Regression Models Course Project

GOAL : Exploring the relationship between a set of variables (Transmission Type) and miles per gallon (MPG Fuel Efficiency)

Report Executive Summary:

- We are interested in determining whether an automatic or manual transmission is better for fuel efficiency (higher MPG), and we want to quantify the MPG difference between automatic and manual transmissions. We address these questions by analyzing the Motor Trend Car Road Tests data set available in R. In general, a car picked at random with a manual transmission has better fuel efficiency than a car picked at random with an automatic transmission. However, once you properly account for the important confounding effect of automobile weight, there is not enough information to answer these questions about the effect of transmission type on fuel efficiency across all weight ranges. But, we can be confident in saying that fuel efficiency decreases more slowly with respect to increased weight for automatic transmissions than for manual transmissions.

Report Exploratory Analysis:

- If we plot just the variables in the question (figure 1), it seems on first look obvious that manual transmissions have a higher average MPG than automatic transmissions. We first build an exploratory linear model using every feature available in the data set (a “Funnel Flow” model):

```
funnel_flow <- lm(mpg ~ ., mtcars)
```

- The given table lists the coefficients of the funnel flow model. With so many variables and so few data points, the p-values for the coefficients are not statistically significant at $p < 0.05$; however, weight (denoted “wt”) does have the most extreme absolute p-value by far of all the variables, and a plot of weight versus MPG indicates that we should account for weight (figure 2). This makes sense physically; it takes more energy to move heavier objects. In fact, our initial understanding to the effect of transmission type on fuel efficiency might even be completely explained by the weight.
- The funnel flow model has a positive coefficient for having a manual transmission (amManual) of 2.5202, indicating that having a manual transmission increases fuel efficiency by 2.5202 MPG when we hold the other variables constant, but the coresponding p-value is only 0.2340, which we reject at a signficance level of 0.05. So we cannot make a statement about a significant effect on fuel efficiency by the transmission type under this model.

Analyzing the Bi-Models of Weight and Transmission Type:

- For model selection, we will search for a model with a very low p-value with respect to the coefficients for transmission type, so we can be more confident about our conclusions. Based on our exploratoration, we will consider a linear model that only takes into account weight and transmission, and one with the interaction term added:

```
fit <- lm(mpg ~ am + wt, mtcars)
interaction_fit <- lm(mpg ~ am * wt, mtcars)
```

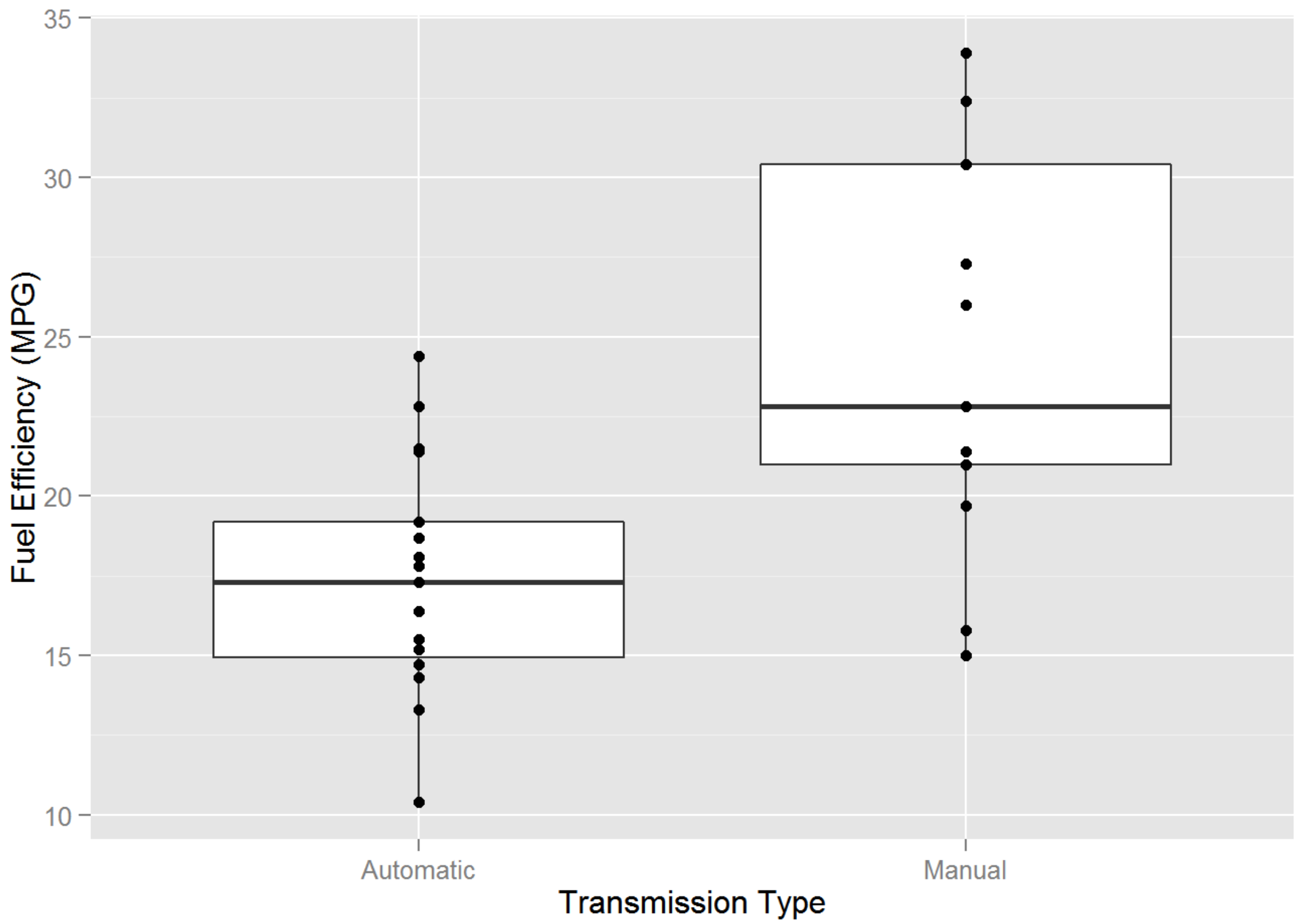
- The given table lists the coefficients of the first model. **After accounting for weight discrepencies, it appears that having a manual transmission might slightly reduce fuel efficiency if we assume the same rate of change in fuel efficiency with respect to weight for the two kinds of transmissions. However, there is a very high p-value, so we would be wrong to make a such a claim on this data alone.** *If* it were signficant, we would have looked at the amManual coefficient and say that, comparing cars of the same weight, the a manual transmission would have 0.02 MPG lesser fuel efficiency.
- The given table lists the coefficients of the second model. **With a p-value of 0.001, we can say that, for a given weight, when we go from an automatic transmission to a manual transmission, fuel efficiency degrades on average 5.2984 MPG faster.** The corresponding plot in is in figure 4. The p-value for the difference in the F statistic from ANOVA between the two models is 0.0010171, indicating that we need to include the interaction term.

Diagnostics and Discussion

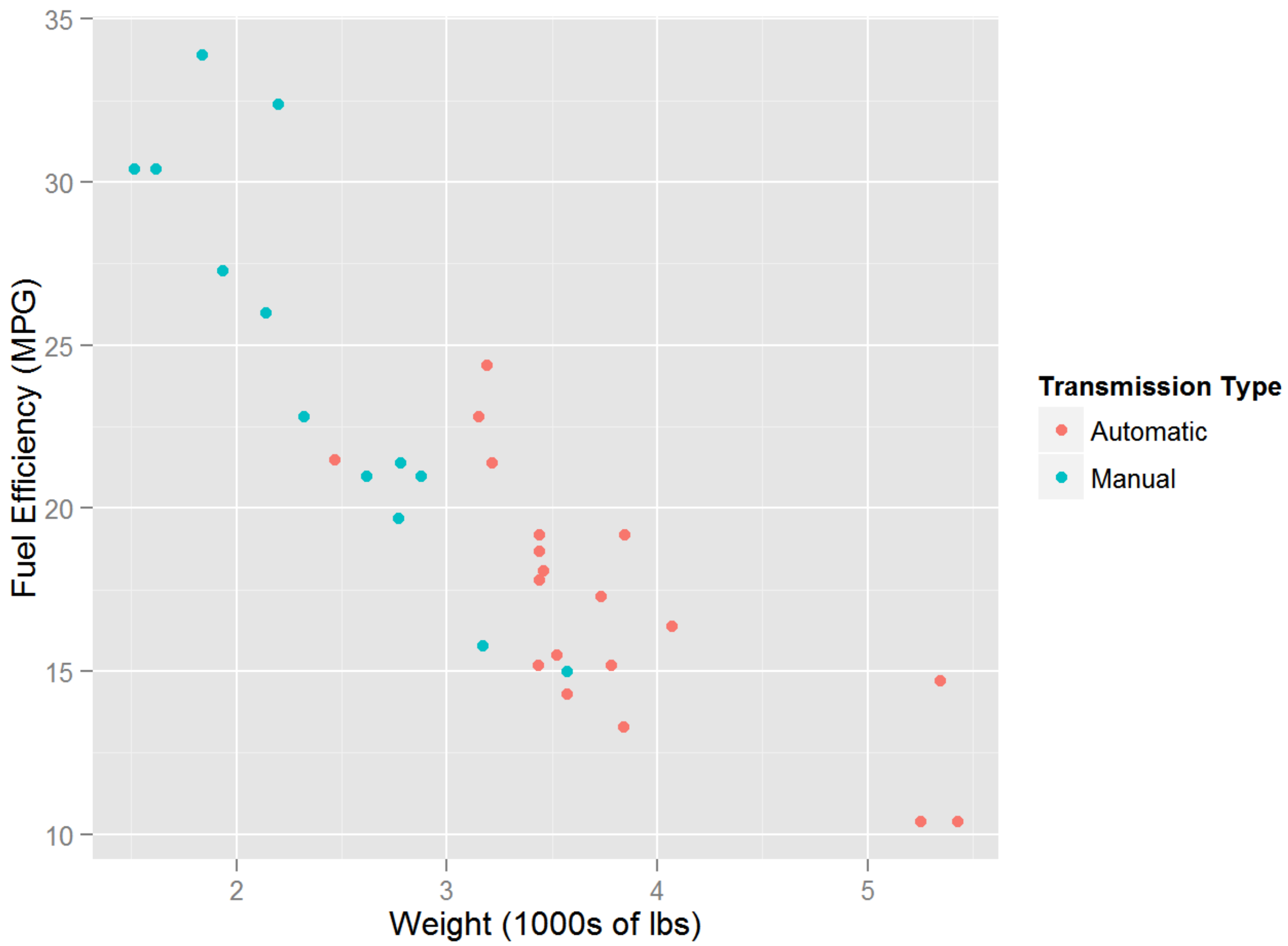
- Figure 4 is a plot of the residuals versus the fitted values under the interaction model, as well as a Normalized Q-Q plot. We see that we have divergent behavior in the tails, and there is almost certainly other variables that could be involved. Finally, if we look at the PRESS residuals to find the car that deviated the most from the interactions model when held out of the model was Fiat 128; it would be worth investigating if there are any particular features of this car that make it stand out.
- Since weight has such an important impact on MPG, and there was relatively little overlap in weight distributions between the two groups (see figure 2 again), the conclusions we make are highly based on the model we use. Because of this small overlap, it would be interesting to consider the hypothesis that companies simply build cars with the more efficient transmission type for their weight.
- A caveat to our analysis is that we are using data from 1974. It would probably serve us better to use data from cars that have benefited from forty years of advancements in automotive engineering, and in particular look at cars that come with a choice of automatic and manual transmission.

Appendix

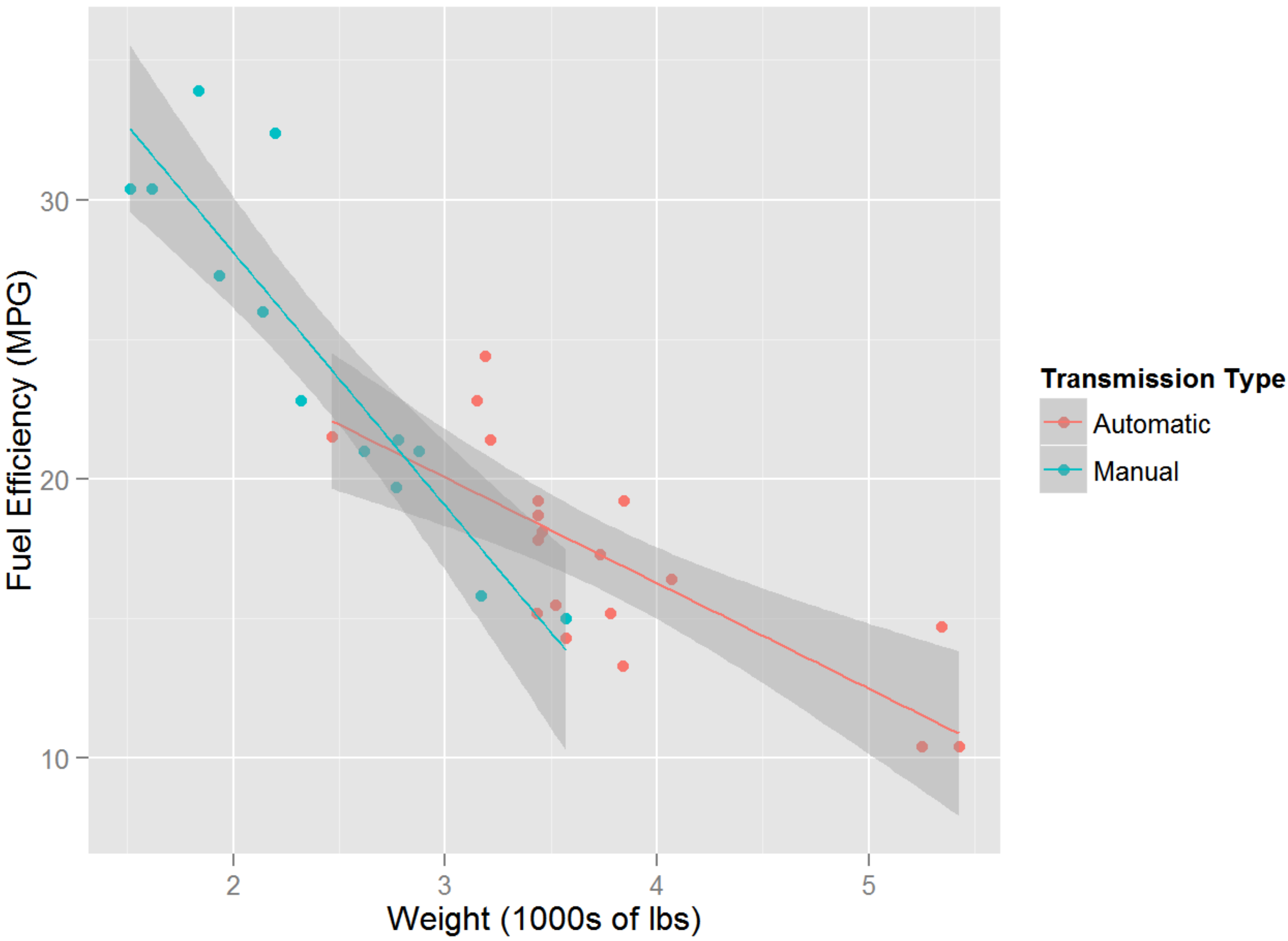
- The raw source for this analysis is at (https://github.com/iyermobile/Regression-Models)https://github.com/iyermobile/Regression-Models (https://github.com/iyermobile/Regression-Models).
- The code for plotting the residuals and Q-Q plot came from http://ygc.name/2011/08/17/ggplot2-version-figures-25-recipes-started-r/ (http://ygc.name/2011/08/17/ggplot2-version-figures-25-recipes-started-r/)](http://ygc.name/2011/08/17/ggplot2-version-figures-25-recipes-started-r/ (http://ygc.name/2011/08/17/ggplot2-version-figures-25-recipes-started-r/)).



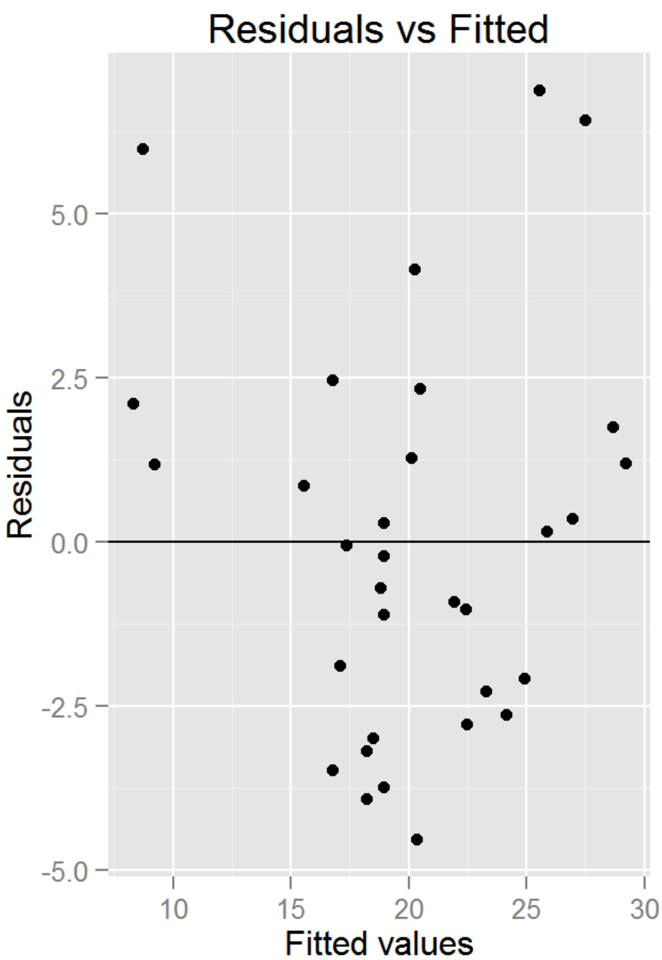
35
30
25
20
15
10
Fuel Efficiency (MPG)
Automatic
Manual
Transmission Type



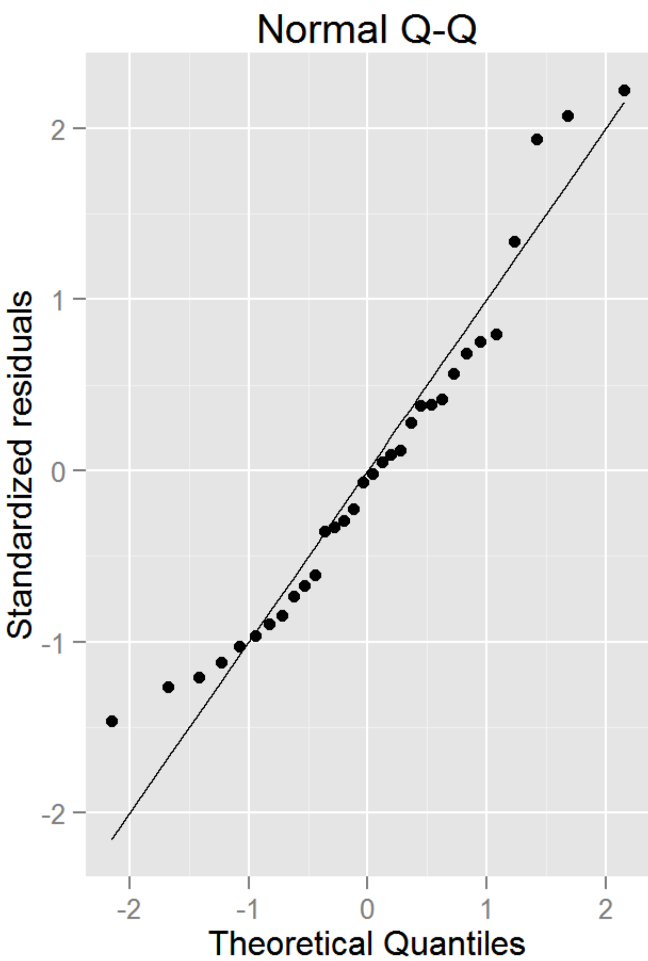
35
30
25
20
15
10
Fuel Efficiency (MPG)
2
3
4
5
Weight (1000s of lbs)
Transmission Type
Automatic
Manual



Fuel Efficiency (MPG)
30
20
10
2
3
4
5
Weight (1000s of lbs)
Transmission Type
Automatic
Manual



Residuals vs Fitted
Residuals
5.0
2.5
0.0
-2.5
-5.0
10
15
20
25
30
Fitted values



Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles