Regression Modeling Project

Brian Linn

January 10, 2017

Executive Summary

This report seeks to perform exploratory analysis against the mtcars dataset to determine if there is any relationship between the miles per gallon a vehicle can drive, versus the type of transmission installed on that vehicle, automatic versus manual. The null hypothesis concludes that there transmission type lends no effect to a vehicle's fuel efficiency as measured by miles per gallon; the alternative hypothesis concludes that the mpg rating for a car is influenced by the transmission type.

Exploratory Data Analysis

```
aggregate(mpg ~ am, data = data, FUN = mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

The initial look into the data indicates that the mean mpg for manual vehicles is 7.245 higher than the mpg for automatic vehicles. Next, a linear regression using only the transmission to predict the mpg is performed.

Single Variable Linear Regression

```
fit <- lm(mpg ~ factor(am), data = data)
summary(fit)$coef
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      17.147368   1.124603 15.247492 1.133983e-15
## factor(am)Manual  7.244939   1.764422  4.106127 2.850207e-04
```

```
confint(fit)
```

```
##                     2.5 %   97.5 %
## (Intercept)      14.85062 19.44411
## factor(am)Manual  3.64151 10.84837
```

In the basic linear model wherein mpg is the outcome as predicted by the factor, transmission type, the automatic transmission is treated as the reference. The intercept for automatic transmissions, beta0 = 17.147, indicating that the average mpg for automatic transmissions is 17.147. The beta1 coeffecient for manual transmissions is 7.245 where am = 1, informs us that the average mpg for manual vehicles is 17.147 + 7.245 = 24.392.

The model's p value indicates enough signifigance to initially reject the null hypothesis and preliminarily conclude that manual transmissions perform better, with respect to miles per gallon. Also, note that the confidence interval does not contain 0, further lending credence to rejecting the null hypothesis. While the r-squared value of .3385 indicates that transmission type is an important part of mpg, it leaves the remaining 66% of variance unexplained.

Multivariate Linear Modeling

```
summary(fitBest)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

We use the step function to determine which additional variables should be included in the multivariate analysis of the data. The results indicate that using cyl, hp, and weight as additive confounders with transmission as the intercept will explain 84% of the variance in mpg.

Multivariate Residual Analysis

In order to ensure that the multivariate analysis should be used, the independence of the residuals must be confirmed. The supporting plot is contained in the appendix.

The plots show us that the residuals and fitted line show no apparent relationship or pattern, and while there are some outliers, the normality plot indicates that the residuals are approximately normally distributed in addition to being independent.

Compare Single to Multivariate Analysis

```
anova(fit, fitBest)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

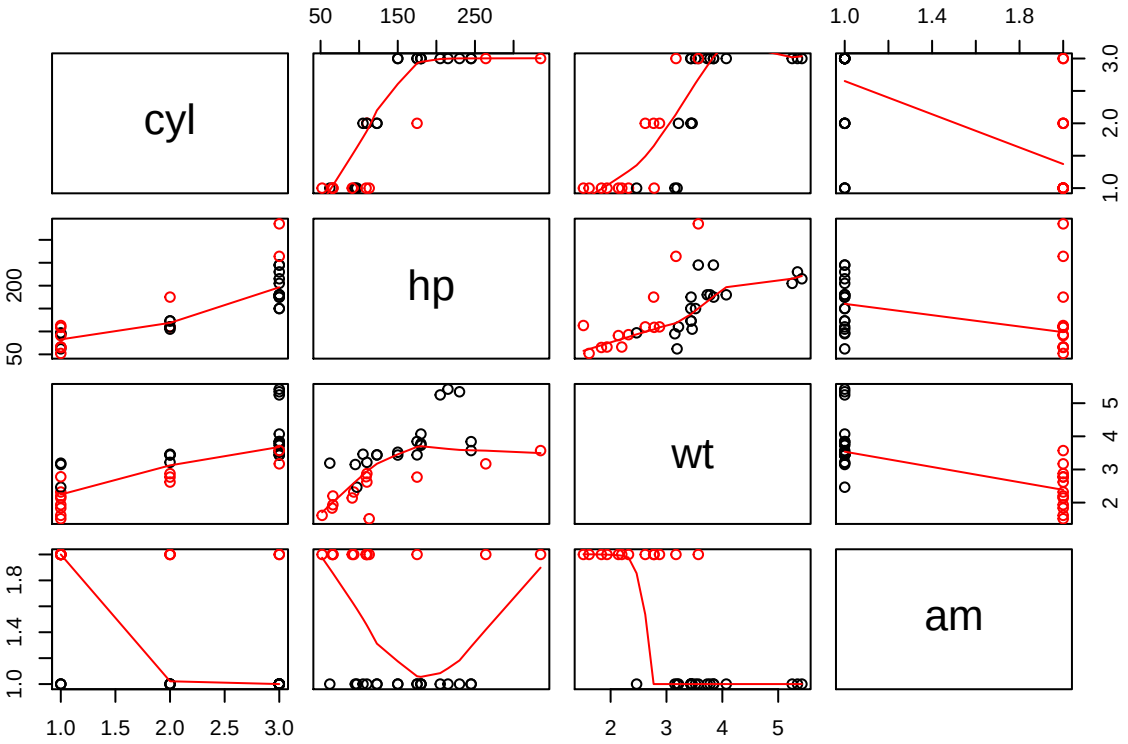
The low p-value indicates that the multivariate analysis explains more of the variance of mpg, the conclusion is that the best model should include transmision type, but also include some confounders to explain more of the variance.

Conclusion

The report concludes that the null hypothesis can be rejected as the transmission type has been shown to be a statistically significant influence on the miles per gallon outcome. However, the alternative hypothesis must specify that the mpg is not affected only by the transmission type, but also by other statistically significant variables, such as the weight of the car, the number of cylinders in the car's engine, and the horsepower produced by the car.

Appendix

Pairs



Residual Analysis