Regression Models Course Project

Yuan Liao

9/20/2020

Executive summary

Looking at a data set of a collection of cars, this project explores the relationship between a set of variables and miles per gallon (MPG) (outcome). Particularly, this project looks into the below two aspects:

1 Is an automatic or manual transmission better for MPG

2 Quantify the MPG difference between automatic and manual transmissions

Q1 Is an automatic or manual transmission better for MPG

Both models output suggest that a manual transmission is better for MPG than an automatic transmission.

Q2 Quantify the MPG difference between automatic and manual transmissions

In the single-variable model, the manual transmission has 7.245 more MPG than the automatic one. However, this quantity is ascribed to the other factors.

In the selected model, the advantage of have a manual transmission on MPG is 1.375 given the same number of cylinders (cyl), displacement (disp), gross horsepower (hp), rear axle ratio (drat), and weight (wt).

Model design

According to the exploratory analysis in the appendix, we have found that 1) Miles/(US) gallon (mpg) is lower for the automatic than the manual (am), and 2) The observed difference between the automatic and the manual can be affected by the other factors such as cyl, disp, hp, drat, and wt.

Therefore, we try two models; one uses mpg vs am (Single-variable regression model) and the other uses mpg vs am among other variables (Multivariate regression model). Single-variable regression model has adjusted R-squared of 0.3385 which is lower than Multivariate regression model's 0.8211. Also by comparing these two models' residuals, we select the multivariate model.

Single-variable regression model

```r
fit.single <- lm(mpg ~ am, data = mtcars)
summary(fit.single)
```

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

Multivariate regression model

```r
fit.multi <- lm(mpg ~ am + cyl + disp + hp + drat + wt, data = mtcars)
summary(fit.multi)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + drat + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.437 -1.574 -0.688  1.310  5.551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.04938    7.60553   4.740 7.31e-05 ***
## am           1.37506    1.56866   0.877  0.38906
## cyl         -1.03335    0.72405  -1.427  0.16590
## disp         0.01257    0.01195   1.052  0.30307
## hp          -0.02887    0.01444  -1.999  0.05658 .
## drat         0.48586    1.49495   0.325  0.74788
## wt          -3.27472    1.15685  -2.831  0.00903 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.549 on 25 degrees of freedom
## Multiple R-squared:  0.8557, Adjusted R-squared:  0.8211
## F-statistic: 24.72 on 6 and 25 DF,  p-value: 2.266e-09
```

Appendix

Dataset

A dataframe with 32 observations on 11 (numeric) variables.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (1000 lbs)

[, 7] qsec 1/4 mile time

[, 8] vs Engine (0 = V-shaped, 1 = straight)

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

Exploratory analysis

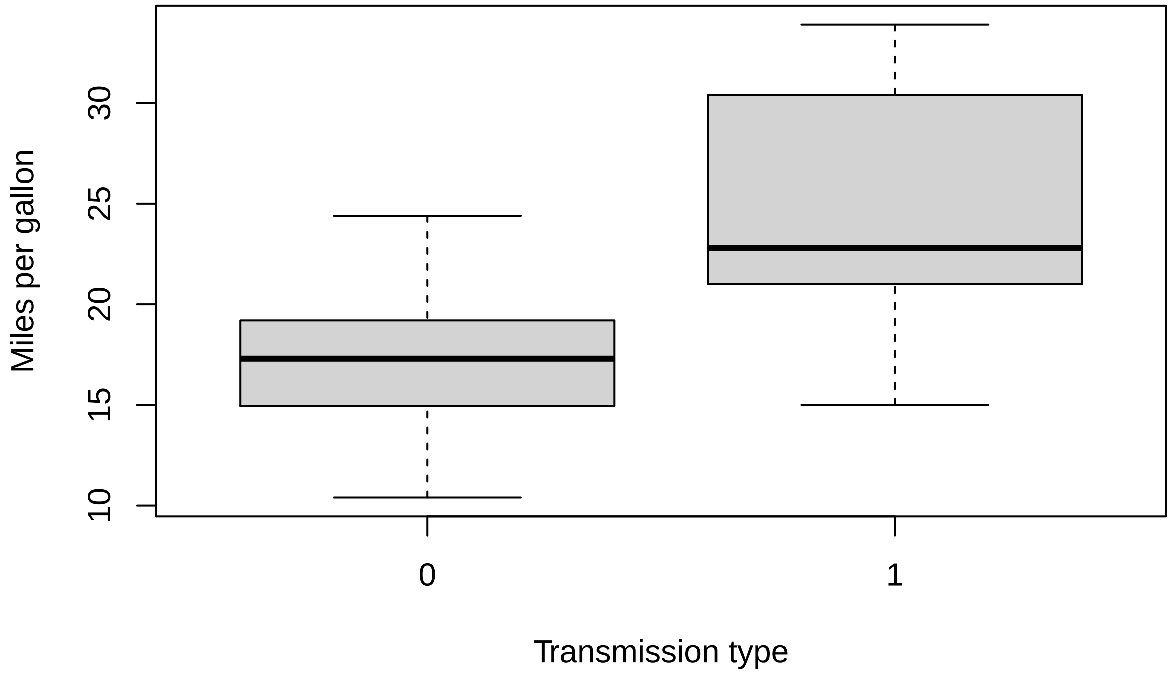
This part takes a look at the original dataset and visualize it with the focus on the mpg vs. am, which might be affected by the other variables. First, let's take a look at the data.

```r
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

How mpg of different transmission types look like is visualized below.

```r
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type", ylab = "Miles per gallon")
```

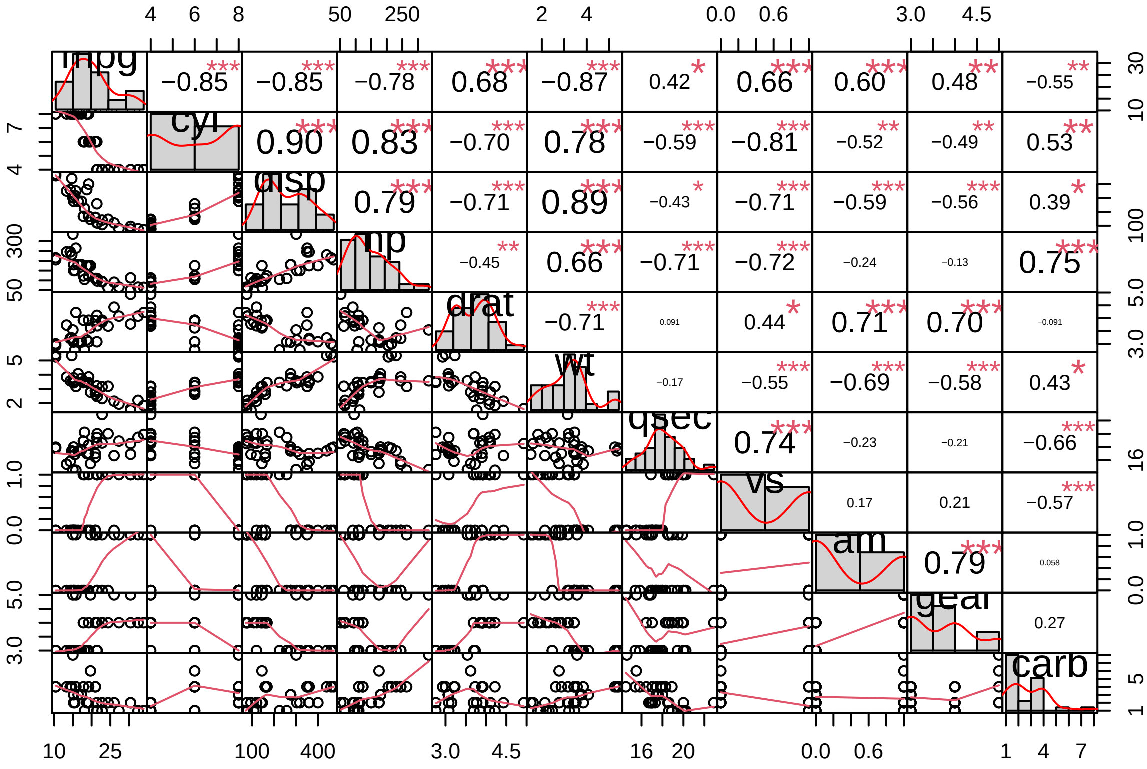


Variables selection

This step explores the relationship between all the variables to identify what variables to include into the modelling.

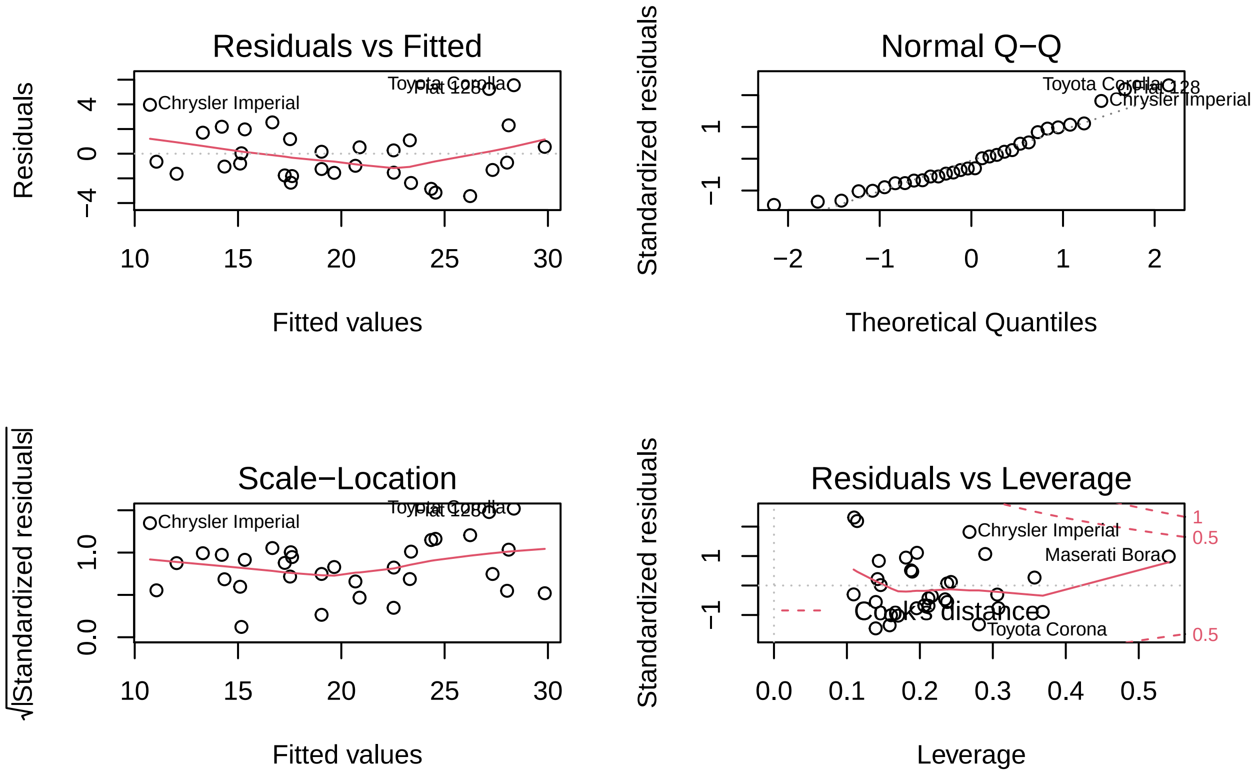
The difference of mpg between the transimission types might be caused by the other factors. Now let's see how these variables are correlated with each other. According to the below figure, the variables involved to answer the target questions are mpg and am, as well as cyl, disp, hp, drat, and wt.

```
chart.Correlation(mtcars, histogram=TRUE, pch=19)
```



Residual plot

```r
par(mfrow = c(2,2))
plot(fit.multi)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
−4
0
4
10
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
1.0
10
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5
−1
1
0.0
0.1
0.2
0.3
0.4
0.5