# Motor Trend Regression Analysis

Georgeanne Purvinis

5/28/2021

## Executive Summary

This analysis uses a *Motor Trend*'s dataset called mtcars to answer two questions about vehicles' miles per gallon (MPG) and the type of transmission, specifically:

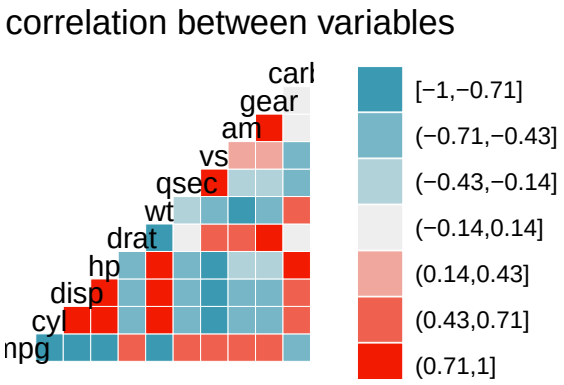
- *"Is an automatic or manual transmission better for MPG?"* The model found that manual transmission is better for mpg, for a given horse-power vehicle.
- *"Quantify the MPG difference between automatic and manual transmissions."* The model found that mpg will increase by 5.2771 when going from automatic (am = 0) to manual transmission (am = 1), for a given (constant) horse-power (hp). Using transmission type alone was insufficient to predict the mpg.

The analysis used multiple regression analysis to find a relationship between MPG and transmission (automatic vs. manual) while taking into account possible confounding variables such as weight (wt), number of cylinders (cyl), horse-power (hp), and more. The validity of each model tested is checked by looking at plots of residuals and checking goodness of fit with p-values. The Principal of Parsimony, i.e. the model should contain the smallest number of variables necessary to fit the data, is also respected. Variables are added in a nested fashion to perform a stepwise regression.

## Data Analysis

### Exploratory Analyses

The mtcars dataset contains 32 observations of 11 variables. Exploring the data involved plotting each variable (predictor) and the MPG outcome, and also calculating the correlation between each predictor and outcome. This has the effect of showing which predictors are most correlated to the outcome. The R package GGally lets us look at this data in a matrix of plots, while also showing the correlation between each variable (i.e. wt and cyl are highly correlated). The plot is shown in the appendix. Another useful graphic just shows the correlations in a matrix, with dark colors showing strong correlations. From this plot (below) we see that cyl, disp, hp have the strongest correlation with mpg but are also correlated to each other. The type of transmission (am) appears to be a weaker correlation to mpg.



### Strategy and Model Selection

The strategy used for model selection consists of adding variables and checking if the estimated coefficients are largely unchanged after addition, and if the p-values are small. This is a stepwise process. Initially, we model mpg ~ am, then add the regressors cyl, disp, hp. The results for mpg ~ am and mpg ~ am + hp are shown below. Other models had higher p-values, and using all the variables as a model had all high p-values and did not make a good model.

```
##             Estimate   Pr(>|t|)
## (Intercept)  17.1470 1.1340e-15
## am            7.2449 2.8502e-04
```

```
##              Estimate   Pr(>|t|)
## (Intercept) 26.585000 1.0740e-17
## am           5.277100 3.4603e-05
## hp          -0.058888 2.9204e-08
```

The best model is fit5 <- lm(mpg ~ am + hp, data = mtcars). The model becomes: mpg = 26.585 + 5.2771am + -0.058888hp.The model means that mpg will increase by 5.2771 when going from automatic to manual transmission, for a given (constant) hp. It also means that mpg will decrease by 0.058888 for each increment of hp, if the transmission type is held constant.

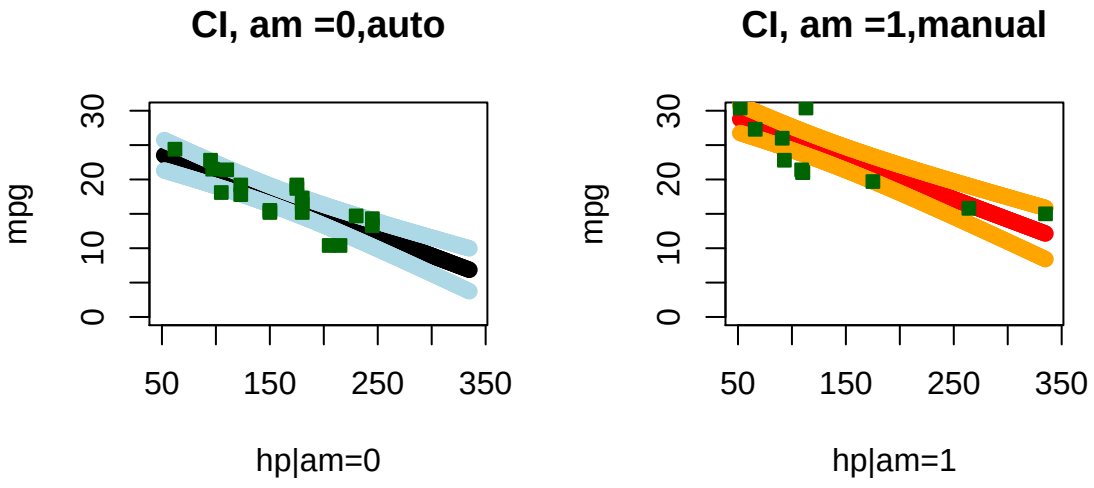
### Diagnostics and Checking Assumptions

Plots of the residuals versus fitted values and residuals versus independent variables are a standard method of checking if model assumptions are violated. The assumptions are 1) the errors are independent, 2) the errors have mean 0, 3) the errors all have about the same variance, and 4) the errors are normally distributed. The plots do not have any serious violations of the model assumptions.The plots may be viewed in the appendix.

### Uncertainty in Results

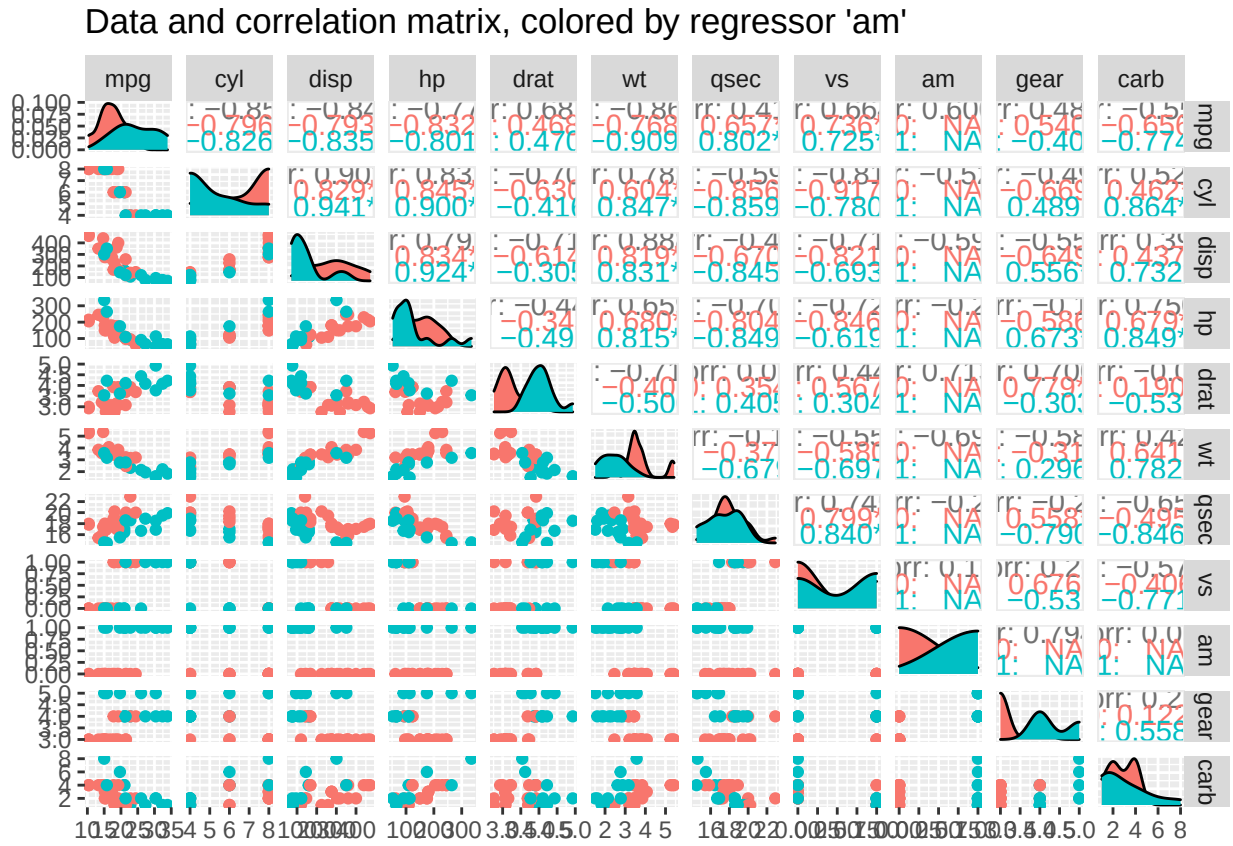
The parameter R^2 explains the percentage of variation in the regression model, which is 78% for this model. It is calculated from summary(fit5)$r.squared.

Lastly, the model is fit to the data and the confidence intervals are shown. Since there are two regressors, am and hp,and am takes on only two values, the model is used twice: for the case when am =0 (auto transmission and when am=1 (manual)). The model and data fit well with very few points outside the confidence interval. The plot for am=1 (manual) is shifted higher than for am =0 auto transmission, indicating higher mpg.



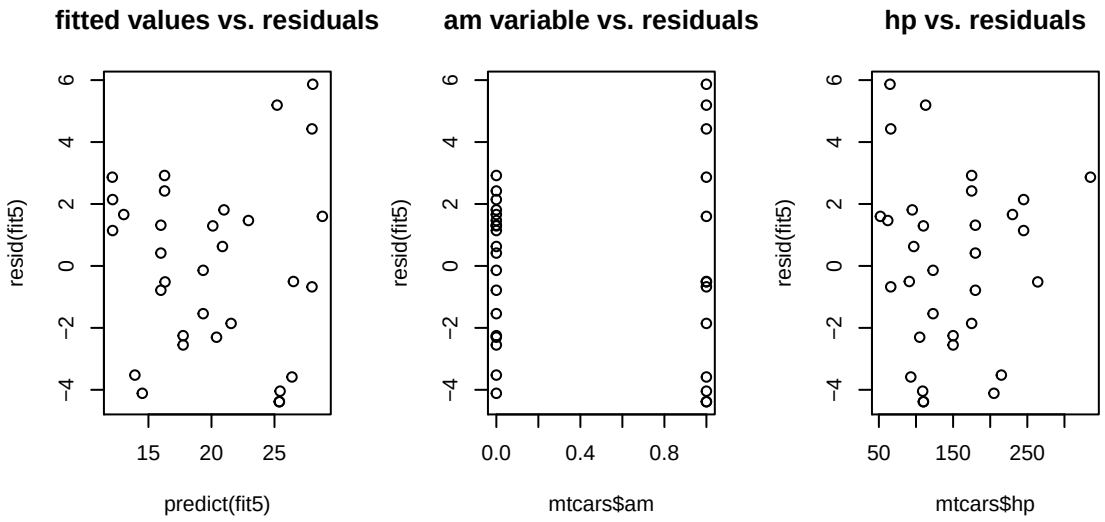
## Appendix

### Total dataset correlation matrix and data scatter plots




### Model Selection Iterations

### Diagnostics and Checking Assumptions



```r
# Errors are independent indicated by p-values near 0
mean(resid(fit5)) # 2) residuals have mean 0
```

```
## [1] -1.700029e-16
```

```r
# 3,4) Not enough data to test the variances at each mpg point
```