MOTOR TREND: MPG vs. predictors

EXECUTIVE SUMMARY

In this analysis we assess the relationship between 2 predictor variables (automatic and manual transmission) and the Miles per gallon variable (outcome) of the "mtcars" dataset by answering 2 questions: 1. Is an automatic or manual transmission better for MPG? 2. Quantify the MPG difference between automatic and manual transmissions Dataset description can be found here: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

LOAD AND PROCESS DATA

```
data(mtcars)
```

Looking at the data with `str(mtcars)` we see that some variables that should be categorical variables are numerical ones. lets correct that:

```
mtcars$cyl <- factor(mtcars$cyl);mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear);mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

EXPLORATORY ANALYSIS

To start we explore the relationships of all possible pairs of variables from the dataset. The matrix plot is shown in the appendix. This plot shows as strong correlation between `mpg` and the variables `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am`.

After having confirmed the existence of a strong relationship with the `am` variable through this first macro-analysis, lets have a closer look at the `mpg~am` relationship. The second plot in the Appendix is a boxplot of `mpg` versus transmission type (`am`), ie. manual or automatic transmission. The plot explicitly shows a significant difference in the distribution of the `mpg` variables depending on whether the car has a manual or automatic transmission.

```
## Mean MPG for Automatic    Mean MPG for Manual
##               17.14737               24.39231
```

We will know find out if this difference is statistically significant.

REGRESSION ANALYSIS

Now that we have some evidence that `mpg` and `am` seems strongly correlated, we proceed to a hypotheses test using linear regression models.

- **H0 (null hypothesis)**: There is no difference in `mpg` for different transmission methods.
- **H1 (alternative hypothesis)**: There is a difference in `mpg` for different transmission methods. To confirm the significance of the relationship we show that H0 is rejected.

Relationship Strength

```
c("p-value"=t.test(mpg ~ am, data = mtcars)$p.value,
"r.squared"=summary(lm(mpg ~ am, data = mtcars))$r.squared)
```

```
##     p-value   r.squared
## 0.001373638 0.359798943
```

The p-value is significantly low, meaning that the there is a significant impact of `am` on `mpg` and so H0 is rejected. Also R-squared is 36% or rather this linear regression explains 36% of `mpg` variability.

Best fit model

`am` only partially explain `mpg` variability, so we will use the `step()` method to automatically choose the best multivariate linear regression model using the AIC algorithm. Then using simple ANOVA we will compare the regression of `mpg~am` and the result of the "step()" function. Finaly we will perform a Residual analysis to assess the independance of the variables.

Model Choice

First we compute the linear regression model using the step function which runs `lm` multiple times to find the best fit regression model

```
model<-lm(mpg ~ ., data= mtcars)
best_model<-step(model, direction="both", trace=0)
summary(best_model)$call;summary(best_model)$adj.r.squared
```

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

```
## [1] 0.8400875
```

The outcome shows that the best regression model uses `cyl`, `wt`, `hp` and `am` as relevant variables. According to the `summary`, **84% of the variability of mpg is explained by this model**.

b. ANOVA

Now lets compare the 2 models. ANOVA will help us determine if the confounder variables are relevant to the model fitting or not:

```
am_only_model<-lm(mpg ~ am, data= mtcars)
anova(best_model, am_only_model)$"Pr(>F)"[2]
```

```
## [1] 1.688435e-08
```

The p-value obtained is clearly significant and indicates that `cyl`, `hp` and `wt` explanatory variables do contribute to the model's accuracy.

c. Residuals Analysis

According to the plot 3 in Appendix, the random distribution of the residuals guarantee independance of the variables, residuals are normally distributed and indicate constant variability.

CONCLUSION

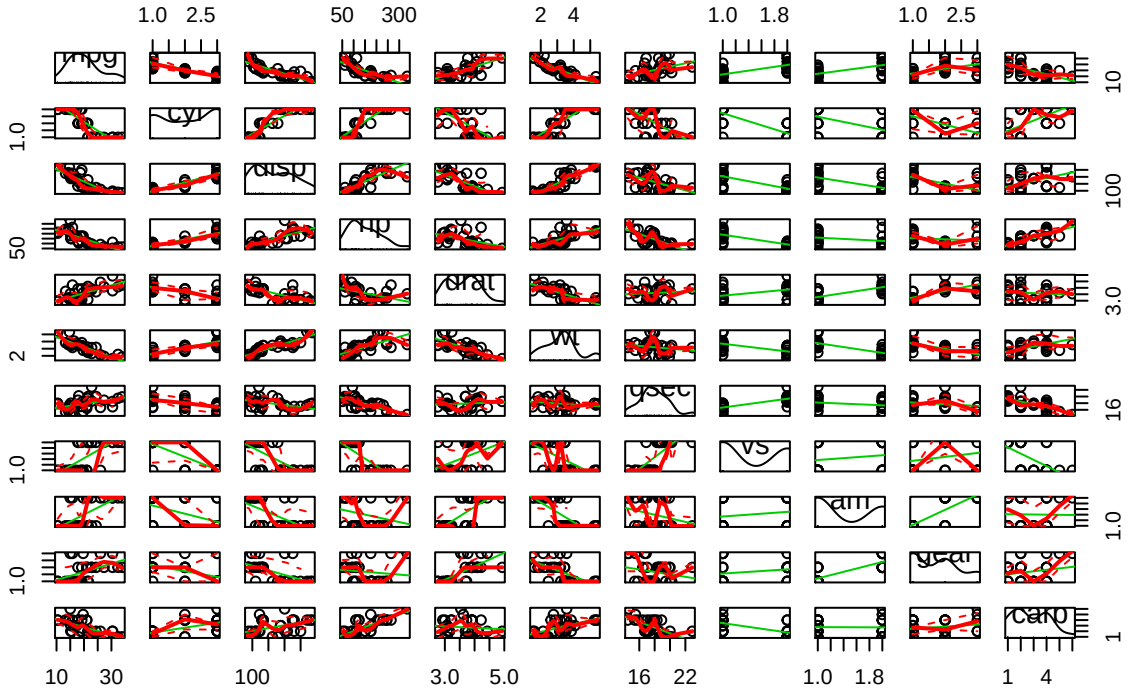
- As we can see by the different results, `mpg` is higher on average for cars with manual transmission compared to cars with automatic transmission. Manual transmission is better than automatic transmission.
- Miles per Gallon (`mpg`) is expected to increase by 1.81 on average for a car with manual transmission compared to a car with automatic transmission (adjusted by `hp`, `cyl` and `wt`).

APPENDIX

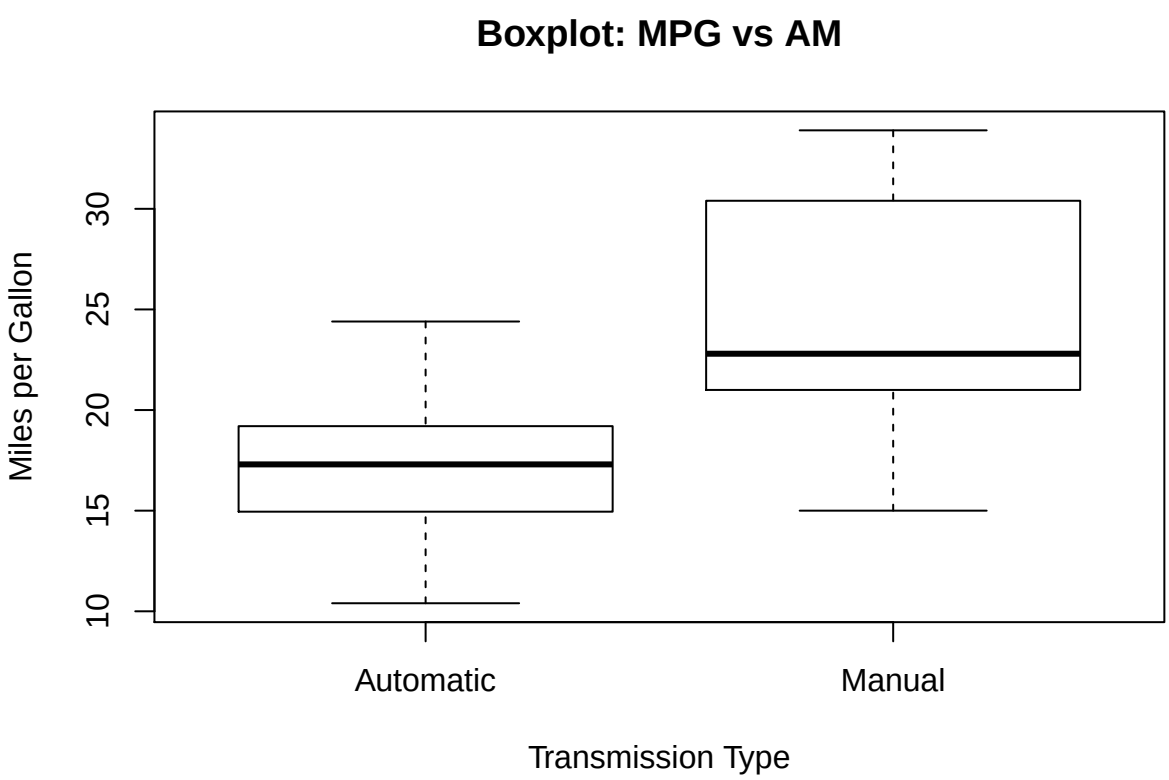
Here is the exploratory analysis of the dataset mtcars. This first plot shows the relationships between all the paired variables.

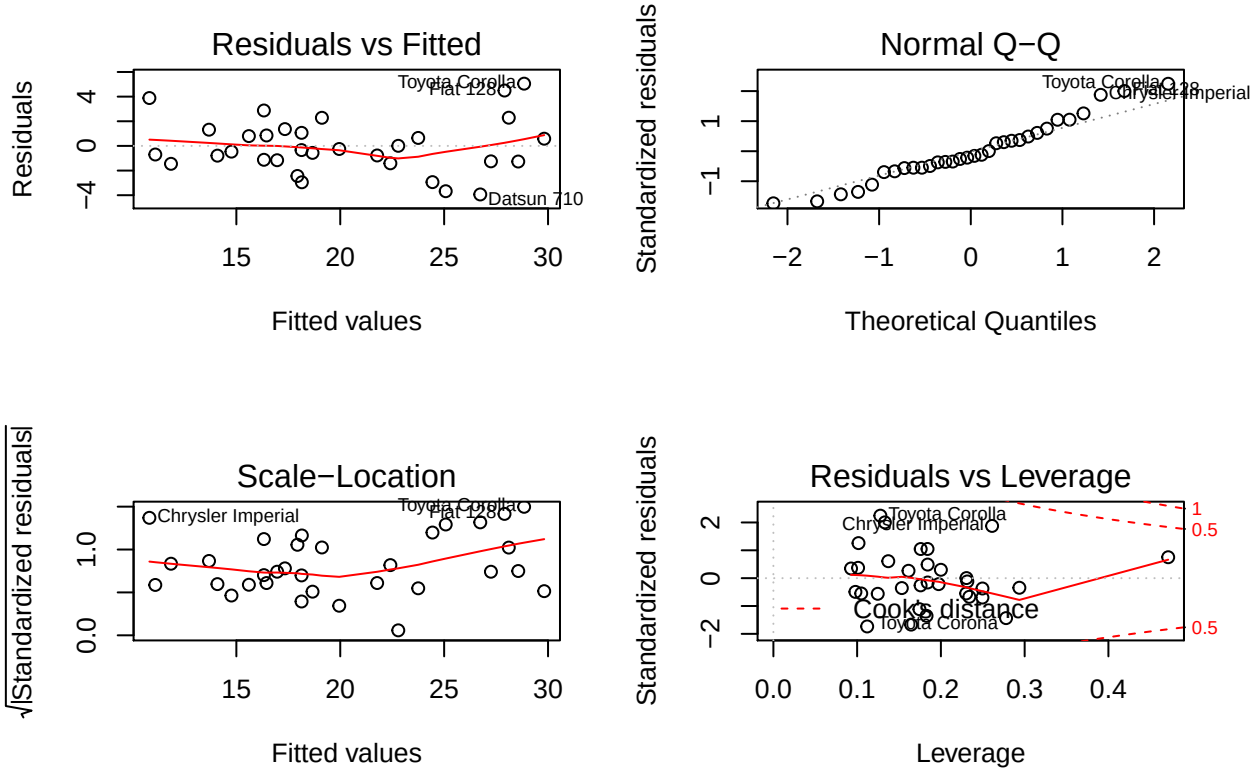
```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```



The second plot shows the significant difference between mpg of automatic and manual cars.


Boxplot: MPG vs AM
Miles per Gallon
10
15
20
25
30
Automatic
Manual
Transmission Type


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona

Plot 3