Regression Models Project

S. Phillips

November 15, 2015

Executive Summary

Motor Trends magazine has asked me to solve which features of an automobile affect fuel efficiency. In particular, they are interested in comparing miles per gallon of automatic vs manual transmissions. This document will walk the reader through the analysis and its conclusion given in the summary.

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
## Warning: package 'knitr' was built under R version 3.2.3
```

Exploratory Analysis of the Data and Their Corellations

Motor Trend's database lists 32 automobile models with 11 variables that could affect fuel efficiency. To bring the characteristics of this data to light, the top rows are given below.

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

To gain a better understanding of the variables, we'll look at the correlation of each one independently against **(mpg)** and observe where **(am)** ranks among them. These results predict a positive correlation that manual transmissions tend to deliver better overall gas mileage. However, **am** is less correlated than other variables. Also, the *qqnorm* plots given in the Appendix show that many of our variables are subject to outliers - especially **(wt)***.

See Appendix: Selecting Predictors (QQNorm Plots)

	Description	Correlation
wt	Weight (lb/1000)	-0.8676594
cyl	Number of cylinders	-0.8521620
disp	Displacement (cu.in.)	-0.8475514
hp	Gross horsepower	-0.7761684
carb	Number of carburetors	-0.5509251
qsec	1/4 mile time	0.4186840
gear	Number of forward gears	0.4802848
am	Transmission (0 = automatic, 1 = manual)	0.5998324
vs	Engine Type 0=V(as in V8) or 1=S(as in straight)	0.6640389
drat	Rear axle ratio	0.6811719
mpg	Miles/(US) gallon	1.0000000

Evaluate Predicters for Model Selection

The notes below gather what we know about the correlations combined with domain knowledge gleaned from researching the data.

- Although **wt** has significant outliers, the manufacturer data confirms accuracy so we accept it as a predictor candidate. *
- We'll exclude **cyl, disp and carb** because they are likely predictors of **hp**. Logically, **hp** could be a predictor of **wt** or high-performance.
- The data in **qsec** seems a good variable for predicting high-performance that has to be reconciled against **wt**. Although **qsec** is near the median of correlations, it is very locigal as a predictor.

**See Appendix: Selecting Predicters (QQNorm Plots)*

The research up to this point indicates **am, wt, drat & qsec** as the starting predictor list. However, we will see changes our in our final selections based upon *Adjusted R Squared* and *P Value* results from exhaustive model permutations.

Model Selection Strategy

From exhaustive testing of Multi-variable Regression Models, we found 3 finalists for best fit. We will eliminate one of them based upon results from *Adjusted R Squared* and *P Values.* Last, we will use a plot of *Residuals* to pick a best fit from our two finalists.

- FIT CANDIDATE 1: **mpg ~ wt + qsec + as.factor(am)** Adj R Squared=0.8335561. The p-values for this predictor all clearly reject the null hypothesis. However, the coefficients from the other two candidates were far stronger. *
- FIT CANDIDATE 2: **mpg ~ hp + qsec + as.factor(am) * wt** Adj R Squared=0.8758576. Highest R2 of all three models, but p values are weaker. **Wt**, in particular, indicates the alternative hypothesis. *
- FIT CANDIDATE 3: **mpg ~ wt + qsec + as.factor(am) * hp** Adj R Squared=0.8444169. High R2, p-values reject Ho except for the hp confounder *hp* which suggests Ha. *

**See Appendix: Coefficients*

Evaluate Residuals to Select Best Fit

The *Residuals vs Fitted* plot in the appendix indicate that FIT CANDIDATE 2's fit line stays closer to the mean and the residuals are tighter than FIT CANDIDATE 3. Also, the *Normal Q-Q* plot shows that FIT CANDIDATE 2's residuals are closest to a normal distribution.

**See Appendix: Residuals for the Two Finalists*

Conclusion

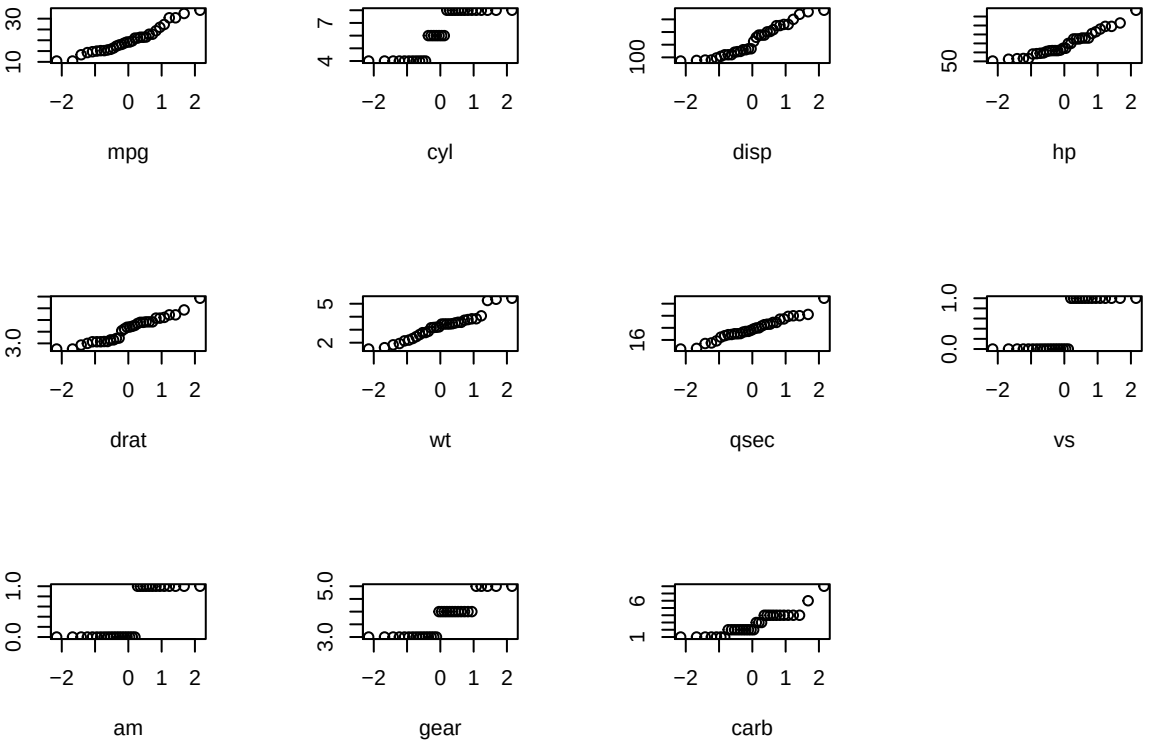
1. The best fit for predicting **mpg** is FIT CANDIDATE 2: **lm(mpg ~ wt + qsec + as.factor(am) * wt)**
2. The 14 mpg in added fuel economy for FIT CANDIDATE 2 cant be trusted because of 3. . . ..

3. All models with reasonably good coefficients predict that manual transmissions deliver better fuel economy than automatics. However, the mpg estimate is highly variable, and therefore, difficult to predict considering many results showing good fit such as High R-Squared, Residuals with Normal Distributions and P-values negating a null hypothesis. Ultimately, an array of statistically valid results with a wide range of estimated outcomes can't be trusted.

Appendix

Selecting Predictors (QQNorm Plots)

Use QQNorm plots to look at each variable to evaluate weights,leverage and distribution - perferring normal distributions and being mindful of outliers.



Output from *Leaps* regsubsets command. Darkest columns are indicated as being predictors for **mpg**.

Coefficients

Candidate 1

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     9.617781  6.9595930  1.381946 1.779152e-01
## wt             -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec            1.225886  0.2886696  4.246676 2.161737e-04
## as.factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
```

Candidate 2

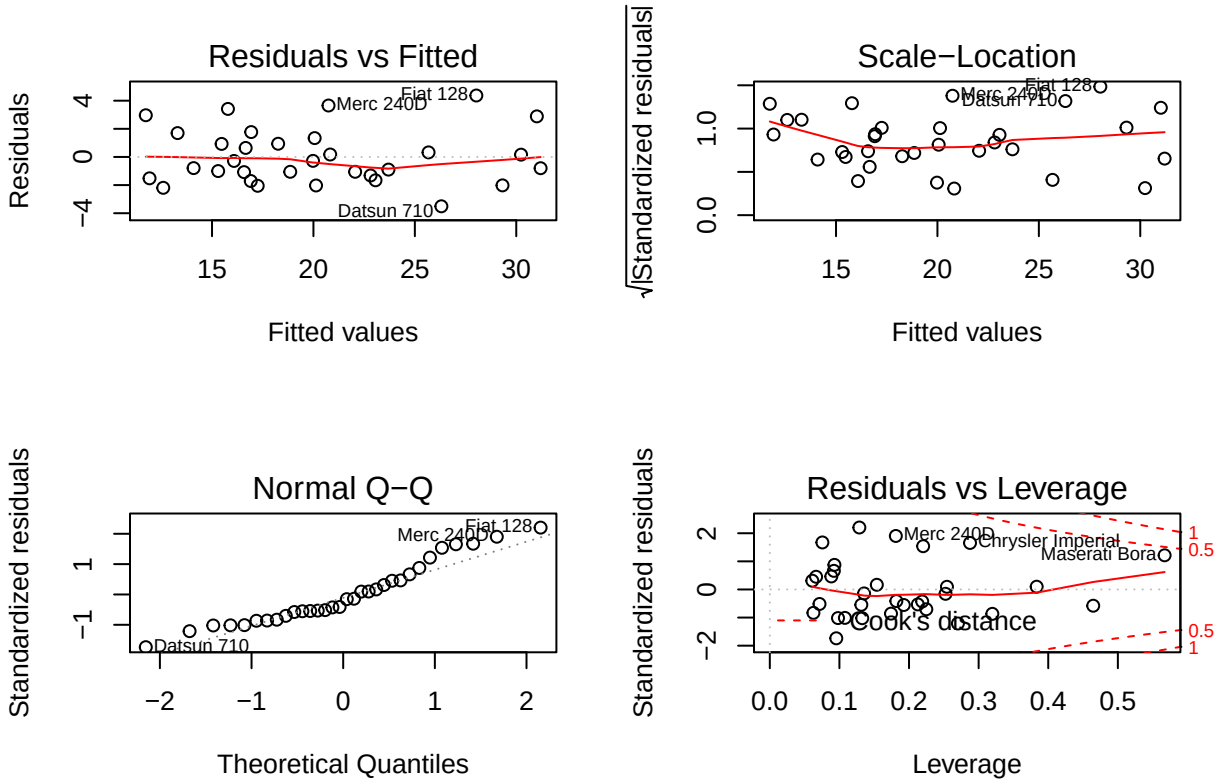
```
##                      Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)       10.23078964 8.45742447  1.2096815 0.237285626
## hp                -0.00114799 0.01345264 -0.0853357 0.932648254
## qsec               0.99223473 0.38727049  2.5621233 0.016545920
## as.factor(am)1    13.95728476 3.78154610  3.6908937 0.001041257
## wt                -2.90307957 0.78369868 -3.7043313 0.001005885
## as.factor(am)1:wt -4.09623331 1.32924057 -3.0816343 0.004822999
```

Candidate 3

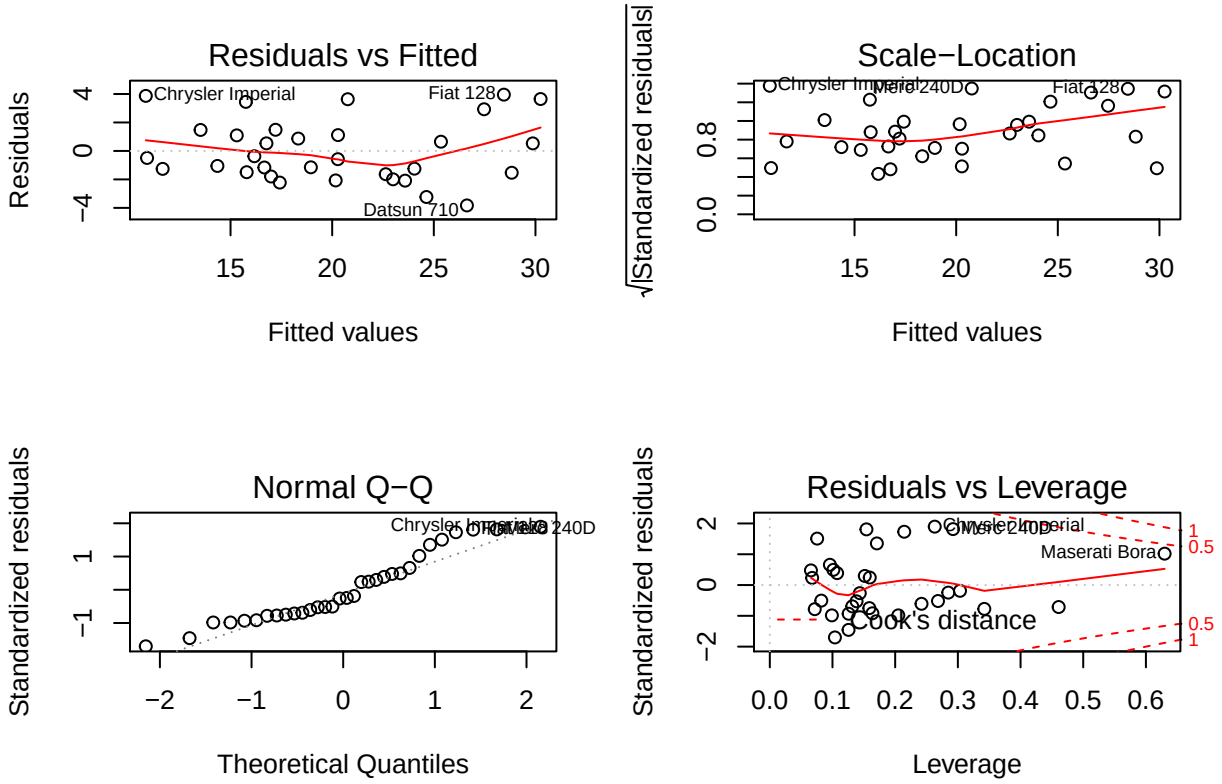
```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       11.574623809 9.87892413  1.1716482 0.2519636174
## wt                -3.683223879 0.91662790 -4.0182323 0.0004457912
## qsec               1.034701303 0.45301776  2.2840193 0.0307804107
## as.factor(am)1     6.015490311 2.44348182  2.4618519 0.0207699521
## hp                 0.003995245 0.01981124  0.2016656 0.8417476860
## as.factor(am)1:hp -0.021969525 0.01441363 -1.5242186 0.1395265854
```

Residuals for the Two Finalists

Candidate 2



Candidate 3


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