MOTOR TREND MAGAZINE

Pleasing Gears & MPG Appearance

Eduardo B. Díez — August 2014

Executive Summary

The debate about the manual or automatic transmissions is open. However, findings sponsored by MOTOR TREND MAGAZINE will help us to make better decisions about whether to use vehicles with manual gearboxes, pleasing for everyone loving the motor sport, as is customary in Europe or automatic transmissions as is usually done in America.

Although the findings of our study are not definitive yet, they indicate that vehicles with automatic transmissions —versus manual version— travel, in round numbers, 3 miles less per gallon of fuel that all of us have to pay from our own pockets. This has surely a considerable impact on the battered economies of many families in our country.

Finally, and wonderfully, abandoning those bland automatic transmissions by the use of manual gearboxes, not only we become into coherent lovers —of the four wheels— but additionally we'll reduce harmful emission and help preserve our planet for future generations.

"*. . . Essentially, all models are wrong, but some are useful.*"
George Box

We'll give clear and concise answers to determine if an automatic transmission is better than a manual one to get more miles per gallon of fuel and whether there is any difference between the two types, what is that amount.

All analyses were performed using *R version 3.1.1 (2014-07-10)*[1] and *RStudio*[2] as IDE, with the default base packages *grid, stats, graphics, grDevices, utils, datasets, methods, base* and additionally to produce this report in PDF the packages *xtable*[3], *gtable*[4] and *knitr*[5].

To find a feasible regression model on the basis of the data available `mtcars` we made on it one change concerning with the variable –predictor– of interest `am`, which was used to create a new variable. The new predictor —`Automa.Trans`— will be used in the regression model instead of the original am. `Automa.Trans` results in the inverse binary copy of `am` where the value `1` means automatic gearbox/transmission and `0` if not. This way the results are easiest to be interpreted, even that the new variable it's really the same as the original but binary negate. Also we're going to stepwise through the candidate predictors with a **backward elimination p-value strategy**[6].

[1] R Core Team (2014). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[2] 2009-2013 RStudio, Inc. *RStudio: Integrated development environment for R.* (Version 0.98.978). [Computer software]. Boston, MA. Retrieved May 20, 2012. http://www.rstudio.org/.

[3] David B. Dahl (2014). *xtable: Export tables to LaTeX or HTML.* R package version 1.7-3. http://CRAN.R-project.org/package=xtable.

[4] Hadley Wickham (2012). *gtable: Arrange grobs in tables.* R package version 0.1.2. http://CRAN.R-project.org/package=gtable

[5] Yihui Xie (2014) *knitr: A Comprehensive Tool for Reproducible Research in R.* R package version 1.6. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

[6] Diez, David M, Christopher D Barr, and Mine Çetinkaya-Rundel. "Introduction to Linear Regression," p 361. In *OpenIntro Statistics*, Second Edition, 2013. http://www.openintro.org/stat/textbook.php.

So we started with a model described with the formula:

$$mpg \sim cyl + disp + hp + drat + wt + qsec + vs + gear + carb + Automa.Trans$$

and after performing the parsimonious backward method, the model was reduce as the table 1 shows.

	Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1				21.00	147.49	70.90
2	- cyl	1.00	0.08	22.00	147.57	68.92
3	- vs	1.00	0.27	23.00	147.84	66.97
4	- carb	1.00	0.69	24.00	148.53	65.12
5	- gear	1.00	1.56	25.00	150.09	63.46
6	- drat	1.00	3.34	26.00	153.44	62.16
7	- disp	1.00	6.63	27.00	160.07	61.52
8	- hp	1.00	9.22	28.00	169.29	61.31

Table 1: The Akaike's Information Criterion (AIC) for the backward parsimonious model where the smaller the AIC, the better the fit. (See: Sakamoto, Y., Ishiguro, M., and Kitagawa G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.)

Finally, we obtained a regression model that explains roughly 83% of the total variance of the response variable —`mpg`— and leaving the rest 17% to be explained by residual variance. The coefficients can be found in table 2, with $R^2_{adj} = 0.8336$ and df = 28.

	Estimate	Std. Error	t value	Pr(>|t|)	ci: 2.5%	ci: 97.5%
(Intercept)	12.55	6.06	2.07	0.0475	0.15	24.96
wt	-3.92	0.71	-5.51	0.0000	-5.37	-2.46
qsec	1.23	0.29	4.25	0.0002	0.63	1.82
Automa.Trans	-2.94	1.41	-2.08	0.0467	-5.83	-0.05

Table 2: Summary of the regression with the addition of confidence intervals.

The data provide convincing evidence that all β_i are significantly different than 0, so the explanatory variables —`wt, qsec` and `Auto.Trans`— are significant predictors of the response variable —`mpg`. With a significance level of $\alpha = 0.05$ we reject H_0: $\beta_i = 0$ favoring H_A: $\beta_i \neq 0$. Also, we are 95% confident that all β_i fall inside the indicated endpoints of their confidence interval.

Thus, we are 95% confident that $mpg \sim wt + qsec + Automa.Trans$ predicts, all else being equal, that the vehicles equipped with automatic gearboxes travel on average **2.94 miles less** per each gallon of fuel than the vehicles equipped with gearboxes that aren't automatics.

A fast diagnostic of our model starts by looking at the residuals vs fitted plot —$e \sim \hat{y}$ allows for considering the entire model with all explanatory variables at once— that indicates a fail in the homoscedasticity, because —in our case— the variability of points —residuals— around the 0 line —the least squares line— should be roughly constant but the points seem to get more distanced from the 0 line as the abscissa value —predicted value— increases. So, seems to be a lack in the condition of *constant variability of residuals*. Also, the condition of *nearly normal residuals with mean 0*, by using the normal Q-Q plot seems with anomalous points in both extremes.

Both situations may be due to weaknesses in the data. A priori, given the extensive amount of variety of technical specifications in vehicles, seems to us very limited to find strong conclusions about the issue by using only 32 cases. Consequently, we should consider our finds as provisional until a deeper study can be realized to fulfill the necessities about the variability and the normality of residuals.

Appendix

Now follows a series of plot resulting from the initial exploration, and the last one, corresponds with the final obtained model. All images has a capstone with a try explanation.

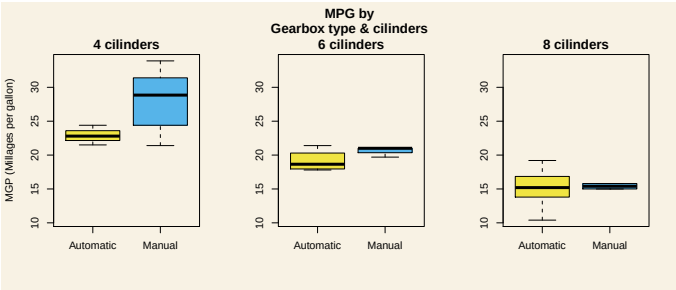

Figure 1: The plot shows that number of cylinders and gearbox type are both variable predictors associated —negatively— with the mpg response.

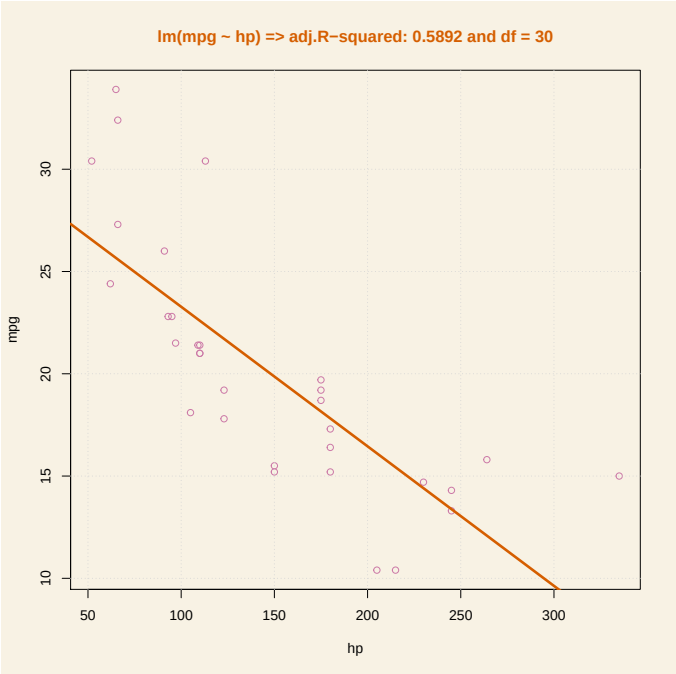

Figure 2: We can see the regression line how follows the scatters point but, also those point seems to follow a second degree curve.

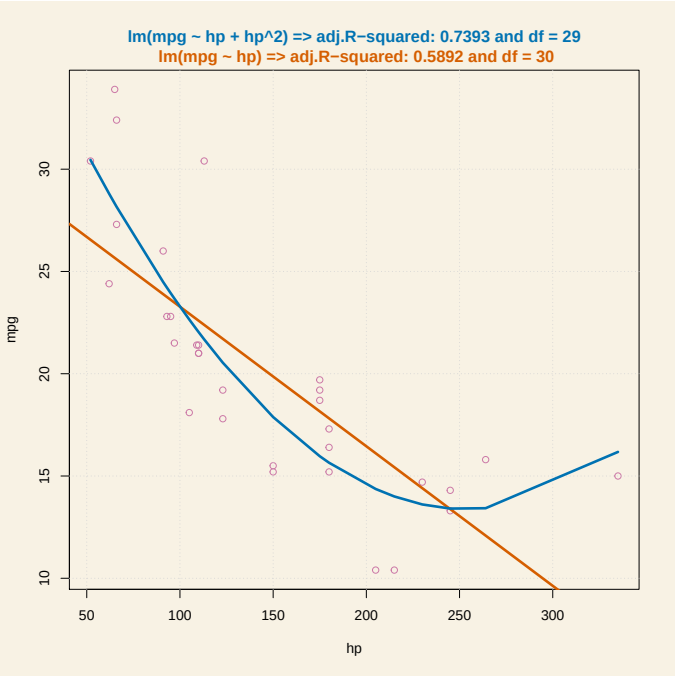

Figure 3: Constructing a second order model we can see the the curve fits better the points also, the adj R-square has improved markedly

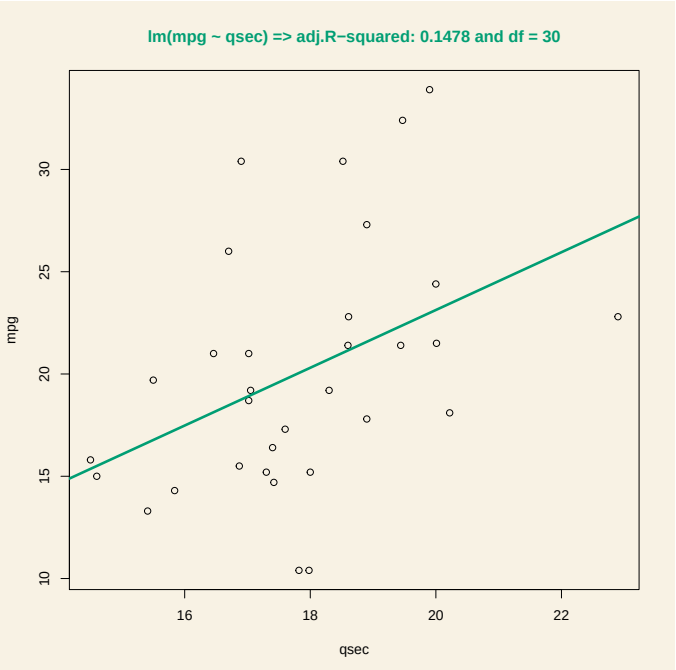

Figure 4: The predictor qsec, the time the vehicle is capable to travel 1/4 of milles as fast as possible, is a good indicator of all the technical variables as one and also, the final model has this predictor and wt additional to the gearbox

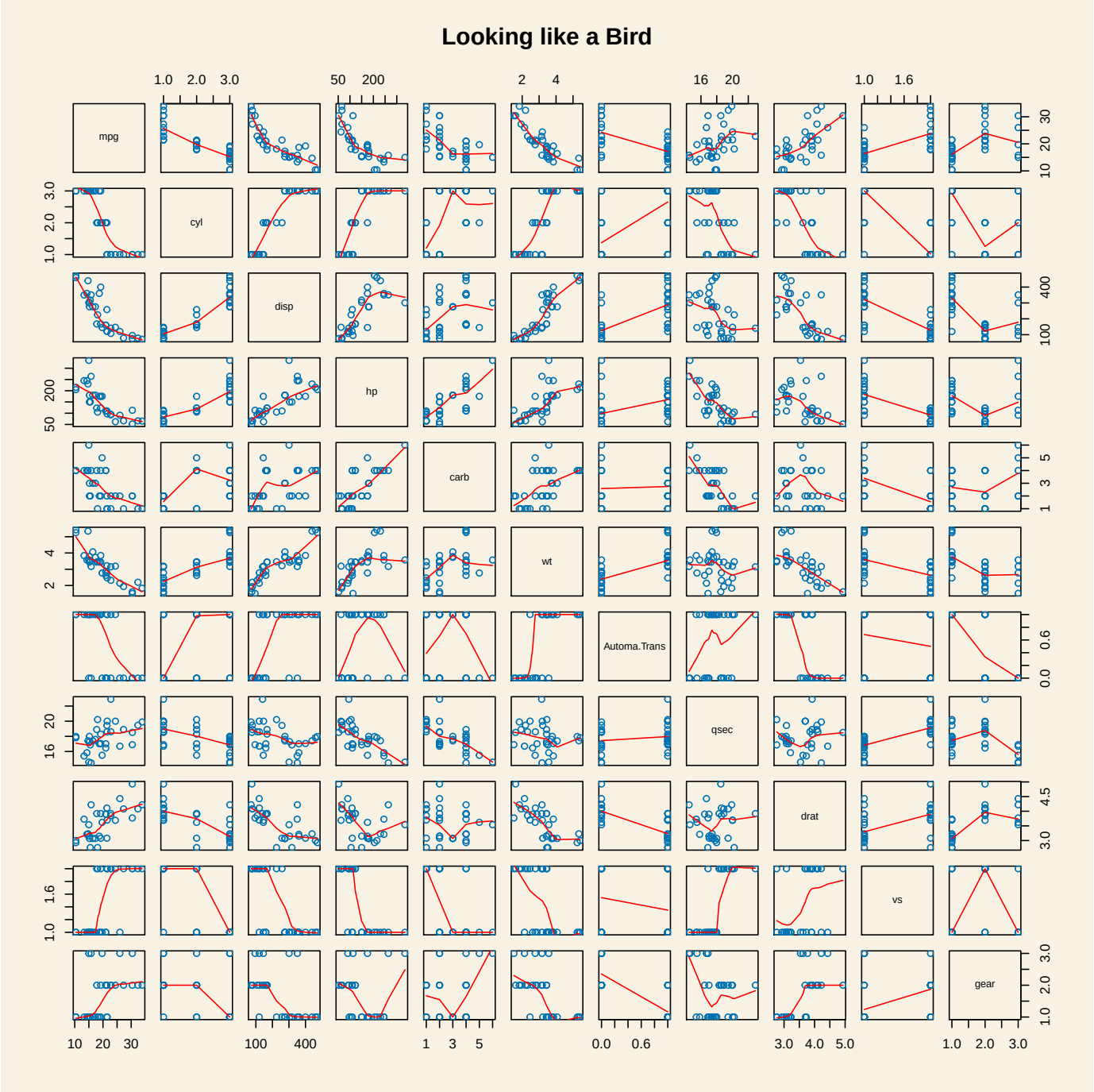

Figure 5: Scatterplot is a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. We note, at the very first row starting with the variable `mpg`, that there is an apparent negative association between `mpg` and the variables `cyl, disp, hp, carb, wt` and `Automa.Trans` while it seems a positive relation with the variables `qsec, drat, vs` and `gear`. With the former set of predictors, the response `mpg` goes down —decrease— with any increase in the value of each predictor —giving less miles per gallon— while it happens the opposite with the latter set, where any increase in the value of the predictors produce an increase in the value of the response —giving more miles per gallon. We should state that a V-motor is less than an S-motor for our study and the rest of predictors should be trivial in the amounts, so, a carburetor with 2 chambers is less than a carburetor with 8 chambers.

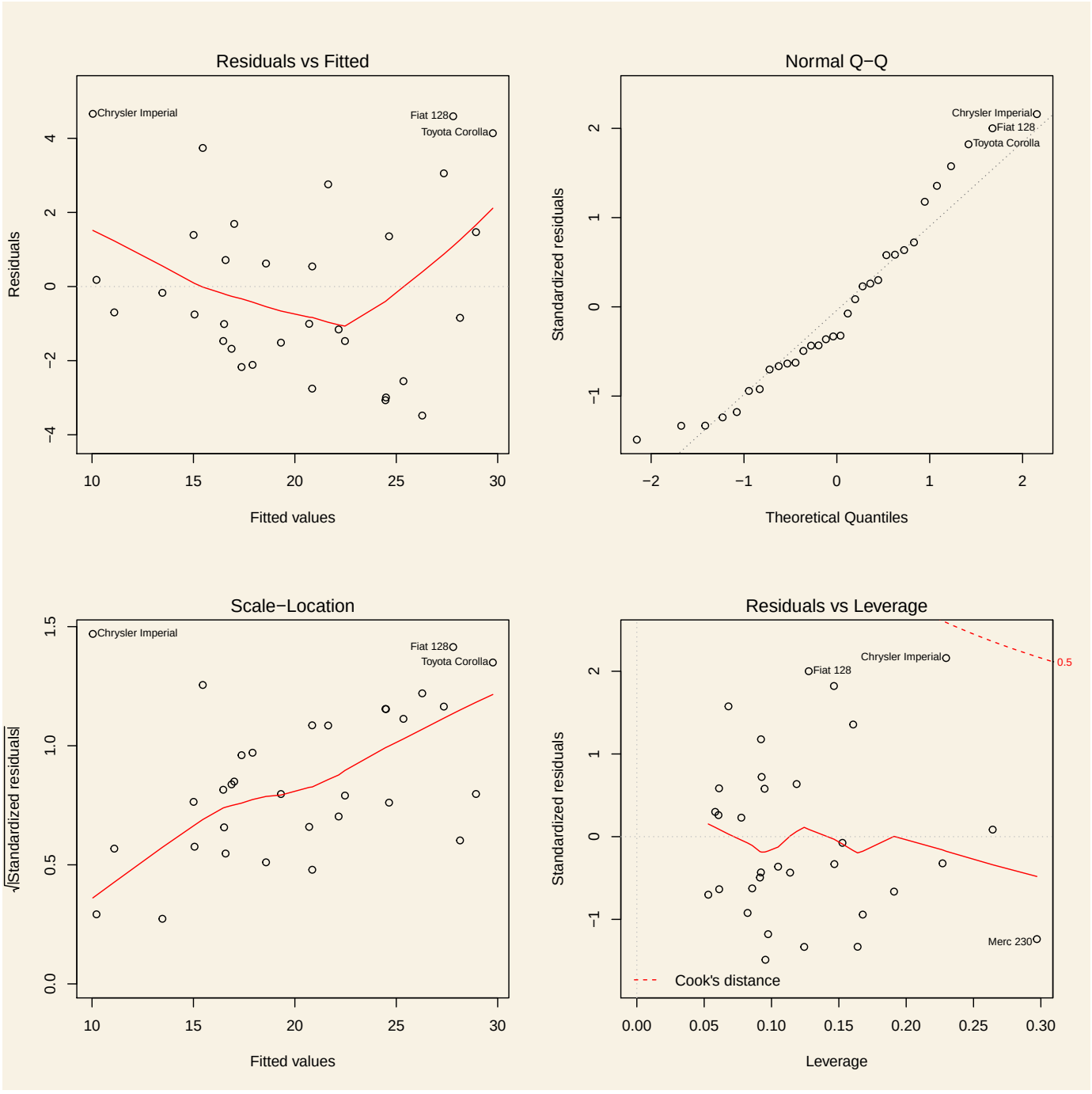

Figure 6: The result of the regression is inspectable graphically. The plot of the second row and the second column indicates that the outliers have Cook's distances less than unity so that we need not worry about them. Our major concerns, in this case, are coming from the two plots from the first row. The first picture shows what is called `fan pattern`, where the residuals are not uniformly distributed but as the fitted values increase, the difference between the observed and predicted increase their value of separation to the axis of abscissae. The second plot shows, at both ends, that values are distanced away from the reference line corresponding to the normal distribution. From what we observed in both plots, there is not normal residues nor evenly distributed.