Regression Models Course Project

Sheila Braun

July 3, 2018

Executive Summary

This report relates to the mtcars data set. It contains linear modeling analysis and responses to the following:

- **Is an automatic or manual transmission better for MPG?** As of 1974 manaul transmissions were better for MPG than automatic transmissions.
- **Quantify the MPG difference between automatic and manual transmissions**. Going from an automatic (am = 0) to a manual transmission (am = 1) increases **mpg** by 11.9385, unless the car also becomes heavier by half a ton, in which case mpg actually goes down, this time by 4.1974. Since 4.1974 (loss from getting a heavier car) is less than 11.9385 (gain from switching to a manual), it would still be worth getting a manual transmission even if it made your car heavier–at least in 1974.

The final model, detailed below, accounts for .8588 of the variance (adjusted R squared) with a p value that is basically 0.

Data

The data come from Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411. They are collected into a data frame with 32 observations on 11 (numeric) variables. The following table is a summary of the data.

```
##       mpg        cyl        disp             hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec        vs             am      gear   carb  
##  Min.   :1.513   Min.   :14.50   V:18   automatic:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   S:14   manual   :13   4:12   2:10  
##  Median :3.325   Median :17.71                         5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                                4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1  
##  Max.   :5.424   Max.   :22.90                                8: 1  
```

See *Figure 1* in the appendix for an initial look at the relationships among the variables. On the face of it, **mpg** and **am** (automatic vs. manual transmission) have a linear relationship. However **mpg** has interesting relationships with all the variables with the possible exceptions of **gear** and **qsec**. We must look at other variables that may affect **mpg** and at any interaction terms in order to understand thoroughly the relationship between **mpg** and **am**.

Check Assumptions

Linear regression modeling relies on normal data for accurate results. See the appendix, Figures 2 and 3, for a histogram and qqplot of **mpg**. Neither of them looks like a lovely normal distribution should, but given a Shapiro-Wilk normality test with a p-value of 0.1228814, for now we fail to reject the null hypothesis that

mpg is normally distributed, bearing in mind that nevertheless the data are less normally distributed than is ideal.

Early Models

First, examine **mpg** as dependent variable with everything else as predictors. In this initial model, the adjusted R squared is high at 0.8066423 and the overall p-value is quite low at 0.0000004. An anova reveals three significant predictors: **cyl**, **disp**, and **wt**. The variable **am** is not significant in this model, but it is our variable of interest and it has a known linear relationship with **mpg** (see Figure 1 in the appendix), so this model is suspect.

The second model is similar to the first one but the two variables suspected of having little or no linear relationship with **mpg**, that is, **gear** and **qsec**, are no longer in the model. Model 2's p-value is 0 and the adjusted R squared is higher than Model 1's at 0.8118215. An anova shows significant influences of **cyl**, **disp**, and **wt** again. This model shows that we can safely exclude **gear** and **qsec**.

The focus of the third model was to finalize a predictor list without considering any interactions. That the variables **cyl**, **disp**, and **wt** belong in the model is possible based on results of models 1 and 2. The variable **hp** correlates so highly with cyl ($r = 0.8324475$; $p = 0$) that it might be possible to leave it out while including **cyl**. Similarly, **cyl** and **disp** are not only theoretically related, but their correlation is even higher ($r = 0.9020329$; $p = 0$). Keeping **cyl** in the model makes sense, and dropping **disp** from the model also makes sense. Weight stays in because it is theoretically different from the number of cylinders or automatic vs. manual transmission and because it has shown as a strong predictor in earlier models. The variables **drat** and **am** correlate highly and could be accounting for the same phenomenon ($r = 0.7127111$; $p = 0.0000047$); we are interested in **am**, so we drop **drat**. The variables **qsec** and **gear** are already out of the model. The number of carburetors (**carb**) is another variable that correlates highly with **hp** ($r = 0.7498125$; $p = 0.0000008$) but not so much with the heretofore **hp** proxy, **cyl** ($r = 0.5269883$; $p = 0.0019423$). For that reason, we will put **hp** back in and leave **carb** out rather than just putting **carb** in. We do not have any reason to leave out **vs**, so it will also be in the model.

Model 3 accounts for a respectable 0.8228405 of the total variance in **mpg** and it has a p value of 0. The fact that the model acounts for more variance than our previous best model, despite having fewer variables, is encouraging.

Final Model

By way of creating a final model, I looked carefully at all the other models and the correlations to arrive at a final predictor list. This model contains a possible interaction term, weight x automatic vs. manual transmission. In addition, I dropped **hp** and **vs** because they performed poorly in early models.

```
fit5 <- lm(mpg ~ wt + cyl + am + am * wt, mtcars)
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + am + am * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4621 -1.4913 -0.7879  1.3959  5.3499
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)  34.2830     2.7965  12.259 0.00000000000152 ***
```

```
## wt           -2.3689     0.8244  -2.874         0.00782 **
## cyl          -1.1814     0.3803  -3.106         0.00442 **
## am           11.9385     3.8453   3.105         0.00444 **
## wt:am        -4.1974     1.3115  -3.200         0.00350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.265 on 27 degrees of freedom
## Multiple R-squared:  0.877,  Adjusted R-squared:  0.8588
## F-statistic: 48.13 on 4 and 27 DF,  p-value: 0.000000000006643
```

According to this model, mean miles per gallon is 34 when holding all other variables equal. However, for every half ton of weight, **mpg** goes down by 2.3689. Every 2 cylinders further reduces **mpg** by 1.1814. Going from an automatic (am = 0) to a manual transmission (am = 1) increases **mpg** by 11.9385, unless the car also becomes heavier by half a ton, in which case **mpg** again goes down, this time by 4.1974. Because the loss of **mpg** per extra half ton is less than the gain in **mpg** due to switching to a manual transmission, it would still be worth making the switch, at least in terms of miles per gallon–and in 1974. See Figure 3 and 4 in the appendix.

Appendix

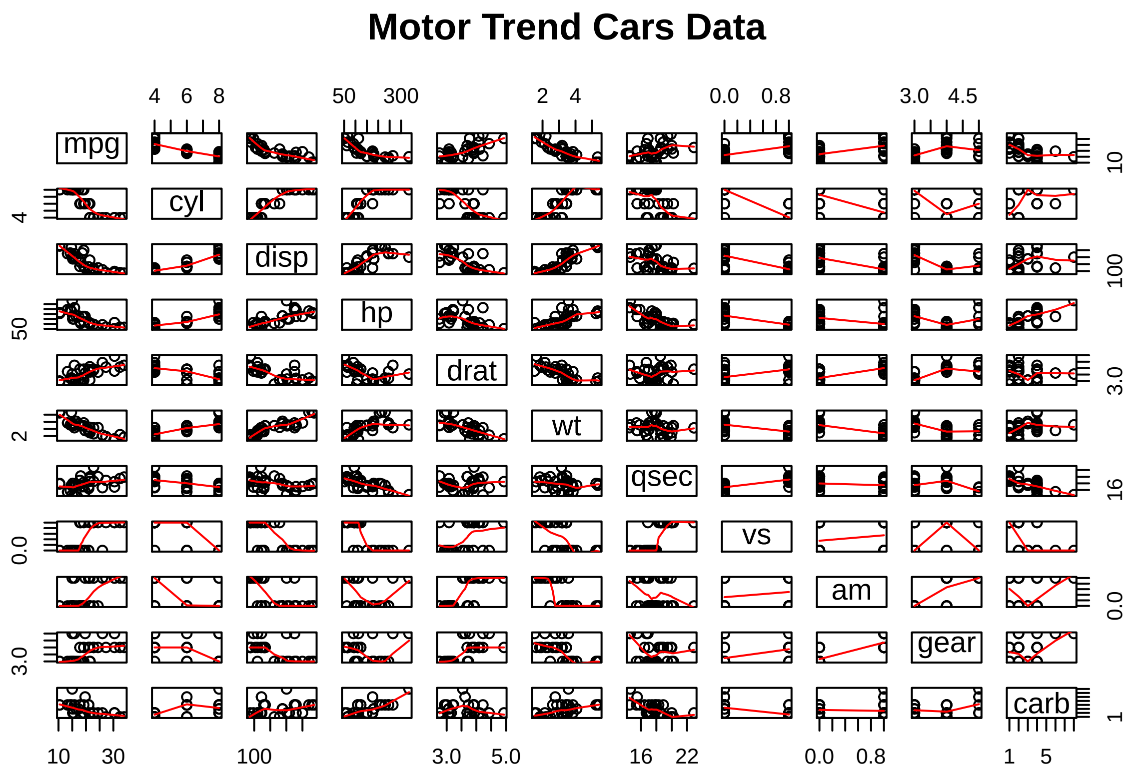

Figure 1. Initial look at relationships among variables.

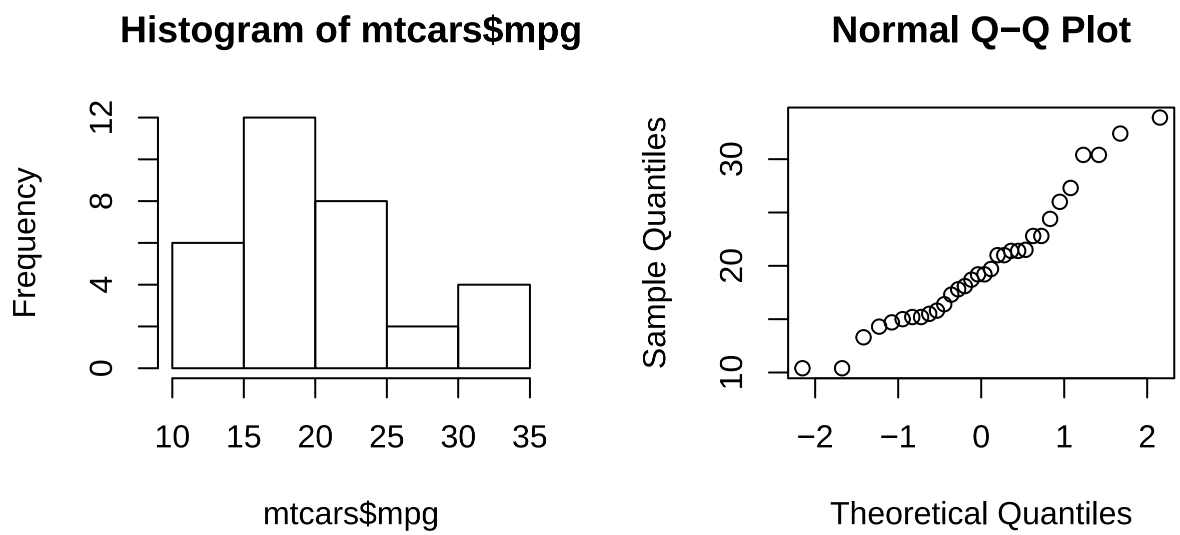

Figure 2. Normality tests for **mpg**.

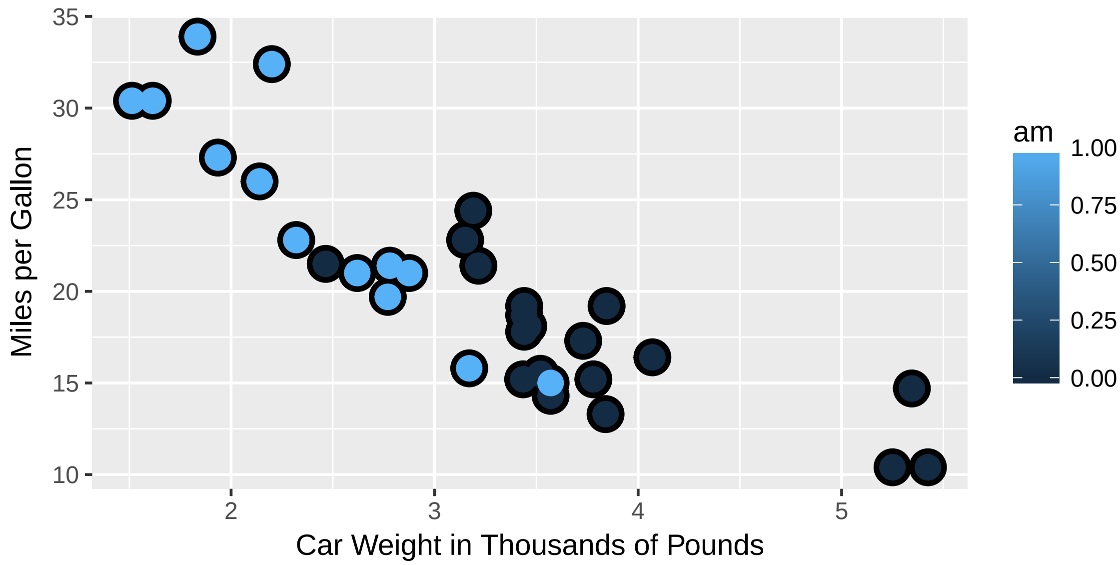

Figure 3. The relationship between **mpg**, **weight**, and **am**.

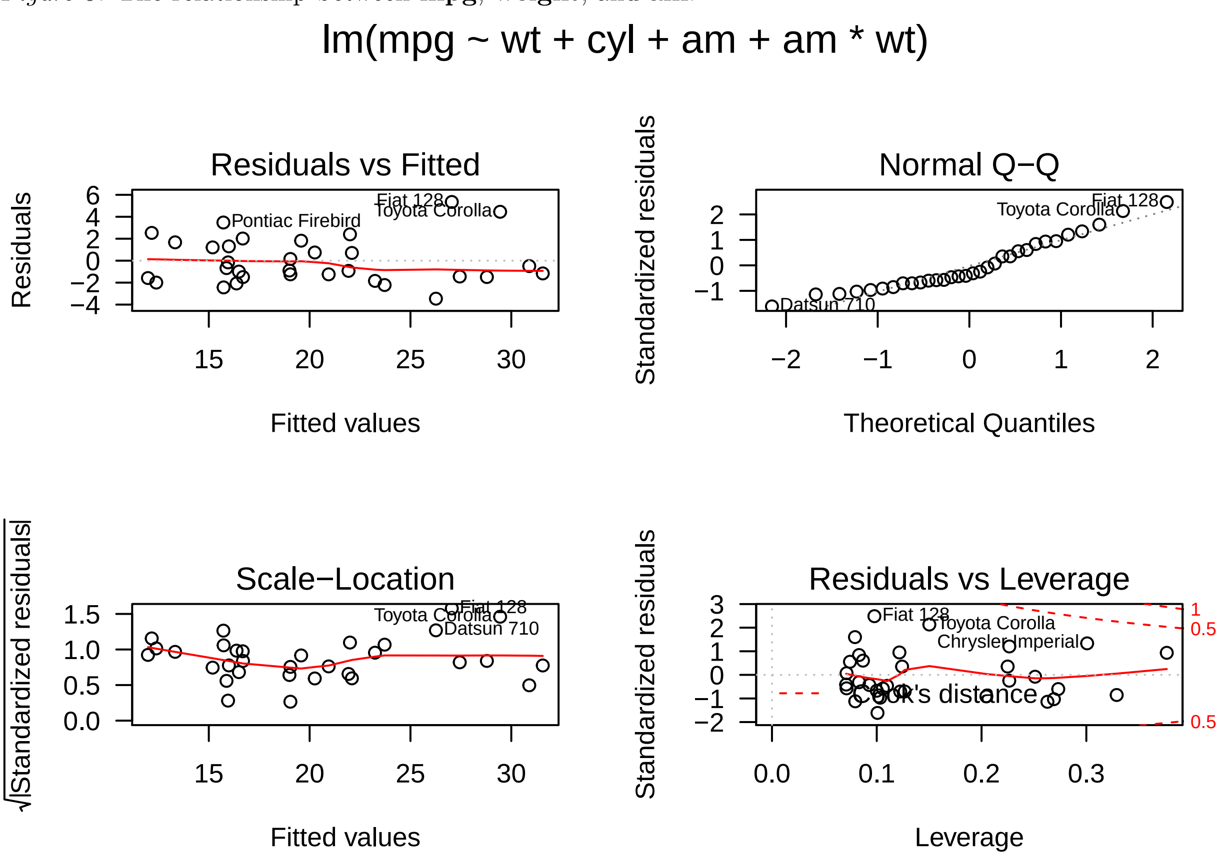

Figure 4. Diagnostic plots.