Regression Models project, Data Science specialization.

Ilyas Aaqaoui

Executive Summary

Motor trend is a magazine about the automobile industry. They are interested in studying the relationship between a set of variables and fuel consumption, measured by Miles per Galon. After conducting a study: cleaning, exploring and applying regression models on the data set provided, we can say that manual transmission is better than automatic transmission in terms of fuel consumption. In the following we will showcase this result and provide a measurment of how much manual is better than automatic.

Study

The data set in hand is called "mtcars", extracted from the 1974 Motor Magazine. it has in total 32 cars (observations) and 11 variables. we interested in the mileage per galon (MPG) as the outcome dependant on the other variables (regressors), in particular, the transmission mode: * Is a manual or automatic transmission better for MPG? * Quantify the MPG difference between auto and manual.

The data has, besides MPG, 10 other variables, classical measures in the auto industry: number of cylinders, displacement, horsepower, weight, transmission etc.

Let's start with some cleaning.

Cleaning Data

We load the data and convert some variables to factors, as they should be!

```r
data(mtcars)
mtcarsOld = mtcars
mtcars$am   = factor(mtcars$am)
mtcars$gear = factor(mtcars$gear)
mtcars$cyl  = factor(mtcars$cyl)
mtcars$carb = factor(mtcars$carb)
mtcars$vs   = factor(mtcars$vs)
```

For more easyness, let's change the am levels names to more significative ones:

```r
levels(mtcars$am) = c("Automatic", "manual")
```

Packages Loading

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.2.5
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.2.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(bitops)
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.2.5
```

Exploratory Analysis

The objective of this part is to explore data and relationship between variables for better modelling.

Relationship/Correlation between variables

We plot mpg against each of the other variables(see appendix - fig.1) to understand how they are correlated with it (just as we did with the Swiss feritility data). We find out that most of these variableas are highly correlated with mpg, individually (mpg with each variable separately).

Let's now try to find the correlation between regressors, excluding the outcome, in order to find which variable should not be included in the model as predictor.

```
c=cor(mtcarsOld)
corIndices = findCorrelation(c, cutoff = 0.75) ; print(corIndices)
```

```
## [1]  2  3  1 10
```

We find that variables of indices 2, 3, 1, 9 which are cyl, disp, mpg, and am are highly correlated. mpg is the output variable. am is the variable under study so we will include both of mpg and am while cyl, and disp should preferably not be included in the model together (just like the Education and Examination variables in Swiss data). We will leave the step function to fine tune the selection process.

The effect of transmission type

Also, to compare the effect of the type of transmission of the fuel consumption, we plot the different distributions of mpg vs transmission type(see appendix - fig.2). The figure shows that manual transmission cars tend to consume less fuel.

Regression Analysis

We build a base regression model that include all variablea as predictors for the MPG then we use the step function to build further linear models to select the most approprate variables as predictors while eliminating those that are not very significant to the model.

```
baseModel = lm(mpg ~., data= mtcars)
bestModel = step(baseModel, direction = "both")
```

As we can see, the algorithm has selected cyl, hp, wt and am as predictors.

```
summary(bestModel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

but let's check the strength of our model through residuals: ###Residuals and diagonistics

To check residuals for normaility and homoscedasticity, we plot the residuals(see appendix - fig.3). We find that the residuals are normally distributed and homoscedastic.

Results of the Model Chosen

The model we used for the regression included the cylinders, horsepower weight and transmission.

Variability of the data measurment

This model, bestModel, has successflly explained more than 83% of the variablilty of the data.

Model coefficients

Transmission control

Miles per gallon increases by a factor of 1.8 (1.8) with manual transmission.

mpg vs hp

Miles per gallon decreases by a factor of 0.03 (-0.03) as horspower increases.

mpg vs hp

Miles per gallon decreases by a factor of 2.5 (-2.5) for every increase of 1000lb in the weight.

mpg vs cyl6/cyl8

Miles per gallon decreases by a factor of 3.03 (-3.03) for 6 cylinders and by a factor of 2.16 (-2.16) for 8 cylinders. Which is normal since more cylinders mean more fuel injection to the pistons chambers.

Intercept

The intercept is at 33.7 mpg.

P-value

The overall p-value is very small (1.506 x 10 ˆ -10), we then reject the null hypothesis that variables are not significant.

Conclusion

“Which one is better for MPG” As illustrated in this study, the manual transmission is better for fuel consumption, miles per gallon. This is demonstrated by the best fit model which estimates the decrease of fuel consumption by a factor of 1.8 for manual transmission.

“Quantifying the difference between automatic and manual transmissions” Fuel consumption, mpg, for manual transmission increases by a factor of 1.8 over automatic transmission.

```r
manualData = mtcarsOld[ mtcarsOld$am == 0, 1]
automaticData = mtcarsOld[mtcarsOld$am == 1, 1]
manualCI = t.test(manualData, mu = 0)$conf.int
automaticCI = t.test(automaticData, mu = 0)$conf.int
print(manualCI);print(automaticCI)
```

```
## [1] 15.29946 18.99528
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 20.66593 28.11869
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval for mpg for automatic transmission (15.3, 19) while it is (20.67, 28.12) for the manual transmission. This is an evidence that manual transmission consumes less fuel. Manual transmission travels around 7.24 more miles per gallon, (24.39mpg), than automatic transmission (17.15 mpg).

Appendix

Exploratory features plot

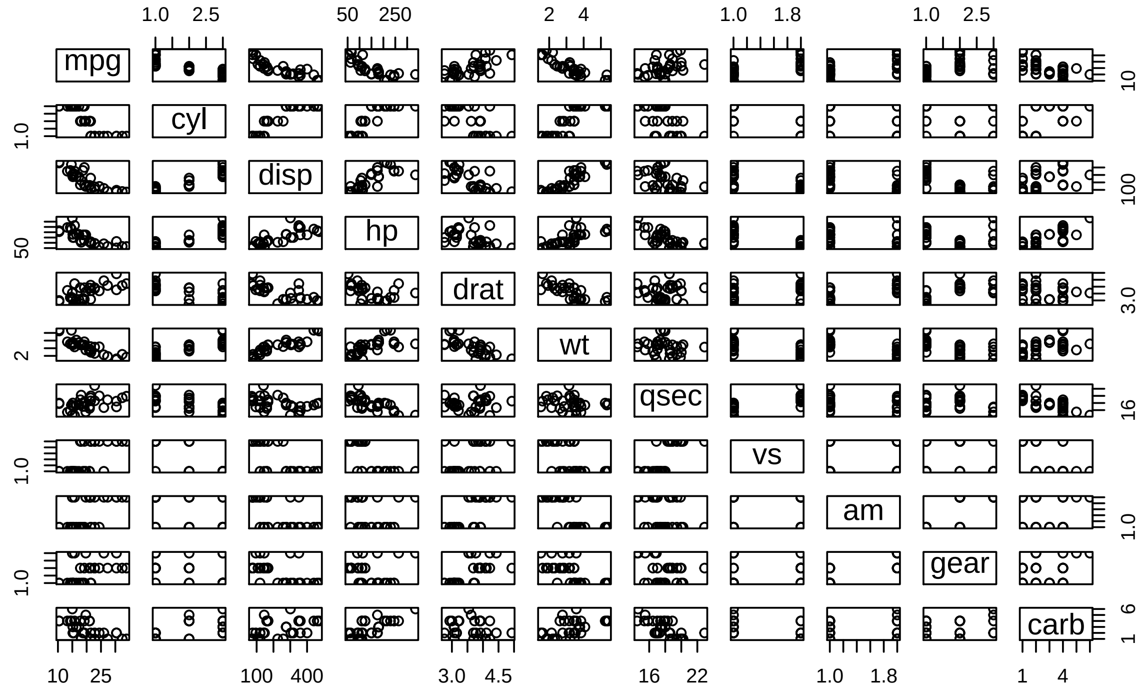


Figure 1: Figure 1. Featues plot for the dataset

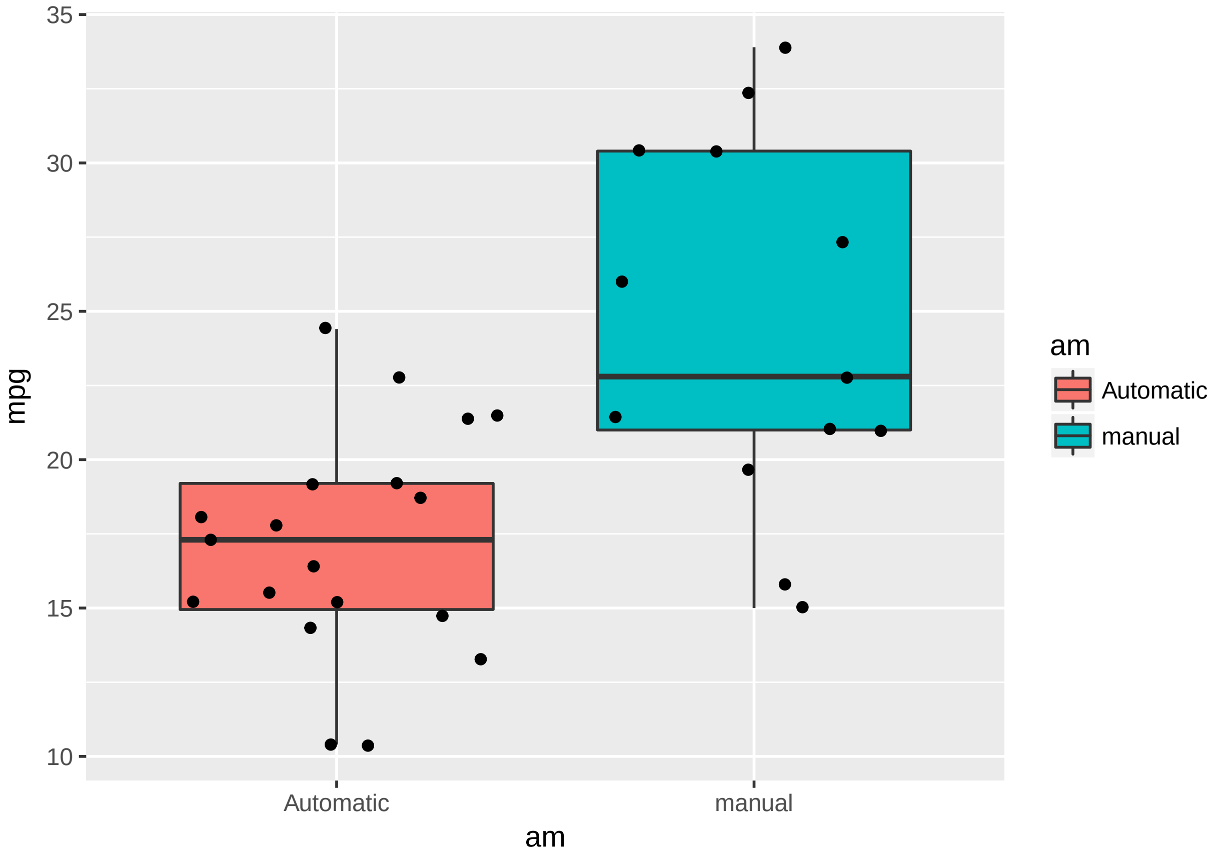
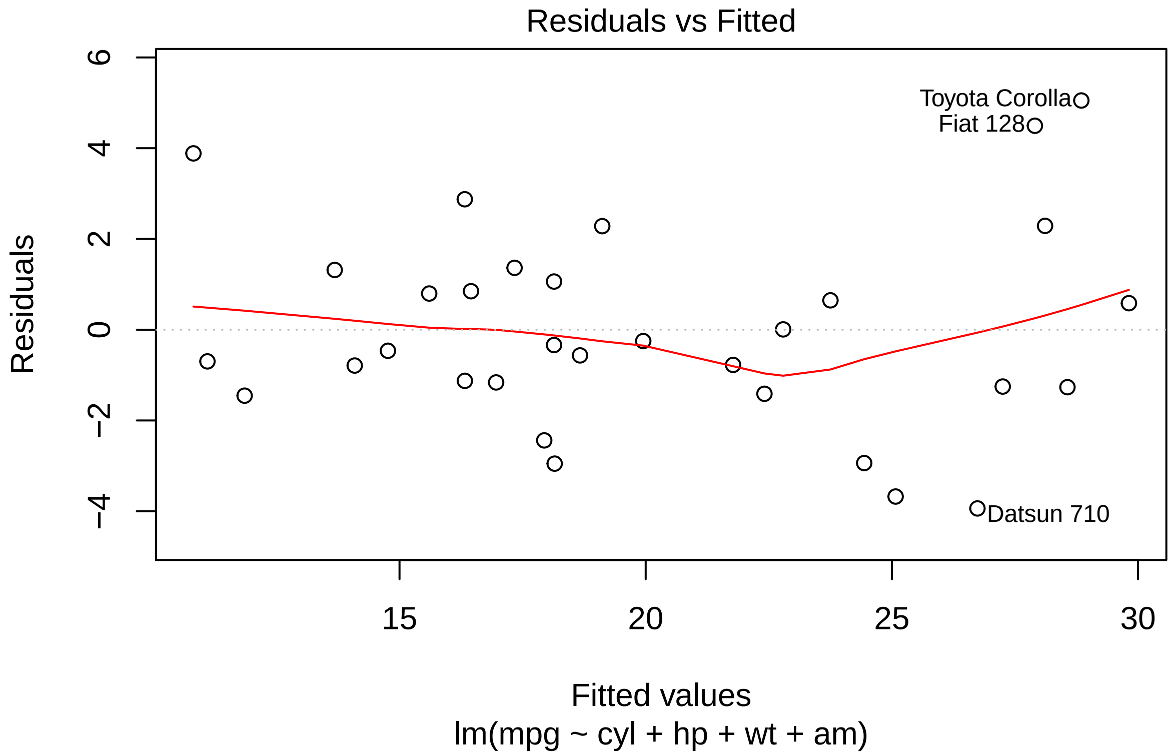
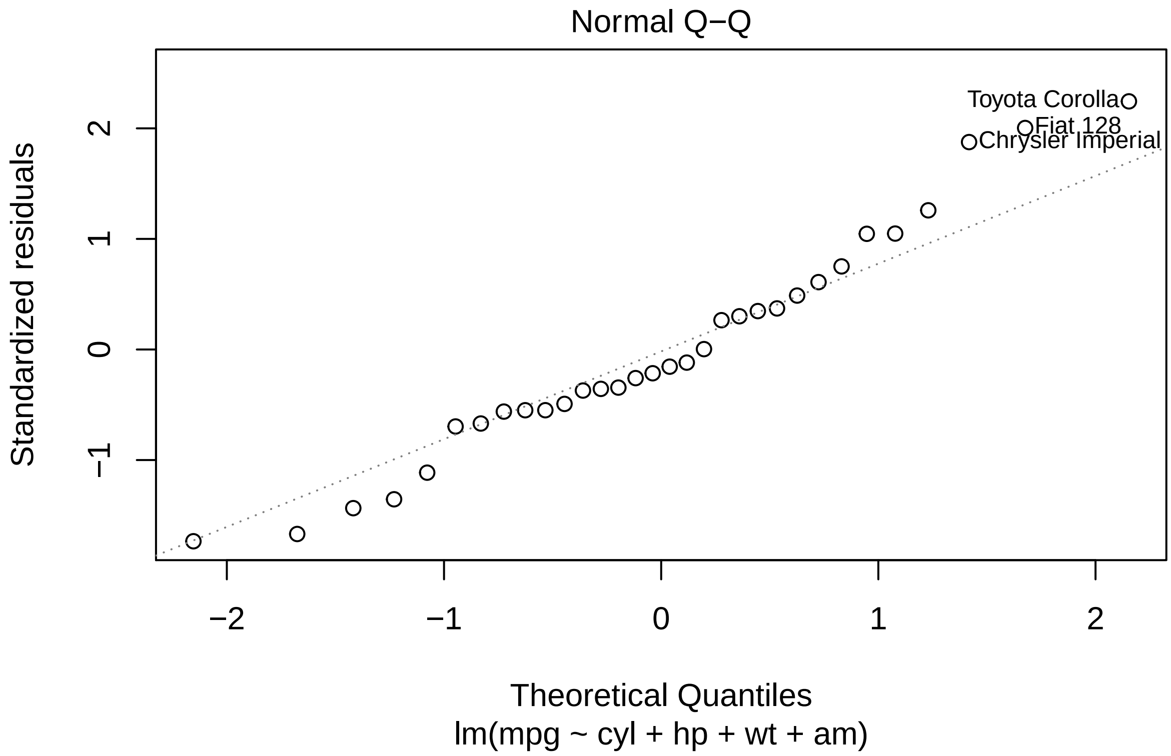


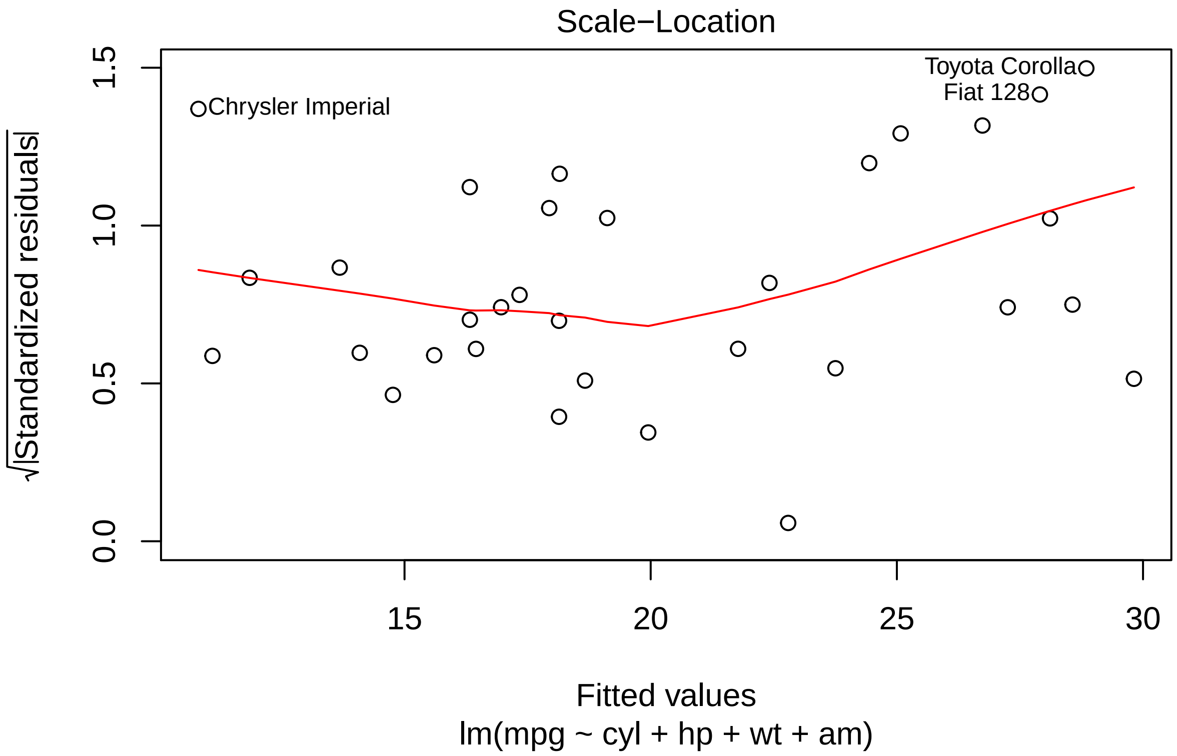
Figure 2: Figure 2. MPG VS. Transmission Type

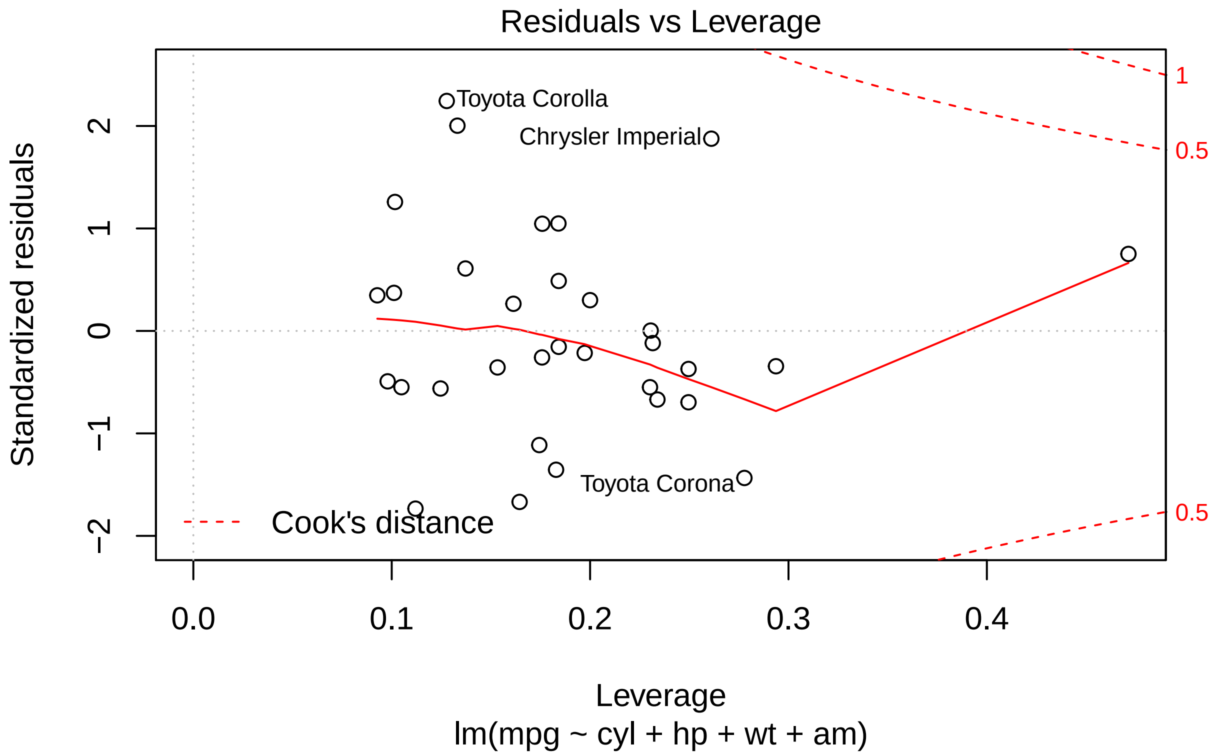
MPG vs. transmission type

Residuals plot

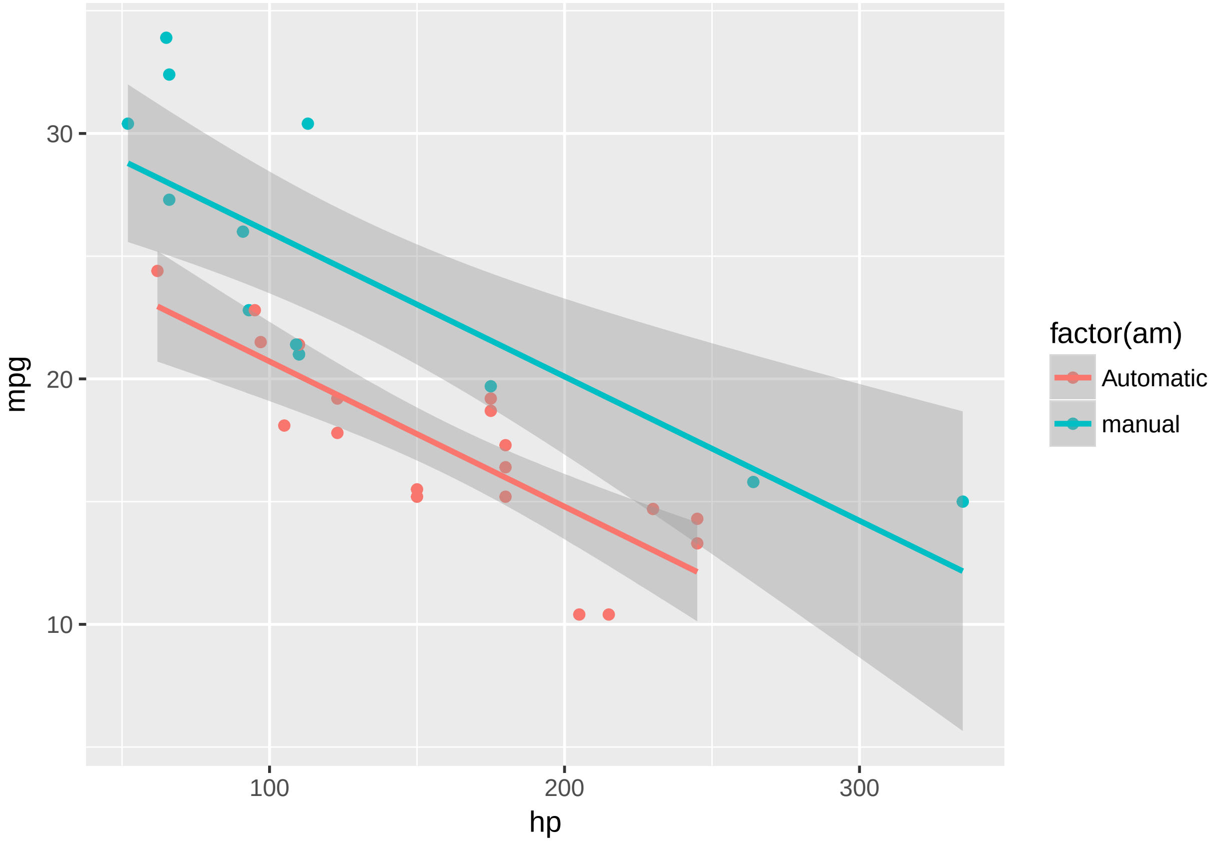


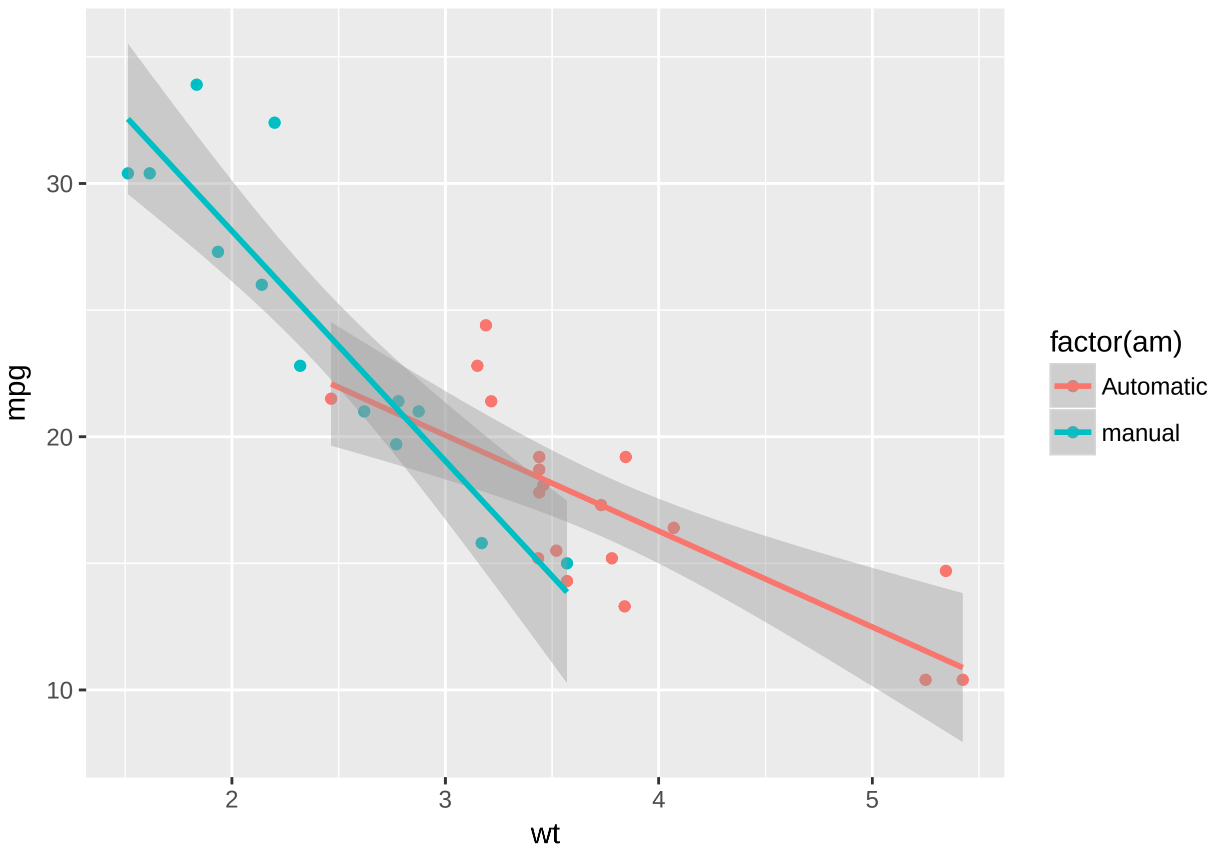

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cyl + hp + wt + am)
Toyota Corolla
Fiat 128
Chrysler Imperial
−2
−1
0
1
2
−1
0
1
2

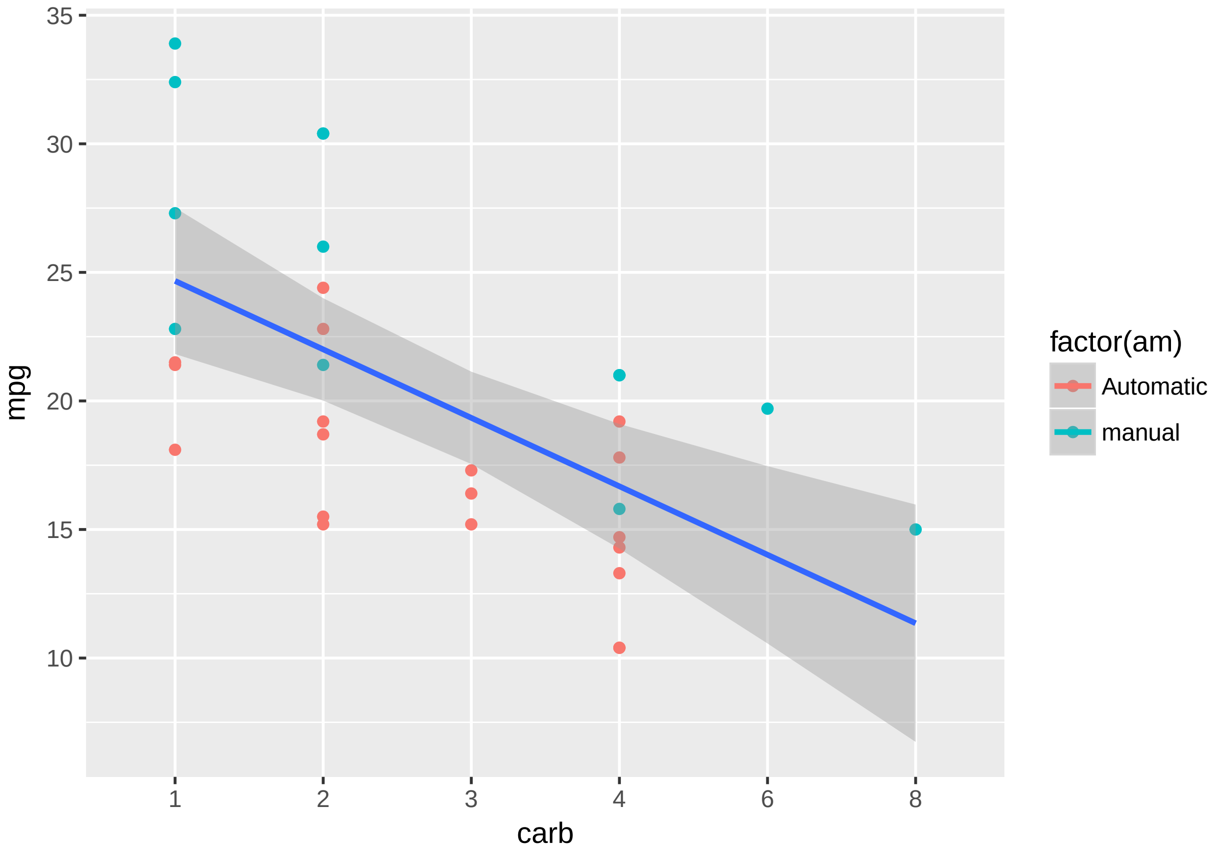

Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Fitted values
lm(mpg ~ cyl + hp + wt + am)


Residuals vs Leverage
Standardized residuals
Leverage
lm(mpg ~ cyl + hp + wt + am)
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
2
1
0
−1
−2
0.0
0.1
0.2
0.3
0.4

Exploration of other linear models




mpg
30
20
10
wt
2
3
4
5
factor(am)
Automatic
manual


35
30
25
20
15
10
mpg
1
2
3
4
6
8
carb
factor(am)
Automatic
manual