RegressionClassProject

Bassam Abdelnabi

9/9/2019

Synopsis

Motor Trend is a magazine that focuses on the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

This study shows that the autmoatic trasnmissions is better with an average saving of 1.4-2. It may be as high as 2.9 mpg. The models showing this conclusion have an adjusted RMS of 0.45-0.83. All the models agree that there is a benefit and saving. However, they disagree in the specific value. The author of this report recommends taking more data for validation. Please check the rmd code in the github repo to check the code.

Loading necessary libraries and data

Exploring and cleaning the data

Data structure is showing a lot of numerical while they should be listed as factors so we will correct that. We will check the data for NA, zero covariates or near zero. We will check collinearity. PLease check the code on github for the details.

```
## [1] 0
```

Prelimanry estimates

We have prepared the data set for analysis. We will start by a simple t test since the number of data is limited. We will do a plot of both groups as a visulaization of the problem. We will also do a plot of all variables to have a feel of the probelm under study. Let us see the results of the t-test. Please check the code on github for details.

```
##
##  Welch Two Sample t-test
##
## data:  datamtcars$mpg by datamtcars$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
```

Building models and exploring them

In the previous section we have seen that there is a difference in the mpg for both groups based on the the test. It is statistically significant. The plots in the appendix show several variables correlate with the mpg, however, the displacment is sufficent to be express the cylinders number so i will remove it from the analysis.

Now, we will build several models and compare them. We will do a linear model, a ridge regression, lasso and elastic net. We will also do a simple physics based model based on my experience as a combustion engineer. Check the details in the code. I will only compare the models here.

```
## $MAE
##                  Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## LineraModel 3.472335 3.573274 3.910761 3.983717 4.171707 4.888976    0
## Ridge       1.683005 2.213993 2.580965 2.481542 2.865176 2.992733    0
## Lassso      1.666780 2.123752 2.180514 2.225096 2.323506 2.846763    0
## ElasticNet  1.523257 2.056244 2.190967 2.175113 2.307276 2.788103    0
##
## $RMSE
##                  Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## LineraModel 3.910259 4.476914 5.472838 5.164610 5.716972 6.188433    0
## Ridge       2.120229 2.719470 3.010408 3.121628 3.572100 4.206900    0
## Lassso      1.977563 2.620160 2.856157 2.791072 3.064148 3.381581    0
## ElasticNet  1.792260 2.521580 2.877348 2.718877 2.970037 3.362381    0
##
## $Rsquared
##                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## LineraModel 0.3753897 0.4254009 0.5689848 0.5215343 0.6076227 0.6178047
## Ridge       0.6808755 0.7605909 0.8363998 0.8194268 0.8947637 0.9133462
## Lassso      0.7488458 0.8130018 0.8247706 0.8309293 0.8696729 0.8934698
## ElasticNet  0.7168003 0.8192341 0.8342059 0.8393412 0.8956547 0.9204539
##             NA's
## LineraModel    0
## Ridge          0
## Lassso         0
## ElasticNet     0

## [1] 1.951657

## [1] 1.419306
```

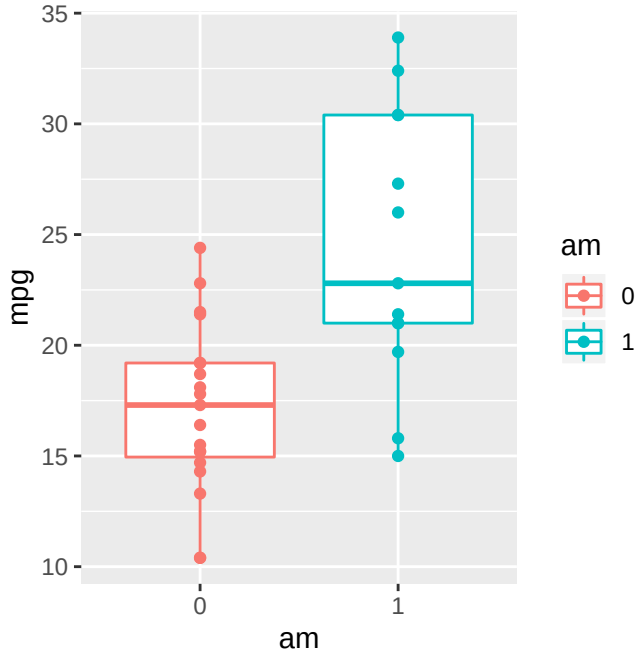
Simplifying the model based on physics

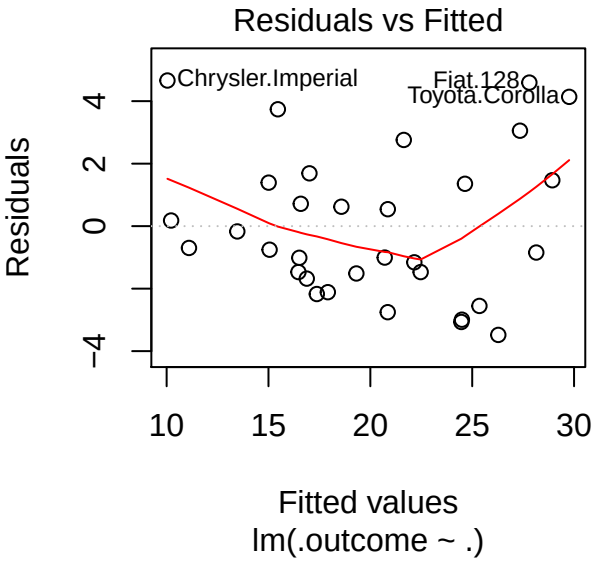
The previous analysis shows that based on several models fitted, the average saving in mpg is between 1.4 and 2 mpg as we go from manual to automatic transmission. The linear model and elastic net both suggest that mpg increase with transmission type. Let us now try a simple model based on physical sense only. In my experience, using the disp, weight, qsec, hp and transmission are enough to build a physically meaningful model.

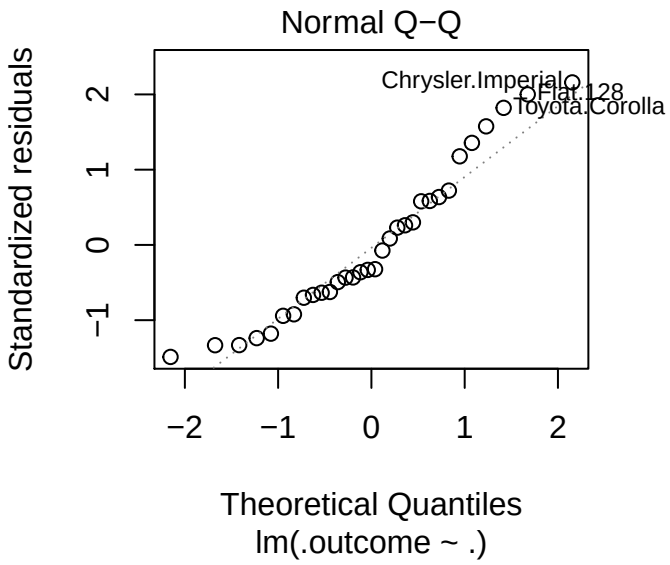
Conlsusion

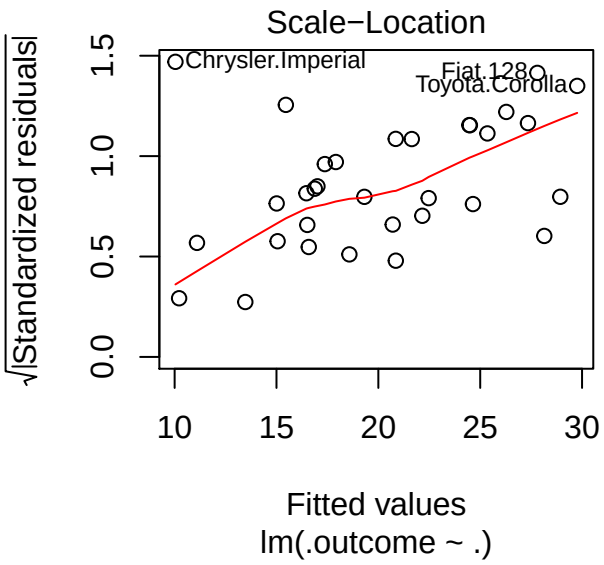
Buying a car with automatic transmission will be more cost saving for the user with about 1.4 to 2 mpg.

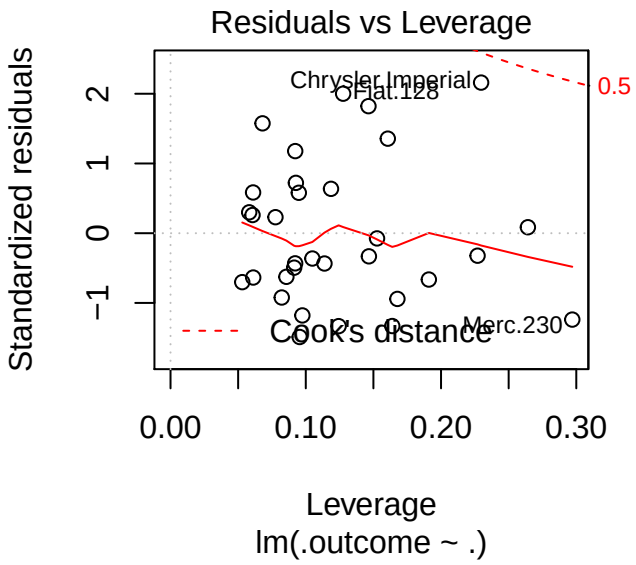
Appendix


mpg
35
30
25
20
15
10
0
1
am
am
0
1


Residuals vs Fitted
Residuals
Chrysler.Imperial
Fiat.128
Toyota.Corolla
4
2
0
−4
10
15
20
25
30
Fitted values
lm(.outcome ~ .)


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


Scale−Location
√|Standardized residuals|
Chrysler.Imperial
Fiat.128
Toyota.Corolla
1.5
1.0
0.5
0.0
10
15
20
25
30
Fitted values
lm(.outcome ~ .)


Residuals vs Leverage
Standardized residuals
Chrysler Imperial
Fiat.128
Merc.230
0.5
Cook's distance
0.00
0.10
0.20
0.30
Leverage
lm(.outcome ~ .)