Regression Models - Course Assignment

Tomás A. Maccor

25/2/2020

Executive Summary

"Motor Trend", the premier automobile industry magazine, once again delivers one of its special reports!

This time we analyse the relationship between Miles per Galon (MPG) in Automatic versus Manual transmission motor vehicles. We reviewed a comprehensive database of representative vehicles around the world, and a after running a series of models to fit the data, we conclude that:

- Vechicles with Manual transmission deliver more MPG than the ones that have Automatic transmission.
- The quantifiable difference is 11.06 more MPG for Manual transmission vehicles, ADJUSTED by the vehicle's weight (see details in "MODELLING" section of this report).

Exploratory Data Analysis

The dataset has 32 vechicles and includes 11 associated variables. All dataset variables are NUMERIC. We rename the variable that containg the tranmission information, to "TRANSMISSION", for easy of understanding at follow up & in plots.

Transmission = 0 equals an AUTOMATIC Transmision
Transmission = 1 equals MANUAL Transmision

We first do an exploratory plot (see Appendices, Plot #1), where we see a visual difference of approximately 5 MPG better performance by MANUAL Tranmission automoviles -corresponding specifically to the vehicles in the dataset.

Statistical Analysis

We now check for a statistically significant difference between the 2 transmission types:

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

And we see there is a statistically significant difference between the means of these 2 groups (p = 0,001374).
Manual Tranmission = 17.147 MPG
Automatic = 24.392 MPG

Modelling

We will try to find the best model that explains the data, and best quantifies the difference in MPG for these 2 vehicle POPULATIONS.

1. Simplest model

We will firstly run a model that only uses the TRANMISSION to explain the data:

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   17.147368   1.124603 15.247492 1.133983e-15
## Transmission1  7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] "R Squared ="       "0.359798943425465"
```

This model only explains explains 36% of total variability (R2 = 0.3598). In addition, the residuals plot (see Plot #2 in Appendices) shows a very clear pattern of aggregation along the X axis, which is not supposed to be.

2. Model with ALL VARIABLES

```
##                  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
## Transmission1  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
```

This model is much better, it explains 87% of total variability: R2 = 0.87, and the RSE is lower (2.65) than for our 1st model. BUT....none of the model's coefficients are statistically significant. There's just too many predictors that have been inserted into the model.

If we run a correlation between all variables in the dataset:

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

We can see that the variables most correlated with MPG are CYL (# of cilynders), DISP (Displacement), wt (WEIGHT) & HP (horsepower).

Thus, our next model will only have these variables + TRANSMISSION type:

This **3rd model** explains 86% of total variability. And RSE (residual variation) is LOWER than for the 2nd model: RSE = 2,50. BUT...the only statistically significant coefficient is the WEIGHT (and TRANSMISSION is NOT)

After researching some of our other sources, we see that most often automatic transmission vehicles are heavier that the ones with manual transmission (due to the characteristics of the transmission itself). Thus, the TRANSMISSION variable interacts with the WEIGHT variable (since if Automatic Transmission vehicle, its weight will most likely be heavier), thus we will run a **4th model** including the INTERACTION TERM between Transmission & Weight:

```
##
## Call:
## lm(formula = mpg ~ Transmission + Transmission:wt + cyl + disp +
##     hp + wt, data = motorcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4690 -1.5654 -0.6272  1.3389  5.1104
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.554662   3.928921   8.540 6.99e-09 ***
## Transmission1    11.058385   4.274869   2.587   0.0159 *
## cyl              -0.880560   0.632187  -1.393   0.1759
## disp              0.001562   0.011741   0.133   0.8952
## hp               -0.013111   0.014342  -0.914   0.3693
## wt               -2.292967   1.132666  -2.024   0.0537 .
## Transmission1:wt -3.642843   1.557480  -2.339   0.0276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.314 on 25 degrees of freedom
## Multiple R-squared:  0.8811, Adjusted R-squared:  0.8526
## F-statistic: 30.89 on 6 and 25 DF,  p-value: 2.124e-10
```

The result is a model with even better fit (R2 = 88%) and better residual fit (RSE = 2,31), PLUS is statistically significant for the "Transmission" & the "Transmission:wt" coefficients. If we do the residuals plot, all seems adequate (no distinct patterns, see Plot #3)

Finally, if we do an ANOVA test (see APPENDICES) the results tell us **we have found the best model** to explain data and obtain insights.

To summarise:
If we only consider Transmission as the sole predictor of MPG, the difference between Automatic vs. Manual tranmission that we can expect to obtain is 7,24 MPG with Manual Transmission.

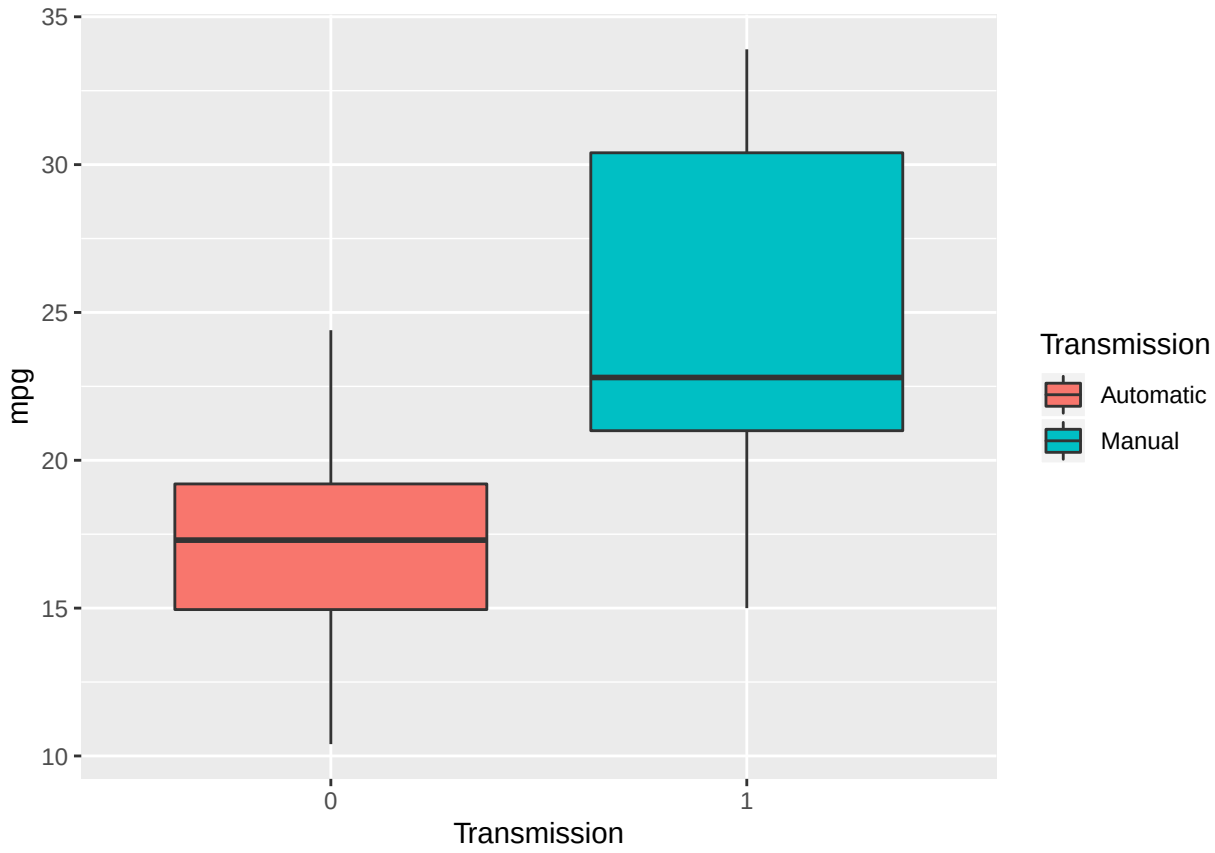
With the more robust model (more accurate) selected, we can infer (for the POPULATIONS of cars with Manual vs. Automatic transmissions), that the increase in MPG when using Manual Transmission is 11.06, adjusted negatively by 3.64 times the car's weight (in pounds / 1000), and with the following 95% confidence intervals:

```
## Transmission =

## [1]  2.254128 19.862643

## Weight =

## [1] -6.8505340 -0.4351523
```

LOGISTIC or POISSON models are not applicable in this case –the outcome variable (MPG) is not binomial neither count data

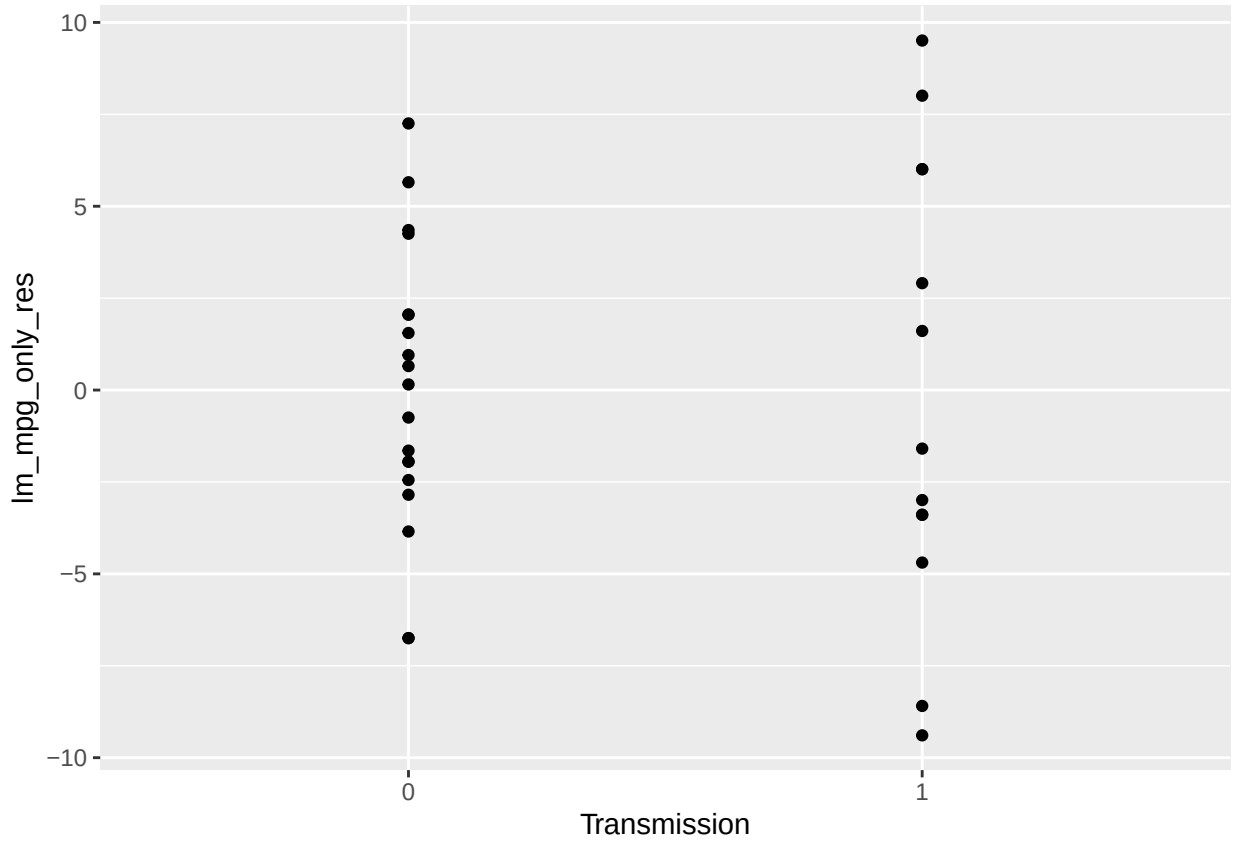
APPENDICES

PLOT #1



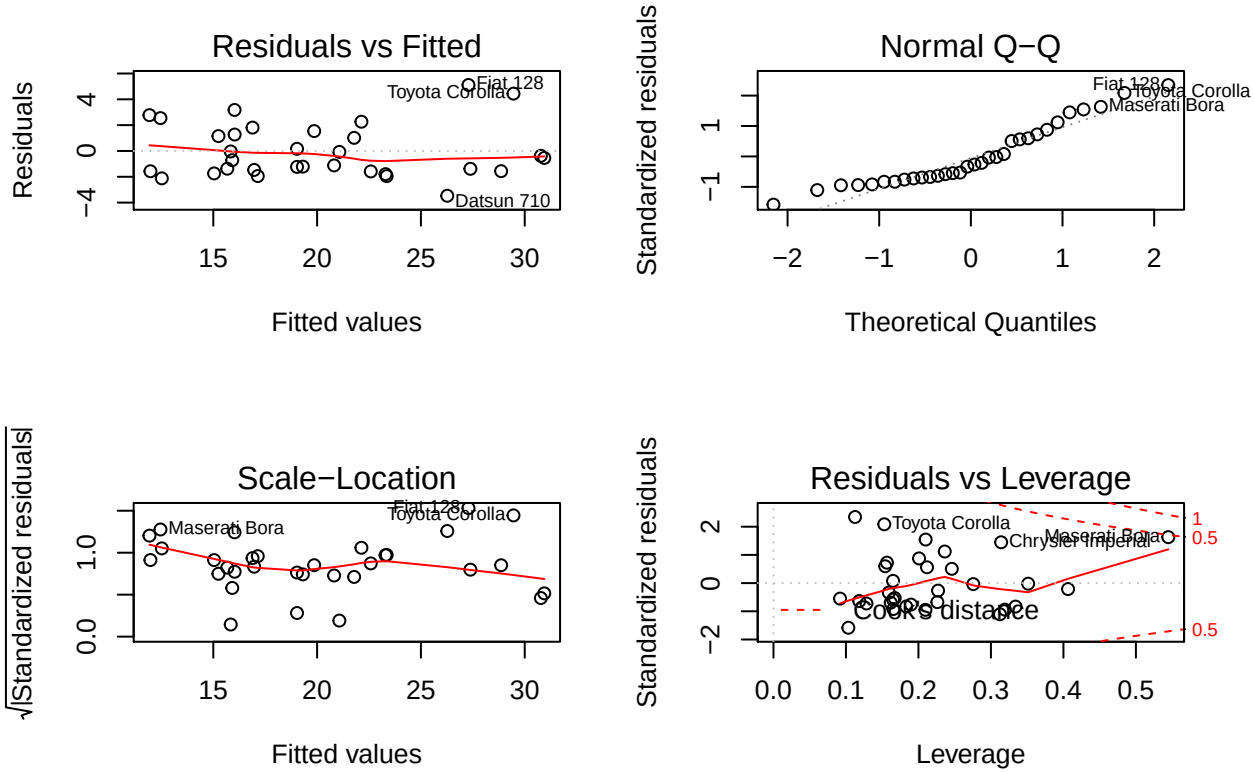
PLOT #2

Residual plot.
Model: only Transmission as predictor



PLOT #3

Residual plot of final model chosen



ANOVA for all models run

```
## Analysis of Variance Table
##
## Model 1: mpg ~ Transmission
## Model 2: mpg ~ Transmission + Transmission:wt + cyl + disp + hp + wt
## Model 3: mpg ~ Transmission + cyl + disp + hp + wt
## Model 4: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + Transmission +
##     gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     25 133.83  5    587.06 16.7170 1.067e-06 ***
## 3     26 163.12 -1    -29.29  4.1697   0.05391 .
## 4     21 147.49  5     15.63  0.4449   0.81206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```