis

Motor Trend, a magazine about the automobile industry is interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Looking at the `mtcars` dataset we will try to answer these two questions. For more information on this dataset, we can have a look on the help and have a description typing `?mtcars`.

Data Exploration

```r
library(dplyr, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
data("mtcars")
```

First we need to transform a bit the data, changing the type of some variables:

```r
mtcars <- mtcars %>%
  mutate(vs = factor(vs)) %>%
  mutate(gear = factor(gear)) %>%
  mutate(carb = factor(carb)) %>%
  mutate(am = factor(am, labels = c("auto", "manual")))
```

A brief summary of the data can be found in appendix using `str`. Regarding effect of transmission type on MPG, we can first make a boxplot (see the figure in appendix) .

The boxplot seems to show that manual car provides more MPG than automatic ones.

Student test

To test this sentence, a Student test can be performed. To do that, we assume that density is normal in each groups (with same standard deviation), that is not really true regarding the plot in Appendix.

```r
t_test <- t.test(mpg ~ am, data=mtcars)
t_test$p.value
```

```
## [1] 0.001373638
```

p-value is under 0.05 that rejects the null hypothesis: the transmission type has significant impact on MPG. To quantify this difference we can perform regression analyse.

Regression analyse

In the first model, all variables are included as predictors of MPG to see how it performs

```
complete_model <- lm(mpg ~ ., data=mtcars)
summary(complete_model)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 25.31994337 23.88164477  1.0602261 0.30478503
## cyl         -1.02343435  1.48131027 -0.6908980 0.49953134
## disp         0.04376554  0.03057568  1.4313841 0.17156359
## hp          -0.04881225  0.03189192 -1.5305523 0.14541042
## drat         1.82084238  2.38100971  0.7647354 0.45556110
## wt          -4.63539945  2.52736612 -1.8340831 0.08530813
## qsec         0.26966987  0.92631150  0.2911222 0.77469794
## vs1          1.04907556  2.70494812  0.3878357 0.70324874
## ammanual     0.96265239  3.19137777  0.3016416 0.76681049
## gear4        1.75359631  3.72533672  0.4707216 0.64419293
## gear5        1.87898502  3.65935137  0.5134749 0.61463655
## carb2       -0.93427482  2.30934499 -0.4045627 0.69115583
## carb3        3.42168886  4.25512809  0.8041330 0.43310616
## carb4       -0.99363962  3.84682616 -0.2583011 0.79946771
## carb6        1.94388997  5.76982873  0.3369060 0.74056649
## carb8        4.36998439  7.75434447  0.5635530 0.58087155
```

According to the results, a lot of variables are not significant so we use `step()` function to use AIC criteria forward and backward and with a fix `k` with the help of the `step()` documentation.

```
best_model <- step(complete_model, k=log(nrow(mtcars)))
```

```
summary(best_model)$call
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

```
summary(best_model)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## ammanual     2.935837  1.4109045  2.080819 4.671551e-02
```

```
print(paste("Adjusted R-Squared =", summary(best_model)$adj.r.squared))
```

```
## [1] "Adjusted R-Squared = 0.833556080257604"
```

The adjusted R-squared of this "best model" is 0.8336, so we tend to explain 83.36% of the MPG variance that is pretty good and all the variables selected are significant. We can rely on this model. And according to it, with `wt`, `qsec` staying constant, manual transmission cars get an average of 2.94 more MPG than with automatic ones. So we can answer the first question: automatic is better for MPG (we go further with automatic cars with the same quantity of gas that with manual cars). A residuals analysis is perform in Appendix section to see how the model perform and if hypothesis of linear regression are not violated.

To quantify this difference we can fit a model with only transmission variable.

```
quantify_model <- lm(mpg ~ am, data = mtcars)
summary(quantify_model)$coef
```

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

```
print(paste("Adjusted R-Squared =", summary(quantify_model)$adj.r.squared))
```

```
## [1] "Adjusted R-Squared = 0.338458908206314"
```

It shows that a automatic cars have an average of 17.147 mpg, and manual ones increase MPG by 7.245. But, this model has an adjusted R-squared of 0.3385, which means the model can only explain about 33.85% of the variance of the MPG, that is not really enough to be consider as a good model. So other variables are important like we demonstrated with the "best model".

Conclusion

With this study, we conclude that manual cars have better performance in term of MPG because you can go in average 7.245 miles further with the same number of gas gallon than with automative cars. We prove it using a student test that shows that there is a statical difference between auto and manual transmission, and then regression analyses confirm this difference and quantify it at 7.245 MPG. We confirm our first assumption from the boxplot after data exploration. So we answer the questions of interest for Motor Trend magazine.

Appendix

Exploratory analysis

```
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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

```
ggplot(mtcars) +
  geom_boxplot(aes(x=am, y=mpg, fill=am)) +
  ggtitle("Miles per gallon according to transmission type") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
ggplot(mtcars, aes(mpg)) +
  geom_density(aes(fill="both"), alpha=0.4) +
  geom_density(aes(fill=am), alpha=0.4) +
  ggtitle("MPG density according to automative type") +
  theme(plot.title = element_text(hjust = 0.5))
```

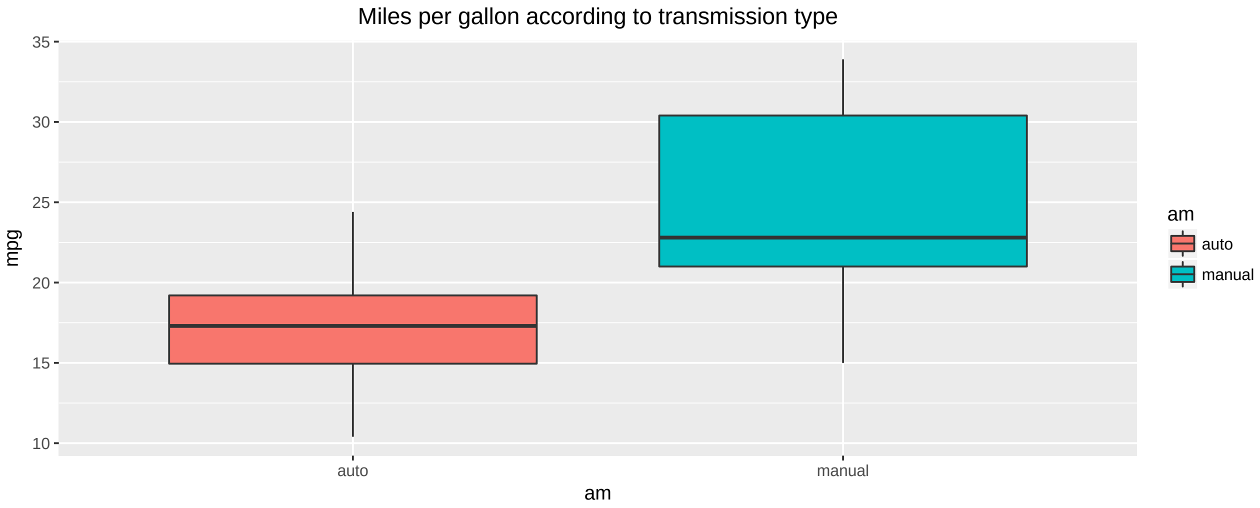

Figure 1: Boxplot MPG vs AM

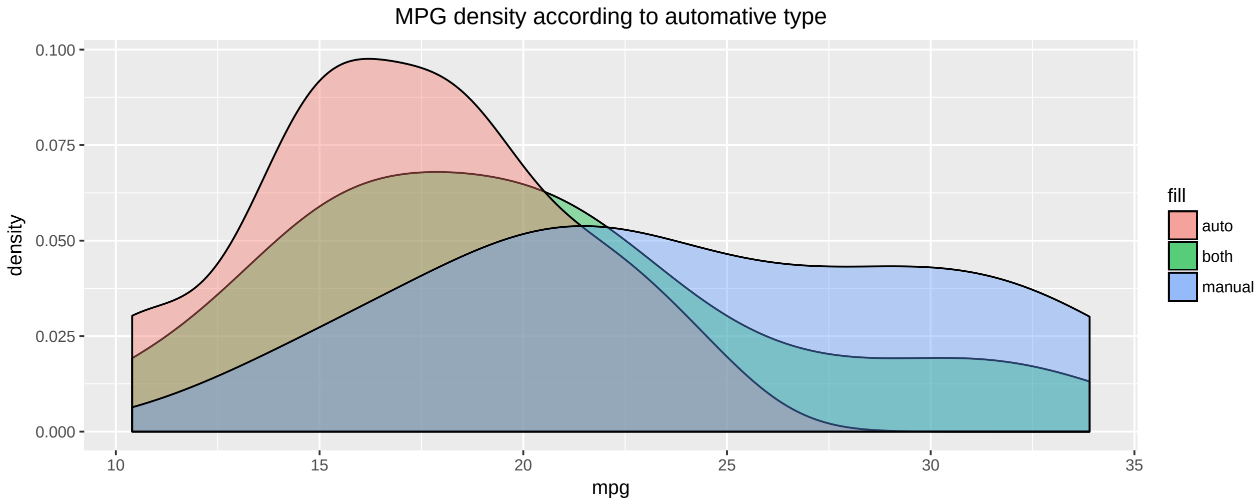

Figure 2: Density of MPG vs AM

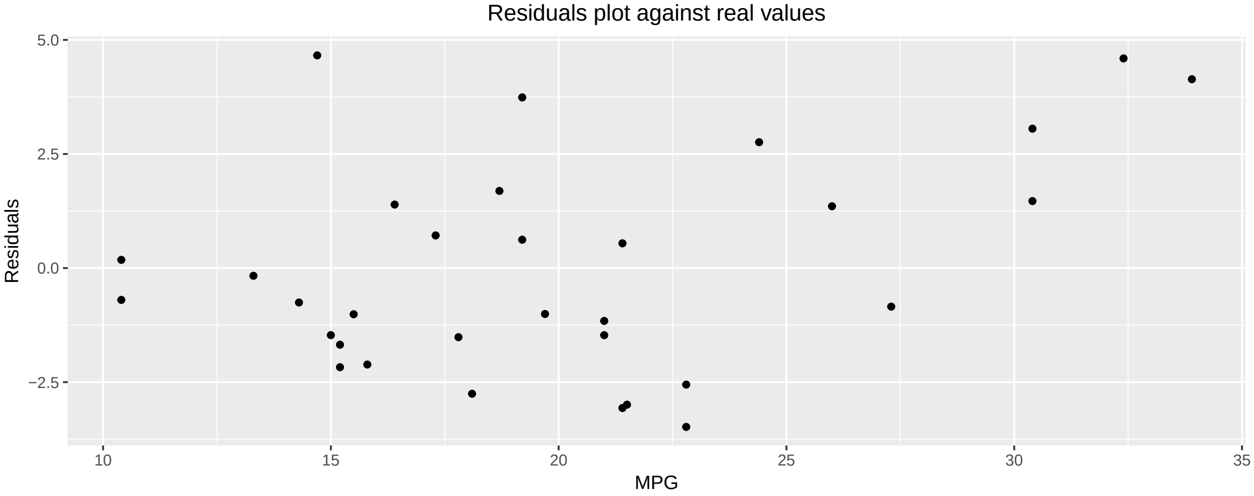

Figure 3: Residuals Analysis

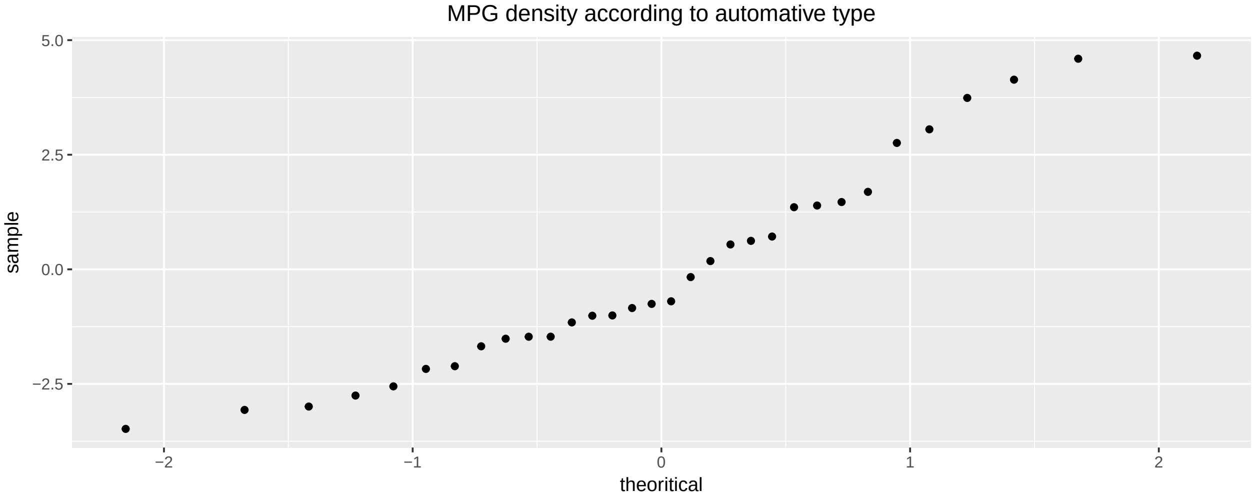

Figure 4: Q-Q plot

Residuals analysis

```
residuals <- data.frame(mpg = mtcars$mpg, error = resid(best_model))
ggplot(residuals, aes(x=mpg, y=error)) +
  geom_point() +
  xlab("MPG") +
  ylab("Residuals") +
  ggtitle("Residuals plot against real values") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
ggplot(residuals, aes(sample = error)) +
  stat_qq() +
  xlab("theoritical") +
  ylab("sample") +
  ggtitle("MPG density according to automative type") +
  theme(plot.title = element_text(hjust = 0.5))
```

Based on these two residuals plots, we can say that:

- The residuals doesn't show any pattern that seems to confirm that observations are independant.
- Q-Q plot shows that errors have likely a normal distribution so that the data fit to the model and linear regression assumptions are not broken.