JOHN´S HOPKINGS REGRESSION COURSE PROJECT

Samuel Bozzi Baco

0 EXECUTIVE SUMMARY

For the final project, the students from John´s Hopkins Regression classes were asked to analyze the **mtcars** dataset and answer two questions:

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

The regression will be done following a framework based on author´s experience: (1) response analysis and transformation, (2) x´s multicollinearity analysis, (3) partial regression plots, (4) model fitting and statistical tests, (5) goodness of fit and resitual analysis and (6) outlier and influential points.

1 RESPONSE ANALYSIS

The response considered for the regression is **mpg** (miles per gallon) for each car model. Although the response seems a little skewed, the normality test that can be seen at appendix A.1 shows no necessity to use a transformation.

2 X´S MULTICOLINEARITY

The X matrix presents (at appendix A.2) several variables with noticeable correlation, as the displacement (*disp*) and weight(*wt*) or power (*hp*) and rear axle ratio (*drat*). For the variable in question, type of transmission (*am*), there is a correlation with weight (*wt*) and rear axle ratio (*drat*). Because the type of transmission is a factor variable, it is not possible to construct a value-added plot (partial correlation) for it. So the model will be constructed iteratively, looking at VIF values for variables.

3 MODEL FITTING

Before fitting the model, a feature engineering will be done to scale the numeric variables and transform the variables from numeric to factor (but only the one that makes sense).

```
mtcars2 <- cbind(mtcars["mpg"],
                 as.data.frame(scale(mtcars[ ,2:7], center = T, scale = T)),
                 mtcars[8:11])

mtcars2 <- mtcars2 %>%
        mutate(vs = factor(vs, levels = c(0, 1), labels = c("V", "S")),
               am = factor(am, levels = c(0, 1), labels = c("A", "M")))
```

The first model that will be created will consider all possible variables.

```
m1 <- lm(mpg ~., data = mtcars2)

vif(m1)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

The variables that present the biggest VIF (variance inflation factor) are the displacement (*disp*), with a value of 21.62 and the cylinders (*cyl*)). So, they will be removed to construct the second model. The variable *qsec* will also be removed.

```
m2 <- lm(mpg ~. - disp - cyl -qsec, data = mtcars2)

vif(m2)
```

```
##       hp     drat       wt       vs       am     gear     carb
## 4.207016 3.109237 4.819161 2.530943 4.124605 4.686962 4.200759
```

After the removal of the variables cited below, the VIF for other variables has much better values. It is interesting to check if the models(*m1* and *m2*) are statistical different.

```
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ (cyl + disp + hp + drat + wt + qsec + vs + am + gear +
##     carb) - disp - cyl - qsec
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     21 147.49
## 2     24 158.32 -3   -10.828 0.5139 0.6772
```

The model with fewer factors is not statistically significant different if compared with the full model. So it will be maintened.

```
summary(m2)
```

```
##
## Call:
## lm(formula = mpg ~ . - disp - cyl - qsec, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9954 -1.5199 -0.4047  1.4574  5.1519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.9595     4.0110   4.478 0.000157 ***
## hp           -1.4721     0.9462  -1.556 0.132845
## drat          0.4977     0.8134   0.612 0.546415
## wt           -2.0318     1.0127  -2.006 0.056209 .
## vsS           1.3178     1.4561   0.905 0.374426
## amM           2.0460     1.8775   1.090 0.286644
```

```
## gear          0.7301     1.3536   0.539 0.594609
## carb         -0.7000     0.5854  -1.196 0.243432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 24 degrees of freedom
## Multiple R-squared:  0.8594, Adjusted R-squared:  0.8184
## F-statistic: 20.96 on 7 and 24 DF,  p-value: 8.87e-09
```

The intended variable, transmission, has a p-value of 0.28, presenting a weak statistical significance. Its estimate, using the *manual* level as a base is 2.04. This means that, in average, change to automatic would increase the mgp average of 2.04.

4 GOODNESS OF FIT

With the predictors selected, the model presents a R-square of 0,8184 and a residual standard error of 2.56 mpg.

The model residuals seem to follow the taken assumptions of NID with zero mean and constant variation (appendix A.3). Some points seem to be on the border of influential limits.

5 OULIER AND INFLUENTIAL

To quick analyze the possible outliers and influential points, a graphic with all information (studentized residuals, hats, and cooks distance will be constructed).

```
outl <- data.frame(rstudent(m2), hatvalues(m2), cooks.distance(m2))

names(outl) <- c("student", "hats", "CooksD")
```

There is no point with considerable high (>3) student residuals (appendix A.4). Although three points has a hat value bigger than suggested (2p/n = 2*7/32 = 0,4375), they are not influential, since all of them has Cooks Distance smaller than 1.

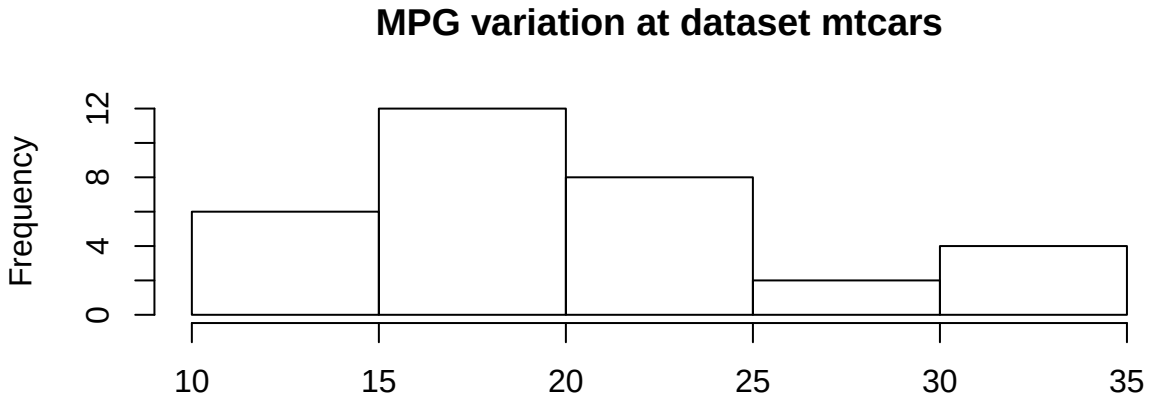
6 CONCLUSIONS

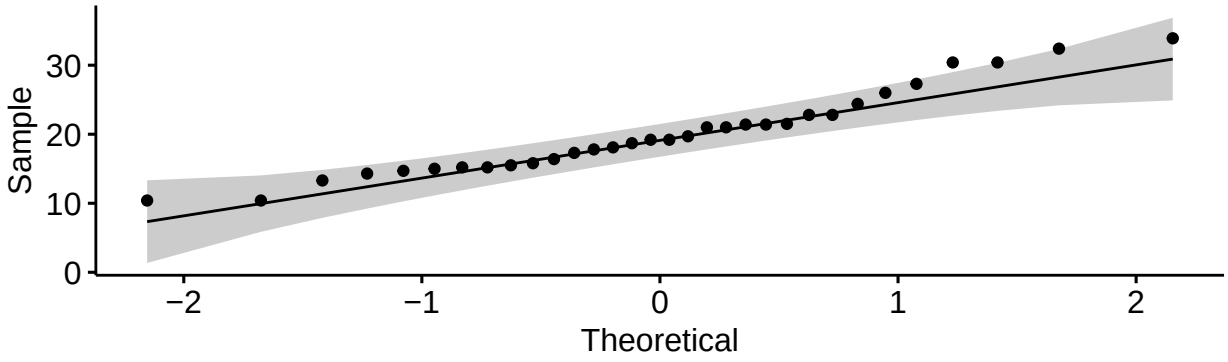
Concluding the study, related to the question (1), with the data presented it is difficult to say that automatic transmission is better (more economic) than manual since the p-value for this regression is bigger than 0.05. Even so, the parameter estimates to change between automatic and manual was 2.04 increase in mpg, in favor of automatic (question 02)

APPENDIX

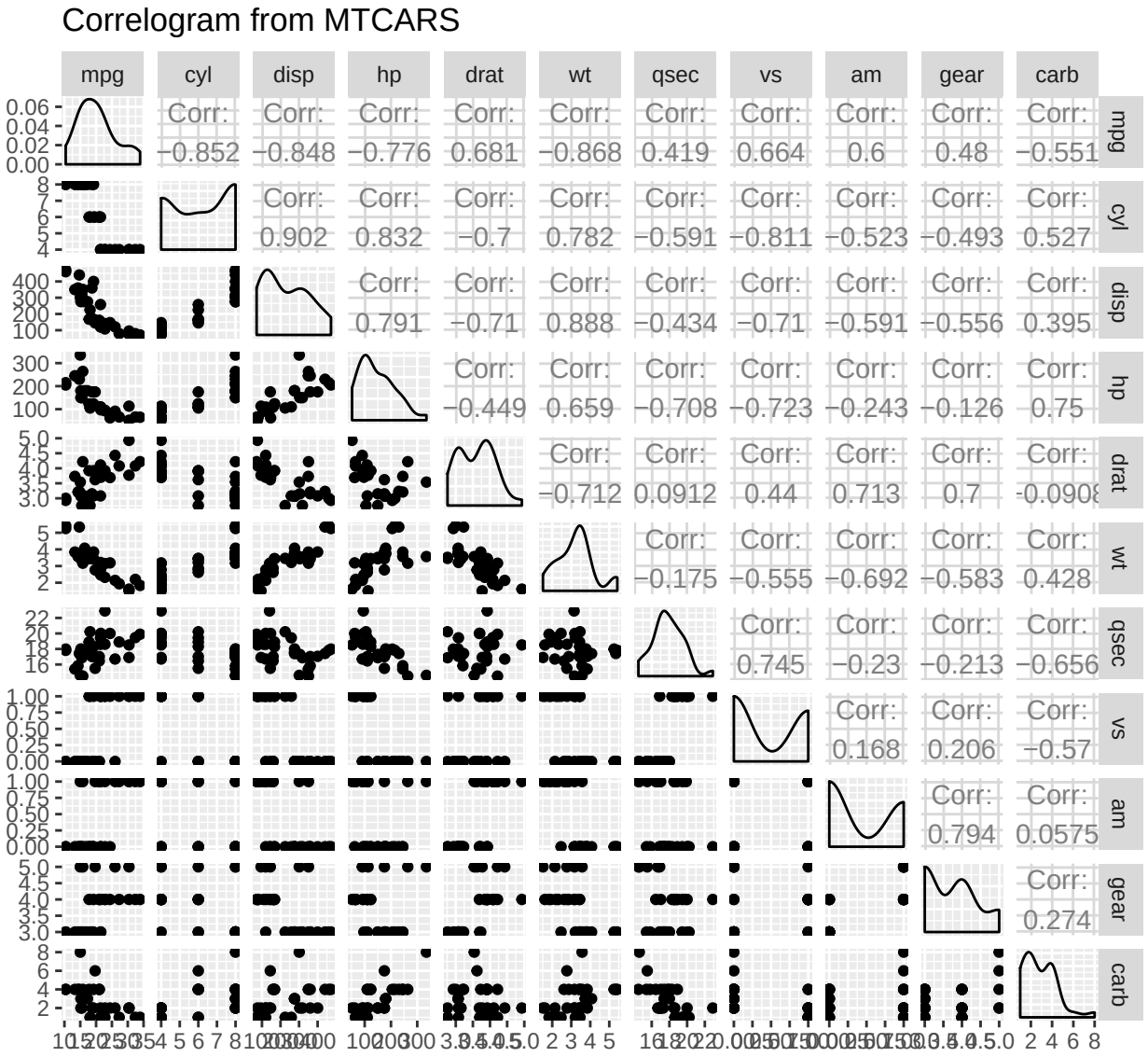
A1 Reponse normality check

```
with(mtcars, hist(mpg, xlab = "Miles per gallon", main = "MPG variation at dataset mtcars"))
```

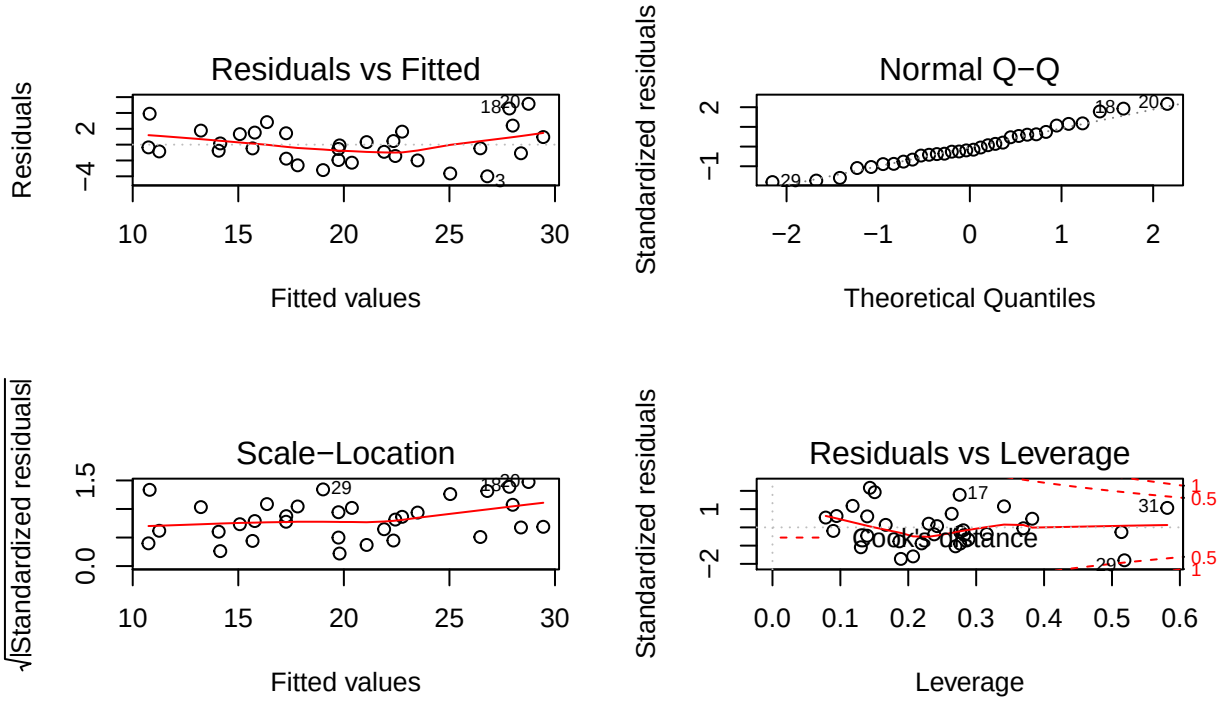




A.2 MULTICOLINEARITY



A.3 RESIDUAL ANALYSIS



A.4 OUTLIERS AND INFLUENTIALS