Regression Models Course Project Submission

Executive Summary

This analysis uses the mtcars dataset to investigate whether or not automatic or manual transmission is better for mileage. This dataset was extracted from the 1974 Motor Trend magazine and includes 10 aspects of automobile design and performance for 32 automobiles. It is found that the transmission type does have an influence on mpg in a simple marginal model, however the difference in average mpg is found to be not statistically significant when the data is fit with a multivariate model.

Variables

The mtcars dataset includes the following variables. For this analysis, am is renamed as transmission. The following variables are treated as categorical: cyl, vs, transmission, gear, and carb.

- mpg – Miles/(US) gallon
- cyl – Number of cylinders
- disp – Displacement (cu.in.)
- hp – Gross horsepower
- drat – Rear axle ratio
- wt – Weight (lb/1000)
- qsec – 1/4 mile time
- vs – V/S
- am – Transmission (0 = automatic, 1 = manual)
- gear – Number of forward gears
- carb – Number of carburetors

The Marginal Model

As seen in Figure 1 in the appendix, mpg does appear to increase with a manual transmission versus an automatic one. The marginal linear model including only the transmission type as a factor reveals that the null hypothesis $H_0 : \beta_{transmissionManual} = 0$ (the difference in mean mpg between manual versus automatic transmissions is zero) is rejected using a p-value of 0.05 as a benchmark. The expected difference in mean mpg between manual and automatic transmissions is 7.24 gallons. We can say with a 95% confidence that the difference in mean mpg between manual and automatic transmissions is between 3.64 and 10.85. The R^2 for this model is 0.3598, suggesting that inclusion of other variables is likely required.

```
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          17.147      1.125  15.247 1.134e-15
## transmissionManual    7.245      1.764   4.106 2.850e-04
```

Examining either the pairwise scatterplot (Figure 2 in the appendix) or the correlation coefficients below shows that the transmission is not independent of other variables in the dataset and mpg has other dependencies. Specifically, transmission shows correlation with the variables drat, wt, and gear and mpg shows correlation with cyl, disp, hp, and wt. It is clear that the marginal linear model is likely not adequate and that more complex models need to be explored.

Select Correlation Coefficients:

```
##                cyl  disp    hp drat    wt  qsec   vs gear  carb
## mpg          -0.85 -0.85 -0.78 0.68 -0.87  0.42 0.66 0.48 -0.55
## transmission -0.52 -0.59 -0.24 0.71 -0.69 -0.23 0.17 0.79  0.06
```

Exploring Multivariate Models

The approach taken to finding a statistical model that includes more variables is to start with a model that includes all the variables, choose the variable with the highest p value, and then do an anova comparison to determine whether or not there is any difference in the model by removing that variable. If there is a significant difference in models, the variable is retained, otherwise it is removed from the model. This is iteratively done through all the variables (except transmission) with order determined by p-values in the model that was inclusive of all the variables.

The final model obtained using this method (fit9) includes the covariates wt, hp, and the transmission type. The coefficients are summarized below and the R^2 is now 0.84, an improvement over the marginal linear model. Note that the p-value for $\beta_{transmissionManual}$ is no longer less than 0.05. By adding the covariates to the model, the null hypothesis that $\beta_{transmissionManual} = 0$ can no longer be rejected. Running an anova comparison between the model of mpg ~ hp + wt + transmission (fit9) and mpg ~ hp + wt (fit9a) shows that transmission does not significantly change the model results and can be removed. For the fit9a, the R^2 is now 0.827, likely lower than that of fit9 due to fewer covariates.

Model9 with mpg ~ hp + wt + transmission:

```
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        34.00288   2.642659  12.867 2.824e-13
## hp                 -0.03748   0.009605  -3.902 5.464e-04
## wt                 -2.87858   0.904971  -3.181 3.574e-03
## transmissionManual  2.08371   1.376420   1.514 1.413e-01
```

Model9a with mpg ~ hp + wt:

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285 2.565e-20
## hp          -0.03177    0.00903  -3.519 1.451e-03
## wt          -3.87783    0.63273  -6.129 1.120e-06
```

For comparison, the same methodology was followed, but this time the order of variable elimination was determined by examining the new p-values after a variable has been elminated. The p-value for transmission now indicates that the difference in average mpg between manual and automatic transmissions is now significant. The expected value of the difference between means is 2.94 with a 95% confidence of being between 0.05 and 5.82. R^2 is now 0.85, slightly higher than that of fit9a, but it includes an additional variable.

Model19 with mpg ~ wt + qsec + transmission:

```
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            9.618     6.9596   1.382 1.779e-01
## wt                    -3.917     0.7112  -5.507 6.953e-06
## qsec                   1.226     0.2887   4.247 2.162e-04
## transmissionManual     2.936     1.4109   2.081 4.672e-02
```

Examine the Residuals

In order to choose between model9a and model19, the residuals are examined. Plots of the residuals for both models are included in the appendix. As can be seen in Figure 3, both models have some tendency for the residuals to be higher for the lower or higher fitted values. Examination of the normal Q-Q plots (Figure 4) shows that the fit19 plot has more deviation from normality for the standardized residuals. In addition, the scale-location plot (Figure 5), shows a clear trend in the square root of the standardized residuals. The Cook's distance appears acceptable for both models. Based on this analysis, model9a is selected and it is concluded that there is no statistically significant dependent of mpg on transmission type when covariates are included in the linear model.

Appendix

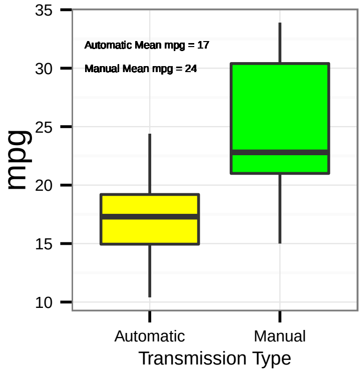

Figure 1: Boxplot of mpg by transmission type.

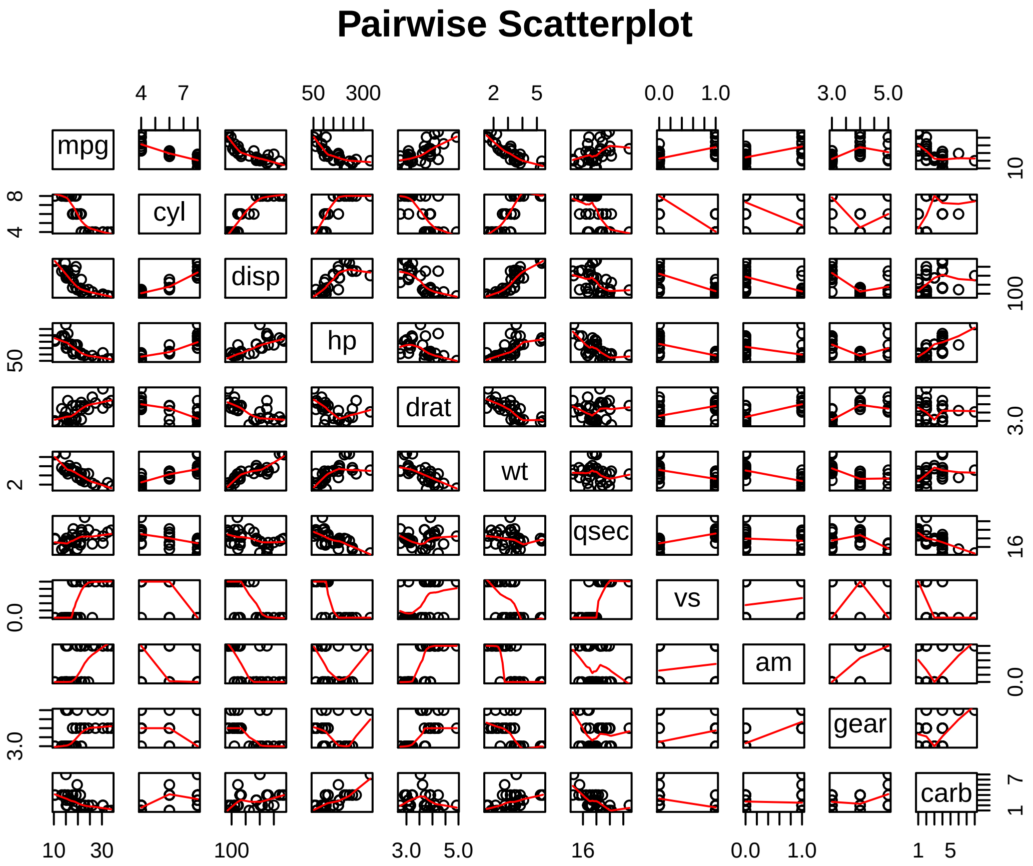

Figure 2: Pairwise scatterplot for all variables in mtcars.

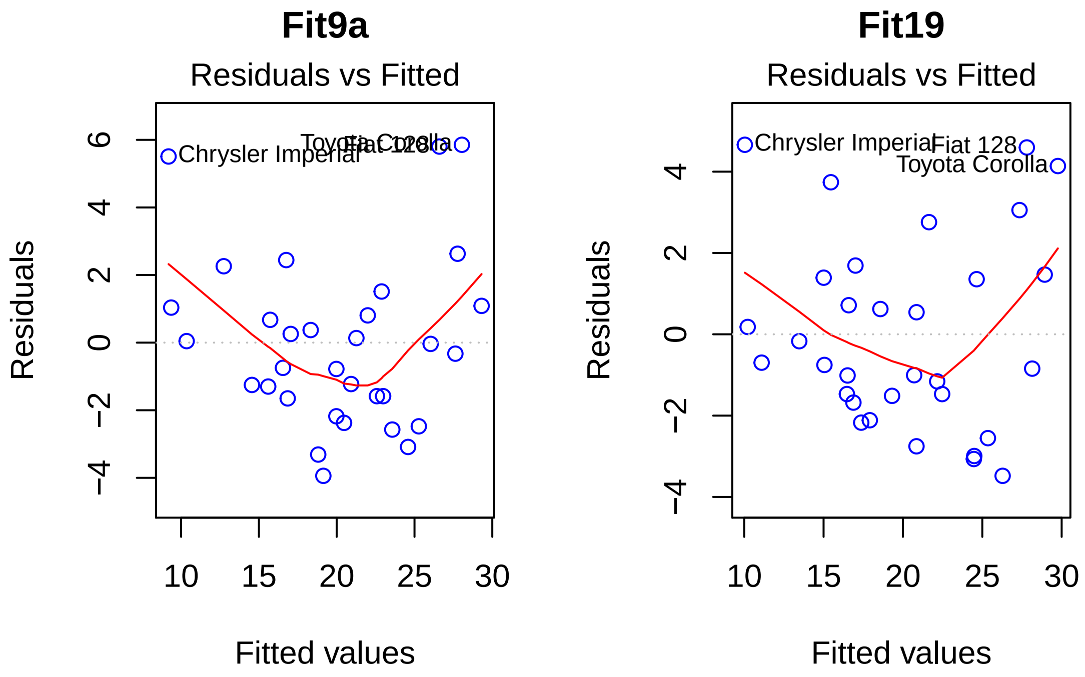

Figure 3: Residuals vs. Fitted for fit9a and fit19.

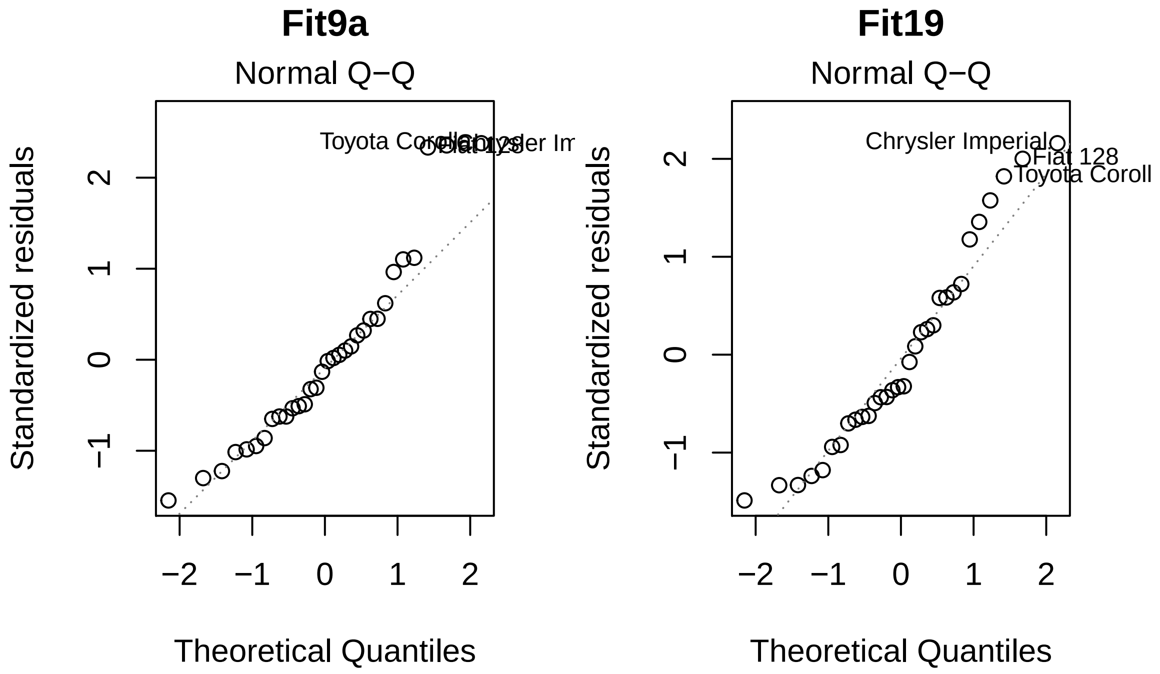

Figure 4: Normal Q-Q Plot for fit9a and fit19.

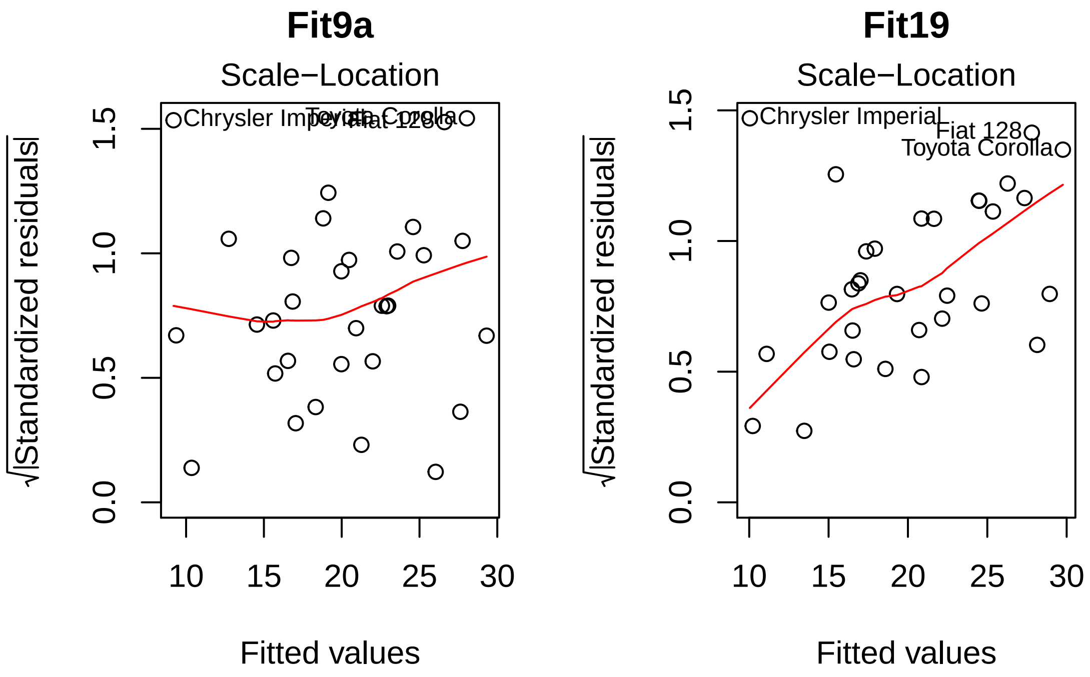

Figure 5: Scale-Location Plot for fit9a and fit19.

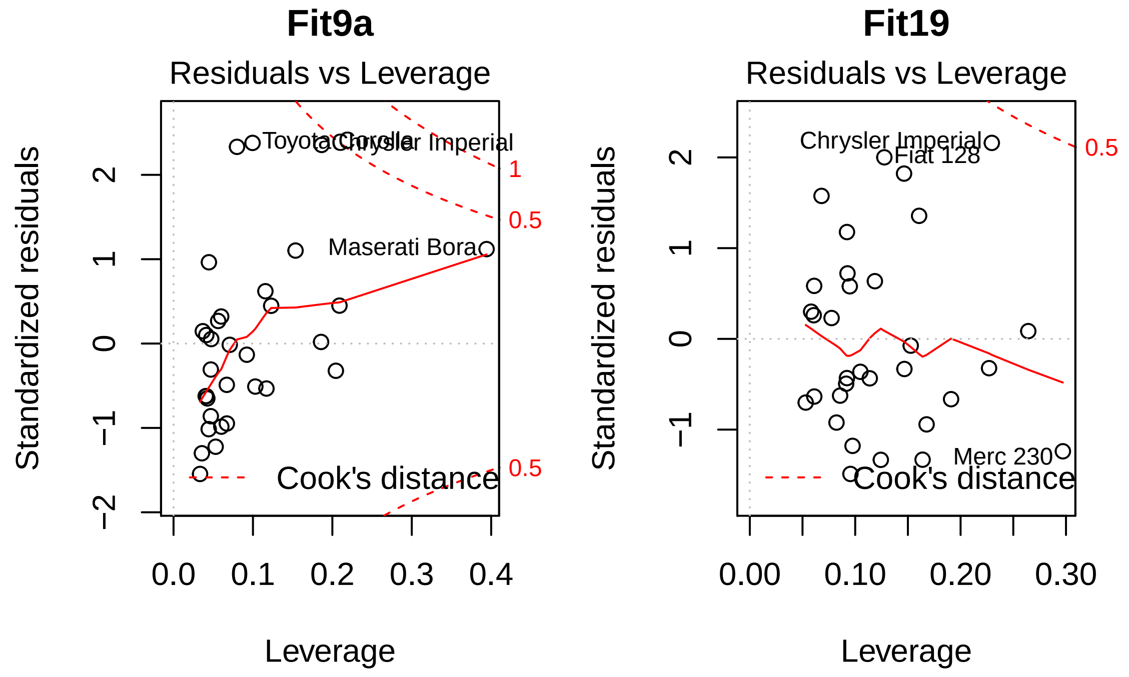

Figure 6: Residuals vs. Leverage Plot for fit9a and fit19.