Regression models

Dr. Chem. Eng. Nikolaos Fotopoulos

Sunday, May 10, 2015

Course Project

The project's assignment for the Regression models Course https://class.coursera.org/regmods-014, consists in answering the following :

Given you work for Motor Trend, a magazine about the automobile industry. you are asked by examining a data set of a collection of cars, to answer first if Is an automatic or manual transmission better for MPG (mileage per gallon) and then to Quantify the MPG difference between automatic and manual transmissions

Source for this analysis can be found on Github at https://github.com/foton263/Regression_Models

Analysis

First we load the data set...

```
library(car)
library(ggplot2)
data(mtcars)
```

We examine the regressors classes, and we change the categorical variables to factors..

```
sapply(mtcars,class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        vs        am      gear      carb
## "numeric" "numeric" "numeric" "numeric"
```

```
mtcars$am<-as.factor(mtcars$am)
levels(mtcars$am)<-c("automatic", "manual")
mtcars$cyl<-as.factor(mtcars$cyl)
mtcars$vs<-as.factor(mtcars$vs)
mtcars$gear<-as.factor(mtcars$gear)
mtcars$carb<-as.factor(mtcars$carb)
```

For answering the first question of the project, about what is better for mileage (mpg), automatic or manual gear tranmission, we use a simple linear model.

```
automan<-lm(mpg~am -1,mtcars)
summary(automan)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## amautomatic   17.147      1.125   15.25 1.13e-15 ***
## ammanual      24.392      1.360   17.94  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

As it seems manual gear is superior / better / has a more positive effect / you drive longer per gallon, effect, than automatic gear. To ensure that this difference is significant and that we are legitimate to make the above inference, we compute the confidence intervals for the coeficients

```
confint(automan)
```

```
##                2.5 %   97.5 %
## amautomatic 14.85062 19.44411
## ammanual    21.61568 27.16894
```

As we can see at 95% confidence level, the coefficients' ranges do not overlap, so the difference between them is signifficant and we can reject the null hypothesis. The boxplots in the appendix support, visually, the above qualitative interpretation.

Next, for quantifying the automatic/manual gear effect on mileage (mpg), we need first to construct a parsimonious model, to ensure normality in residuals and then to examine the quantified effect of the manual gear transmision in the multivariate context.

We define a global model (mdl) on which we perform an anova and variance inflation factor analysis to decide about what variables to keep in our final model (fit)

```
mdl<-lm(mpg~. ,mtcars)
anova(mdl)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 51.3766 1.943e-07 ***
## disp       1  57.64   57.64  7.1813   0.01714 *
## hp         1  18.50   18.50  2.3050   0.14975
## drat       1  11.91   11.91  1.4843   0.24191
## wt         1  55.79   55.79  6.9500   0.01870 *
## qsec       1   1.52    1.52  0.1899   0.66918
## vs         1   0.30    0.30  0.0376   0.84878
## am         1  16.57   16.57  2.0639   0.17135
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(mdl)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

Significant variables are cyl (as intercept), disp and wt. Since we are interested in quantifying the effect of am on mpg we have to also include am and hp in our model and reexamine how variance inflation is modified for different models, before we choose the best one.

```
fit1<-lm(mpg~cyl+disp+wt+am ,mtcars)
fit2<-lm(mpg~cyl+disp+wt+am+hp ,mtcars)
fit3<-lm(mpg~cyl+wt+am+hp ,mtcars)
fit4<-lm(mpg~cyl+disp+am+hp ,mtcars)
fit5<-lm(mpg~cyl+wt+am ,mtcars)
anova(fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + wt + am
## Model 2: mpg ~ cyl + disp + wt + am + hp
## Model 3: mpg ~ cyl + wt + am + hp
## Model 4: mpg ~ cyl + disp + am + hp
## Model 5: mpg ~ cyl + wt + am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 182.87
## 2     25 150.41  1    32.461 5.3954 0.02862 *
## 3     26 151.03 -1    -0.617 0.1025 0.75149
## 4     26 183.04  0   -32.014
## 5     27 182.97 -1     0.071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(fit1)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## cyl   6.971234  2        1.624903
## disp 12.813023  1        3.579528
## wt    6.583720  1        2.565876
## am    1.931783  1        1.389886
```

```
vif(fit2)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## cyl   9.765272  2        1.767751
## disp 12.901490  1        3.591864
## wt    6.821979  1        2.611892
## am    2.590898  1        1.609627
## hp    4.736101  1        2.176258
```

The best model seems to be the fit2. We examine the summary of this model and we test for normality with shapiro test.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ammanual     1.806099   1.421079   1.271   0.2155
## hp          -0.032480   0.013983  -2.323   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

```
shapiro.test(fit2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.971, p-value = 0.5274
```

We can conclude that residuals follow normal distribution In the appendix, residual diagram and Cook distances are given for spoting outliers and leverage points (figures 2-7).

Finally, we calculate the confidence intervals for the models coefficients so to have an estimate of the quantitative effect the manual gear transmission has on mpg.

```
confint(fit2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 28.31296354 39.415588581
## cyl6        -6.16171468 -0.110418430
## cyl8        -8.68663174  3.251069157
## disp        -0.02220684  0.030382623
## wt          -5.16066572 -0.316723497
## ammanual    -1.12066818  4.732867169
## hp          -0.06127916 -0.003681192
```

Result

So, answering the second question we can state that the average overall effect of manual gear transmision system over mpg is +1.8 miles / gallon within [-1.1, 4.7] 95% confidence interval.

References

P. Teetor, **R Cookbook**, O'Reilly, 2011.

W. Chang, **R Graphics Cookbook**, O'Reilly, 2012.

J. Adler, **R In A Nutshell**, O'Reilly, 2012.

J. Faraway, **Practical Regression and Anova using R**,2002

may the R be with you...

Appentix

```
boxplot(mpg ~ am, data=mtcars,ylab = "MPG (miles per gallon)",col=c("cyan","yellow"))
```

```
plot(fit2$fitted.values,rstudent(fit2)); abline(0, 0)
```

```
plot(fit2,which=1)
```

```
plot(fit2,which=2)
```

```
plot(fit2,which=3)
```

```
plot(fit2,which=4)
```

```
plot(fit2,which=5)
```

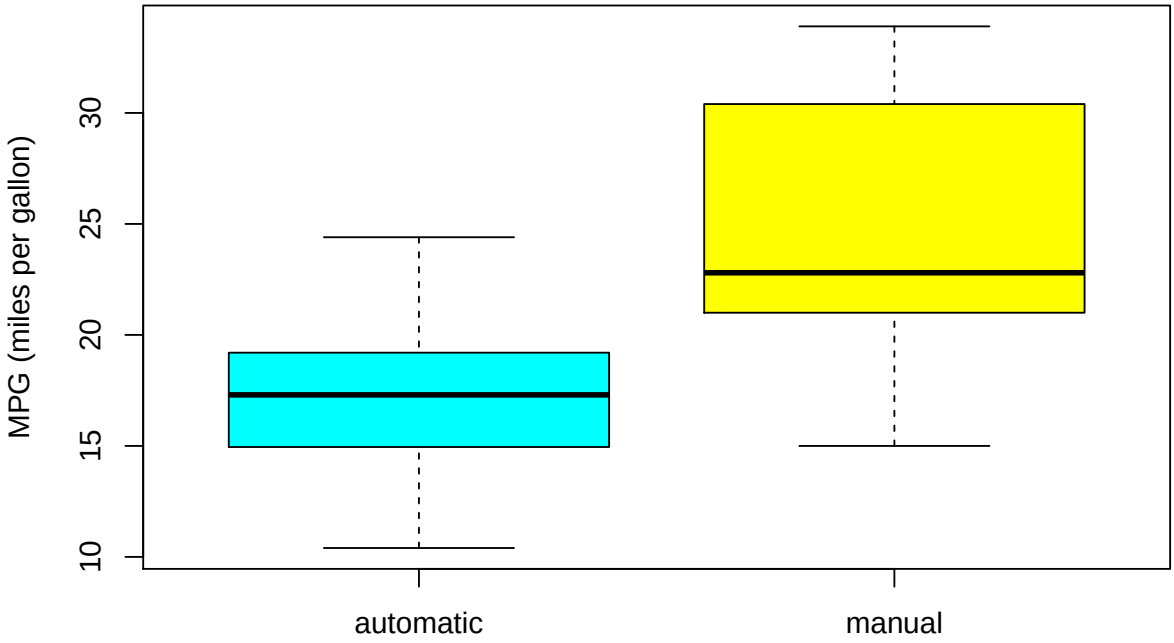

Figure 1: Gear transmission type effect on MPG

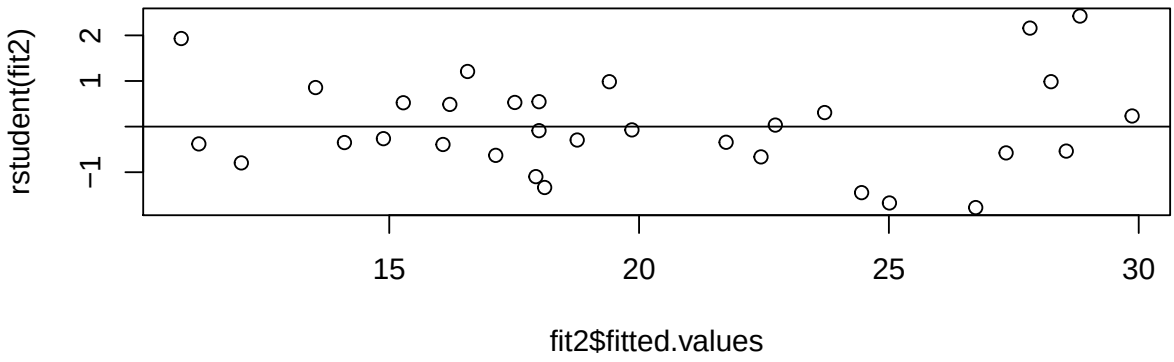

Figure 2: studentised residuals

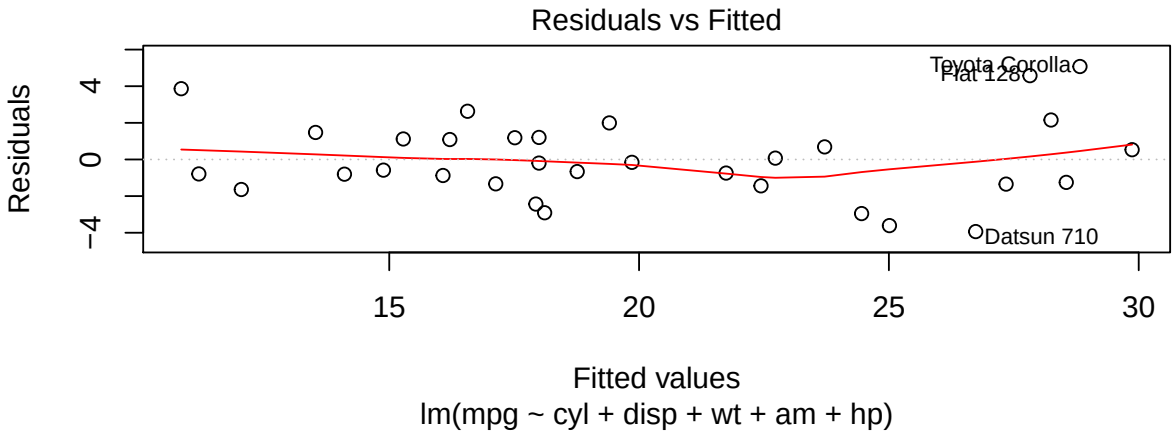

Figure 3: residuals vs fitted

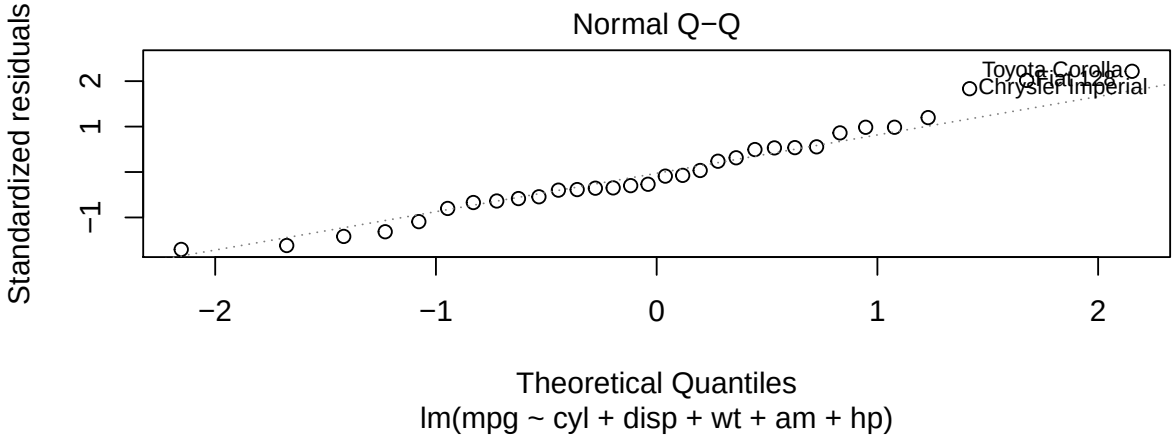

Figure 4: Q-Q diagram

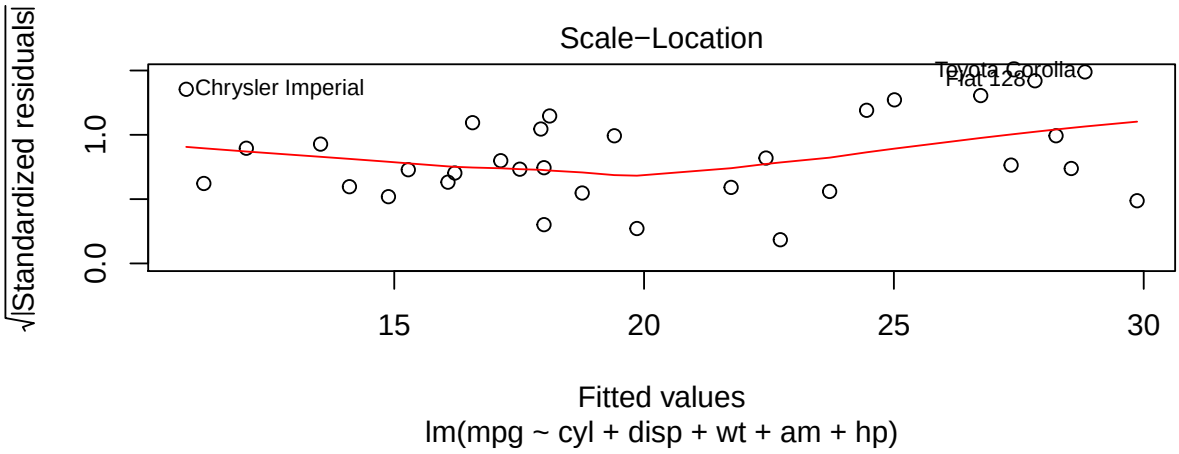

Figure 5: standardized residuals (rstandard function)

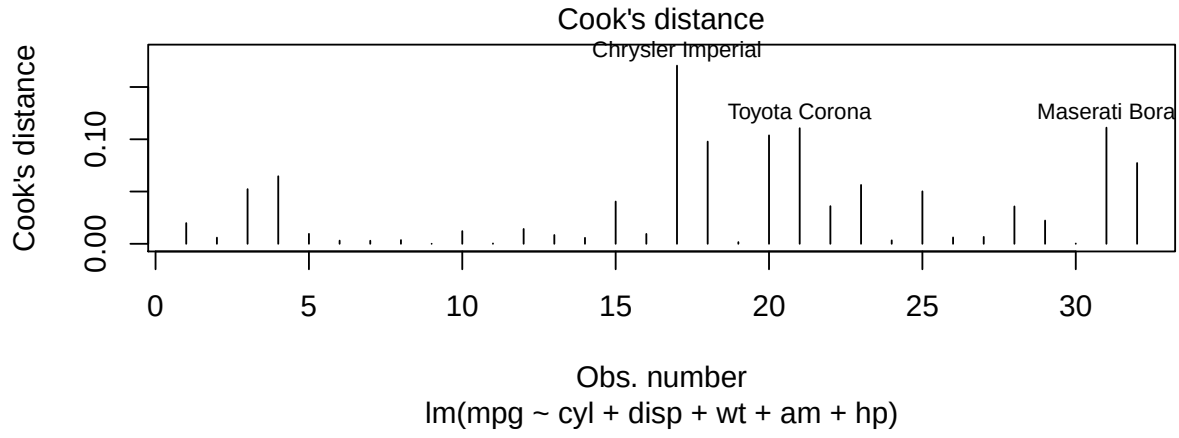

Figure 6: Cook's distances

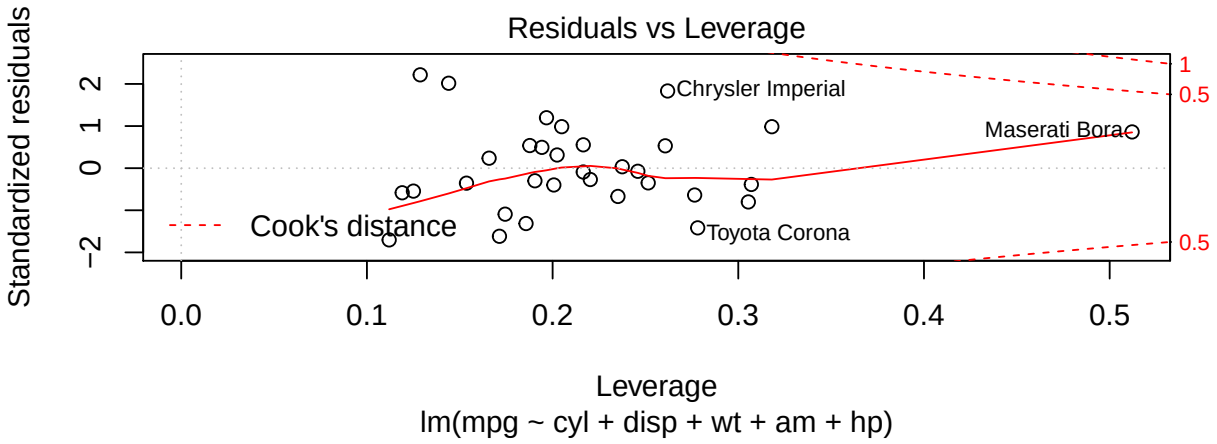

Figure 7: Outliers and leverage