Regression Model - Course Proj.

thecapacity

February 22, 2015

Summary

Using a linear regression model, based on the `mtcars` dataset provided by R, we found that a car with manual transmission yields more miles per gallon than a car with automatic transmission (`27.362` vs. `24.802`) based on our model. It was also discovered that the number of cylinders (`cyl`) plays an importat role as well, but the layout of the engine represented by the `vs` variable is not as significant as initially assessed in a second model.

We have only superficially explored the relationship across other vaiables, based on the assumption that a number of 'power-type' indicators are strongly correlated with `cyl` count. Potential areas for further exploratory data analysis would include quantifying the correlation across variables, such as whether a measure like `hp` is a better predictor thatn `cyl` or `disp`.

Analysis

Looking at a data set of a collection of cars, I am interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

We are particularly interested in the following two questions:

- *Is an automatic or manual transmission better for MPG*
- *Quantify the MPG difference between automatic and manual transmissions*

Approach

This analysis will consider the interpretation of the model coefficients including a review and plot of Residuals. Some alternative models for fit will be explored including discussion of the selection strategy. Conclusions and potential uncertainty and confounders are included throughout as well as potential areas for future exploratory data analysis

Based on my *'expertise'* in cars, I believe that the most significant factor to consider, in addition to the transmission type, may be the number of cylinders `cyl`. We will first begin the analysis with the most simplistic relationship and then investigate the impact of other variables.

Pre-processing is minimal, serving to load useful defaults and convert key variables to factors, but not shown to reduce overall page count.

Our most parsimonious model will be to model the relationship between `mpg` and `am`:

```
fit1 <- lm(mpg ~ am, mtcars)
```

The model shows that the transmission factor is significant (with a `t value` of `4.106127`) in predicting `mpg` , by `7.2449393` mpg. The residuals also have what looks to be a relatively balanced magnitude (`-9.392` - `9.508`), but we will save more thorough review of the residuals once we feel more confidant in the completed model.

Based on this initial analysis, we feel it would be useful to compare this directed approach to a generic 'kitchen sink' approach, modeling the `mpg` impact as a function of all variables via the following model:

```
fit2 <- lm(mpg ~ ., mtcars)
summary(fit2)$coef
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 17.81984325 16.30602324  1.09283809 0.28745417
## cyl6        -1.66030673  2.26229662 -0.73390320 0.47152449
## cyl8         1.63743980  4.31573451  0.37941162 0.70838075
## disp         0.01391241  0.01740176  0.79948296 0.43340363
## hp          -0.04612835  0.02712018 -1.70088685 0.10446190
## drat         0.02635025  1.67648954  0.01571752 0.98761549
## wt          -3.80624757  1.84664309 -2.06117121 0.05252853
## qsec         0.64695710  0.72195025  0.89612421 0.38084614
## vs1          1.74738689  2.27267212  0.76886889 0.45095593
## am1          2.61726546  2.00474936  1.30553251 0.20653091
## gear         0.76402917  1.45668015  0.52450029 0.60569589
## carb         0.50935118  0.94244181  0.54045902 0.59484874
```

Here we see that `wt` and `cyl` have a measurable (>1) estimated effect on `mpg`, and `vs` - an indication of the car's engine shape (http://stackoverflow.com/questions/18617174/r-mtcars-dataset-meaning-of-vs-variable). Based on this 'cut-off', other variables like `gear` are excluded from further analysis but may be valubel in future work.

What is interesting, is the impact by number of cylinders seems to be reversed (by sign) from the expected result, i.e. `cyl6` shows a negative impact where as `cyl8` shows a positive relationship. This iseems counter to intuition, but **we believe this 'questionable outcome' is an important exploration, and ultimate indicator that this model does not represent q 'best fit'**.

A third model was used to review these selected coefficients, but is not echoed here to reduce page space.

From the coefficients table of that model `fit3`, we saw that `am` and `cyl` were deemed significant, but the V/S status was not to the same degree. **We also saw support for a more intuitive relationship between `mpg` and `cyl`, specifically that mpg decreases as the number of cylinders increase**.

Therefore our final model is simplified in final form to:

```
fit4_final <- lm(mpg ~ am + cyl, mtcars)
fit4_final$coef
```

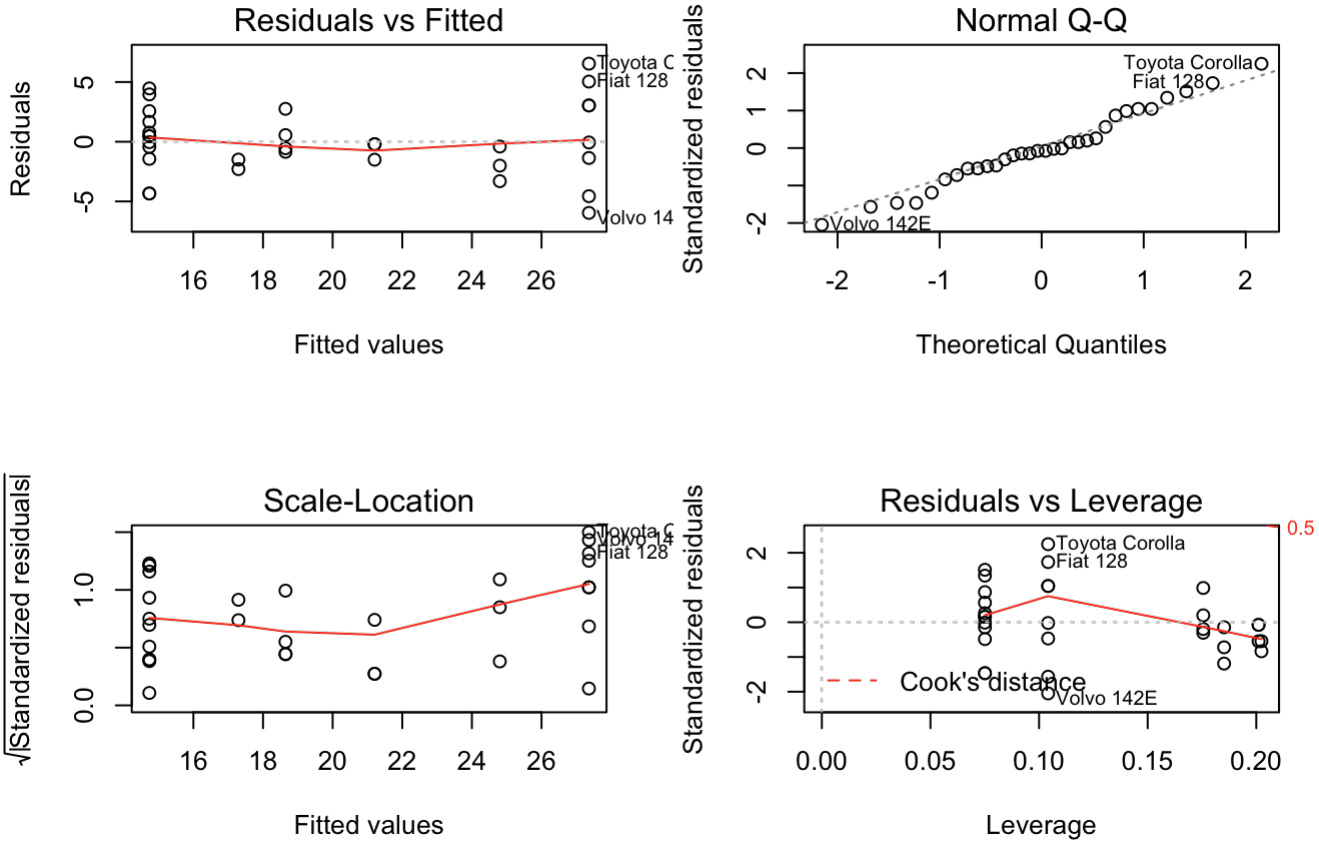
```
## (Intercept)         am1        cyl6        cyl8
##   24.801852    2.559954   -6.156118  -10.067560
```

From this model we see the relationships people familiar with automotives would expect; namely that MPG goes up for manual cars, i.e. `am == 1` and down as the number of cylinders increase. With a residual variance of `3.0734792` this model seems to best balance simplicity, for building an understanding, with usefulness, for future predicions of mpg.

Note: I find it useful to plot the model's values *without an intercept* to facilitate reporting on the final model, and this is included below, but not echoed, for completeness. Also the `echo = FALSE` parameter has also been added to the code chunks in the appendix to prevent printing of the R code that generated the plot to reduce extraneous text and meet the assignment's demands.

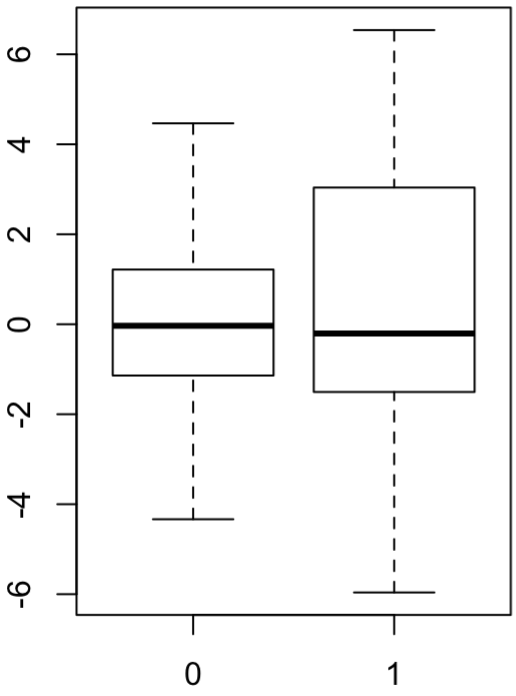
Appendix

Best Fit Plot

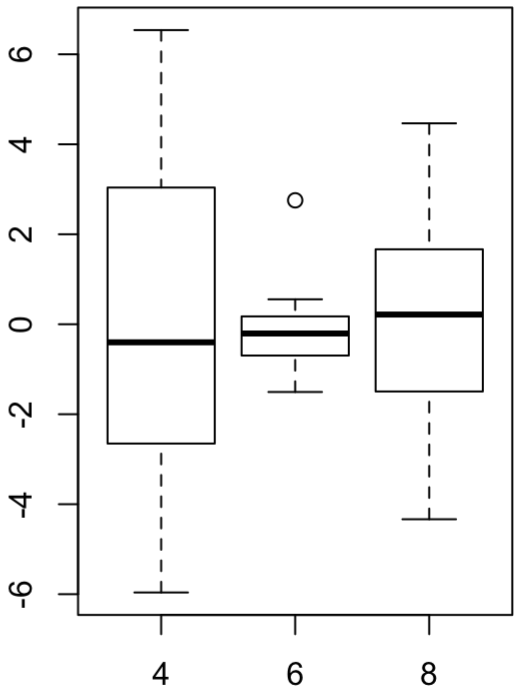


Residuals vs. X's in Final Model

Residuals by Transmission

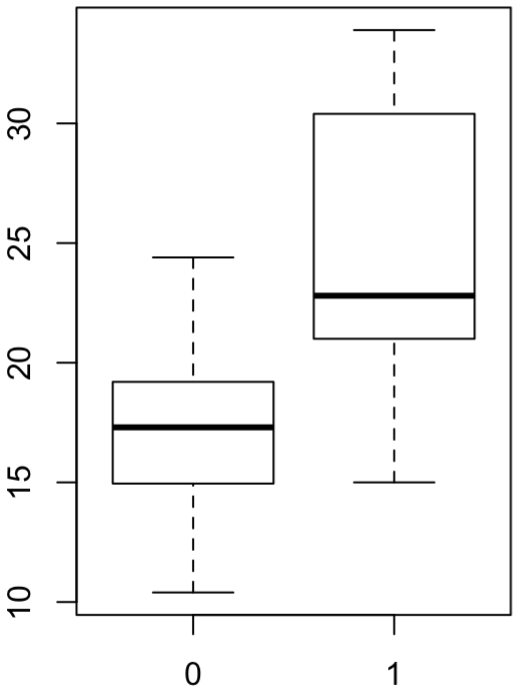


Residuals by # of Cylinders



Plot of `mpg` vs our final key factors, `am` and `cyl`.

MPG by Transmission



MPG by # of Cylinders