Regression Models Course Project

1 Executive Summary

We analyze the mtcars data set in order to answer the question of whether an automatic or manual transmission is better for miles per gallon (MPG), and to quantify the difference. Our exploratory analysis shows that transmission type is correlated with characteristics such as weight and horsepower. We fit a variety of linear models to adjust for the confounding factors and estimate the effect of transmission. When we adjust for the confounding factors we are unable to estimate the effect of transmission reliably. Specifically, weight is the predictor that is most highly correlated with MPG, and weight is also very highly correlated with transmission type. This means that our data set is not able to tell us what the effect of transmission type is on MPG.

2 Data Exploration

Figure 1. shows the pair-wise relationships between the continuous variables. We see that MPG is strongly inversely correlated with horsepower (`hp`), displacement (`disp`), and weight (`wt`), and correlated with quarter mile time (`qsec`) and rear axle ratio (`drat`). Also, we see that cars with manual transmissions (manuals) are mostly light weight. The heaviest manual car is lighter than the 9 heaviest automatics.

The data exploration has uncovered a basic problem: manuals and automatics differ greatly in the other variables. Ideally we want to compare two cars that are identical in all aspects *except* transmission type. Since we do not have any cars like that in our data set the effect of transmission will largely be determined by our model. If we are highly confident in our model then that might not be a problem, but uncertainty about the model will make us unsure of the effect of transmission on MPG.

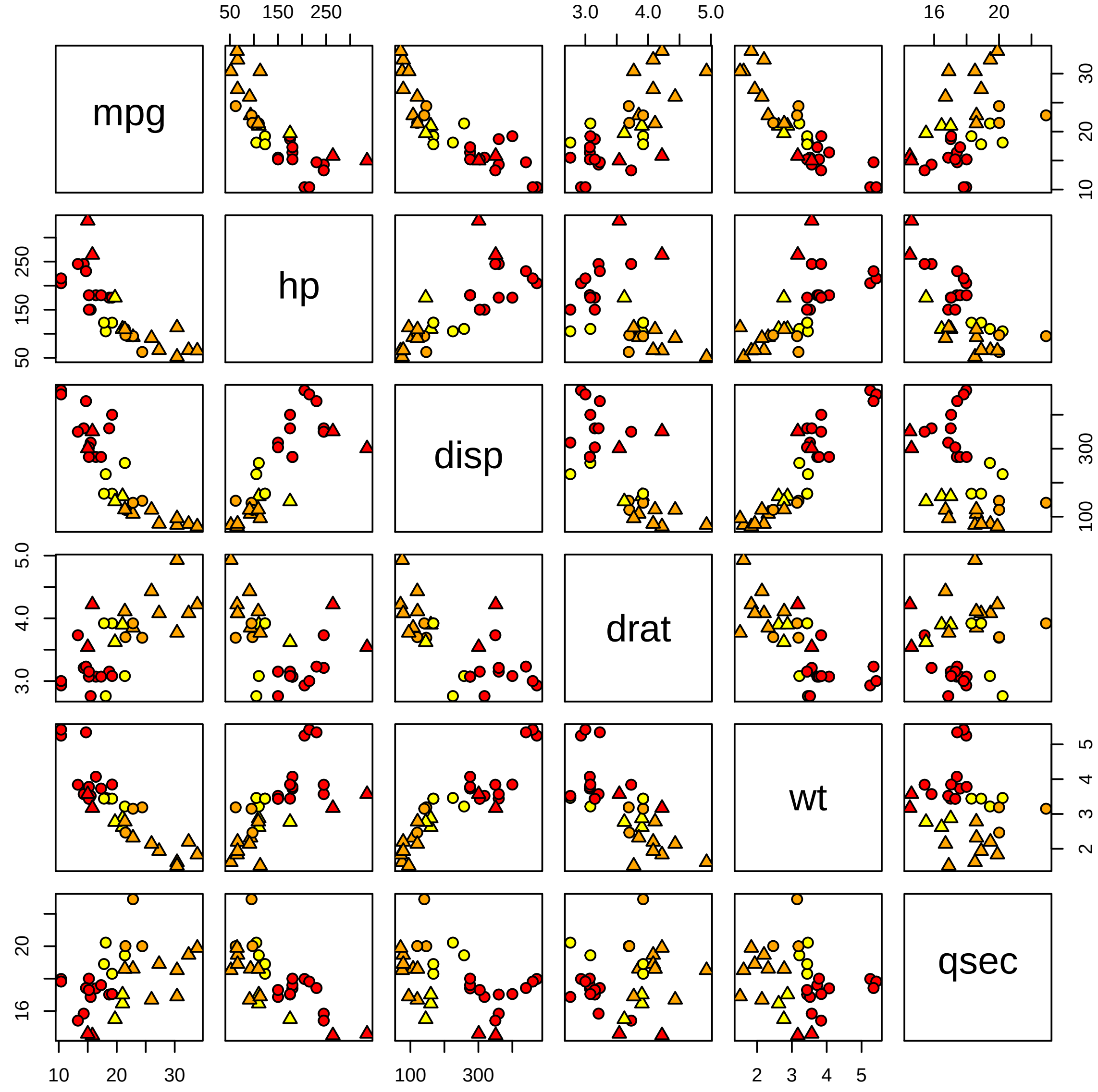
3 Modeling

We start by taking the most basic model of averaging MPG over automatic and manual transmissions. We have renamed the variable `am` from the `mtcars` data set to `auto` where `auto = 1` indicates an automatic transmission.

$$mpg = 24.392 + (-7.245) * auto \tag{1}$$

The coefficient for `auto` is -7.245 which means that automatics have an average MPG of `17.147 = 24.392 +` -7.245. However, from our data exploration we saw that there are confounding variables that may explain away the effects of transmission.

Figure 1: Pair-wise plots of continuous variables. Automatics are shown as triangles. Four, six, and eight cylinders are shown in orange, yellow and red respectively.



For our second model we will include weight as a predictor.

$$mpg = 37.298 + -5.353 * wt + 0.024 * auto \tag{2}$$

By accounting for weight we have reduced our estimate of the `auto` coefficient to 0.024 with a standard error of 1.546, meaning that we cannot estimate the impact of transmission accurately. The residual standard error is 3.098 which gives an approximate measure of the precision of our model. The residual plot for model (2) (not shown) shows no bias by transmission, but cylinder shows bias in that 4 cylinders are estimated too low and 8 and 6 cylinders too high.

$$mpg = 32.358 + (-3.153) * wt + (-0.022) * hp + (3.359) * cyl.4 \tag{3}$$

$$mpg = 30.163 + (-2.428) * wt + (-0.025) * hp + (11.143) * cyl.4 + (-3.071) * wt * cyl.4 \tag{4}$$

$$mpg = 30.032 + (-2.006) * wt + (-0.028) * hp + (10.301) * cyl.4 + (-1.094) * wt * cyl.4 + (-2.817) * auto \tag{5}$$

Models (3) and (4) and (5) use the predictor `cyl.4` which is 1 for four cylinders and 0 otherwise. These models account for the bias by cylinder in model (2).

We preform a likelihood ratio test to determine which model to use. The results support including the interaction term `wt*cyl.4 (p = 0.042)`, but excluding `auto (p = 0.398)`. Thus we take model (4) as our final model.

Figure 2. displays the fitted lines and residuals for model (4). Visually the slopes of weight for four cylinder cars is much greater than for cars with six or more cylinders. This can be seen by the value of the coefficient for `wt*cyl.4` of -3.071. Also, the residual plots show slightly higher variation at low weights. **Figure 3.** shows diagnostic plots for model (4) and **Figure 4.** displays the coefficients and their standard errors. The QQ plot shows a deviation from the 45 degree line indicating that MPG is higher for the most fuel efficient cars than our model predicts. The residual standard error of the model is 2.255, which means that we can predict mpg to a precision of about 2.25 MPG.

4 Conclusion

Our final model, model (4), is able to predict MPG reasonably well using only 4 predictors and an intercept term. We did not include `auto` as a predictor because we were not able to estimate it reliably from the data. It is possible that with more data we would be able to estimate the effects of transmission reliably and that the effects could be significant. However, given our current data the effect of transmission cannot be determined.

Figure 2: fitted line and residual plot for Model 2. As before manuals are shown as triangles and number of cylinders is shown by color

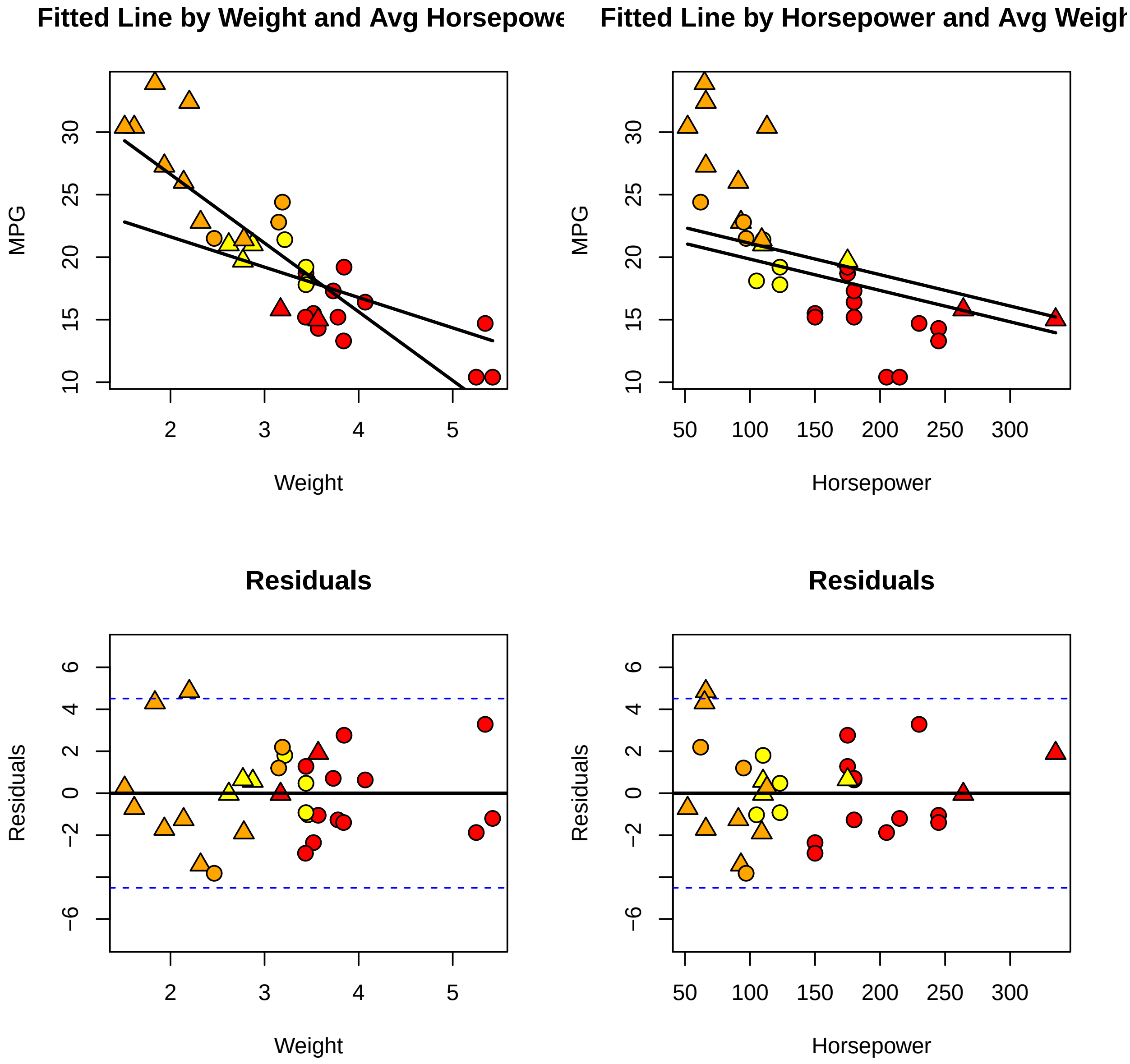


Figure 3: Diagnostic plots for final selected model - Model 4

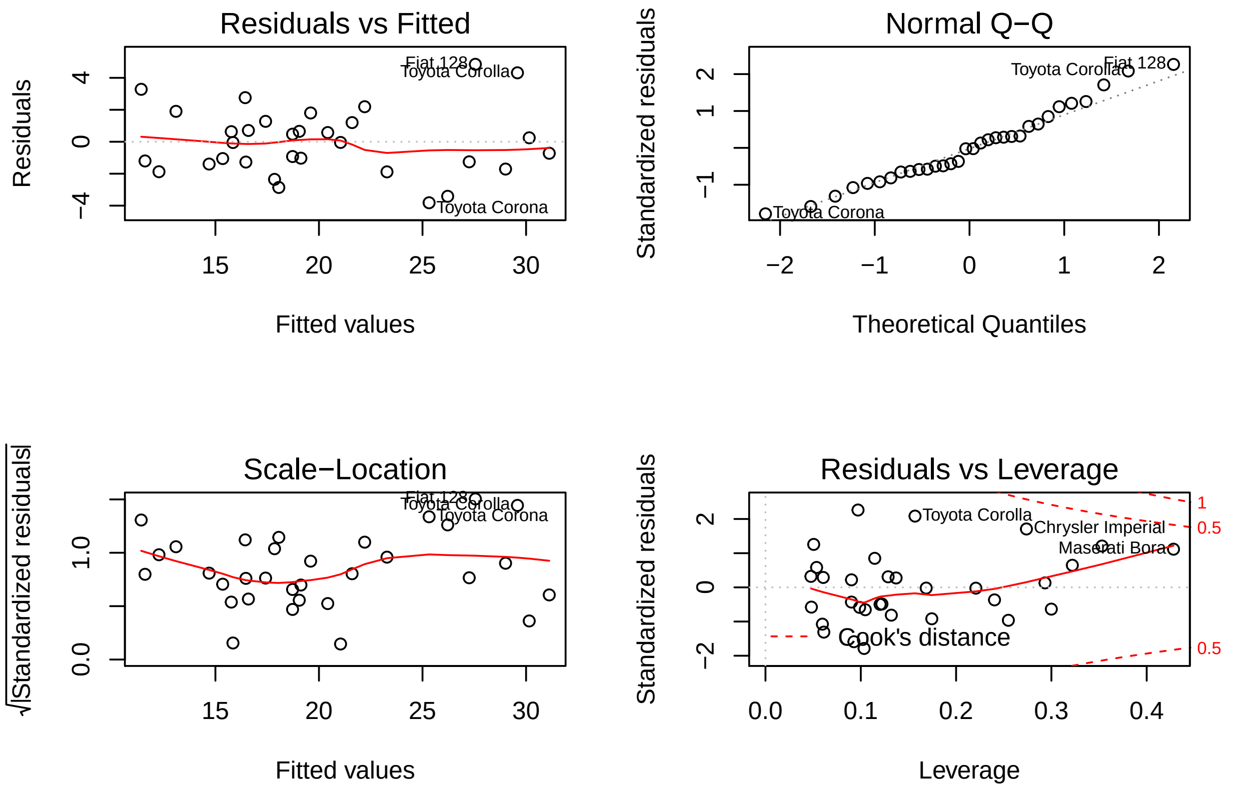


Figure 4: Coefficients for Model 4

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       30.163      2.549  11.832    0.000
## wt                -2.428      0.703  -3.454    0.002
## hp                -0.025      0.009  -2.866    0.008
## cyl.4             11.143      3.843   2.900    0.007
## wt:cyl.4          -3.071      1.428  -2.151    0.041
##   Resid.Std.Error R.Squared
## 1           2.255     0.878
```