Analysis of the Impact of Transmission Type on Fuel Efficiency

Executive Summary:

Motor Trend is interested in understanding the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

By optimizing the linear regression model exploring the relationship between MPG and transmission, I found:

- Yes, a manual transmission is better for MPG than an automatic transmission.
- The MPG of cars with manual transmission is 2.94 higher than that of cars with automatic transmission.

This document does not contain any of the R code used to perform the analysis. The .Rmd file can be found on `https://github.com/hoho1109/RegMod_Project`.

Data Description:

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The dataset contains a data frame with 32 observations on 11 variables: mpg (miles/(US) gallon), cyl (# of cylinders), disp (displacement in cu.in.), hp (gross horsepower), drat (rear axle ratio), wt (weight in lb/1000), qsec (1/4 mile time), vs (engine V/S), am (transmission), gear (# of forwarded gears), carb (# of carburetors).

The only data processing I perform is to transform the transmission (`am`) variable into a `factor` variable with two discrete levels: `automatic = 0` and `manual = 1`.

Data Analysis:

The flow of this analysis is as follows:

1. Visualize fuel efficiency (MPG) against transmission types (automatic vs. manual)
2. Select an appropriate regression for comparing the effect of transmission (automatic vs. manual) on fuel efficiency (MPG).
3. Perform statistical analysis based on the best model and quantify the MPG difference between automatic and manual transmission given there is a statistically signficant difference.

Part I: MPG vs. Transmission Types

I plot fuel efficiency (MPG) by transmission types (automatic vs. manual) (see ***Appendix 1***). Just by eye insepction, manual transmission could be better for fuel efficiency compared to automatic transmission. Now, I perform statistical analysis to determine whether this is true. If manual is indeed significantly better, then by how much?

Part II: Model Selection

To select the appropriate statistical model, I first examine the pairs plot for the mtcars dataset to observe how each variable relates to one another (see ***Appendix 2***). Observations from the plot: 1) There appears to be MPG dependence on transmission type; 2) MPG appears to depend on more than just the transmission type and 3) There is co-dependence between other variables. Together, these observations suggest that a model with just the transmission type as the predictor may underfit the data while a model that considers all variables may overfit the data and that an improved model can be constructed by optimizing the list of predictors.

Model 1: Transmission (`am`) *ONLY*

This model has an R^2 of 0.3385, meaning that only 33.85% of the data variation is explained by this model (see ***Appendix 3***). The low R^2 value suggests that MPG is confounded by other variables in the dataset and that the model under-fits the data as I have hypothesized earlier. Under-fitting a dataset introduces bias in an analysis. This is probably not very surprising given it is logical that the type of engine, the number of cylinder, weight of the car…etc. may also impact fuel efficiency. And this is also consistent with the pairs plot (see ***Appendix 2***).

So next I evaluate a model where all variables are included.

Model 2: *ALL* measured variables

This model has an R^2 of 0.8066, meaning that 80.66% of the data variation is explained by this model (see ***Appendix 4***). The second diagnostic test on this model that I am to run is to plot the residual and other variations of this fit (see ***Appendix 5***). The data appears to be normal and the residual plots do not contain obvious trends that indicate bias. The third and final diagnostic test is determining the inflation factors (***Appendix 6***). Many variables have high inflation factor, especially cyl, disp, and wt. This indicates that this model is probably over-fitting the data and therefore introudcing variance inflation as hypotheized.

Next, I try to improve the model by reducing the number of predictors.

Model 3: *OPTIMIZING* the list of predictors

To begin to choose which predictors to include, I inspect the p-values for the previous model fit. More signficant variables would have lower p-values. I construct 4 nested models with the 5 variables that have the least p-values (wt, am, qsec, hp and disp) (see ***Appendix 4***). And I compare all models constructed using the `anova()` function (see ***Appendix 7***). From this, I conclude that the best model has three predictors: `wt`, `am`, and `qsec`. Inclusion of any more variables do not provide statistica benefit in fitting the data.

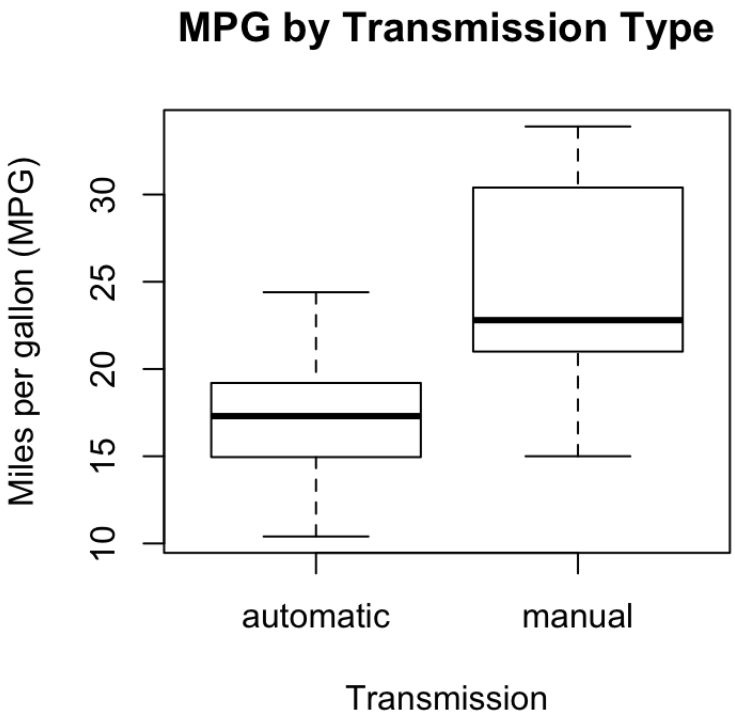
This model explains83.36% of the data variance (***Appendix 8***). The residual plots and variations do not contain obvious trends, indicating the model is not under-fitting the data so that there is no obvious bias (***Appendix 9***). The inflation factors for all predictors in this model are low, indicating that the model is not over-fitting the data (***Appendix 10***).

Part III: Statistical Analysis/Conclusions

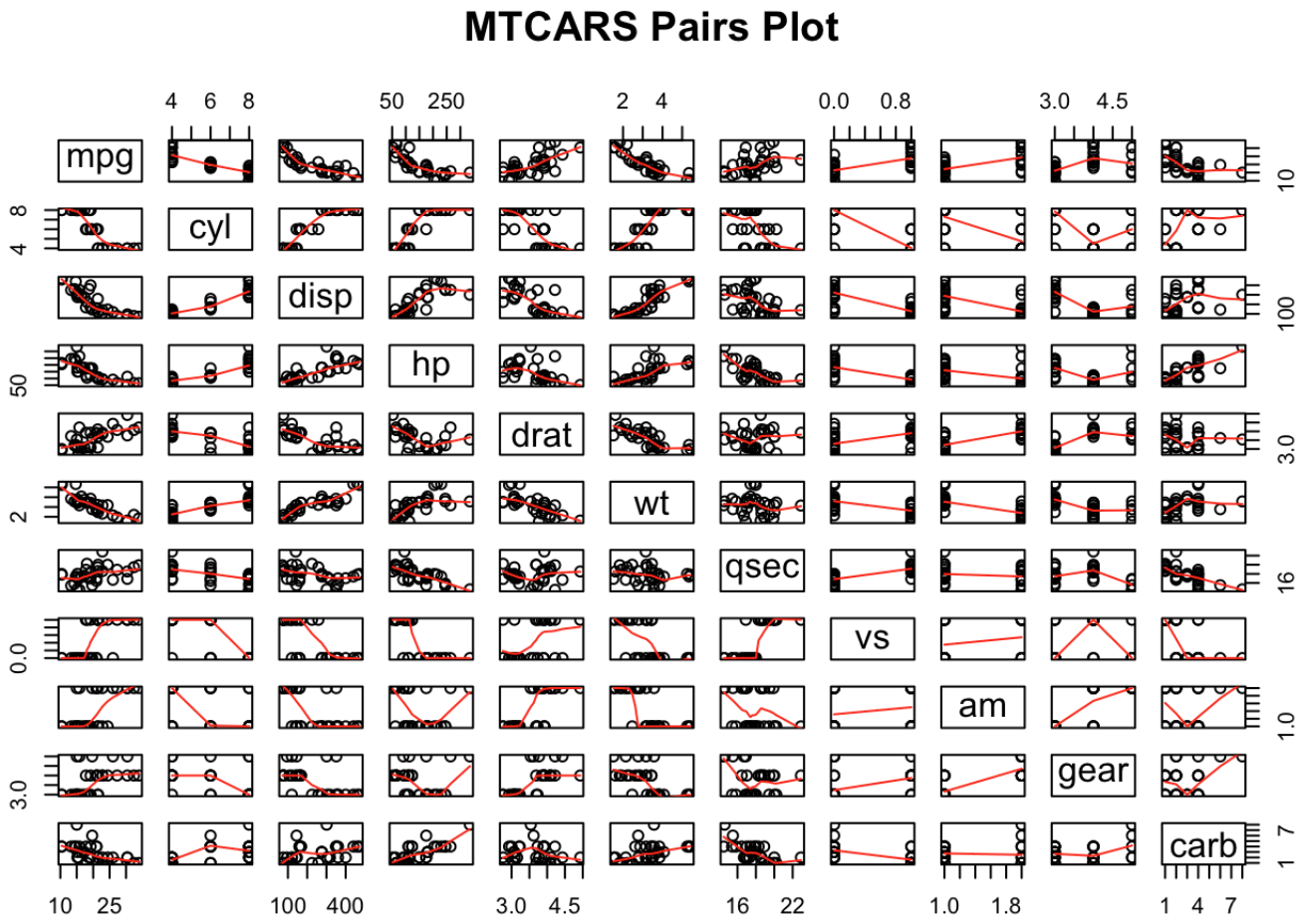
Based on the best model (refer to *Appendix 8* for the summary), MANUAL transmission is BETTER for fuel efficiency compared to AUTOMATIC transmission by 2.94 MPG.

Appendices:

Appendix 1: MPG by Transmission Type



Appendix 2: Pairs Plot



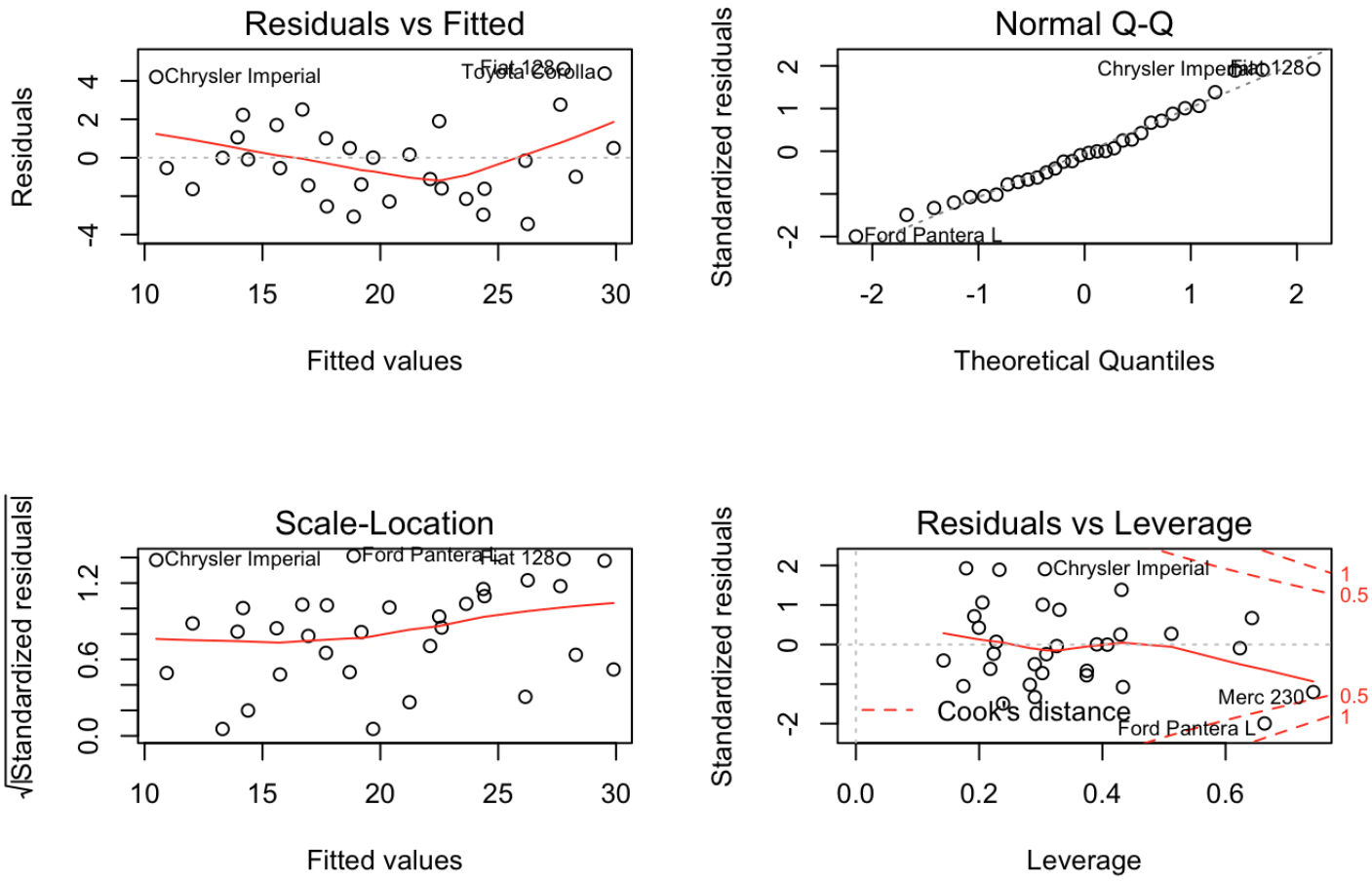
Appendix 3: Model Summary for Model 1 - Trasmission Only

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Appendix 4: Model Summary for Model 2 - All Variables

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Appendix 5: Diagnosing Fit Bias for Model 2



Appendix 6: Diagnosing Variance Inflation for Model 2

```
##       vif.fit2.
## cyl   15.373833
## disp  21.620241
## hp     9.832037
## drat   3.374620
## wt    15.164887
## qsec   7.527958
## vs     4.965873
## am     4.648487
## gear   5.357452
## carb   7.908747
```

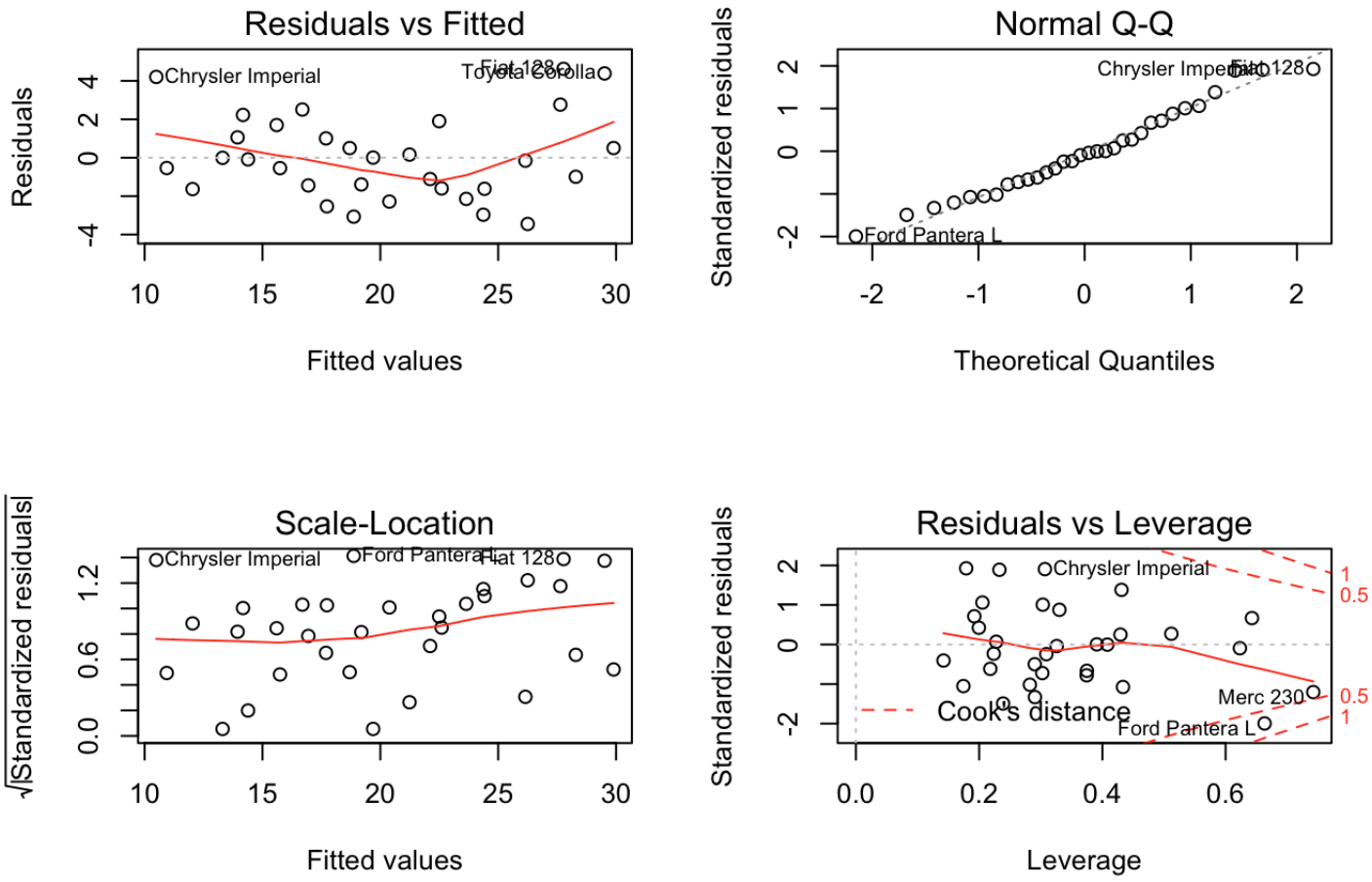
Appendix 7: Optimizing Predictors

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt + am + qsec
## Model 4: mpg ~ wt + am + qsec + hp
## Model 5: mpg ~ wt + am + qsec + hp + disp
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 169.29  1    109.03 15.5240 0.0007497 ***
## 4     27 160.07  1      9.22  1.3127 0.2648040
## 5     26 153.44  1      6.63  0.9438 0.3423662
## 6     21 147.49  5      5.94  0.1692 0.9711466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 8: Summary of Best Model

```
##
## Call:
## lm(formula = mpg ~ wt + am + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Appendix 9: Diagnosing Fit Bias for the Best Model



Appendix 10: Diagnosing Variance Inflation for the Best Model

```
##      vif.fit4.
## wt    2.482952
## am    2.541437
## qsec  1.364339
```