Motor Trend Magazine mpg Transmission Analysis

Ann Crawford

October 2, 2017

# Executive Summary

This paper uses regression models and exploratory data analysis on the mtcars dataset to examine transmision type (mannual or automatic) as a predictor for miles per gallon. Specifically address these questions"" 1. Is an automatic or manual transmission better for MPG? In this case better means higher value for miles per gallon.  
2. Quantify the MPG difference between automatic and manual transmissions.

## Steps to Analysis

1. Examine the DataSet
2. Analyze a simple linear regression mpg to transmission: automatic vs manual
3. Review other perdictors using multivariate analysis to quantify transmission as a predictor.

## Conclusion

The strongest indication of mpg is defined by the inverse realation of the weight to mpg. Transmission and weight are correlated. Including transmission in regression analysis add little information for explaining the variance for mpg.

### Examine the DataSet

The mtcars data frame contains 32 observations on 11 variables. vs,am, gear, carb and cyl can be treated as factors. For details consult help **?mtcars**

y <- mtcars$mpg  
x <- mtcars$am  
# check for nas   
sapply(mtcars,function(x) sum(is.na(x)))

## mpg cyl disp hp drat wt qsec vs am gear carb   
## 0 0 0 0 0 0 0 0 0 0 0

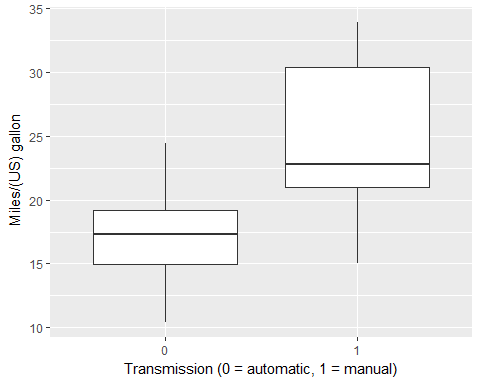
sapply(mtcars, function(x) length(unique(x)))

## mpg cyl disp hp drat wt qsec vs am gear carb   
## 25 3 27 22 22 29 30 2 2 3 6

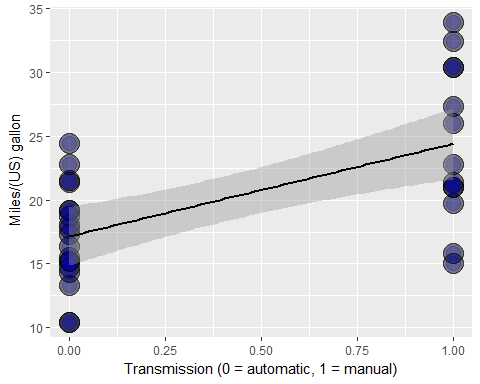
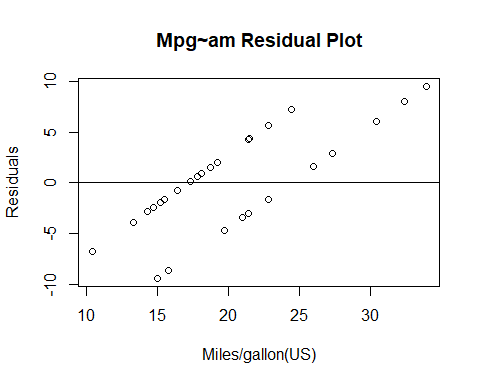
##   
## Call:  
## lm(formula = mpg ~ am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3923 -3.0923 -0.2974 3.2439 9.5077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*  
## am 7.245 1.764 4.106 0.000285 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.902 on 30 degrees of freedom  
## Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385   
## F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

### Analyze a simple linear regression mpg to transmission: automatic vs manual

Since question 1 specifcally asks to compare mpg to tranmission, the box plot shows that mannual transmissions have higher mpg in general with some variance. To further explore the variance and the value of am as a predictor, start with simple linear regression using transmisison as the single perdictor for mpg and review the residuals. fit1 <- lm(mpg ~ am , data= mtcars) e <- resid(fit1) The model fit returns a pvalue of 2.850207410^{-4} indicating statical significance at the 95% level. R squared = 0.3597989 indicating transmission accounts for roughly 36% of the variablity in mpg.



### Residual Plot shows errors with the model design

The residual plot shows, the relationship is **not a simple linear regression**. The pattern in residual plot reveals errors with the model design. Residuals represent variation left unexplained by our model. Points should be radomly clustered around the 0. 

# Multvariate Analysis

This section use multi-variate linear regression to explore how the other variables in the mtcars dataset may be related to mpg.

First fit a model with all the variables. But the results contradict the initial findings, where the p values show no statisical significance at the 95% level, all greater than 0.05, with wt having the lowest p value. This indicates and interation between the variables.

## Checking the correlation of perdictors

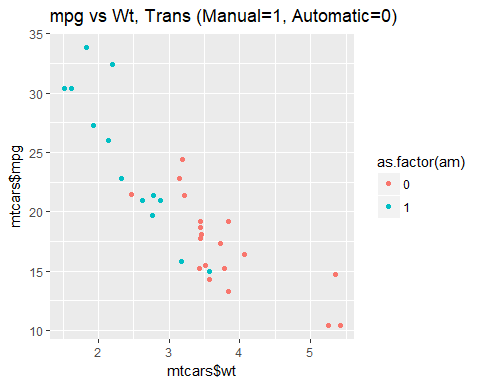
The correction matrix confirms the amount of correlation in the mtcars varaibles. It seems that wt is a strong predictor for the mpg response variable.

Ploting weight mpg and color coding the transmission shows a strong inverse relationship between wt and am. cor(mtcars[,1:11])

fit2 <- lm(mpg ~ ., data = mtcars)  
summary(fit2)

##   
## Call:  
## lm(formula = mpg ~ ., data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4506 -1.6044 -0.1196 1.2193 4.6271   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.30337 18.71788 0.657 0.5181   
## cyl -0.11144 1.04502 -0.107 0.9161   
## disp 0.01334 0.01786 0.747 0.4635   
## hp -0.02148 0.02177 -0.987 0.3350   
## drat 0.78711 1.63537 0.481 0.6353   
## wt -3.71530 1.89441 -1.961 0.0633 .  
## qsec 0.82104 0.73084 1.123 0.2739   
## vs 0.31776 2.10451 0.151 0.8814   
## am 2.52023 2.05665 1.225 0.2340   
## gear 0.65541 1.49326 0.439 0.6652   
## carb -0.19942 0.82875 -0.241 0.8122   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.65 on 21 degrees of freedom  
## Multiple R-squared: 0.869, Adjusted R-squared: 0.8066   
## F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

## Multiple Linear regression

 # Is weight or weight as an interaction variable with transsmission a better model for predicting mpg? Given the correlation betwwen wt and mpg is ??

fitwtam <- lm(mpg~am\*wt, mtcars)  
summary(fitwtam)

##   
## Call:  
## lm(formula = mpg ~ am \* wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6004 -1.5446 -0.5325 0.9012 6.0909   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.4161 3.0201 10.402 4.00e-11 \*\*\*  
## am 14.8784 4.2640 3.489 0.00162 \*\*   
## wt -3.7859 0.7856 -4.819 4.55e-05 \*\*\*  
## am:wt -5.2984 1.4447 -3.667 0.00102 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.591 on 28 degrees of freedom  
## Multiple R-squared: 0.833, Adjusted R-squared: 0.8151   
## F-statistic: 46.57 on 3 and 28 DF, p-value: 5.209e-11

fitwt <- lm(mpg~wt, mtcars)  
summary(fitwt)

##   
## Call:  
## lm(formula = mpg ~ wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5432 -2.3647 -0.1252 1.4096 6.8727   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.2851 1.8776 19.858 < 2e-16 \*\*\*  
## wt -5.3445 0.5591 -9.559 1.29e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.046 on 30 degrees of freedom  
## Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446   
## F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10