Regressions Model Final Project

## Introduction

Thanks for tuning in for this month's edition of Motor Trends magazine! In this issue, we'll be focusing in on the relationship between the type of transmision a vehicle has and the associated miles per gallon. Overall, we will want to answer two questions:

“Is an automatic or manual transmission better for MPG?” "Quantify the MPG difference between automatic and manual transmissions?"

## Exploratory Analysis

First, let's take a look at the data to understand what we're dealing with.

head(mtcars)  
summary(mtcars)

It looks like the transmission variable (am) is not being treated as a factor; let's change that! Then lets look at the correlation between the transmission and mpg.

mtcars$am <- as.factor(mtcars$am)

summary(lm(mpg ~ ., data=mtcars))

Looking at all of the variables it looks like there is no significance with any of the specific variables, so let's look at the mpg and transmission specficially.

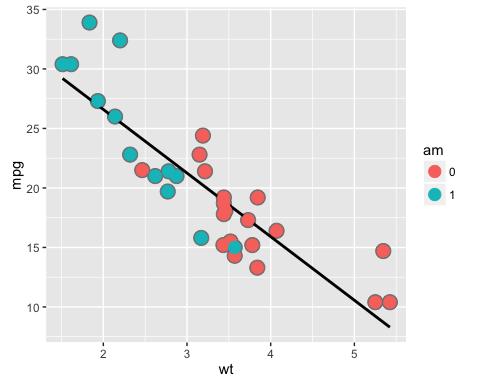
summary(lm(mtcars$mpg ~ mtcars$am))

While the correlation is statistically significant given the p-value of 0.000285, let's plot the mtcars data set. Given that many people's first inclination is that heavier cars tend to weigh more, I will plot miles per gallon vs weight, while coloring the data with the transmission factor.

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.3

g=ggplot(mtcars, aes(y=mpg, x=wt, colour=am))  
g= g + geom\_point(colour = "grey50", size=5) + geom\_smooth(method = lm, se = FALSE, colour= "black") + geom\_point(size=4)  
g



From this plot, it looks like manual transmissions do have higher miles per gallon, but it also looks like the manual vehicles tend to weigh less as well. For this reason, we must look at other variables as well for this analysis.

In the appendix, you can see the various linear models we put together; however, the linear model below shows the variables with the most statistical significance for multiple variables.

final <- lm(mtcars$mpg ~ mtcars$am+mtcars$wt+mtcars$hp)  
summary(final)

##   
## Call:  
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$hp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4221 -1.7924 -0.3788 1.2249 5.5317   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.002875 2.642659 12.867 2.82e-13 \*\*\*  
## mtcars$am1 2.083710 1.376420 1.514 0.141268   
## mtcars$wt -2.878575 0.904971 -3.181 0.003574 \*\*   
## mtcars$hp -0.037479 0.009605 -3.902 0.000546 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.538 on 28 degrees of freedom  
## Multiple R-squared: 0.8399, Adjusted R-squared: 0.8227   
## F-statistic: 48.96 on 3 and 28 DF, p-value: 2.908e-11

## Residual Analysis

The residuals of this analysis support this analysis (which can be viewed in the appendix) due to several reasons:

1. In the Residuals vs Fitted, there is no consistency in the pattern
2. Given that the Normal Q-Q points are close to the line, we can assume that the model is normally distributed
3. We get 0 for the dfbeta test below

sum((abs(dfbetas(final)))>1)

## [1] 0

## Conclusion

Based on the analysis, we cannot say that an automatic or manual transmission has a significant effect on the miles per gallon. Rather, we should be looking at other variables such as weight and horsepower to answer this question.

However, based on this dataset alone, a move from automatic to manual car ownership would likely result in an increase of 7.245 miles per gallon.

## Appendix: Exploratory Data Analysis and Residuals

#### Model Fitting

Knowing that weight was a significant varable, we looked at linear models for other variables as well, to see if we could find models with other statistically significant variables.

The two models with the highest R Squared values were those with:

1. am, wt, cyl, and hp
2. am, wt, hp

However, given that the R-Squared values were very similar for the two, I chose the second model since both weight and horsepower had statistically significant values.

##   
## Call:  
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5295 -2.3619 -0.1317 1.4025 6.8782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.32155 3.05464 12.218 5.84e-13 \*\*\*  
## mtcars$am1 -0.02362 1.54565 -0.015 0.988   
## mtcars$wt -5.35281 0.78824 -6.791 1.87e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.098 on 29 degrees of freedom  
## Multiple R-squared: 0.7528, Adjusted R-squared: 0.7358   
## F-statistic: 44.17 on 2 and 29 DF, p-value: 1.579e-09

##   
## Call:  
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$cyl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.1735 -1.5340 -0.5386 1.5864 6.0812   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.4179 2.6415 14.923 7.42e-15 \*\*\*  
## mtcars$am1 0.1765 1.3045 0.135 0.89334   
## mtcars$wt -3.1251 0.9109 -3.431 0.00189 \*\*   
## mtcars$cyl -1.5102 0.4223 -3.576 0.00129 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.612 on 28 degrees of freedom  
## Multiple R-squared: 0.8303, Adjusted R-squared: 0.8122   
## F-statistic: 45.68 on 3 and 28 DF, p-value: 6.51e-11

##   
## Call:  
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$hp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4221 -1.7924 -0.3788 1.2249 5.5317   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.002875 2.642659 12.867 2.82e-13 \*\*\*  
## mtcars$am1 2.083710 1.376420 1.514 0.141268   
## mtcars$wt -2.878575 0.904971 -3.181 0.003574 \*\*   
## mtcars$hp -0.037479 0.009605 -3.902 0.000546 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.538 on 28 degrees of freedom  
## Multiple R-squared: 0.8399, Adjusted R-squared: 0.8227   
## F-statistic: 48.96 on 3 and 28 DF, p-value: 2.908e-11

##   
## Call:  
## lm(formula = mtcars$mpg ~ mtcars$am + mtcars$wt + mtcars$cyl +   
## mtcars$hp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4765 -1.8471 -0.5544 1.2758 5.6608   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.14654 3.10478 11.642 4.94e-12 \*\*\*  
## mtcars$am1 1.47805 1.44115 1.026 0.3142   
## mtcars$wt -2.60648 0.91984 -2.834 0.0086 \*\*   
## mtcars$cyl -0.74516 0.58279 -1.279 0.2119   
## mtcars$hp -0.02495 0.01365 -1.828 0.0786 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.509 on 27 degrees of freedom  
## Multiple R-squared: 0.849, Adjusted R-squared: 0.8267   
## F-statistic: 37.96 on 4 and 27 DF, p-value: 1.025e-10

#### Residuals

