Assignment Regression Model: Automatic or Manual gearbox?

Duy Nguyen

Introduction

This analysis will look at the mtcars dataset and answering two questions of interest:
(1) Is an automatic or manual transmission better for MPG
(2) Quantify the MPG difference between automatic and manual transmissions

Exploratory data analysis

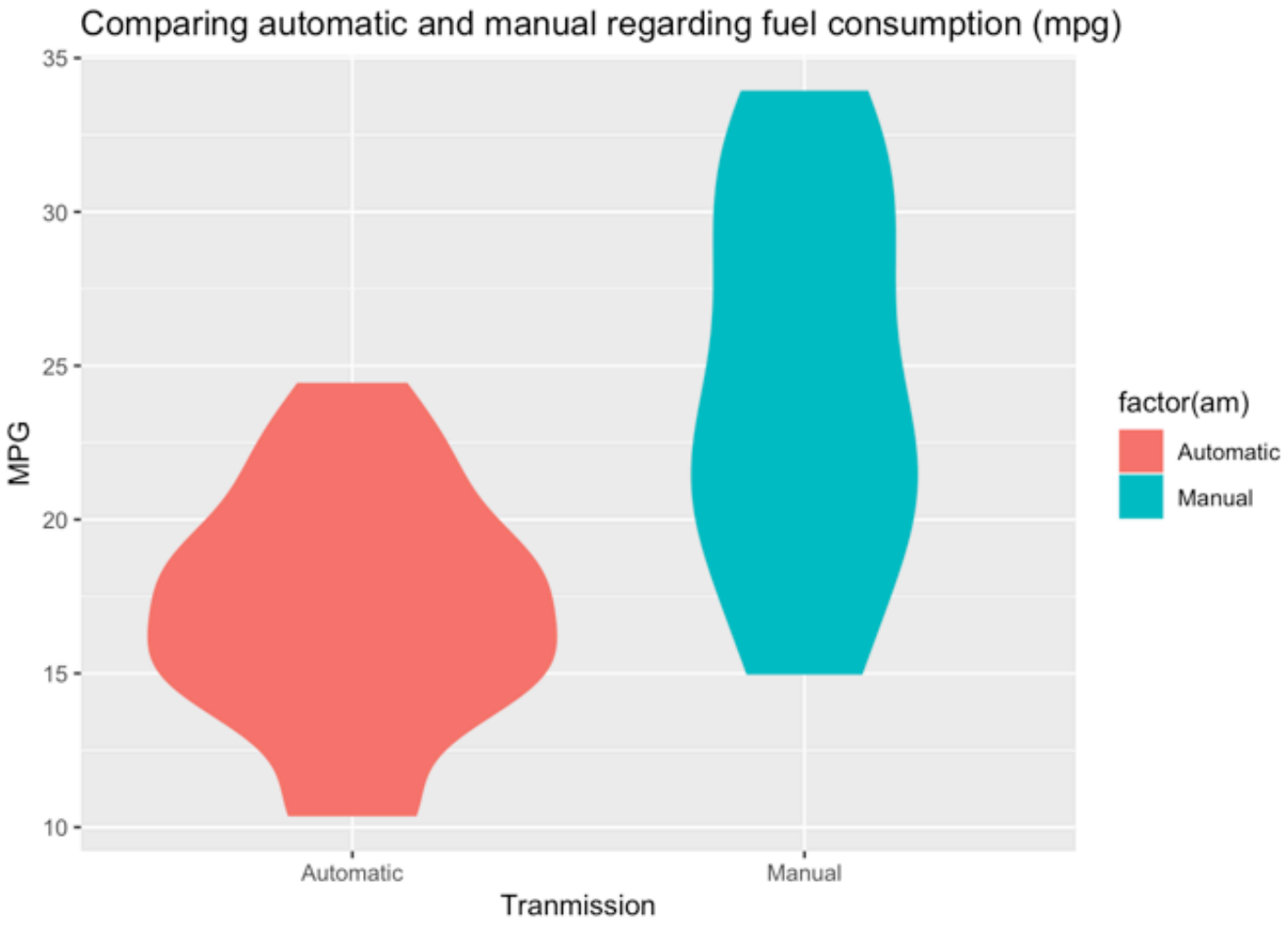
The correlations between MPG and other variables show that all variables have influence on the fuel consumptions. As expected, cyl, disp, hp, wt, and carb have a negative relationship with MPG. At the fist glance, tranmission (AM where automatic = 0, manual = 1) seems to be in a positve relationship with MPG. The violin plot comparing the types of tranmission regarding the fuel consumption show that, on average, the MANUAL cars consume more fuel than the AUTOMATIC cars. In fact, this is influenced by other variables and will be further addressed by linear regression model.

```
library(datasets)

data(mtcars)
kable_styling(kable(round(cor(mtcars$mpg,mtcars[,-1]),3),caption = "Correlation bet
ween MPG and other variables"))
```

Correlation between MPG and other variables

cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.6	0.48	-0.551

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) = c("Automatic","Manual")

ggplot(mtcars, aes(y=mpg, x = factor(am), fill = factor(am), color = factor(am))) +
geom_violin() + labs(x= "Tranmission", y = "MPG", title = "Comparing automatic and
manual regarding fuel consumption (mpg)")
```



Hypothesis testing

T test for automatic and manual gear shows that with the p-value less than 5% and the confidence interval do not contain 0, we would reject the null hypothesis. This mean that the manual seems to consume more fuel than the automatic.

```
t.test(mtcars$mpg~as.factor(mtcars$am))
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by as.factor(mtcars$am)
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Regression model

Use tranmission type as the only predictor

To answer the two primary questions, the analysis will, in the first attempt, model the MPG with only the type of tranmission as a linear predictor. The summary of the regression shows that with P-value < 0.05, we can reject the NULL hypothesis. This means that the type of tranmission has statiscal significant influence on MPG. The coefficient of MANUAL is +7.2 which indicates that MANUAL has greater postive effect on MPG. In other words, MANUAL cars consume more fuel.
The explained variance by this linear model, however, is just about 36% (indicated by the R squared). This motivates the need for further model selection considering adjusting the effects of other variables.

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)Manual    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
```

Model selection by adjusting the effect of other variables

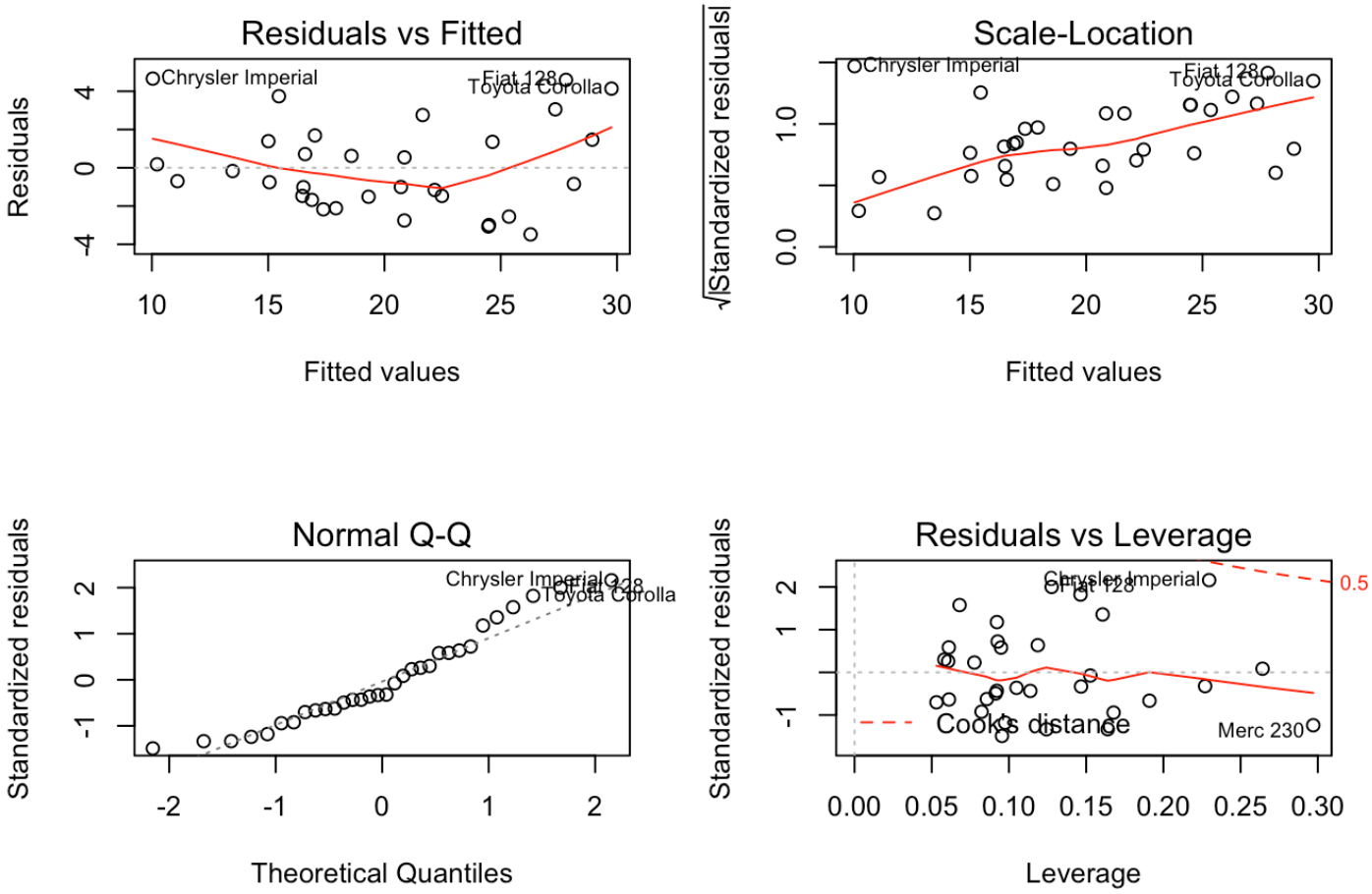
Model development could be done in different ways in R, this analysis will employ the function “step” to find the best fitted linear model.
As seen, the model with three variables **wt**, **qsec**, **am** can explain 85% percent of the variance in the dataset. The P-value are less than 5% so we would reject the NULL hypothesis. The residual plot shows that the residual do not have any pattern. The Q_Q plot shows good agreement between theoretical quantiles and the standardised residuals.

The analysis will choose this as the final model. However, it should be noted that the model could be further developed to address uncertainty by considering the interaction between variables.

```
#mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
fit2 <- step(lm(mpg~.,data = mtcars), trace = 0)
summary(fit2)
```

```
##
## 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 ***
## amManual      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
```

```
par(mfcol = c(2,2))
plot(fit2)
```



Quantifying the difference between MANUAL and AUTOMATIC gearbox

From the output of the linear model with three regressors **wt**, **qsec**, **am** , it is shown that the MANUAL consume about **3** mpg more than the AUTOMATIC cars

CONCLUSION

The MPG is unsuprisingly influenced by all variables. However, a linear model with three variables **wt**, **qsec**, **am** can explain 85% of the variance.
Answer to the two questions:
+ AUTOMATIC is better than MANUAL regarding fuel consumption (MPG)
+ AUTOMATIC consumes about 3 mpg less than MANUAL (results extracted from linear model)