Peer-graded Assignment: Regression Models Course Project

Alejandro Osorio

July 12, 2018

Excecutive Summary

Fuel efficiency may depend on many variables, out of which the type of transmission may not be statistically relevant among them. In spite of that, and given the need to have a reasonable understanding of the impact that specific variable has in gas consumption, the problem was approached through linear regression, imposing 'am' (type of transmission) as a binary variable and 'mpg' as outcome. A statistically significant expected increase on intercept values due to a specific type of transmission would answer the question "Is an automatic or manual transmission better for MPG?". Furthermore, its confidence interval would answer the question "Quantify the MPG difference between automatic and manual transmissions".

Based on a brief automotive industry research, plus ANOVA analysis, 2 extra main regressors were added to the equation. Those were 'wt' and 'hp'. Interestingly enough, both 'wt' and 'hp', build up the well known 'Weight/Power' KPI from the automotive industry.

Two types of models were used for the analysis (with and without interaction between the binary factor 'am' and the other regressors). The results obtained from each model, were as follows:

1. Model 1: Eventhough the result suggested an increase on intercept values when the transmission was manual, its confidence interval included the value 0. Therefore this model didn't help answer the central question.
2. Model 2: Suggested a statistically significant positive impact of a manual transmission on the outcome (an expected increase on intercept values when the transmission is manual). Refer to Confidence Intervals for detailed results.

Further leverage analysis is recommended though, due to the probable impact of a couple of outliers in the model. Additionally, given the fact that the regression was based on linear models, further analysis with non linear models is also recommended. Finally, the small size of the dataset must also be taken into account before jumping into more serious conclusions.

More detailed analysis and results are presented for each of the following steps that were taken in order to obtain the final results: Preliminary data analysis - Model selection - Preliminary regressors - Regression modeling - Regression diagnostics - Confidence intervals.

Preliminary Data Analysis

We're talking about a small sample of only 32 observations, with 11 variables. In spite of that, visual correlations were clearly observed between the outcome and many potential regressors (such as the inverse relation between 'mpg' and both 'hp' or 'wt') are pretty clear, as seen in Appendix 1.

Model Selection

Given the nature of the analysis (continuous outcome, obtained from discrete and continuous regressors), the small size of the dataset, plus the many visual correlations observed between the outcome and its potential regressors, the model chosen was of a linear (lm) type, with 'am' (type of transmission) as the main binary-factor variable, among any other additional statistically significant regressors.

Preliminary Regressors

A first analysis considered all mtcars variables as potential regressors. With variables 'cyl', 'vs', 'am', 'gear' and 'carb' as factors, the lm function "fitAll <- lm(mpg ~ ., mtcars)" was analysed using summary(fitAll)$coef.

With a P-value based criteria (refer to Appendix 2), no variable would make the cut (with 'hp' and 'wt' the closest, though). Therefore, further industry research (such as http://www.driverside.com/auto-library/top_10_factors_contributing_to_fuel_economy-317 and https://www.quora.com/On-what-factors-does-mileage-of-a-vehicle-depend) was required in order to determine the best regressor candidates for 'mpg' outcome. That way, the variables obtained as candidates for main regressors, were: displacement ('disp'), power ('hp'), weight ('wt') and number of forward gears ('gear').

Regression Modeling

For the regression analysis, two types of models were used.

1. Model 1: Consisted on different intersection points but the same slope for all regression lines.
2. Model 2: The second one, with different intersects AND slopes for regression lines (such as factor(am):regressor type interactions).

Additionally, ANOVA comparisons (incrementally adding regressors with all possible combinations) were used in combination with P-Value analysis of their coefficients (refer to Appendix 3). The final models obtained were:

1. Model 1: lm(formula = mpg ~ hp + wt + factor(am), data = mtcars)
2. Model 2: lm(formula = mpg ~ hp + wt + factor(am) + factor(am):wt, data = mtcars)

Regression Diagnostics

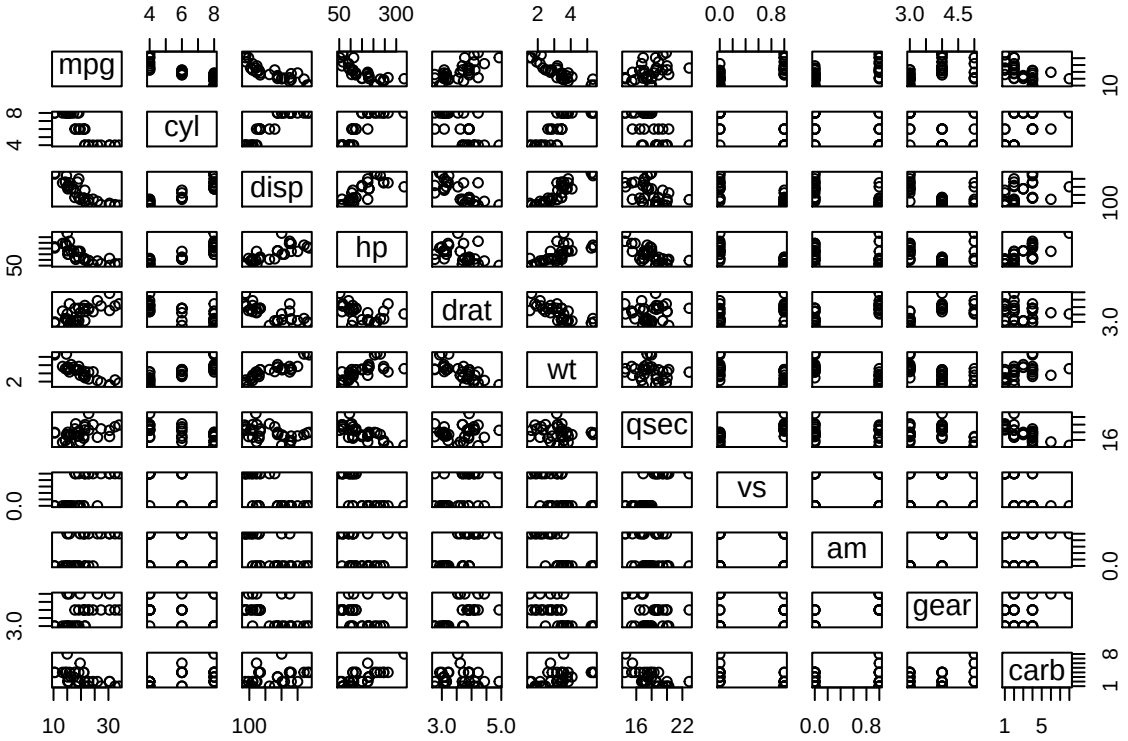
Both Model 1 and Model 2 redisual analysis (refer to Appendix 4) look good enough (no visible unbalanced patterns). Normal distributions of residuals, as well as their leverages, look a bit off though. Further analysis of cases such as Chrysler Imperial (Model 1) and Masseratti Bora (Model 2) is recommended.

Confidence Intervals

1. Model 1: As seen in Appendix 5, with a 95% confidence, we estimate that cars with a manual transmission result in a -0.74 to 4.9 general impact in MPG, compared to those with automatic transmission, at the average value of any of this model's regressors ('wt' or 'hp').
2. Model 2:

- In this case, with a 95% confidence, we estimate that cars with a manual transmission result in a 3.3 to 19.8 general impact in MPG, compared to those with automatic transmission, at the average value of any of this model's regressors ('wt' or 'hp').
- With a 95% confidence, we also estimate that cars with a manual transmission result in an extra -6.53 to -0.618 impact in MPG for each 1000 lbs increase in weight.

Appendix 1: Preliminary Data Analysis



Appendix 2: Preliminary Regressors

```
fitAll <- lm(mpg ~ ., mtcars2)
summary(fitAll)$coef
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## am1          1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
```

```
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

Appendix 3: Regression Modeling

Model 1: Different intersection points, with same slopes

ANOVA Analysis:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ wt + factor(am)
## Model 3: mpg ~ wt + disp + factor(am)
## Model 4: mpg ~ wt + hp + disp + factor(am)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 66.421 9.394e-09 ***
## 3     28 246.56  1     31.76  4.767  0.037878 *
## 4     27 179.91  1     66.65 10.002  0.003842 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficients' Analysis:

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 34.209443370 2.82282610 12.1188632 1.979953e-12
## wt          -3.046747000 1.15711931 -2.6330448 1.382936e-02
## hp          -0.039323213 0.01243358 -3.1626624 3.842032e-03
## disp         0.002489354 0.01037681  0.2398959 8.122229e-01
## factor(am)1  2.159270737 1.43517565  1.5045341 1.440531e-01
```

Model 2: Different intersection points and slopes

ANOVA Analysis:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ wt + factor(am) + factor(am):wt
## Model 3: mpg ~ hp + wt + factor(am) + factor(am):hp + factor(am):wt
## Model 4: mpg ~ wt + hp + factor(gear) + factor(am) + factor(am):wt + factor(am):hp +
##     factor(am):factor(gear)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 188.01  2    532.89 49.5549 3.013e-09 ***
## 3     26 135.90  2     52.11  4.8456   0.01707 *
## 4     24 129.04  2      6.86  0.6378   0.53720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

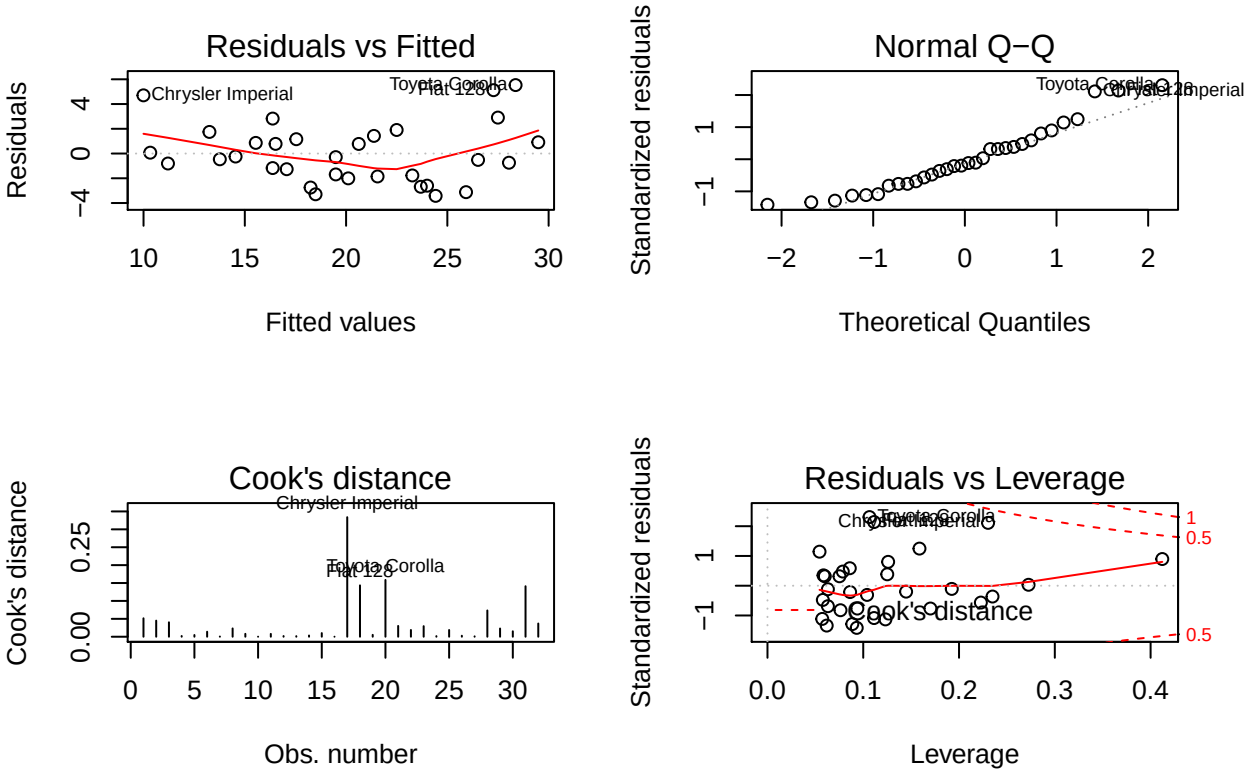
Coefficients' Analysis:

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    30.70392721 2.67515435 11.477441 1.117089e-11
## hp             -0.04094406 0.01362921 -3.004142 5.826559e-03
```

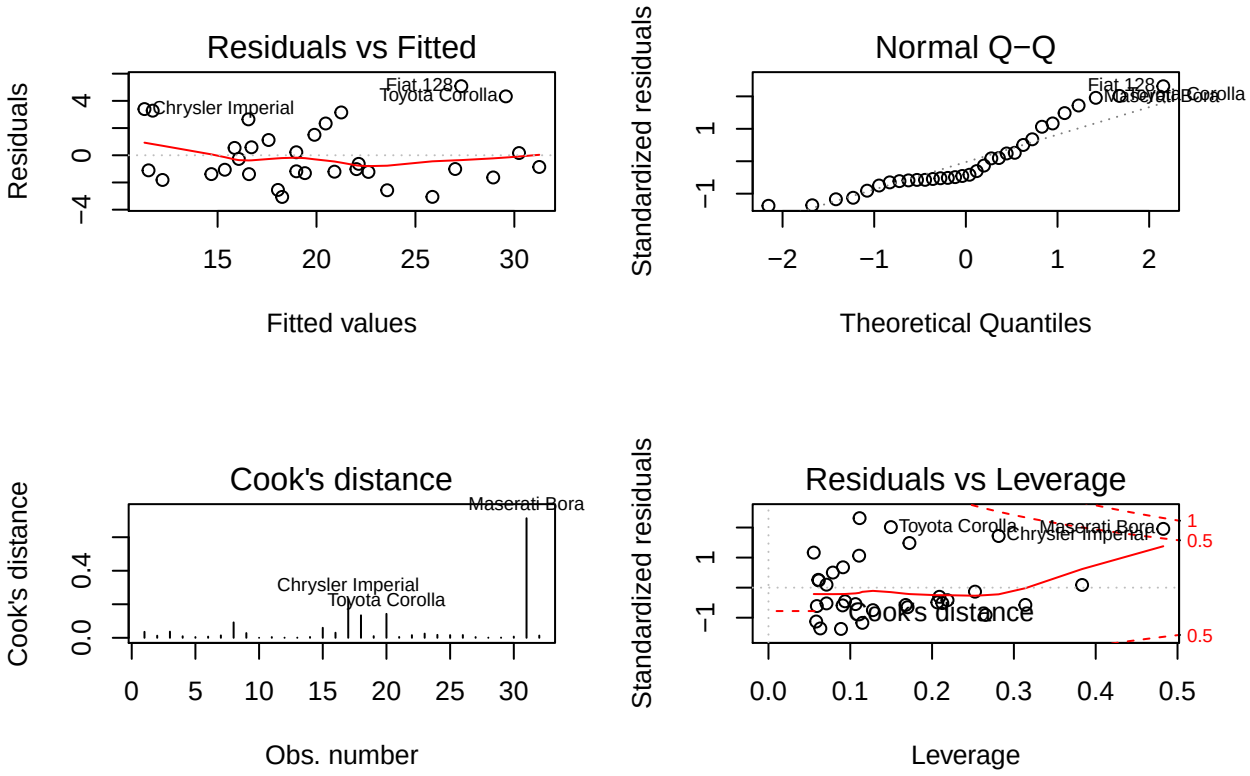
```
## wt             -1.85591121 0.94510642 -1.963706 6.034159e-02
## factor(am)1    13.74000384 4.22337051  3.253327 3.155621e-03
## hp:factor(am)1  0.02779357 0.01920705  1.447050 1.598330e-01
## wt:factor(am)1 -5.76894729 2.07200930 -2.784228 9.870579e-03
```

Appendix 4: Regression Diagnostics

Model 1: Different intersection points, with same slopes



Model 2: Different intersection points and slopes



Appendix 5: Confidence Intervals

Model 1: Different intersection points, with same slopes

Coefficients:

```
sumCoef1 <- summary(model1)$coef
sumCoef1
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## factor(am)1  2.08371013 1.376420152  1.513862 1.412682e-01
```

Confidence interval for the intercept once we include the factor that the transmission is manual (without changing the slopes), is:

```
sumCoef1[4,1] + c(-1,1) * qt(.975, df = model1$df) * sumCoef1[4,2]
```

```
## [1] -0.7357587  4.9031790
```

Model 2: Different intersection points and slopes

Coefficients:

```
sumCoef2 <- summary(model2)$coef
sumCoef2
```

```
##                   Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)    30.94733319 2.723410935 11.363446 8.546944e-12
## hp             -0.02694935 0.009795903 -2.751084 1.047673e-02
## wt             -2.51558550 0.844496532 -2.978799 6.051842e-03
## factor(am)1    11.55481296 4.023276579  2.871991 7.844579e-03
## wt:factor(am)1 -3.57790980 1.442795585 -2.479845 1.967639e-02
```

Confidence interval for the intercept once we include the factor that the transmission is manual (without changing the slopes), is:

```
sumCoef2[4,1] + c(-1,1) * qt(.975, df = model2$df) * sumCoef2[4,2]
```

```
## [1]  3.299731 19.809895
```

Additionally, the confidence interval for the wt:factor(am)1 interaction, is:

```
sumCoef2[5,1] + c(-1,1) * qt(.975, df = model2$df) * sumCoef2[5,2]
```

```
## [1] -6.5382818 -0.6175378
```