Fuel Economy and Transmission Type in `mtcars`

Executive Summary

A simple, two variable model of fuel economy generated from the 'mtcars' data set produces remarkably good diagnostic results. However, the data appear to be ill-suited to determining whether transmission type has any influence on fuel economy. A more effective data set for answering that question would include pairwise comparisons of the various car models, each using both transmissions.

Analysis

The `mtcars` data set was extracted from the 1974 *Motor Trend* magazine and is comprised of fuel consumption and ten other aspects of car design and performance for 32 automobiles (1973-74 models). The selection of cars skews to imported (not-American), sports, and luxury cars.

Weight (`wt`) is the variable most correlated with `mpg`, as shown in Table 1. So the exploratory analysis begins with a plot of these two variables.

Table 1. `mtcars` (Abbrevieated) Correlations with mpg

```
##          cyl    disp      hp   drat      wt    vs     am   gear    carb
## [1,] -0.8522 -0.8476 -0.7762 0.6812 -0.8677 0.664 0.5998 0.4803 -0.5509
```

The curved shape of the scatter in fig. 1 suggests that the variation of `mpg` with respect to `wt` is not linear. However, a transformation of the mpg variable to `gpm` (gallons of fuel consumed per 100 miles) appears to be linear with respect to weight (see Fig. 2) and is more likely to satisfy error assumptions since it is in-line with the original measurement – fuel consumption over a 73-mile route (Henderson & Velleman, 397).

```r
#new gallons per mile variable
mtcars$gpm <- 100 / mtcars$mpg
```

There is very little overlap between the automatic and manual transmission cars with respect to weight, as described by Fig. 2. Nearly all the light cars have manual drive trains while the heavy cars use automatics. This lack of heterogeneity makes me skeptical of a model derived from this particular data set that uses the transmission variable.

Seeing the differentiation between the low-gpm Civic and Fiat 128 and the higher-gpm sports cars, it does appear that a measure of how under- or over-powered a car is would add description to the model. Henderson and Velleman choose `hp` divided by `wt`. The new hp/wt (`hpwt`) variable is not correlated with `wt`, so it is acceptable to add in the first model.

```r
# hp per weight measure of over/under-poweredness
mtcars$hpwt <- mtcars$hp/mtcars$wt
cor(mtcars$hpwt, mtcars$wt)
```

```
## [1] 0.05406
```

```r
# Model 1
fit1 <- lm(gpm ~ hpwt + wt, mtcars)
```

A complete summary of the model is provided in Table A-1. `hpwt` is statistically significant at the 0.01 level. `wt` is significant at the 0.001 level. `gpm` increases 1.47 gallons with an increase in `wt` of 1 (1,000 lbs). The Adjusted R-squared value for Model 1 is 0.8379. VIFs are near 1:

```r
vif(fit1)
```

```
##  hpwt    wt
## 1.003 1.003
```

Model 2 is identical to Model 1 with the addition of the transmission variable `am`.

```r
# include a/m transmission variable
fit2 <- lm(gpm ~ hpwt + wt + am, mtcars)
```

A complete summary of the model is provided in Table A-2. The transmission variable `am` is not statistically significant in Model 2. And a likelihood ratio test suggests that the `am` variable is not necessary, Pr>Chisq = 0.81 (> 0.05).

```
## Likelihood ratio test
##
## Model 1: gpm ~ hpwt + wt
## Model 2: gpm ~ hpwt + wt + am
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   4  -30.6
## 2   5  -30.6  1  0.06       0.81
```

A review of the diagnostic plots for Model 1 shows that the residuals are approximately normally-distributed (fig. 3) and that three American luxury cars have a high influence on the model fit according to their Cook's Distance (fig. 4).

Discussion

A more effective method for measuring the possible influence of transmission type on fuel economy would be a pairwise comparison of the various car models, each with both transmissions. For example, the fuel consumption measures for a Toyota Corolla with a manual transmission could be compared with that of a Corolla with an automatic transmission.

References

Henderson, Harold V. and Paul F. Velleman. "Building Multiple Regression Models Interactively" in *Biometrics*, Vol. 37, No. 2 (Jun., 1981), pp. 391-411.

Appendix

Fig. 1. Miles per Gallon vs. Weight

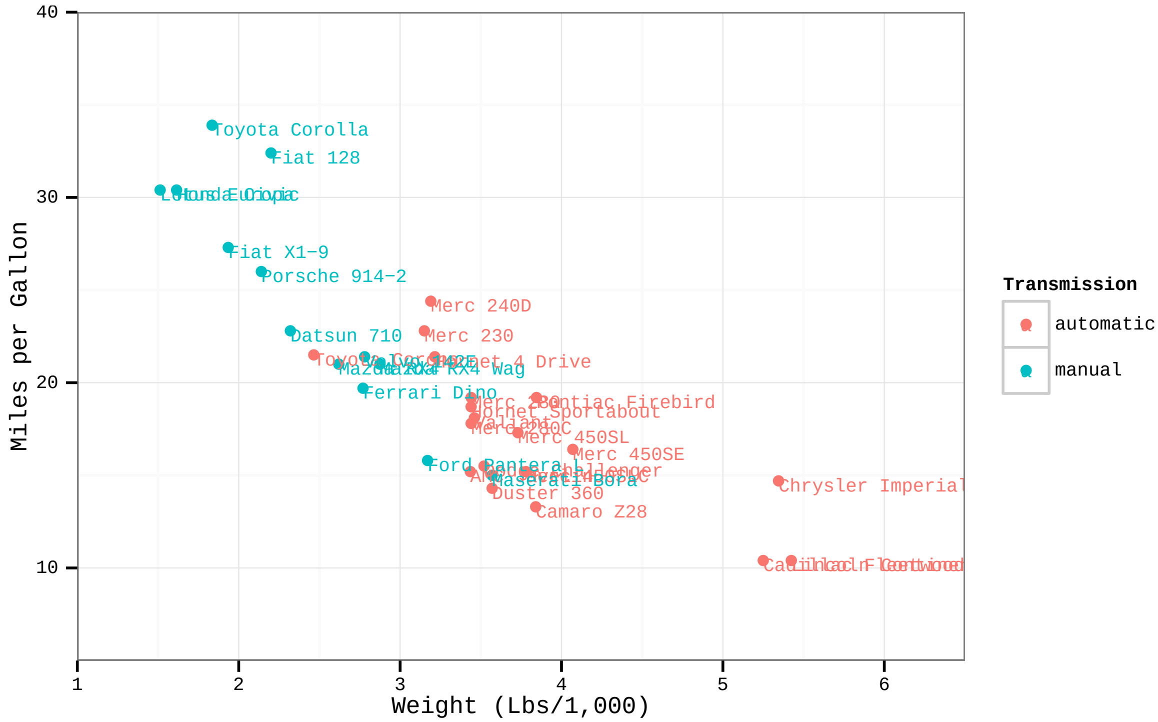

Fig. 2. Gallons per 100 Miles vs. Weight

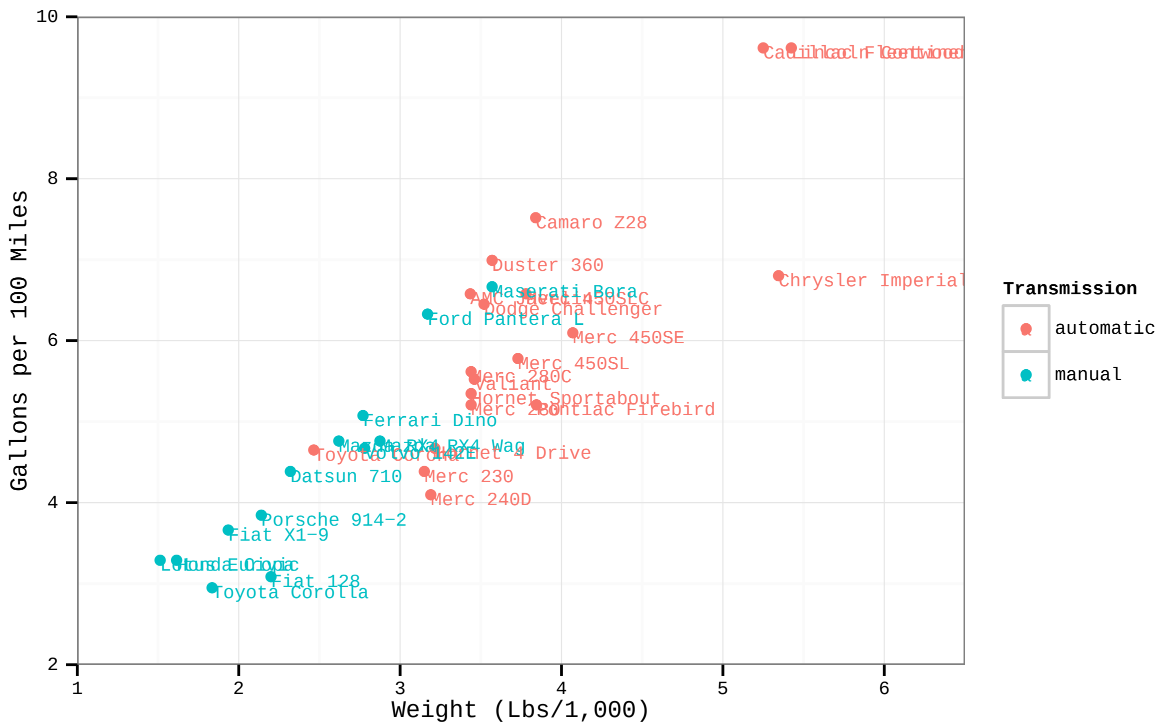

Table A-1. Model 1 Summary

```
##
## Call:
## lm(formula = gpm ~ hpwt + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6971 -0.4682  0.0531  0.4274  1.3510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.4015     0.5120   -0.78   0.4393
## hpwt          0.0240     0.0073    3.29   0.0027 **
## wt            1.4722     0.1216   12.11  7.2e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.661 on 29 degrees of freedom
## Multiple R-squared:  0.848,  Adjusted R-squared:  0.838
## F-statistic: 81.1 on 2 and 29 DF,  p-value: 1.32e-12
```

Table A-2. Model 2 Summary

```
##
## Call:
## lm(formula = gpm ~ hpwt + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7290 -0.4564  0.0219  0.4204  1.3192
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.50080    0.67522   -0.74   0.4645
## hpwt         0.02329    0.00802    2.90   0.0071 **
## wt           1.50236    0.17993    8.35  4.4e-09 ***
## ammanual     0.08369    0.36253    0.23   0.8191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.672 on 28 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.832
## F-statistic: 52.3 on 3 and 28 DF,  p-value: 1.33e-11
```

Fig. 3. Model 1: Residuals vs. Fitted Values

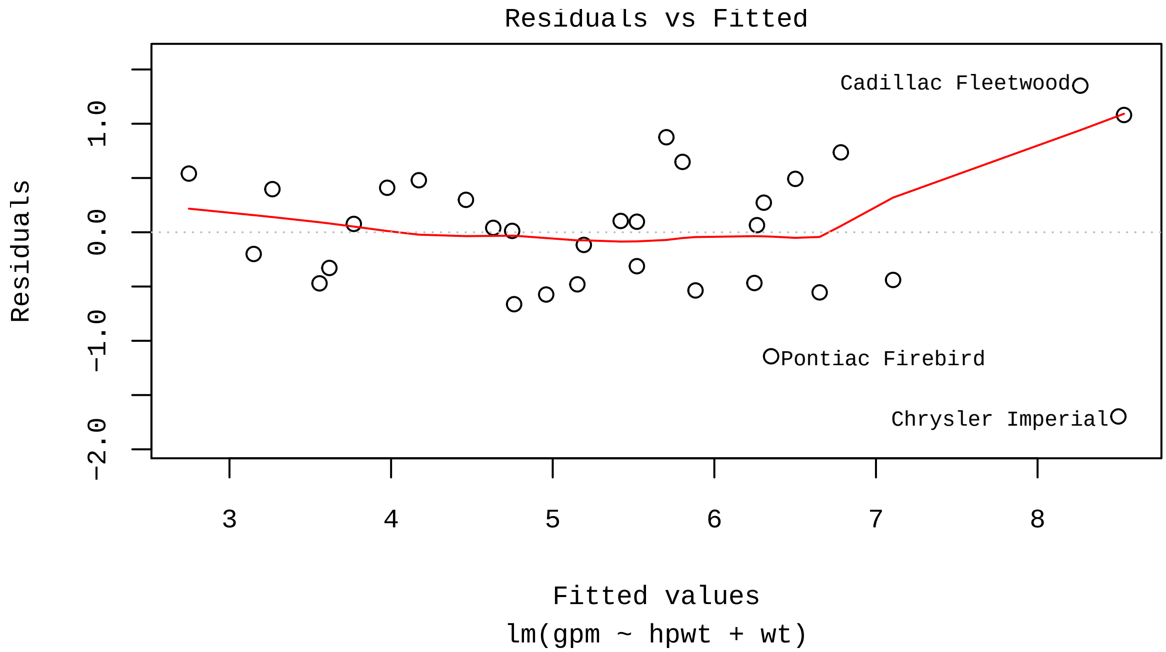

Fig. 4. Model 1: Cook's Distance