Regression Model - Course Proj.

thecapacity

February 22, 2015

Summary

1974 data from Motor Trend, a magazine about the automobile industry, was used to explore the relationship between a set of variables and the outcome they might have on miles per gallon (`mpg`). Of particular focus was whether an automatic or manual transmission resulted in better MPG, and quantifying how different that impact may be.

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`), given that most other variables are excluded from our model. It was also discovered that the number of cylinders (`cyl`) plays an importat role as well, but the layout of the engine (straight vs. V) represented by the `vs` variable is not as significant as initially assessed in our 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

In ths assignment I will pretend to work for a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

They 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 tnterpretation 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`, which is expected to be covariant with other measures of engine power (e.g. `disp`, `hp`, `carb`, etc). We will first begin the analysis with the most simplistic relationship (`mpg` by `am`) and then investigate the impact of other variables.

Processing

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

Our most parsimonious model will simply be to model the relationship between `mpg` and `am`, where the transmission status is `0` for automatic and `1` for manual.

Therefore the model and summary for this fit is as follows:

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

```
## (Intercept)         am1
##   17.147368    7.244939
```

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 more generic 'kitchen sink' approach, i.e. modeling the impact on `mpg` as a function of all variables in the dataset.

For this approach we can operate as follows:

```
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` also have a measurable (`> 1`)) estimated effect on `mpg` as well as `vs`, which is believed to be an indication of whether the car has a V- or Straight- engine (http://stackoverflow.com/questions/18617174/r-mtcars-dataset-meaning-of-vs-variable). Other variables, such as `gear`, are excluded based on this 'cut-off' from further analysis but seem as though they may have value in future modeling and analysis.

What seems most interesting from this dataset, aside from possibly identifying a number of 'less significant' variables, is that the estimate for the impact of the number of cylinders seems to be reversed (by sign) from the expected result, i.e. `cyl6` shows a negative (`-1.66`) impact on mpg where as `cyl8` shows a positive `1.64` relationship - something that seems counter to intuition. Based on experience, one would assume that smaller-cylinder engines would be more fuel efficient. **We believe this 'questionable outcome' is an important part of exploration, and ultimately an indicator that this particular model does not represent the 'best fit'.**

So let's look at the coefficients for a third model, highlighting the fit from these key variables:

```
fit3 <- lm(mpg ~ am + cyl + vs, mtcars)
summary(fit3)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 22.809113   2.928225  7.789398 2.240449e-08
## am1          3.164892   1.528283  2.070882 4.805371e-02
## cyl6        -5.398674   1.837466 -2.938107 6.683364e-03
## cyl8        -8.161240   2.891755 -2.822245 8.841429e-03
## vs1          1.708062   2.234961  0.764247 4.513478e-01
```

From the coefficients table we see that `am` and `cyl` are all deemed significant, but the V/S status is not to the same degree. **We can also see support for what seems to be a more intuitive relationship between `mpg` and `cyl`, specifically that mpg decreases as the number of cylinders increase**, as opposed to the earlier sign issues.

Therefore our final model is simplified in final form to:

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

```
##
## Call:
## lm(formula = mpg ~ am + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9618 -1.4971 -0.2057  1.8907  6.5382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.802      1.323  18.752  < 2e-16 ***
## am1            2.560      1.298   1.973 0.058457 .
## cyl6          -6.156      1.536  -4.009 0.000411 ***
## cyl8         -10.068      1.452  -6.933 1.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.073 on 28 degrees of freedom
## Multiple R-squared:  0.7651, Adjusted R-squared:  0.7399
## F-statistic:  30.4 on 3 and 28 DF,  p-value: 5.959e-09
```

From this model we can clearly see the relationships most people more than casually 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. (Please see Appendix for supporting charts). 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.

```
##
## Call:
## lm(formula = mpg ~ am + cyl - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9618 -1.4971 -0.2057  1.8907  6.5382
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## am0    24.802      1.323  18.752  < 2e-16 ***
## am1    27.362      0.992  27.584  < 2e-16 ***
## cyl6   -6.156      1.536  -4.009 0.000411 ***
## cyl8  -10.068      1.452  -6.933 1.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.073 on 28 degrees of freedom
## Multiple R-squared:  0.9812, Adjusted R-squared:  0.9785
## F-statistic: 364.6 on 4 and 28 DF,  p-value: < 2.2e-16
```

Note that 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.