Regression Models Project

Plato Karageorgis

6/11/2021

A Quick Intro

According to the exercise, we work for Motor Trend, a magazine about the automobile industry. So we are about to run some tests and analyze the given mtcars dataset and hopefully when we are done we will be able to jump into concrete conclusions. The topic of the analysis is the comparison of automatic versus manual transmission and finding out which works best for the MPG and measuring the exact impact on those two on MPG. Without further ado, lets get to it.

To begin with, lets take a look at our data before we start the analysis.

```
##                    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
```

Now, the first part is simple if you have completed the Statistical Inference class. We will run a two-sided t.test where the null hypothesis is H0: x=0 where x says that am not is influential on mpg and HA: x!=0 stating that am is affecting the mpg.

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## 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:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1 
##        17.14737        24.39231
```

The p-value is p=0.001374 so if we set a=0.01 we are 99% confident that we can reject the null hypothesis. This means that am is not affecting the mpg.

As for the quantity difference and whether manual or automatic is better we can calculate a confidence interval that will illustrate the margin we are looking for. But we will jump to that as soon as we have concluded to our final model.

Models

Moreover, we will attempt to fit some models in order to get a clearer view.

First Model The first model will contain every value in order for us to see the general view of our data.

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

We can see that values like cyl, disp, hp and carb are really the opposite of influential on mpg so we choose to remove them and fit the data again.

Second Model

```
##               Estimate Std. Error     t value    Pr(>|t|)
## (Intercept)  8.3685509 10.7727400  0.77682659 0.444549626
## drat         0.8066314  1.5041594  0.53626723 0.596513845
## wt          -3.8141916  0.8594533 -4.43792786 0.000159957
## qsec         1.1884785  0.4820250  2.46559531 0.020891852
## vs          -0.0526668  1.8866110 -0.02791609 0.977950733
## am           2.8346643  1.9083734  1.48538242 0.149944455
## gear        -0.3387703  1.1373401 -0.29786186 0.768269954
```

Similarly, we will remove drat, vs and gear.

Third Model

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

```
## [1] 0.8335561
```

If we remove the least influential value which is qsec here, the am value becomes biased and therefore this is the most precise model we can have with this technique. We can also see that the adjusted R-squared value is 83% which means that the total variance is explained well by this model.

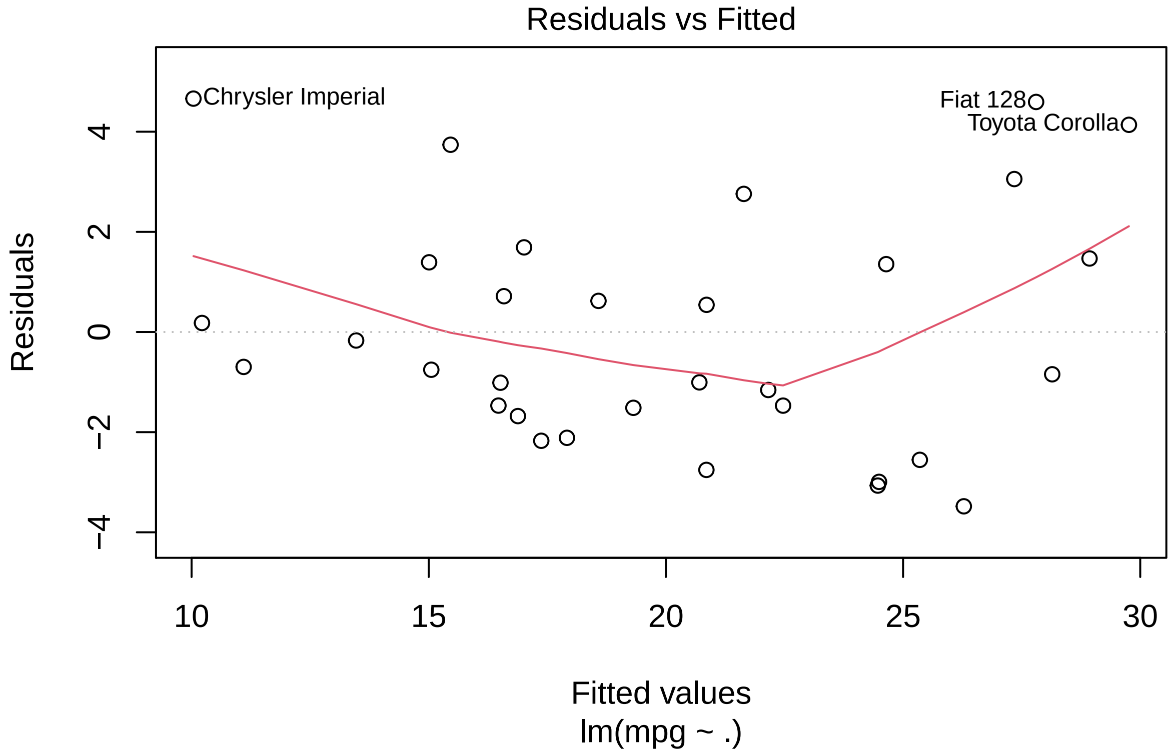
Now that we have our final model, lets do the manual-automatic comparison and the quantity difference of the MPG between manual and automatic transmissions.

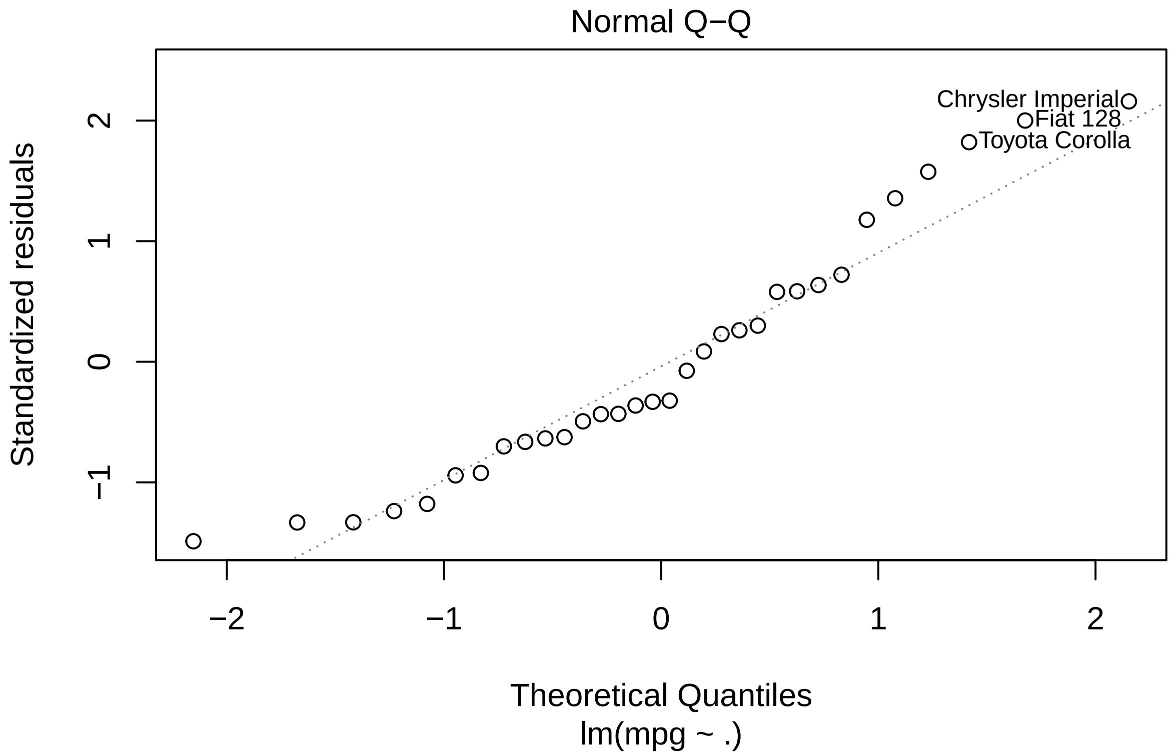
```
## [1] 0.04573031 5.82594408
```

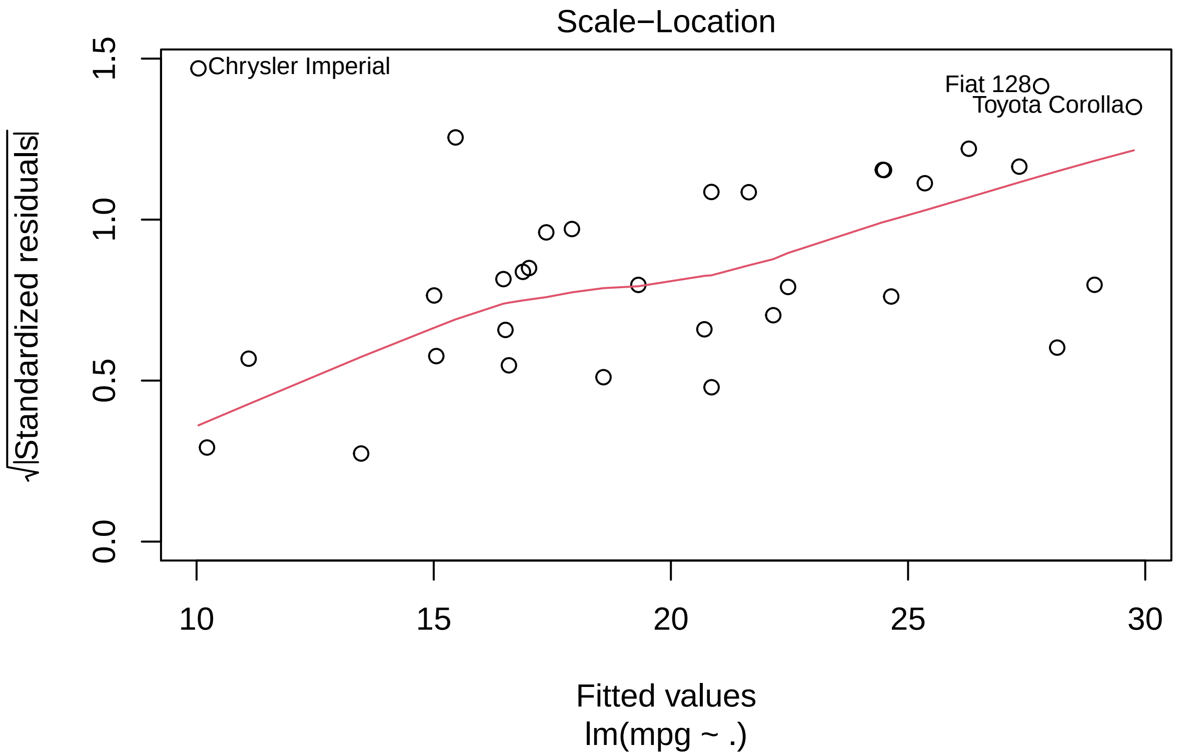
Our confidence interval states that we are 95% confident that a shift from automatic = "0" to manual = "1" will increase the mpg by at least 0.04573031 and at most 5.82594408. So, the manual transmissions seem to be better than the automatic ones.

Exploratory Data analysis

Even though we have pretty much already analyzed along with the models our data, lets add some plots to visualize it. We'll focus on the residuals.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
Chrysler Imperial
Fiat 128
Toyota Corolla
−2
−1
0
1
2
−1
0
1
2



On the first plot there is a clear relation with the fitted regression line but from my point of view it could be more precise. But it is definitely showing what we expected. The second plot is depicting the obvious linearity of the data with the QQ-Plot and the third plot similarly with the first one could be a more precise but is still on point.

Summary

After our analysis we have concluded that there is relation between am and mpg even though its not the strongest one. The comparison between automatic and manual transmissions had manual as a winner and the quantity difference is shown by the confidence interval we calculated before.