Assessment of the relationship between the type of the front of the car and fuel consumption

ps7391

7 January 2019

Executive Summary

Consider a dataset of a collection of cars, and interested in exploring the relationship between a set of variables and miles per gallon. In particularly the following two questions: 1. Is an automatic or manual transmission better for MPG 2. Quantify the MPG difference berween automatic and manual transmissions Perform the following sequence of actions: 1. Process the data, for use of this project 2. Explore the data, especially focussing on the two paramaters we are interested in transmission and MPG 3. Model selection, where we try different models to help answer our questions 4. Model examination, to see wether our best model holds up to our standards 5. Conclusion where we answer the questions based on the data

Processing

In the first place change 'am' to factor (0 = automatic, 1 = manual) and make cylinders a factor as well (since it is not continious)

```
library(ggplot2)
library(GGally)
library(dplyr)
library(ggfortify)
data("mtcars")
mtcars_factors <- mtcars
mtcars_factors$am <- as.factor(mtcars_factors$am)
levels(mtcars_factors$am) <- c("automatic", "manual")
mtcars_factors$cyl <- as.factor(mtcars_factors$cyl)
mtcars_factors$gear <- as.factor(mtcars_factors$gear)
mtcars_factors$vs <- as.factor(mtcars_factors$vs)
levels(mtcars_factors$vs) <- c("V", "S")
```

Exploratory data analyses

Look at the dimensions and head of the dataset:

```
dim(mtcars_factors)
```

```
## [1] 32 11
```

```
head(mtcars_factors, 3) #N2
```

```
##                mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  V manual    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  V manual    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  S manual    4    1
```

Find the relationship between the two parameters of interest.

```
g <- ggplot(mtcars_factors, aes(x = am, y = mpg, fill = am))
g + geom_boxplot() +
scale_fill_manual(name = "am", values = c("yellow", "green")) +
theme(plot.title = element_text(face = "bold", size = 12))+ theme_dark()
```

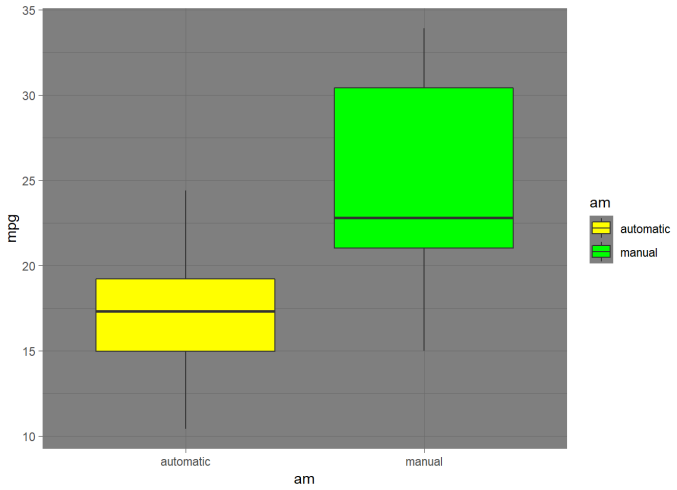

Figure 1. Relationship between MPG and automatic or manual geer

Even this shows clearly that the manual transmissions have higher MPG there could be a bias in the data that we are overlooking. Before creating a model we should look at which parameters to include besides ‘am’. We look at all correlations of parameters and take only those higher then the ‘am’ correlation.

```
cors <- cor(mtcars$mpg, mtcars) # correlation
ordered_cors <- cors[, order(-abs(cors[1,]))]
ordered_cors
```

```
##        mpg         wt        cyl       disp         hp       drat
##  1.0000000 -0.8676594 -0.8521620 -0.8475514 -0.7761684  0.6811719
##         vs         am       carb       gear       qsec
##  0.6640389  0.5998324 -0.5509251  0.4802848  0.4186840
```

```
am_pos <- which(names(ordered_cors)=="am") # N4
subset_columns <- names(ordered_cors)[1:am_pos]
subset_columns
```

```
## [1] "mpg"  "wt"   "cyl"  "disp" "hp"   "drat" "vs"   "am"
```

```
mtcars_factors[, subset_columns] %>%
        ggpairs(
                mapping = ggplot2::aes(color = am),
                upper = list(continuous = wrap("cor", size = 3)),
                lower = list(continuous = wrap("smooth", alpha=0.4, size=1),
                             combo = wrap("dot"))) + theme_dark()
```

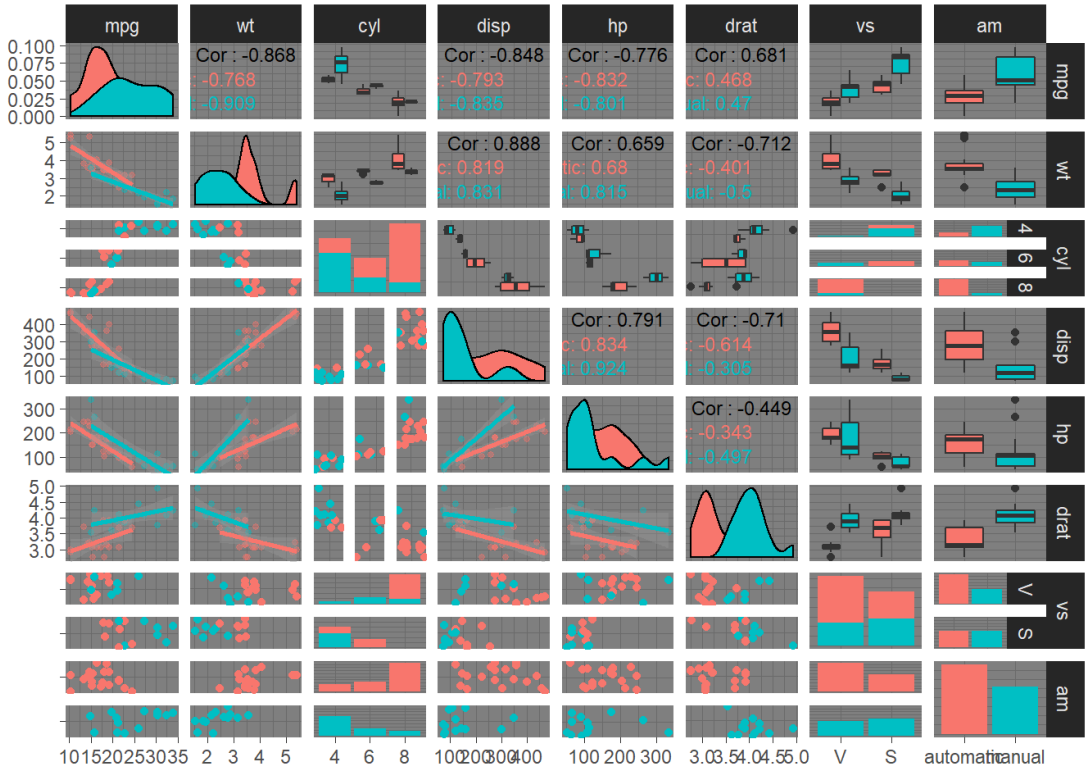

Figure 2. Matrix of scatter diagrams of dependent and independent variables

Model selection

Now we seen that MPG has many other (stronger) correlations than just 'am' we can guess that a model predicting the MPG solely on this parameter will not be the most accurate model. Check this out. Let's start with the basic model.

```
basic_fit <- lm(mpg ~ am, mtcars_factors)
summary(basic_fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_factors)
##
## 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 ***
## ammanual       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
```

Total p-values are actually quite low, the `R^2` is problematic however. Now go to the other side of the spectrum by fitting all parameters of mtcars.

```
total_fit <- lm(mpg ~ ., mtcars_factors)
summary(total_fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars_factors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2015 -1.2319  0.1033  1.1953  4.3085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.09262   17.13627   0.881   0.3895
## cyl6        -1.19940    2.38736  -0.502   0.6212
## cyl8         3.05492    4.82987   0.633   0.5346
## disp         0.01257    0.01774   0.708   0.4873
## hp          -0.05712    0.03175  -1.799   0.0879 .
## drat         0.73577    1.98461   0.371   0.7149
## wt          -3.54512    1.90895  -1.857   0.0789 .
## qsec         0.76801    0.75222   1.021   0.3201
## vsS          2.48849    2.54015   0.980   0.3396
## ammanual     3.34736    2.28948   1.462   0.1601
## gear4       -0.99922    2.94658  -0.339   0.7382
## gear5        1.06455    3.02730   0.352   0.7290
## carb         0.78703    1.03599   0.760   0.4568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

The `R^2` has improved, but the p-values hardly show any significance anymore. Perhaps this is due to overfitting. We now have to meet somewhere in the middle. Let's iterare using the step method.

```
best_fit <- step(total_fit, direction = "both", trace = FALSE)
summary(best_fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars_factors)
##
## 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
```

Examination model

The resulting best model mpg ~ wt + qsec + am is actually dependent on the transmission (am), but also weight (wt) and 1/4 mile time (qsec). All have significant p-values the `R^2` is pretty good to (0.85). Now let's look (amongst others) at the residuals VS fitted.

```
autoplot(best_fit)
```

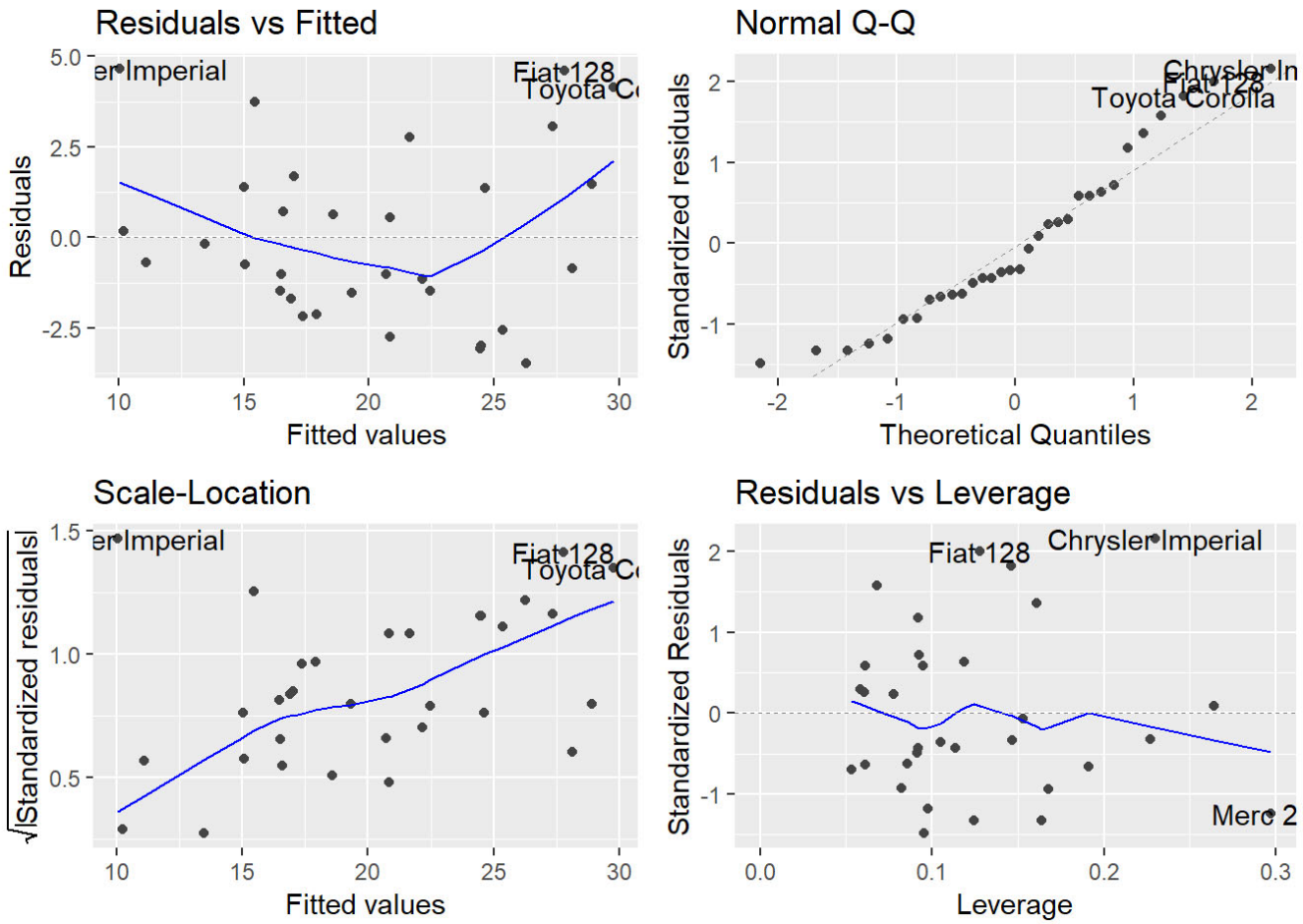

Figure 3. The resulting best model

The 'Normal Q-Q' plot looks good, but the 'Residuals VS Fitted' and 'Scale-Location' both show worrysome trends.

Conclusion

Question - "Is an automatic or manual trancmission better for MPG" may be answered because all models (N5, N6 and N7) show that, holding all other parameters constant, manual transmission will increase your MPG. Question - "Quantify the MPG difference between automatic and manual transmissions" is harder to answer. Based on the best_fir (N7) model mpg ~ wt + qsec + am we could conclude that (with a $p<0.05$ confidence) cars with manual transmission have 2.9358 (~ 3) more miles per gallon than automatic transmissions. The model seems clean with a $p<0.05$ and `R^2` of 0.85. The residuals VS fitted chart however warns us that there is something missing in this model. The real problem is available only 32 observations to train on (N1) and that observations hardly have overlap on the parameters 'wt' and 'qsec' (amongst others) if we look at the diagonal in the matrix of scatter diagrams (Fig. 2). Although the conclusion of ~ 3 mpg better performance on manual transmissions seems fealsible, can't with confidence conclude that this model will fit all future observations.