

Peer-graded Assignment: Regression Models Course Project

Alejandro Osorio

June 29, 2018

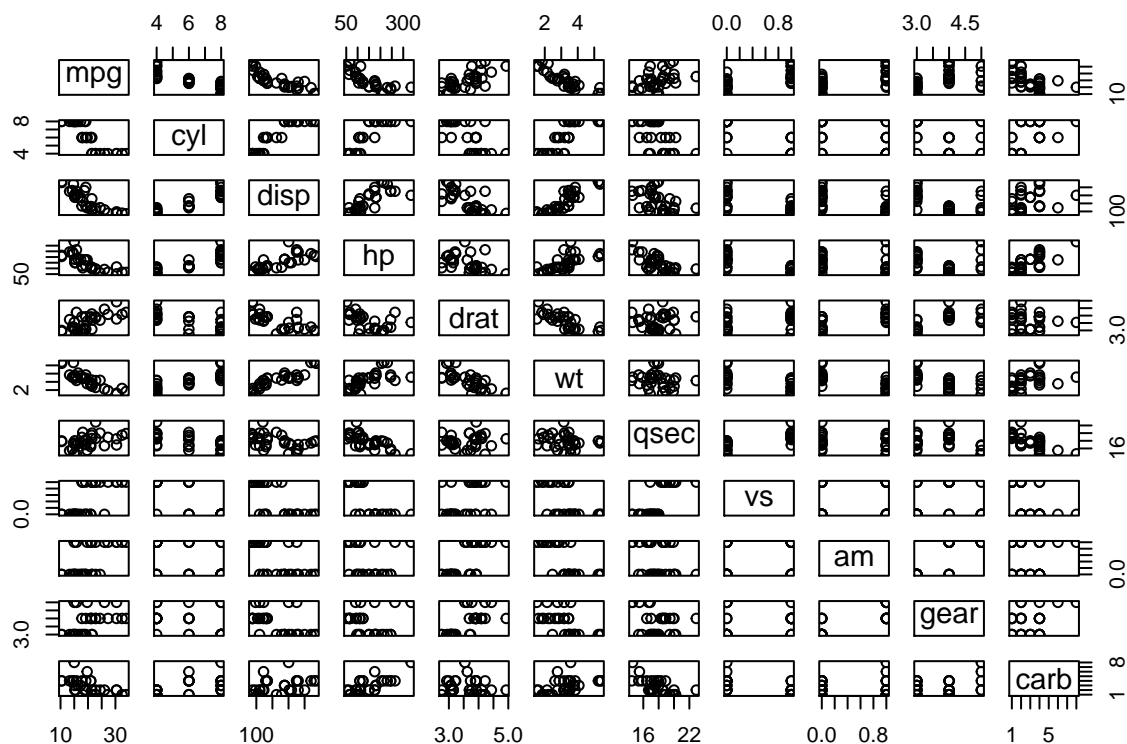
Appendix 1: Preliminary Data Analysis

Basic properties of the 'mtcars' dataset:

```
## '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 ...
```

So, we're talking of a small sample of only 32 observations.

Additionally, visual correlations among pairs of potential variables can be seen as follows:



Appendix 2: Model Selection

Given the nature of the analysis (continuous outcome, obtained from discrete and continuous regressors), the size of the dataset (only 32 observations), plus the visual correlations observed between ‘mpg’ (the outcome) and its potential regressors, the model to be used will be linear (lm type), with ‘am’ as a not-so-dummy binary-factor variable.

Appendix 3: Choosing Preliminary Regressors

A preliminary analysis considered all mtcars variables as potential regressors. With variables ‘cyl’, ‘vs’, ‘am’, ‘gear’ and ‘carb’ as factors, the following lm function was analysed:

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

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	23.87913244	20.06582026	1.19004018	0.25252548
## cyl16	-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

With a P-value based criteria, no variable makes the cut (with ‘hp’ and ‘wt’ the closest, though). Therefore, the first conclusion was that some industry research was required in order to determine the best regressor candidates for ‘mpg’ outcome.

After some web 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>), the main variables suggested (and therefore candidates for main regressors), were: displacement (‘disp’), power (‘hp’), aerodynamics, weight (‘wt’) and number of forward gears (‘gear’).

Besides aerodynamics (not available in the dataset), the visual correlations between them and ‘mpg’ are pretty clear, as seen in Appendix 1 (such as the inverse relation between ‘mpg’ and both ‘hp’ or ‘wt’ -both of which had the lowest P-values when including all variables in the model, as seen before-). Therefore, including ‘am’ (required in order to answer the main question), they were selected as candidates for regression modeling in the next appendix.

Finally, given the lowest P-value results obtained by ‘wt’ and ‘hp’ regressors (in comparison with the rest), plus the well known Weight/Power KPI (Key Performance Indicator) from the automotive industry, in further analysis both ‘wt’ and ‘hp’ will be considered primary potential regressors while ‘disp’ and ‘gear’ will be considered secondary potential regressors (just to prioritize further analysis. Quantitative analysis will finally determine which regressors to keep).

Appendix 4: Regression Modeling

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

The first one with different intersection points but the same slope for all regression lines (no interaction between ‘am’ -the binary factor regressor- and the rest of the regressors), such as:

```
## lm(formula = mpg ~ hp + wt + factor(am), data = mtcars)
```

The second one, with different intersects AND slopes for regression lines (factor(am):regressor type interactions), such as:

```
## lm(formula = mpg ~ hp + wt + factor(am) + factor(am):hp + factor(am):wt,
## data = mtcars)
```

Additionally, ANOVA comparisons (incrementally adding regressors from the preliminary set, defined in Appendix 3) were used in order to determine the best regressor mix for each model. The order in which regressor candidates were added, followed the primary and secondary criteria specified in Appendix 3. Therefore, the regressor sequences to be considered, were the following (all with ‘am’ included):

1. ‘wt’ - ‘hp’ - ‘disp’
2. ‘wt’ - ‘hp’ - ‘gear’
3. ‘hp’ - ‘disp’ - ‘gear’
4. ‘wt’ - ‘gear’ - ‘disp’

Model 1: Different intersection points, with same slopes

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ wt + factor(am)
## Model 3: mpg ~ hp + wt + 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 66.4206 9.394e-09 ***
## 3      28 180.29  1     98.03 14.7118 0.0006826 ***
## 4      27 179.91  1      0.38  0.0576 0.8122229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test with 2nd sequence (replacing 'disp' by 'gear') resulted in almost same outcome.

Just in case, ANOVA analysis of the first sequence was carried on, reordering the sequence as follows: 'wt' - 'disp' - 'hp'

```
## 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
```

Interestingly enough, the result suggests that all 4 regressors are relevant for the outcome. Same test was carried on with the sequence number 2, with less promising results.

Sequences 3 and 4 were not relevant as soon as regressors 'disp' or 'gear' were included (including sequence reordering).

Therefore, model 4 passed the test for final P-value analysis of its coefficients, as follows:

```
##               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
```

Given the P-value > 0.05 for the 'hp:factor(am)'disp' regressor ('am' regressor is included anyway, being required as binary factor), the final model is as follows:

```
## lm(formula = mpg ~ hp + wt + factor(am), data = mtcars)
```

Model 2: Different intersection points and slopes

ANOVA analysis to sequences 1 and 2 suggests 'wt', 'hp' and 'am' are the only relevant regressors (including 'disp' or 'gear' in any order, didn't make a difference) when including variable slopes.

```
## 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
```

Therefore, model 3 passed the test for final P-value analysis of its coefficients, as follows:

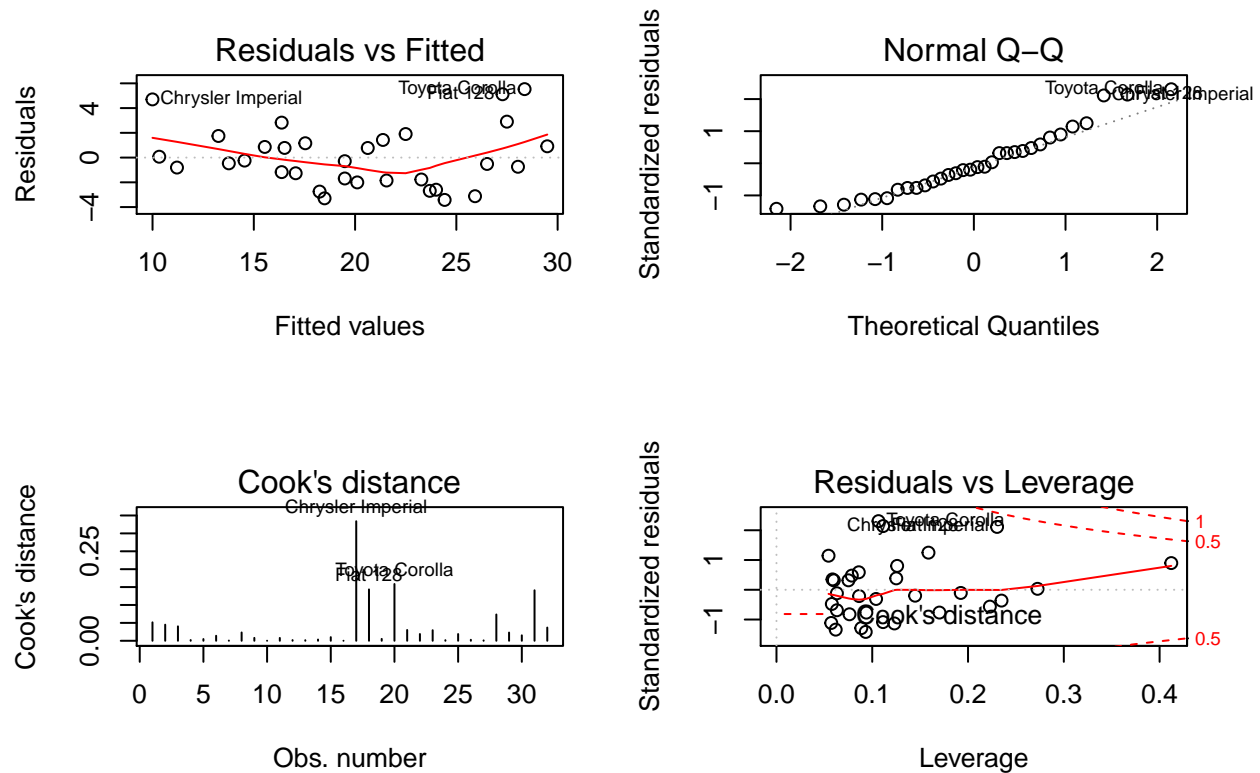
```
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  30.70392721 2.67515435 11.477441 1.117089e-11
## hp          -0.04094406 0.01362921 -3.004142 5.826559e-03
## wt          -1.85591121 0.94510642 -1.963706 6.034159e-02
## factor(am)1  13.74000384 4.22337051  3.253327 3.155621e-03
## 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
```

Given the P-value > 0.05 for the hp:factor(am) coefficient, the final model is as follows:

```
## lm(formula = mpg ~ hp + wt + factor(am) + factor(am):wt, data = mtcars)
```

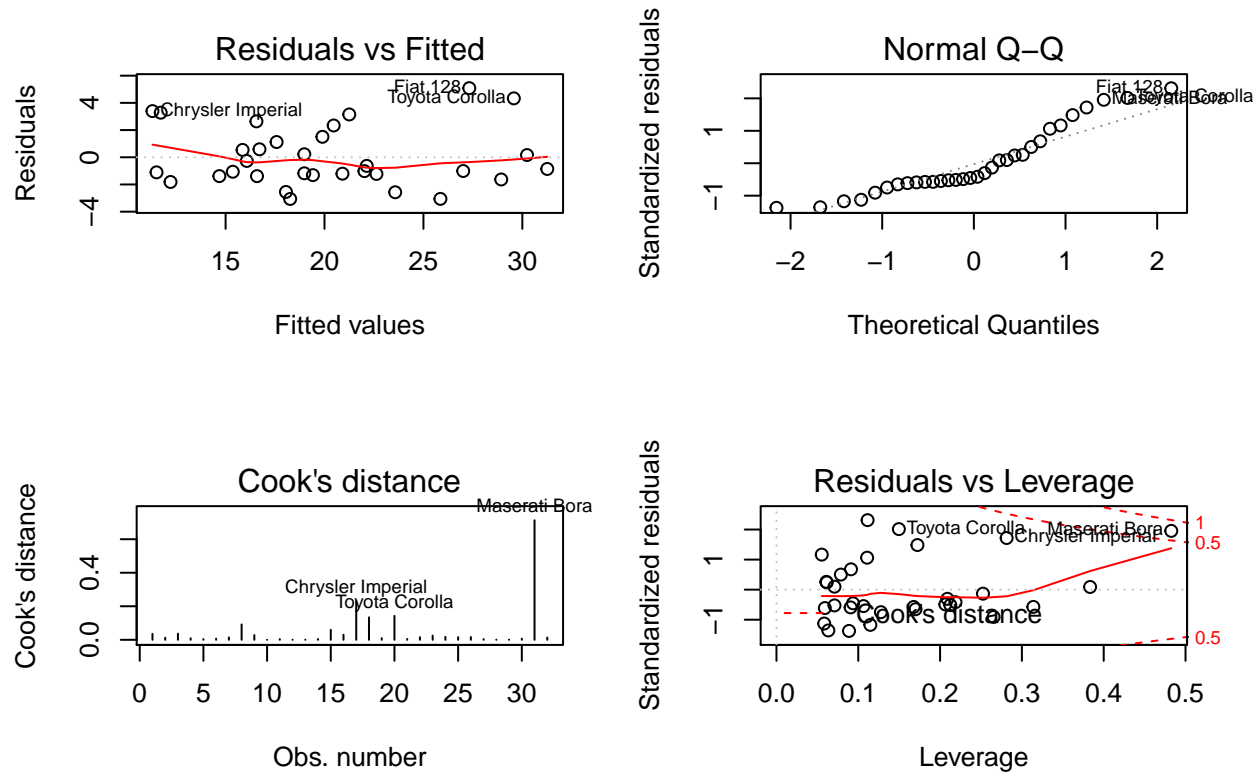
Appendix 5: Regression Diagnostics

Model 1: Different intersection points, with same slopes



Residual analysis looks good enough (no visible unbalanced patterns). Normal distribution of residuals, as well as their leverage, look a bit off though. It probably requires further analysis of cases such as Chrysler Imperial.

Model 2: Different intersection points and slopes



Residual analysis looks good enough (no visible patterns). Normal distribution of residuals, as well as their leverage, look a bit off though. It probably requires further analysis of cases such as Maserati Bora.

Appendix 6: Confidence Intervals

Model 1: Different intersection points, with same slopes

```
##           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
```

Considering that the 4th row represents the impact on the intercept once we include the factor that the transmission is manual (without changing the slopes), the confidence interval for that impact is:

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

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

Which means that, 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').

Model 2: Different intersection points and slopes

```
##              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
```

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

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

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'). Additionally, considering the wt:factor(am)1 interaction:

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

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

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.