Regression Models Extended Project

Ignacio Ojea

July 28, 2018

Executive Summary

This report is a course project within the Regression Models course on the Data Science Specialization by Johns Hopkins University on Coursera. I will examine the mtcars data set and explore how miles per gallon (MPG) is affected by different variables. In particularly, we will answer the following two questions: (1) Is an automatic or manual transmission better for MPG, and (2) Quantify the MPG difference between automatic and manual transmissions.

From the analysis I can show that manual transmission has an mpg **1.8** greater than an automatic transmission. Nevertheless, results show that it is **not** significant.

Exploratory data analisis

```
data("mtcars")
?mtcars
# am variable accounts for transmission, and mpg for miles per gallon.
# am, cyl, vs, gear and carb are factors
mtcars2 <- mtcars #para la correlation table later
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
cols <- c("cyl", "vs", "gear", "carb")
mtcars[cols] <- lapply(mtcars[cols], factor)
fit <- lm(mpg ~am, mtcars)
summary(fit)$coefficients
```

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

```
summary(fit)$r.squared
```

```
## [1] 0.3597989
```

See appendix for exploratory analysis plot. From the plot, as well as the first exploratory linear regression **fit**, the answer to the first question seems positive (coefficient is positive, and p-value indicates significance). But we need to include other variables in order to avoid bias, since the R squared value is 0.36 thus telling us this model only explains us 36% of the variance.

Regression Models

Strategy for model selection

To begin, the type of data does not suggest the need to use either binomial or Poisson generalized linear models, since no variables correspond to binary (or sets of binary) outcomes, nor to counting, number of occurences, or rates.

Hence we need to do a regular multivariable regression. The question now is which are the relevant predictor variables to take into account. Too few will lead to bias, too many to an increase in variance. We need a strategy for model selection. We start with an initial model **fitall** that takes into account all variables, and perfom stepwise model selection to select significant predictors for the final model which is the best model **fitbest** using the *step* function in R, which builds several regression models and makes a selection using the AIC algorithm. I will hide the results for the sake of simplicity, the code is below.

```
fitall <- lm(mpg ~ ., data = mtcars)
fitbest <- step(fitall, direction = "both")
```

Since the assignment needs to be short, I will avoid studying variance inflation factors (via the *vif* function). Also, the *step* function implicitly exhausts a nested model strategy. Rather, I will *anova* to compare the three models we explored:

```
anova(fit,fitbest,fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3     15 120.40 11     30.62  0.3468    0.9588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for **fitbest** is significant, therefore we reject the null hypothesis that the added variables cyl, hp and wp are unnecessary. On the contrary, we see that **fitall** has a p-value larger than 0.05, which suggests that the added predictors may not be necessary.

Since the **fitbest** model is accurate, we want to explore how much of the variance it explicates. We do this computing R squared.

```
summary(fitbest)$r.squared
```

```
## [1] 0.8658799
```

```
summary(fitbest)$adj.r.squared
```

```
## [1] 0.8400875
```

We therefore see that the linear relationship explains about 86% percent of the variance.

Another strategy for model selection

Let's have a closer look to the corelations of 'mpg' to the other variables of "mtcars":

```
res <- cor(mtcars2)
round(res, 2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
```

```
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

According to the correlation table, there are at least four variables with a high correlation to our outcome variable "mpg". The highest value comes from the weight variable "wt". Pero ojo porque aca algunas variables estan consideradas como numericas y no como factors. Pero con esta estrategia un poco podes ver cuales son las variables mas relevantes.

Another strategy

We performed ANOVA (analysis of variance) to identify superfluous variables. The analysis pointed to cylinders, displacement, and weight as the only significant terms. Estos son las Pr(>F) de todas las variables.

```
anova(fitall) %>% .[5] %>% t
```

```
##                 cyl       disp        hp      drat         wt      qsec
## Pr(>F) 1.942616e-07 0.01713914 0.1497477 0.2419149 0.01869624 0.6691803
##               vs       am      gear      carb Residuals
## Pr(>F) 0.8487815 0.171354 0.7360567 0.8814442        NA
```

Inference

We also perform a t-test and we see that the manual and automatic transmissions are significatively different (by looking at p-value and confidence interval).

```
Automatic <- mtcars[mtcars$am == "Automatic",]
Manual <- mtcars[mtcars$am == "Manual",]
t.test(Automatic$mpg, Manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  Automatic$mpg and Manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

Residual diagnostics

See the appendix for Residual Plot.

```
cov(fitbest$residuals, hatvalues(fitbest))
```

```
## [1] -0.008314563
```

The points in the Residuals vs. Fitted plot, as well as the analysis of covariance suggest that the variables are independent, namely that there is not correlation between them, as desired. From the rest of the plots we also see that residuales all normally distributed and homoskedastic.

We also see that there are some outliers. I leave the study leverage (by looking at *hatvalue(fitbest)*) and influence (by looking at *dfbetas(fitbest)*) too avoid making the project too long.

More residual diagnostics

```
#One idea
leverage <- hatvalues(fitbest)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
influential <- dfbetas(fitbest)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

```
#Other idea
#Consistently with the residual plots, those points are the main responsibles for the
#deviation from the residual normality assumption which is fullfilled within a confidence
#interval of 0.05 as it can be seen below
influence <- sort(dffits(fitbest),decreasing=TRUE)
shapiro.test(fitbest$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitbest$residuals
## W = 0.96807, p-value = 0.4479
```

For this reason (the shapiro test is not significant), those higher-influence points do not pose particular problems as they do not invalidate our conclusions about the final model.

Conclusions

```
summary(fitbest)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
```

```
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

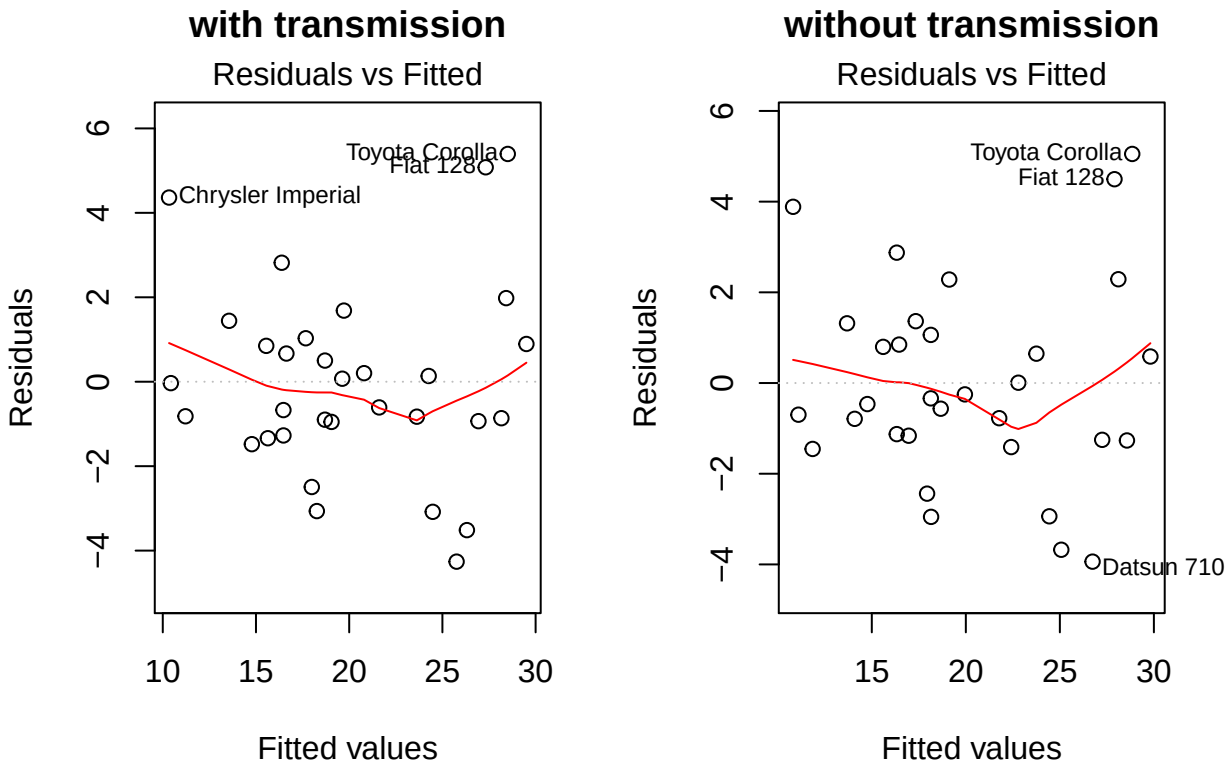
Based on the observations from our **fitbest** model, we can conclude the following,

- Cars with Manual transmission get more miles per gallon compared aganist cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt). mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
- mpg decreases negligibly with increase of hp.
- If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

A case for its irrelevance

To confirm and illustrate the limited impact that transmission has on mpg, we can make two regression models, one modeling mpg with cylinder number, weight, displacement, and transmission type, and one without transmission, and then graph them below. The rear axle ratio is excluded because, despite being accepted earlier in the analysis, it has a p-value well above the 5% threshold needed to be genuinely considered relevant.

```
fitnoam <- lm(mpg~ cyl + hp + wt,mtcars)
par(mfrow=c(1,2))
plot(fitnoam,which=1,main="with transmission")
plot(fitbest,which=1,main="without transmission")
```




```
summary(fitbest)$adj.r.squared
```

```
## [1] 0.8400875
```

```
summary(fitnoam)$adj.r.squared
```

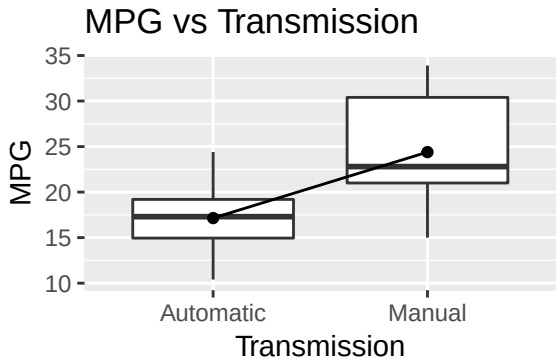
```
## [1] 0.8360668
```

On inspection, these models are virtually identical, so we can confidently say that transmission type has no significant impact on a car's mileage.

Appendix with plots

Exploratory analysis plot:

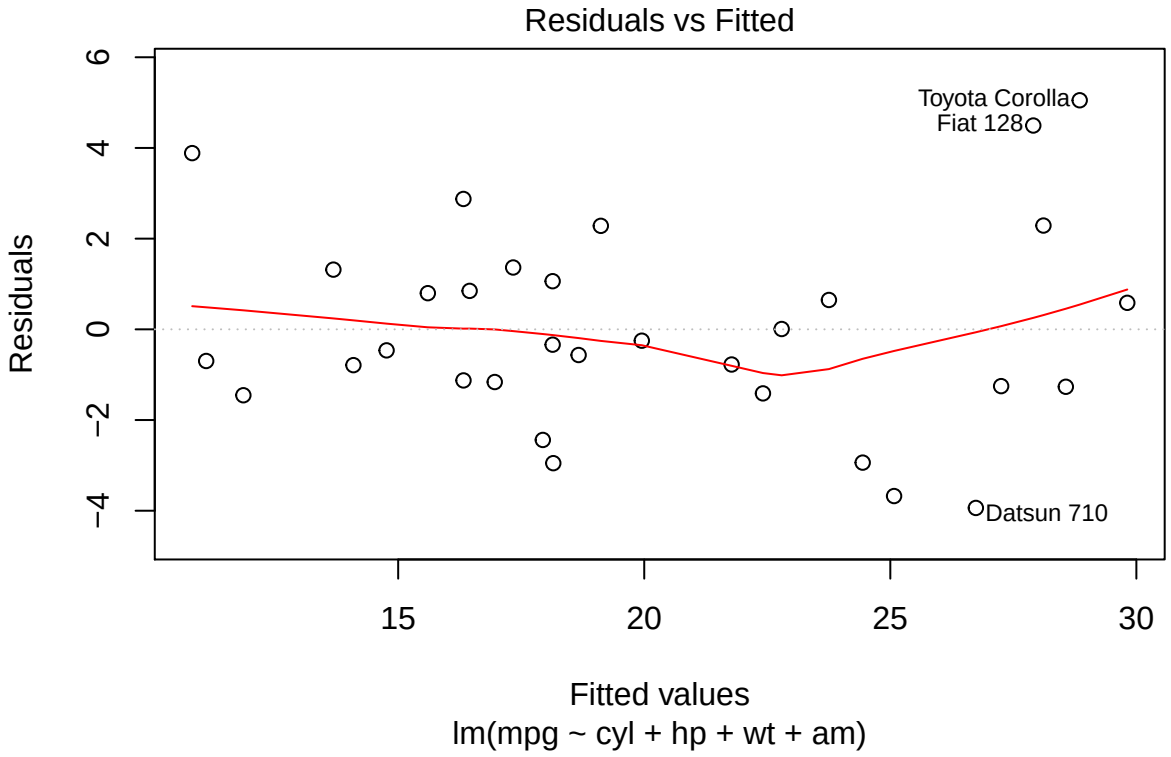
```
g <- ggplot(mtcars) + aes(x = factor(am), y = mpg) + geom_boxplot()
g <- g + stat_summary(fun.y=mean, geom="line", aes(group=1)) + stat_summary(fun.y=mean, geom="point") +
g
```



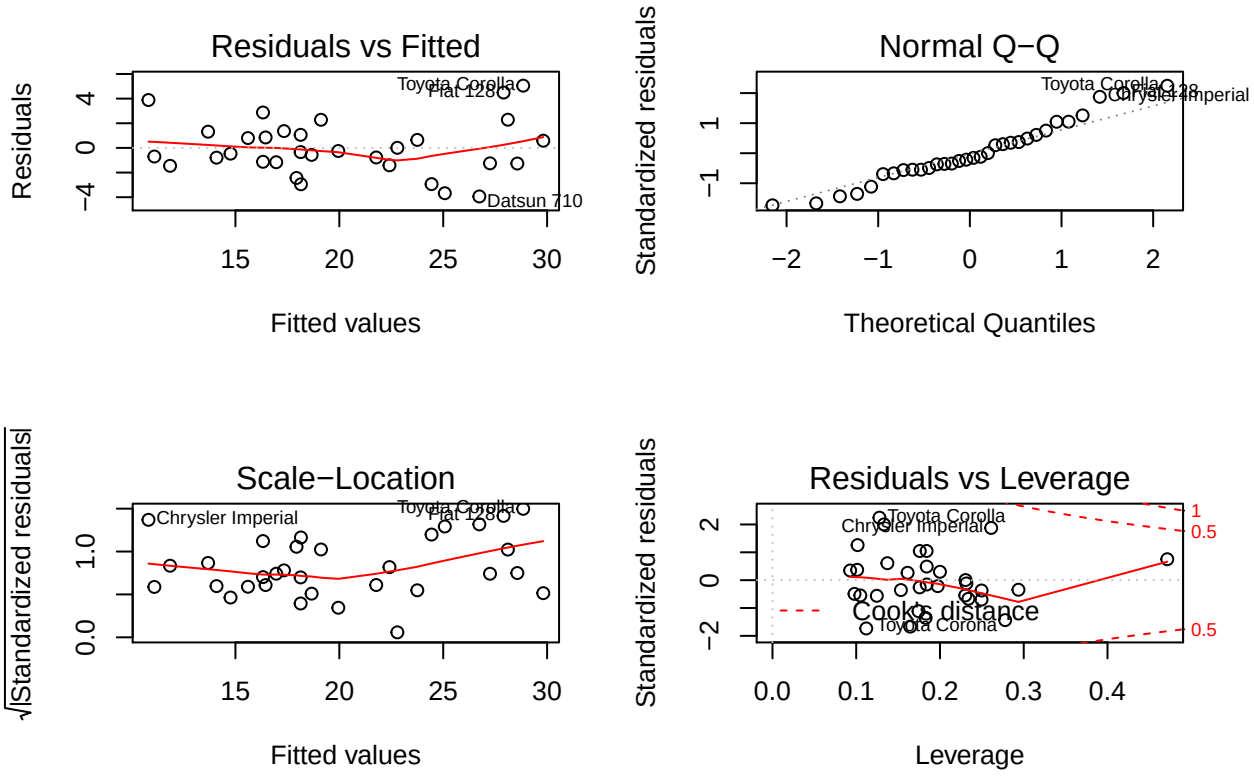
Residual Plots

Para plotear algunos y no todos los plots

```
plot(fitbest,which=1)
```




```
par(mfrow = c(2, 2))
plot(fitbest)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona