Regression Model Course Project

A. Johnson

26 August 2019

Executive Summary

As a (fictional) employee of Motor Trend, a magazine about the automobile industry, I was interested in exploring the relationships between automobile design characteristics and miles per gallon (MPG). My analysis described below details that choosing a car with an automatic versus a manual transmission does not significantly impact the miles per gallon after adjusting for number of cylinders, displacement, gross horsepower, rear axel ratio, and weight of the car. However, uncertainty remains in our conclusion due to unknown correlates of MPG not available to us in the dataset and therefore, not included in the final adjusted model.

Exploratory Data Analysis

The goal of the below code was to examine the variables included in the “mtcars” dataset, recode factor variables as desired, and examine the distribution of the MPG variable by transmission type and engine type.

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Evaluating the correlation table, the following variables have at least moderate correlation to mpg: cyl, disp, hp, drat, wt, vs, and am (correlation > 0.6 or < -0.6). The distribution of MPG (overall, by transmission type, and by engine type) was also evaluated during exploratory data analysis. See the appendix for figures. Looking at the box plots, the ranges of MPG overlap by transmission type and engine type when evaluating the distributions of whiskers. There is also a possible outlier in the v-shaped engine type.

Bivariate and Multivariable Models

```
##              Estimate Std. Error   t value                 Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 0.00000000000000001133983
## factor(am)1  7.244939   1.764422  4.106127 0.00028502074393506776 9
```

In the bivariate model, transmission type is significantly associated with MPG (p=0.000285) with a significance level set at p=0.0001. On average, cars with manual transmission have a significantly higher amount of miles per gallon (24.39) than cars with automatic transmission (17.15 MPG). Typically, it's good practice to avoid selecting the bivariate model as the final model because it fails to take into account the effects of other variables on the outcome. Omitting variables can result in bias in the coefficients of interest if the regressors are correlated with the omitted variables. In fact, when all other covariates were included in the model, the relationship between transmission type and MPG was no longer significant (p=0.23).

Now, let's check to see if I included any redundant variables in the model and explore whether reducing the number of variables improves model fit. Including variables that are unrelated to the model can increase standard errors. Based on the correlation matrix reported above, I see that cyl, disp, hp, drat, and wt have the strongest correlations with mpg. Let's include these in our model with transmission type and see if model fit improves:

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 36.04938339 7.60552882  4.7398918 0.00007311525
## cyl         -1.03334930 0.72404890 -1.4271816 0.16589933102
## disp         0.01256722 0.01195130  1.0515352 0.30307136202
## hp          -0.02887012 0.01444162 -1.9990916 0.05658027715
## drat         0.48585981 1.49494905  0.3250009 0.74788453983
## wt          -3.27472256 1.15684830 -2.8307277 0.00903308854
## factor(am)1  1.37506396 1.56866152  0.8765842 0.38905816255
```

The adjusted R^2 statistic that describes model fit in the adjusted, final model (adjusted R^2=0.82) improves from the model fit in the bivariate model (adjusted R^2=0.34) and the full model (adjusted R^2=0.81). This means that in the final model, 82% of the total variance is described by the model. Transmission type is not significantly associated with MPG (p=0.39) and weight of the car is significantly associated with MPG (p=0.01).

Examining Model Fit

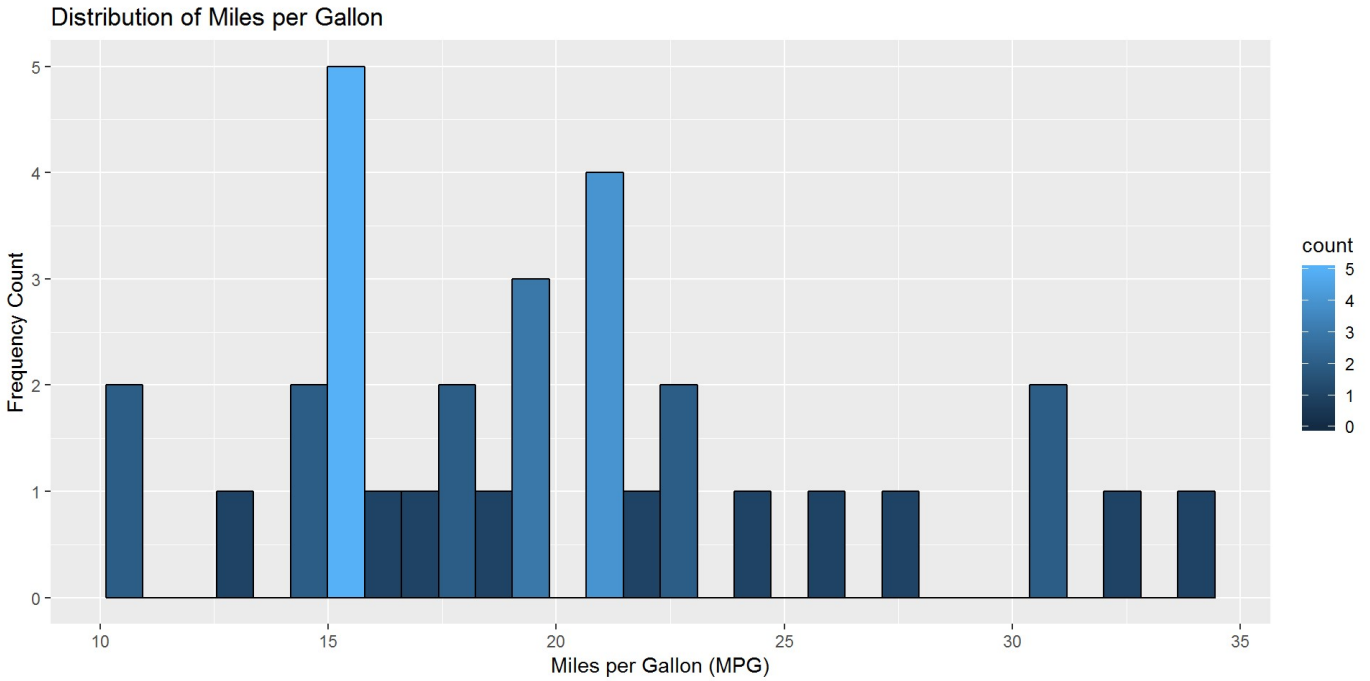
Lastly, I examined residuals, leverage, and influence measures in the final model. For the sake of brevity, the figures are included in the appendix. The plots below are diagnostic tools to evaluate systemic patterns in the data. Viewing the residual vs. fitted plot, we can see that our residuals are homoscedastic (i.e., have equal variance). Viewing the QQ plot, we can see that the errors are approximately normal and the residuals follow the plotted line. The other figures examine the influence of car types on our model within individual coefficients and as a collective.

Appendix

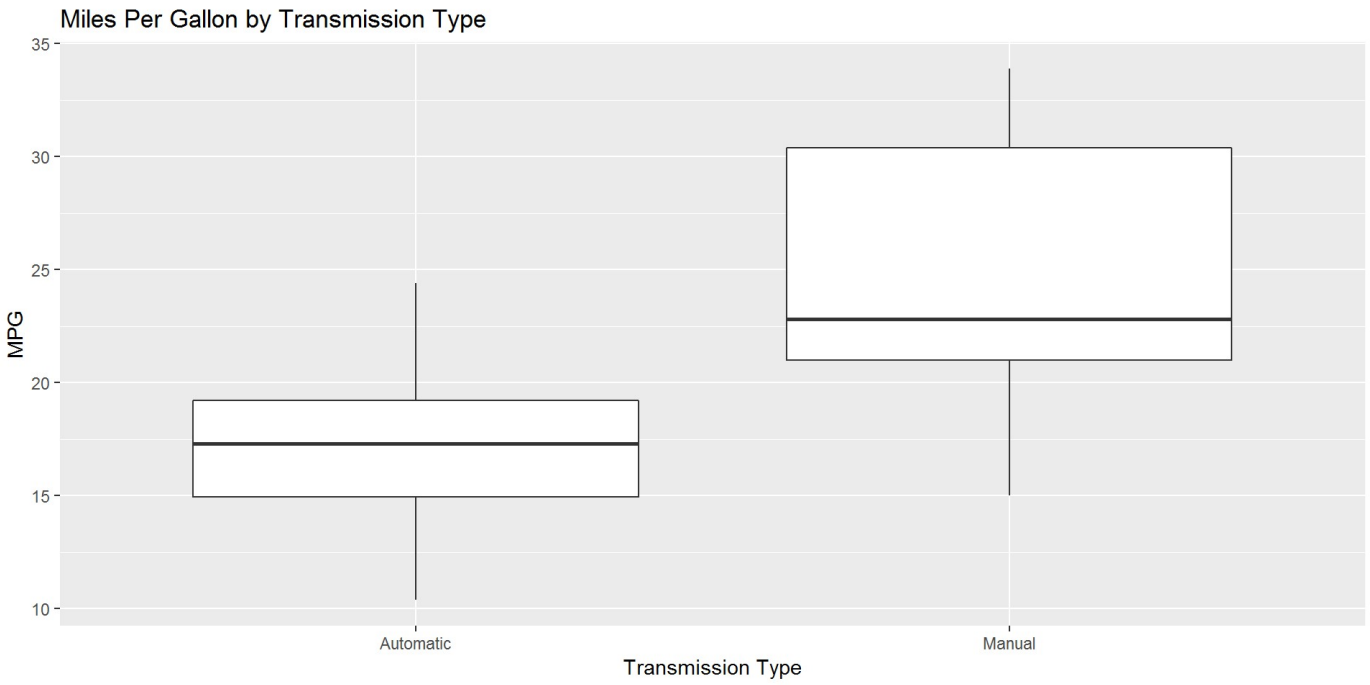
For figures, let's recode the "vs"" and "am"" as factor variables with labels.

```
ggplot(mtcars, aes(x=mpg)) +
        geom_histogram(aes(y=..count.., fill=..count..),
                       color = "black") +
        labs(title="Distribution of Miles per Gallon",
             y="Frequency Count", x="Miles per Gallon (MPG)")
```

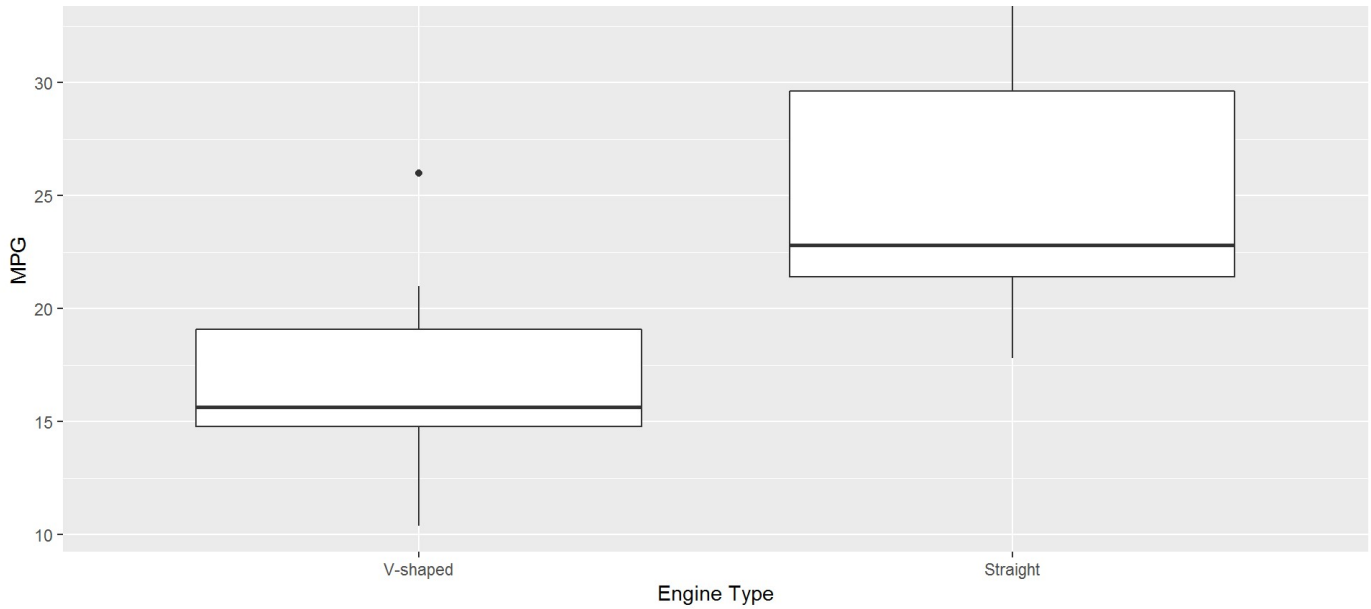
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(mtcars, aes(x=am, y=mpg)) +
        geom_boxplot() +
        labs(title="Miles Per Gallon by Transmission Type",
             x="Transmission Type", y = "MPG")
```




```
ggplot(mtcars, aes(x=vs, y=mpg)) +
        geom_boxplot() +
        labs(title="Miles Per Gallon by Engine Type",
             x="Engine Type", y = "MPG")
```




```
#Plot the fit figures:
par(mfrow = c(2, 2))
plot(finalmodel)
```