Regression Models Project-Coursera

WNH

January 20, 2020

This is a regression analysis that tries to respond to the following 2 questions: 1.Is an automatic or manual transmission better for MPG ? 2.Quantify the MPG difference between automatic and manual transmissions

Executive Summary

Based on mtcars small dataset analysis we can conclude: on average, automatic transmission cars consume more fuel then manual transmission ones, with 7.24 gallons more (24.39 - 17.15, the 2 Manual Transmission and Automatic means) this estimation has a confidence interval of [3.21 , 11.28] the adjusted estimate for the expected change in mpg from Automatic to Manual Transmission is +0.1765 gallons

```
knitr::opts_chunk$set(echo = FALSE,message=FALSE,warnings=FALSE)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.5.3
```

Loading mtcars dataset

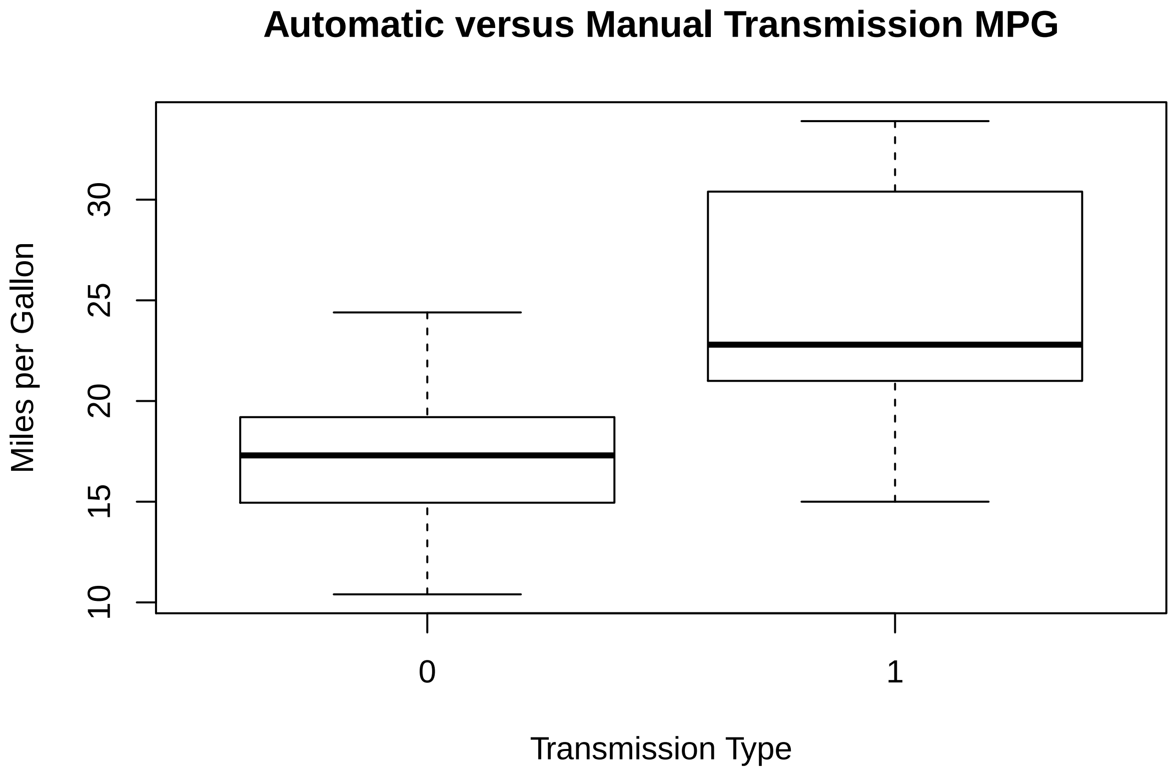
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

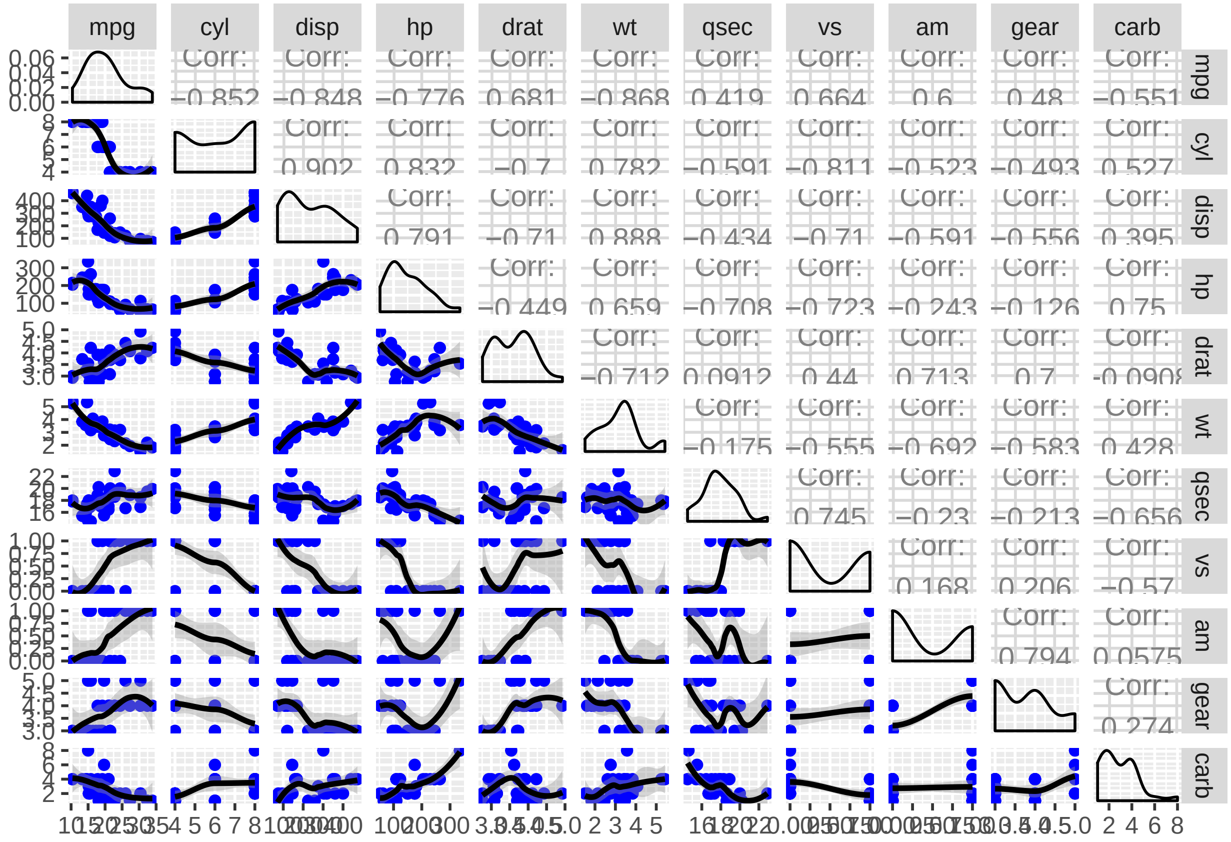
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Data Analysis

```
##
##  Welch Two Sample t-test
##
## data:  mpg.manual and mpg.auto
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

As the p-value is 0.001374 well bellow 5% or 1%, the alternative hypothesis is true: the difference in means is not equal to 0. So, the mean mileage of automatic transmission is 17.15 mpg and the manual transmission is 24.39 mpg. The 95% confidence interval of the difference in mean gas mileage is between 3.21 and 11.28 mpg. We could say that manual transmission could be better than automatic transmission for MPG.





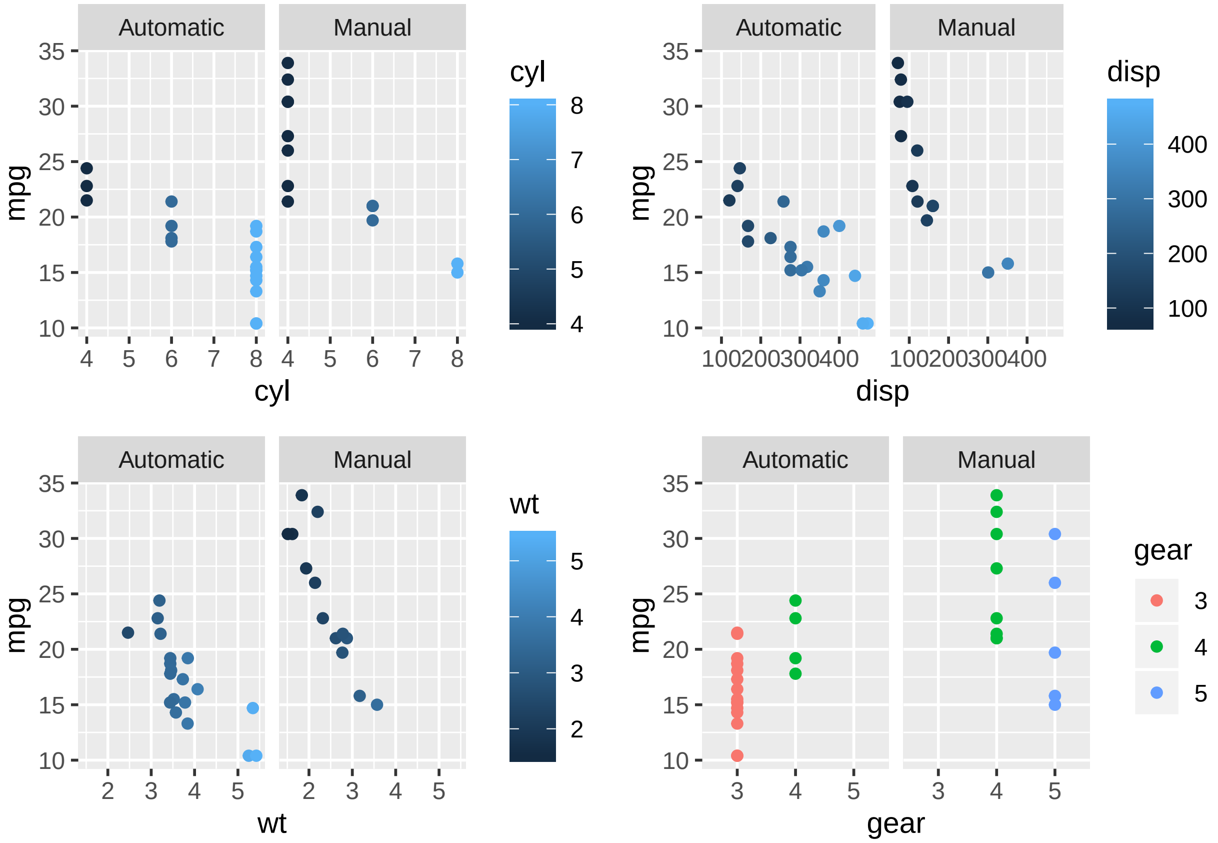
Creating labelled factor variables for the categorical variables

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

We want to explain the data in the simplest way - redundant predictors should be removed. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, thecsimplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is best.

Model Selection

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```



Predictor Selection Forward Selection

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than ??crit. 3.Continue until no new predictors can be added.

```
##
## 0.338458908206314 0.735788906182185 0.742393789059248 0.752150855824599
##                 1                 1                 1                 1
## 0.802926571399959 0.807875947013112 0.812160279934348 0.815148648598381
##                 1                 1                 1                 1
## 0.853553398875962
##                 1
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + wt
## Model 5: mpg ~ am + cyl + disp + wt + gear
## Model 6: mpg ~ am + wt + am * wt
## Model 7: mpg ~ am + cyl + disp + wt + am * wt
## Model 8: mpg ~ am + wt
## Model 9: mpg ~ am + cyl + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                    
## 2     29 271.36  1    449.53 62.7972 2.792e-08 ***
## 3     28 252.08  1     19.28  2.6934  0.113285    
## 4     27 188.43  1     63.66  8.8923  0.006305 ** 
## 5     25 178.96  2      9.46  0.6610  0.525149    
## 6     28 188.01 -3     -9.04  0.4212  0.739427    
## 7     26 138.31  2     49.70  3.4714  0.046731 *  
## 8     29 278.32 -3   -140.01  6.5196  0.002074 ** 
## 9     28 191.05  1     87.27 12.1914  0.001803 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final addidtion of the am*wt variable is a close call. We may want to consider including this variable if interpretation is aided. Notice that the R2 for the lm(mpg~am) model of 0.360 is increased greatly to 0.878 in the final model. Thus the addition of two predictors causes major improvement in fit.

```
## 
## 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 ***
## 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
```

```
## 
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.1735 -1.5340 -0.5386  1.5864  6.0812 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## amManual      0.1765     1.3045   0.135  0.89334    
## cyl          -1.5102     0.4223  -3.576  0.00129 ** 
## wt           -3.1251     0.9109  -3.431  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Conclusion

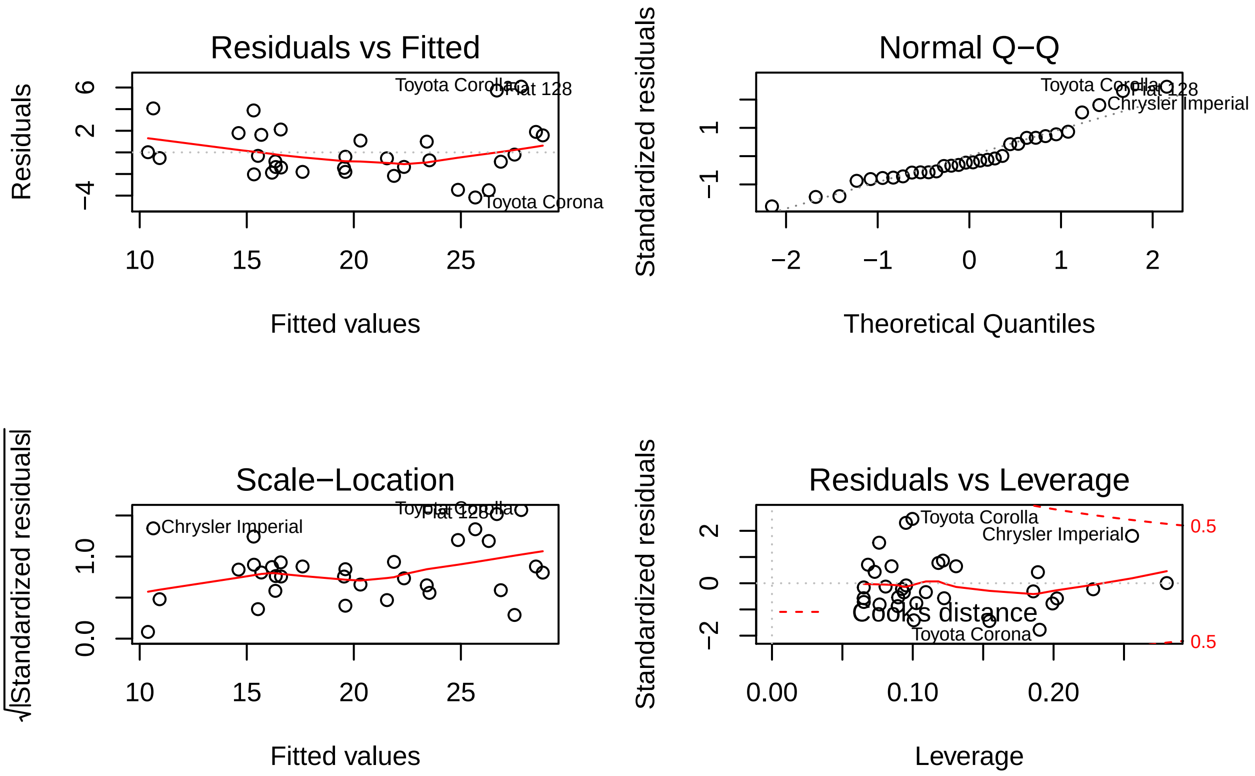
In this model, Pr(>|t|) are very close to zero,it shows that there are small p-value for the intercept and the slope,indicating that we there is a relationship between in miles per gallon (mpg) number of cylinders(cyl),and transmision(am). Now when we read the coefficient for am, we say that, on average, manual transmission cars have 2.56 MPGs more than automatic transmission cars,holding that other are constant.

The "Occam's razor" model explains 83% of mpg variance and contains only 3 predictors: formula = mpg ~ am + cyl + wt amManual estimated coefficient equals now to 0.1765 and represents the adjusted estimate for the expected change in mpg comparing Auto versus Manual for this new model containing 2 other predictors besides am.

amMaual estimated coefficient is the answer to the second question.

Best model residuals are depicted in Regression Dignostics First graphic, "Residuals vs. Fitted values" is not quite a straight line, proof of some outliers.

Regression Dignostics


Residuals vs Fitted
Residuals
6
2
−4
10
15
20
25
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
0.0
1.0
10
15
20
25
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
2
0
−2
0.00
0.10
0.20
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
0.5