Determinants of MPG Ratings

jhsdatascience

06/20/2014

Is an automatic or manual transmission better for MPG? How much better? I use data on 32 vehicles to estimate the effect of transmission type on gas mileage. After controlling for physical characteristics of the cars and measurements of engine power I find little evidence that the type of transmission has an effect on gas mileage. Specifically, though the data show that a vehicle with manual transmission recieves `.176` better gas mileage than a similar vehicle with an automatic transmission, the standard error for this estimate is `1.304` and we cannot conclude that the effect is statistically different from zero.

Exploratory Analysis

The data used for this analysis are from the `mtcars` dataset in the `R` library `datasets`. The data consist of 32 observations of 11 variables. Each observation corresponds to a given make of car. Of the 32 cars for which I have data, 13 have manual transmissions. The mean `mpg` for these cars is 24.3923. The mean `mpg` for those with automatic transmissions is 17.1474. The relationship between transmission type and gas mileage is more complicated than that. Figure 1, for example, shows that the number of cylinders may be confounding the relationship. Cars with automatic transmissions tend to have more cylinders and the number of cylinders is also highly (negatively) corelated with the miles per gallon. To build intuition for the models in the next section, I look more closely at the correlataions between each of the variables (see Figure 2 for a visualization):

```
##        mpg      wt     cyl    disp      hp   drat     vs     am     carb
## mpg 1.0000 -0.8677 -0.8522 -0.8476 -0.7762 0.6812 0.6640 0.5998 -0.55093
## am  0.5998 -0.6925 -0.5226 -0.5912 -0.2432 0.7127 0.1683 1.0000  0.05753
##       gear    qsec
## mpg 0.4803  0.4187
## am  0.7941 -0.2299
```

The following observations will help with model selection:

1. It appears the variables related to the size of the car (`wt` and `disp`) and those related to the power of the engine (`cyl` and `hp`) have the strongest correlation with `mpg`.
2. The strong correlation between `mpg` and `am` disappears once we have controlled for `wt` (Figure 1).
3. `am` is highly correlated with both `drat` and `gear` while these, `drat` especially, are also correlated with `mpg`.
4. `am` is not very correlated with `qsec`, `vs`, and `carb`.

Model Selection

I use these observations to construct a small set of models to test agaist each other. These are:

1. `mpg ~ cyl + wt + am`
2. `mpg ~ cyl + hp + wt + disp + am`
3. `mpg ~ cyl + wt + drat + am`
4. `mpg ~ cyl + wt + gear + am`
5. `mpg ~ .`

I will refer to these as M1-M5. M1 should be thought of as the baseline, building on the intuition from Figure 1. M2 checks this intuition against the first observation above: is it necessary to include all measures of vehicle size and power, or is a subset enough? M3 and M4 are included because of the potential for omitted variable bias (observation 3). M5 is a catch all against which to test each of the others, M1 in particular.

For models M3 and M4, the coefficient on the added regressor is not significant at any of the standard levels: the p-value for the coefficient on drat in M3 is 0.9328; for the coefficient on gear in M4, the p-value is 0.3158. The F-statistics for the nested models M1, M2, and M5 imply that M1 is sufficient for explaining most of the variation in mpg: the p-value for the null hypothesis that M2 does not differ from M1 is 0.1282; the p-value for testing M5 against M1 is 0.5344.

The question remains whether the linear model is even appropriate for this problem. Residual plots and other diagnostics are in Figure 3 in the appendix. It is worth noting that that the errors appear roughly normal and that the scale-location plot indicates no problems with a failure of the homoskedasticity assumption. The residuals plot indicates a less than ideal fit and we should perhaps be concerned about too high leverage amongst some observations (Toyota Corolla, Toyota Corona, Chrysler Imperial have particularaly high dfbetas for some coefficients) but there is little pattern amongst these observations and dropping them does not drastically improve the overall quality of the residuals plot (Figure 4).

Results

Here are the full results for *M1*:

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.173 -1.534 -0.539  1.586  6.081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.418      2.641   14.92  7.4e-15 ***
## cyl           -1.510      0.422   -3.58   0.0013 **
## wt            -3.125      0.911   -3.43   0.0019 **
## am             0.176      1.304    0.14   0.8933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.61 on 28 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.812
## F-statistic: 45.7 on 3 and 28 DF,  p-value: 6.51e-11
```

We cannot conclude that the type of transmission has a significant effect on gas mileage. Though the value of the coefficient on am indicates a positive effect, its p-value of 0.8933 is not significant at any of the standard levels. In particular, the 95% confidence interval for the coefficient on am, (-2.4956, 2.8485), includes zero, implying that we cannot reject the null hypothesis that the effect of am on mpg is equal to zero. For this reason, I hesitate to treat the slight positive increase in gas mileage implied by the coefficient on am in *M1* as a real effect. The coefficients on cyl and wt are significant, implying that increasing the number of cylinders for a car while holding all else constant leads to 1.5102 fewer miles per gallon and that increasing the weight of a car by 1000 pounds, holding all else constant, decreases miles per gallon by 3.1251.

Appendix of figures

Figure 1: A−M Transmission and MPG by Number of Cylinders

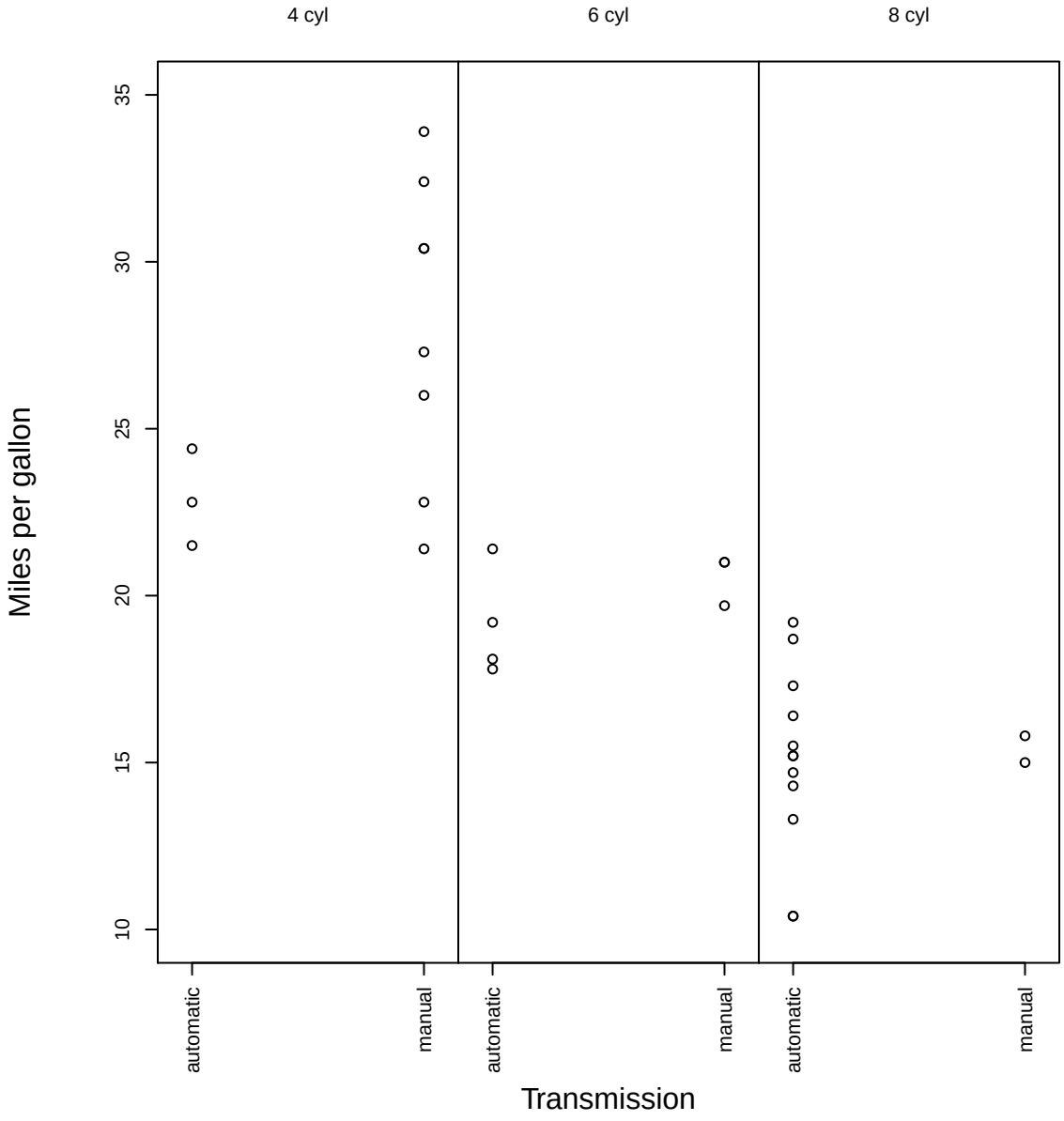

Figure 2: Correlations amonsgst the mtcars data

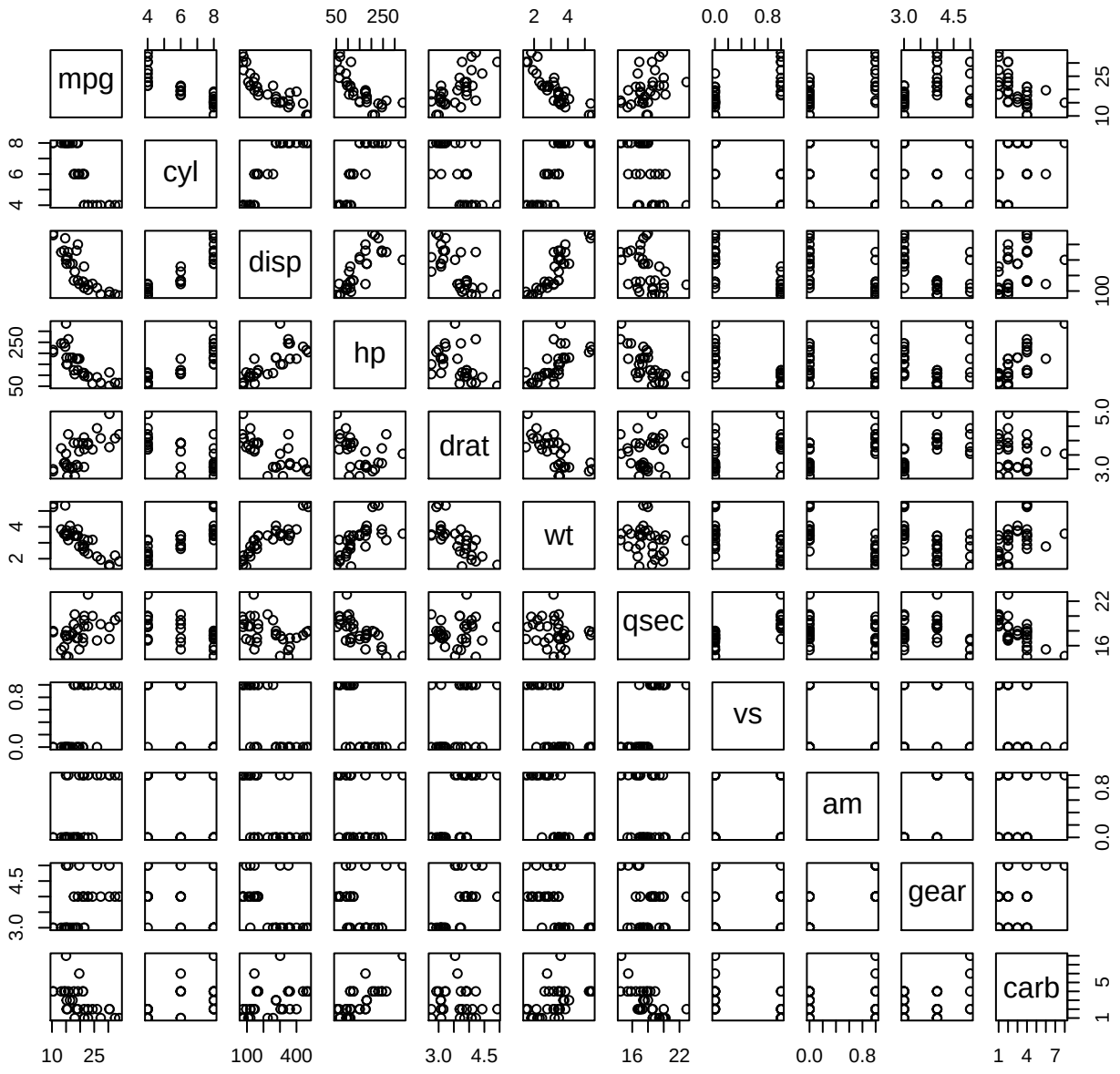

Figure 3: Regression diagnostics for model M1

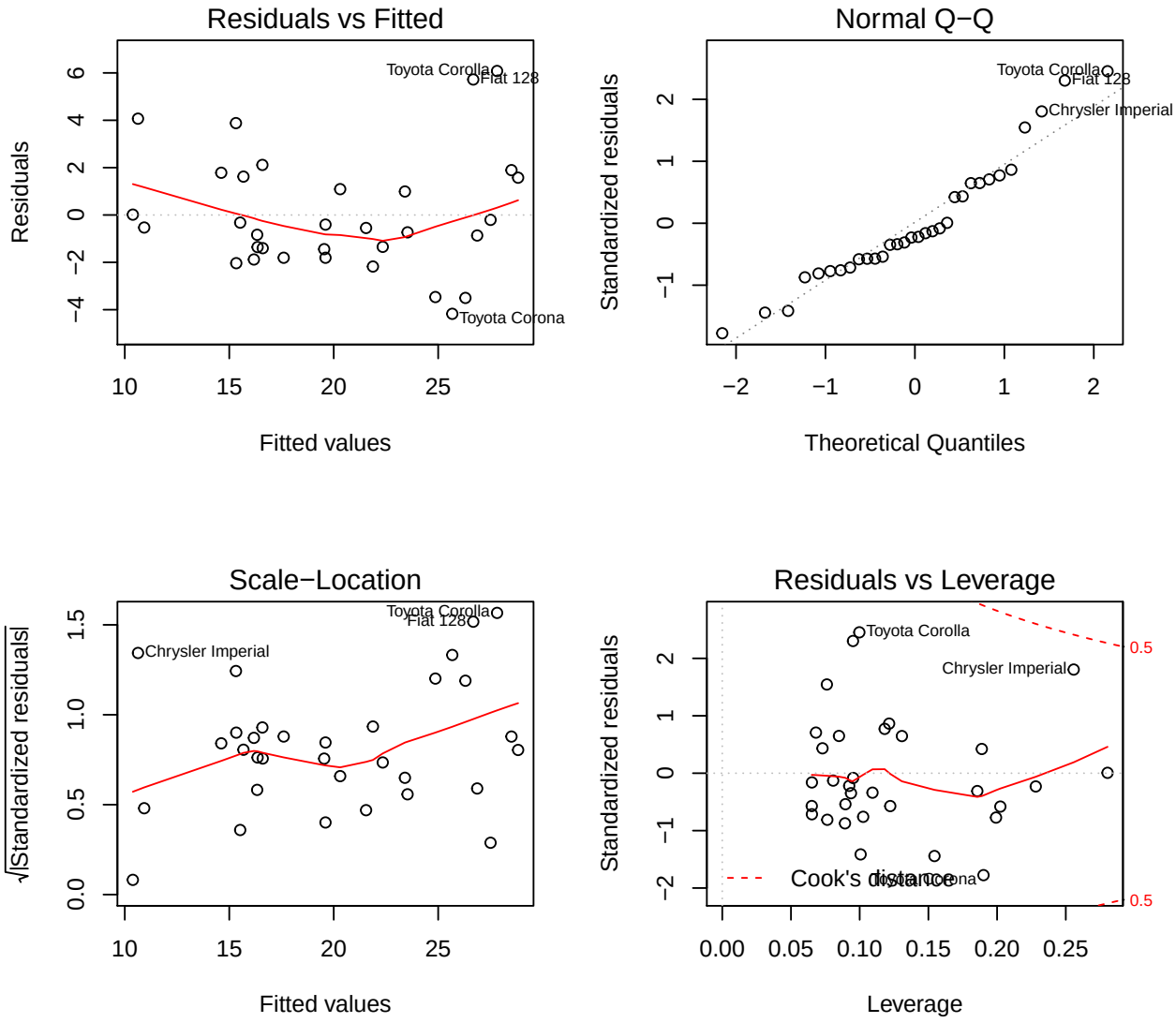

Figure 4: Residuals vs. Fitted after dropping high leverage observations