Regression analysis of mtcars

Andrew J. Dyck

Outline

This report investigates the question of whether an automatic or manual transmission vehicle is more fuel efficient using a sample of 32 different vehicles. An initial analysis is done to explore some features of the dataset and look for possible relationships that may explain MPG in our data sample. Next three models are defined to attempt to explain MPG in the samplem, a champion model is chosen and some model diagnostics completed. Lastly, a conclusion summarizes the analysis and determines that there is no statistical difference in fuel efficiency between automatic and manual transmission vehicles.

Exploratory Data Analysis

First, let's look at the average MPG for automatic and manual transmission vehicles. It's worthy to note that this data is discrete, not continuous, so it must be handled appropriately in this report (ie. as a dummy variable in multiple regression analysis).

```
aggregate(mtcars[, c('mpg', 'cyl', 'disp', 'hp', 'wt')], by=list(mtcars$am), FUN=mean)
```

```
##   Group.1      mpg      cyl     disp       hp       wt
## 1       0 17.14737 6.947368 290.3789 160.2632 3.768895
## 2       1 24.39231 5.076923 143.5308 126.8462 2.411000
```

A quick plot of some features in the dataset (Appendix 1) show that there may be strong relationships between several of these features and fuel efficiency (MPG). It wouldn't be accurate to simply take the means from the table above and conclude that manual transmission vehicles have better MPG without controlling for some addtional features.

Modeling

In this section I attempt to fit a handful of models to the data and search for a model among these that appears to be reasonably sound from a statistical and qualitative perspective. This model will then be the champion model for further analysis. Before diving into running regressions, I'll do a quick correlation check to ensure that multi-collinearity isn't a big problem.

```
cor(mtcars[, c('mpg', 'wt', 'disp', 'hp')])
```

```
##             mpg         wt       disp         hp
## mpg   1.0000000 -0.8676594 -0.8475514 -0.7761684
## wt   -0.8676594  1.0000000  0.8879799  0.6587479
## disp -0.8475514  0.8879799  1.0000000  0.7909486
## hp   -0.7761684  0.6587479  0.7909486  1.0000000
```

General rule-of-thumb is that multi-colinearity tends to become a problem in linear regression when pairwise correlations are over 0.75 to 0.8. This analysis will restrict the features that we can include in regression to Weight and Horsepower since the other variables are too correlated with one another and will cause violations of OLS. Instead of using engine displacement, I'll use some dummy variables for the number of cylinders.

With this outline, I will test the following three models:

```
model1 <- 'mpg ~ am + wt'
model2 <- 'mpg ~ am + wt + hp'
model3 <- 'mpg ~ am + wt + hp + cyl6 + cyl8'
```

Model Results

```
##                  Model1    Model2    Model3
## Estimate    -0.02361522 2.0837101 1.8092114
## P.Value      0.98791459 0.1412682 0.2064597
## AdjRSquared  0.73578891 0.8227357 0.8400875
```

Champion model

I've determined that Model 2 should be the champion model. Although model 3 provides a slighly higher adjusted R^2, I suspect that the dummy variables I created to capture the number of cylinders is co-linear with a car's horsepower, and the addition of these dummy variables complicates the quantification of the automatic vs. manual transmission effect since the intercept is now defined as automatic 4-cylinder cars.

Residual diagnostics of champion model

An image of 4 residual diagnostic plots in Appendix 2 reveal that residuals from the champion model are likely normally distributed with a mean of zero. Based on this analysis, it seems that this model does not break the assumptions of OLS.

Conclusion

Accounting for vehicle weight and horsepower, our champion model reveals that, while it is possible that manual transmissions use less fuel per mile driven (higher MPG) than vehicles with an automatic transmission, this effect is not statistically different from zero at the $\alpha = 0.05$ level of significance.

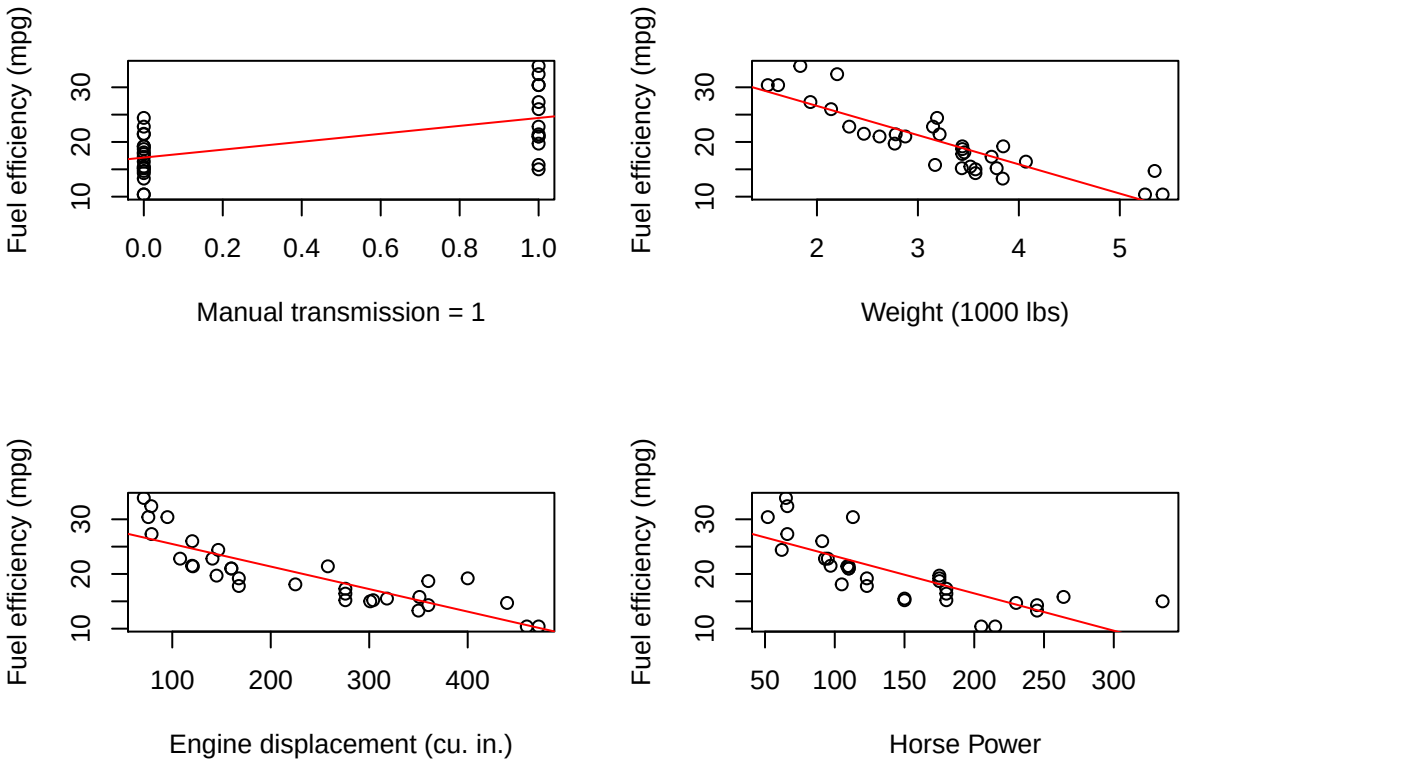
Rather, it seems that vehicle weight is a strong predicter of fuel-efficiency. Our model suggests that a car that weight 1,000 lbs more than the average vehicle will have a -2.88 lower MPG, all else being equal. The horsepower of a vehicle, an instrument for engine size and displacement, also has a negative effect on fuel efficiency. A vehicle with an additional 10 HP of engine power suggesting a reduction in MPG of 0.37. Both of these effects are statistically significant at the $\alpha = 0.05$ level of significance.

However, given the small sample size, if we make the strong assumption that the coefficient on the transmission type is correct at 2.08 in order to quantitatively answer the main question behind this assignment, we would say that all else being equal, a vehicle with an automatic transmission will have a reduced fuel efficiency of 2.08 compared to a manual transmission vehicle. That said, I would again suggest that this analysis cannot answer the question of quantifying the MPG difference between automatic and manual transmission vehicles.

Appendix

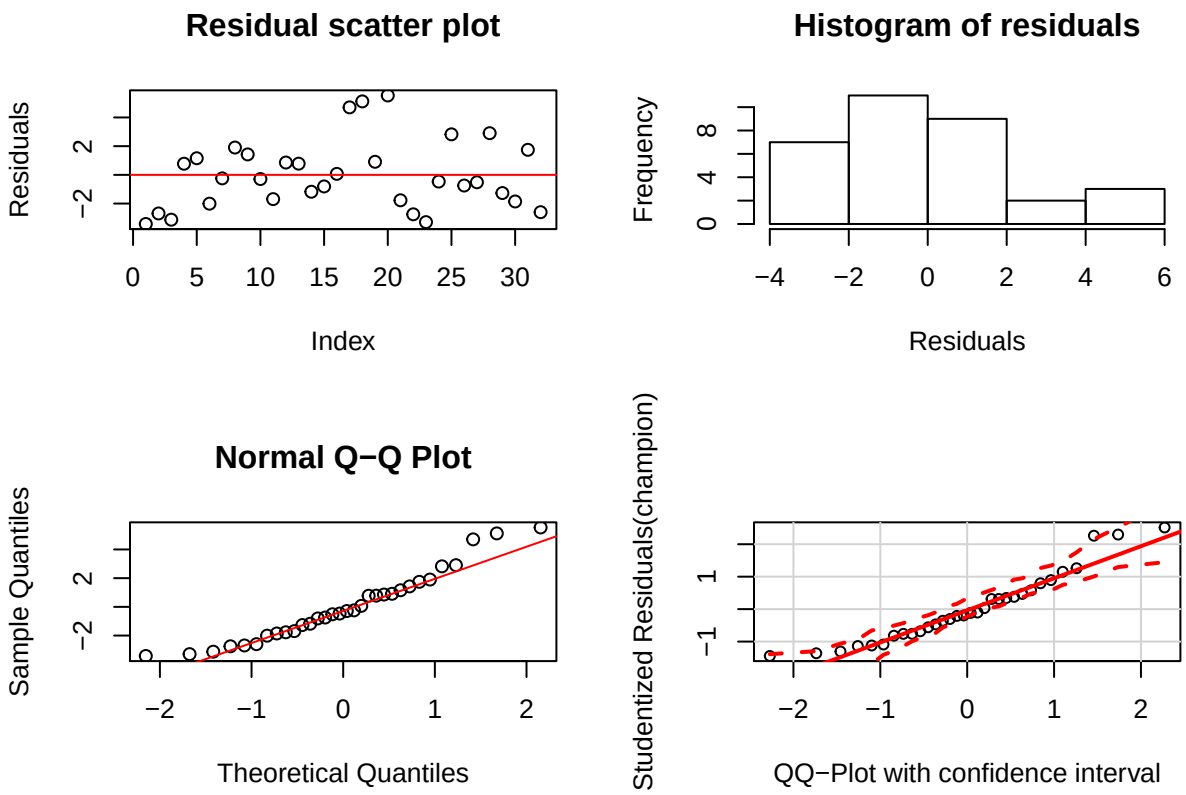
Appendix 1: Exploratory plots

```
par(mfrow=c(2,2))
plot(mtcars$am, mtcars$mpg, xlab = 'Manual transmission = 1', ylab = 'Fuel efficiency (mpg)')
abline(lm(mpg~am, mtcars), col="red")
plot(mtcars$wt, mtcars$mpg, xlab = 'Weight (1000 lbs)', ylab = 'Fuel efficiency (mpg)')
abline(lm(mpg~wt, mtcars), col="red")
plot(mtcars$disp, mtcars$mpg, xlab = 'Engine displacement (cu. in.)', ylab = 'Fuel efficiency (mpg)')
abline(lm(mpg~disp, mtcars), col="red")
plot(mtcars$hp, mtcars$mpg, xlab = 'Horse Power', ylab = 'Fuel efficiency (mpg)')
abline(lm(mpg~hp, mtcars), col="red")
```



Appendix 2: Residual diagnostic plots

```
par(mfrow=c(2,2))
plot(champion$residuals, main='Residual scatter plot', ylab='Residuals')
abline(0, 0, col='red')
hist(champion$residuals, main='Histogram of residuals', xlab='Residuals')
qqnorm(champion$residuals)
qqline(champion$residuals, col='red')
qqPlot(champion, 'QQ-Plot with confidence interval')
```



Appendix 3: Full model summary

```
summary(champion)
```

```
##
## Call:
## lm(formula = model2, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```