Modeling MPG with Regression

Jesus Dacuma

Executive Summary

This report used the `mtcars` dataset from the 1974 Motor Trend US magazine to answer:

- Is an automatic or manual transmission better for MPG?
- What is the quantifiable difference in MPG between automatic and manual transmissions?

Effects of transmission were determined using linear regression models dynamically selected by analyzing residuals and variance inflation factors. Results show that a manual transmission causes an 2.9358 MPG increase in fuel efficiency when all other factors are equal. While transmission type can significantly affect MPG, the report concludes that other aspects of automobile design such as weight and quarter-mile time can more significantly explain changes in MPG.

Exploratory Analysis

Figure 1 is a scatterplot of the `mtcars` dataset showing how two types of transmission may affect MPG. A fitted regression line shows that there is a lot of variance in MPG, even within each transmission style. Explaining this variance requires exploration of other variables in the dataset.

Model Selection

The model selection process used in this report begins with a linear model including all variables. The model is examined using `vif` and the non-significant regressor with the highest variance-inflation is eliminated. This process is repeated, eliminating another regressor with each iteration until the final resulting model only contains variables significant to the 95% confidence level. The goal of this strategy is to minimize variance by only including uncorrelated, orthogonal variables and removing variables without significant effects on MPG.

Using variance-inflation analysis, one can easily eliminate four variables (`carb` , `cyl` , `gear` , `disp`) that contribute to obscenely high variance. The result is the first model that this report will explore.

```
fit1 <- lm(mpg ~ . - carb - cyl - gear - disp, data = mtcars) # Model 1
vif(fit1)
```

```
##       hp     drat       wt     qsec       vs       am
## 5.070665 2.709905 5.105979 5.776361 4.120656 3.272177
```

Continuing to eliminate factors through the model selection process, one will arrive at the final model, which only includes three significant regressors (`wt` , `qsec` and `am`) and has low variance-inflation.

```
fit2 <- lm(mpg ~ am + wt + qsec, data = mtcars) # Model 2
vif(fit2)
```

```
##       am       wt     qsec
## 2.541437 2.482952 1.364339
```

Figure 2 shows residual plots of the two models. Neither plot exhibits any obvious pattern that would suggest other confounders have been omitted. However, one can determine that the second model is a better model - not only because it introduces less variation, but its predicted residual sum of squares (PRESS) statistic is lower when cross-validating the points in the dataset.

```
##               Model_1  Model_2
## PRESS Values 254.7793 231.3035
```

Lastly, the Q-Q plot of Model 2 in Figure 3 shows that the residuals are somewhat linear with theoretical quantiles, confirming the original assumptions of normality.

Understanding the Model

Below are the coefficients and R^2 value of the chosen model (Model #2). For reference, the R^2 value for a regression model including every variable is 0.89, only slightly higher than the R^2 value of the chosen model. The regression model coefficient for `am` estimates an increase of 2.9358 MPG for a manual transmission over an automatic transmission, when all other factors (weight and quarter-mile time) are held constant. With 95% confidence, that increase in gas mileage will range between 0.05 to 5.83 MPG. Thus, a manual transmission is better for MPG.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

```
## R-Squared:  0.8496636
```

```
## Manual Transmission Confidence Interval:  0.04573031 5.825944
```

While transmission type is a significant regressor of MPG, the p-values show that weight and quarter-mile time are even stronger predictors of MPG (significant to the 99.9% confidence level as opposed to 95%). Using general intuition and subject knowledge, the inclusion of weight and quarter-mile time make sense; a heavier automobile will require more gasoline to run the same distance, and higher-performance cars with lower quarter-mile times were designed for speed and not fuel efficiency.

Conclusion

Using a regression model that only includes significant variables and minimizes variance-inflation, one can predict MPG based on automobile transmission type, weight and quarter-mile time (the same variables chosen by R.R. Hocking in his 1976 analysis of the Motor Trend dataset). The model suggests that a manual transmission is better for gas mileage, increasing MPG by an expected 2.9358 miles per gallon over an automatic transmission (with weight and quarter-mile time held constant).

Appendix

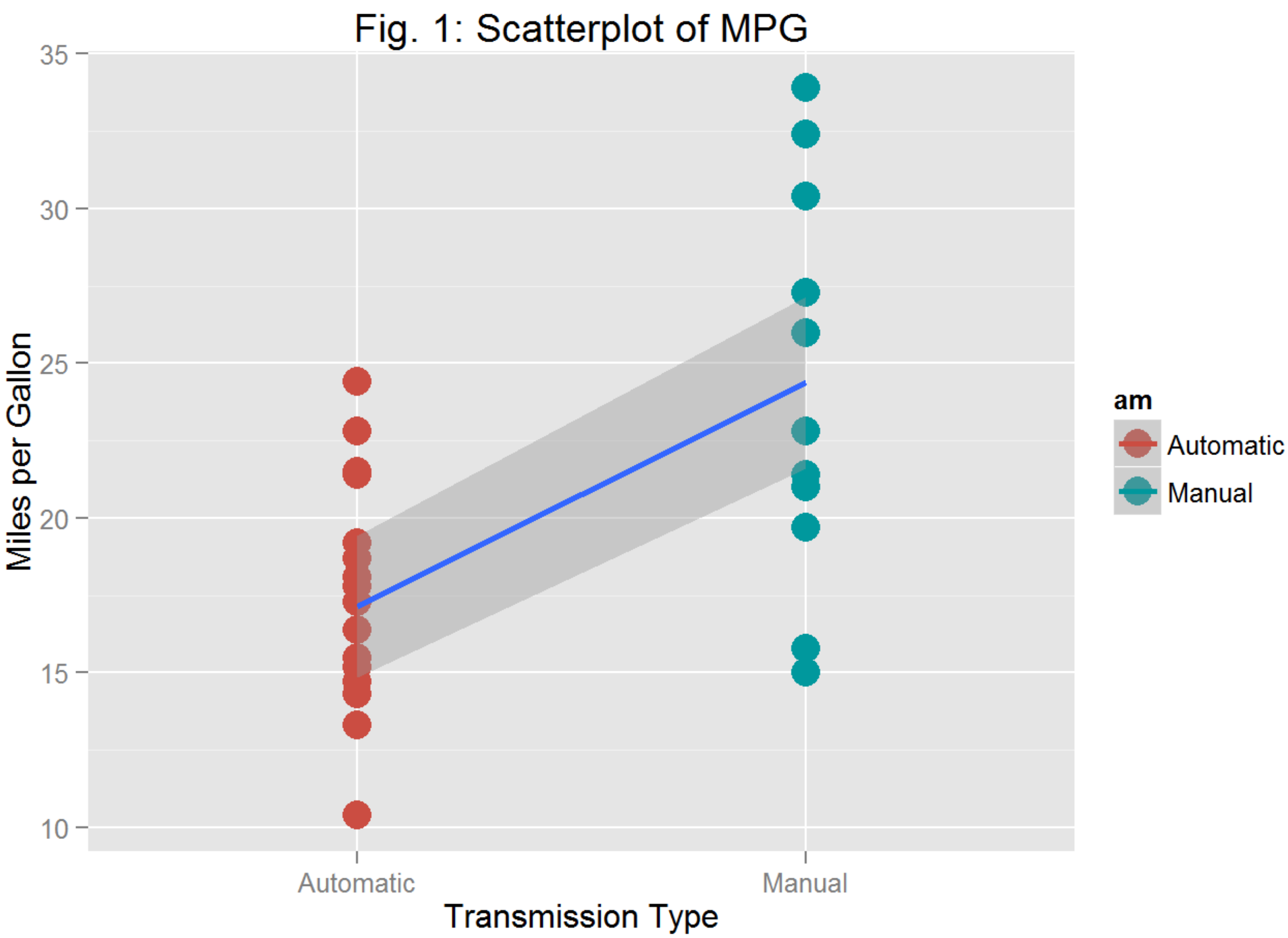
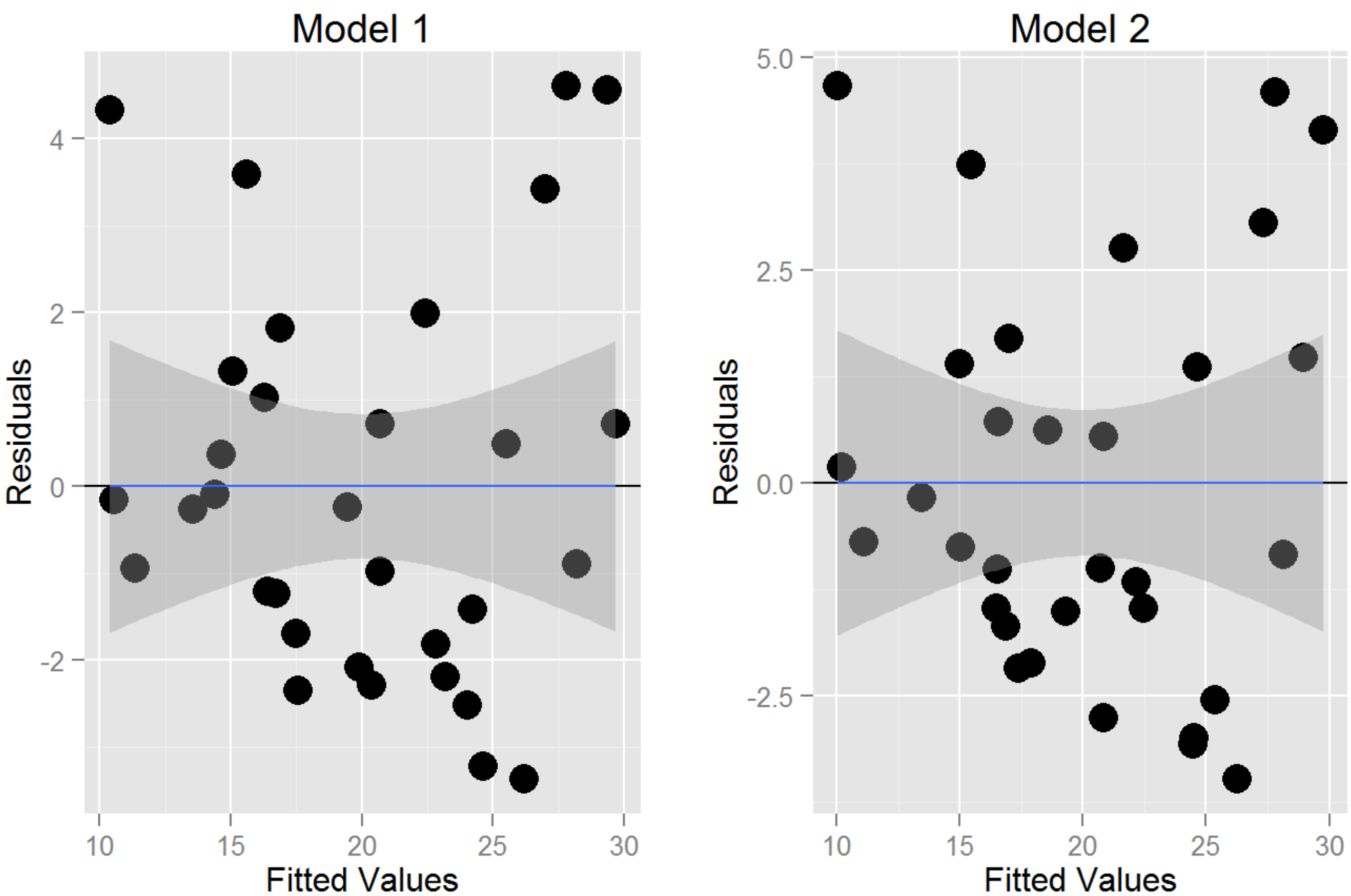


Fig. 1: Scatterplot of MPG
Miles per Gallon
35
30
25
20
15
10
Automatic
Manual
Transmission Type
am
Automatic
Manual

Fig. 2: Residuals vs. Fitted



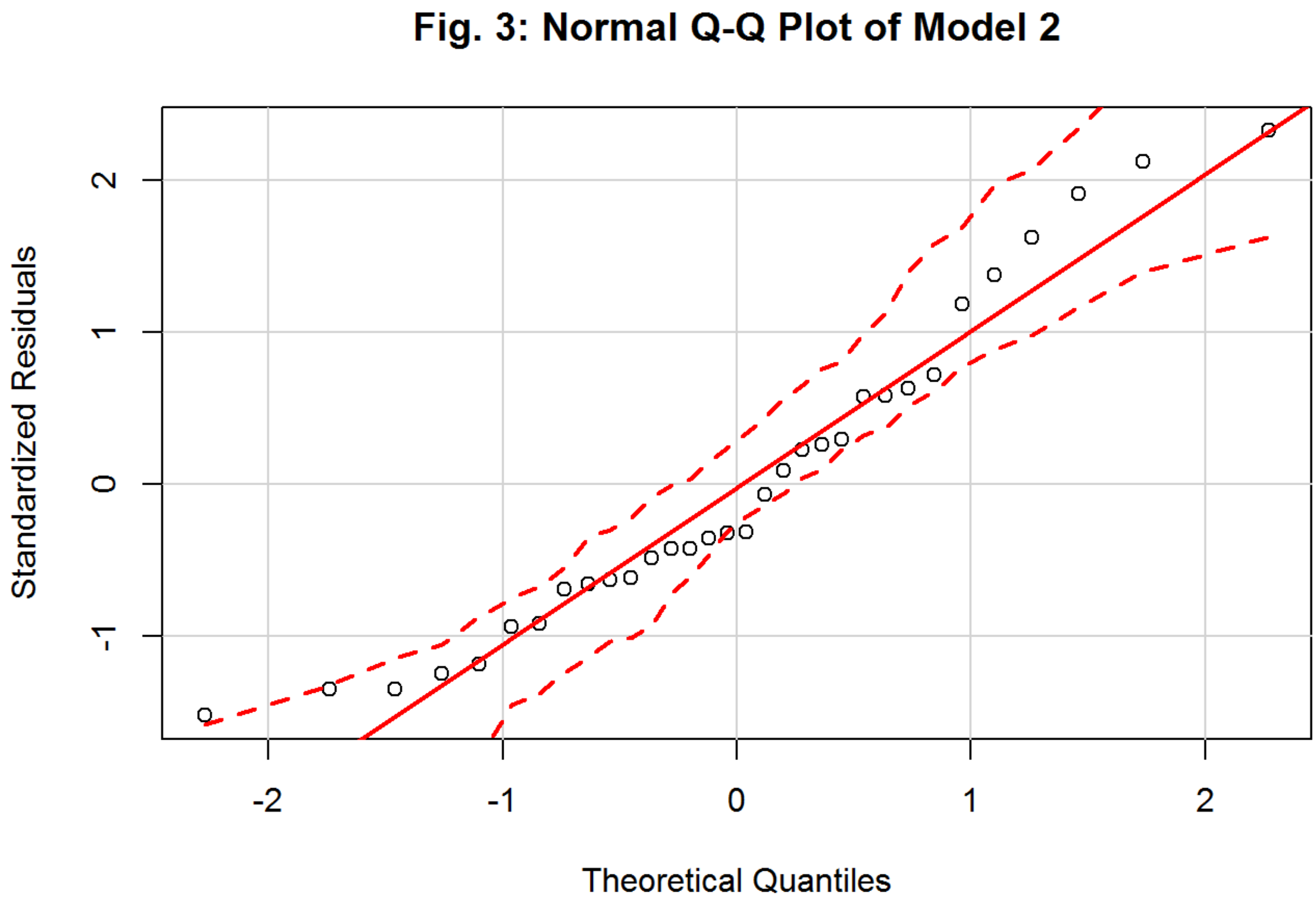


Fig. 3: Normal Q-Q Plot of Model 2
Standardized Residuals
Theoretical Quantiles
-2
-1
0
1
2
-1
0
1
2