Variation in Motor Vehicles' MPG Explored with Linear Regression

J. Thatcher

1.0 Executive Summary

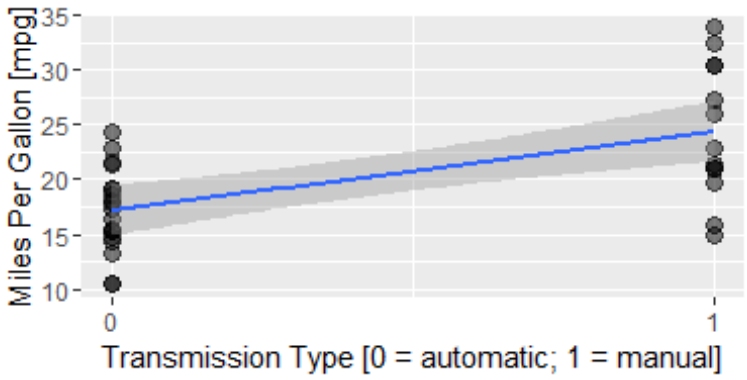
In this project we consider the `mtcars` dataset. We will explore the relationship between Transmission Type (automatic or manual) and Miles Per Gallon. The results of this analysis answer the two main questions asked. One, "Is an automatic or manual transmission better for MPG?" Answer, the manual transmission is on average better for MPG. Two, "Quantify the MPG difference between automatic and manual transmissions." We will demonstrate that using a linear regression model there is an average increase of 7.25 mpg for cars with a manual transmission. This effect is tapered by the addition of an additional regressor, number of cylinders, which is also explored in this report.

2.0 Data Analysis

2.1 Exploratory Data Analysis

In `mtcars` we find that the first column variable is `mpg` (Miles/US Gallon) and the nineth column variable `am` is Transmission Type (0 = automatic, 1 = manual).

Fist, an exploratoty graph of the response and predictor using `ggplot2`.



The graph shows that there may be a significant correlation between MPG and Transmission Type.

2.2 Relationship Between MPG and Transmission Type

Initially we are asked: is MPG better for manual or automatic transmission vehicles? First, we will calculate the coefficients of a linear model using the `lm` function.

```
##
## Call:
## lm(formula = mpg ~ I(factor(am)), data = DF)
##
## 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 ***
## I(factor(am))1    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
```

2.3 Interpretation of Correlation Coefficient, R-Squared

Reviewing the `lm` funciton results, we find **R-squared for the model is 0.36. Which means that transmission type explains 36% of the variance in vehicle MPG.**

2.4 Interpretation of the Intercept, Beta0

The y-intercept, **Beta0 = 17.15, is the mean MPG for automatic transmission vehicles.**

2.5 Interpretation of the Slope, Beta1

Beta1's value will answer our initial quesiton, "is MPG better for manual or automatic transmission vehicles?" Beta1 is +7.25, and is interpreted as the change in the average vehicle's MPG for one unit increase in the predictor variable, Transmission Type. Beta1 is positive, which means that there is an increase in MPG for manual transmission vehicles. **The average MPG of manual transmission vehicles is 7.25 MPG greater than that of automatic transmission vehicles.**

2.6 Is the slope significant?

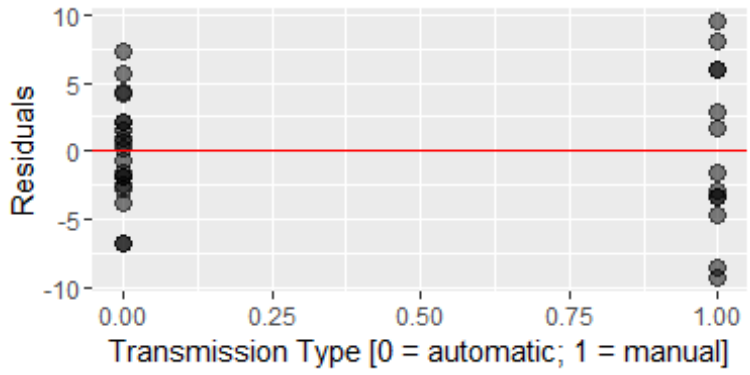
We can perform a hypothesis test on the Beta1 coefficient to determine that the slope is actually significant. The hypotheses are:

H0: Beta1 = 0
H1: Beta1 != 17.2

To determine whether we can reject the null hypothesis, H0, we will go back to our results from the `lm` function. **We find the probability (P) that the the slope is equal to zero is P = 0.000285.** This is much lower than our level of statistical significance, alpha = 0.05. Therefore, **we can reject H0 and determine that the difference in MPG between transmission types is significant.**

2.7 Residuals

In a residual plot, we are looking for the residual points to be spread in a random fashion on either side of the horizantle line. This will help confirm that the linear model was an appropriate model for this data.

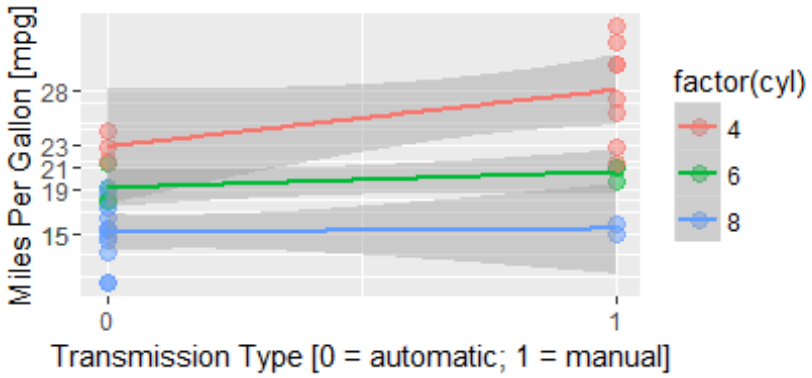


The plot does show that the **residuals are spread evenly above and below the line at 0. However,** the graph shows that the data slightly resembles a property called Heteroscedasticity. In this case **the variance of MPG appears to be slightly higher in cars with manual transmissions compared to automatic transmissions.**

2.8 Alternative Models

We cannot assume that the transmission type will have the same effect for all vehicle models, engines, weights, etc. Therefore, it makes sense to explore the effect of other variables in the model. To do this, we will adjust our model by adding more regressors from the datased. In this case we will adjust the model by adding the variable cylinders `cyl` as a factor variable.

```
##                                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                     22.900000   1.750674 13.080673 6.057324e-13
## I(factor(am))1                   5.175000   2.052848  2.520888 1.817605e-02
## I(factor(cyl))6                 -3.775000   2.315925 -1.630018 1.151546e-01
## I(factor(cyl))8                 -7.850000   1.957314 -4.010599 4.547583e-04
## I(factor(am))1:I(factor(cyl))6 -3.733333   3.094784 -1.206331 2.385526e-01
## I(factor(am))1:I(factor(cyl))8 -4.825000   3.094784 -1.559075 1.310693e-01
```



In this case **the addition of the third variable Cylinders distored the relationship between MPG and Transmission type**. Compared to an automatic transmission vehicle, the **effect of having a manual transmission is now +5.2 MPG for 4-cylinder vehicles, +1.4 MPG for 6-cylinder vehicles, and +0.4 MPG for 8-cylinder vehicles.** Unfortunatly, this is not a well-balaned dataset to explore this effect, and many of the coefficents in this model are not significant.

3.0 Code

```
###Get data
library(ggplot2)
DF <- data.frame(mtcars) #grab data as data.frame

###exploratoty graph of the response and predictor
#plot the data and a linear fit
ggplot(DF, aes(x = am, y = mpg)) +
        geom_point(alpha = 0.5, cex = 3) +
        geom_smooth(method = "lm") +
        scale_x_continuous(breaks = c(0,1)) +
        ylab(label = "Miles Per Gallon [mpg]") +
        xlab(label = "Transmission Type [0 = automatic; 1 = manual]")

###2.2 Relationship Between MPG and Transmission Type
#fit a linear model
fit <- lm(mpg ~ I(factor(am)), data = DF)
summary(fit)

###2.7 Residuals
#plot the residuals
ggplot(fit, aes(x = DF$am, y = resid(fit))) +
        geom_point(alpha = 0.5, cex = 3) +
        geom_hline(yintercept=0, color="red") +
        ylab(label = "Resifuals") +
        xlab(label = "Transmission Type [0 = automatic; 1 = manual]")

###2.8 Alternative Models
```

```r
fit2 <- lm(mpg ~ I(factor(am)) * I(factor(cyl)), data = DF)
summary(fit2)$coef
ggplot(DF, aes(x = am, y = mpg, color = factor(cyl))) +
        geom_point(alpha = 0.5, cex = 3) +
        geom_smooth(method = "lm") +
        scale_x_continuous(breaks = c(0,1)) +
        scale_y_continuous(breaks = round(c(coef(fit2)[[1]],
coef(fit2)[[1]]+coef(fit2)[[3]],
                                            coef(fit2)[[1]]+coef(fit2)[[4]],
                                            coef(fit2)[[1]]+coef(fit2)[[2]],

coef(fit2)[[1]]+coef(fit2)[[2]]+coef(fit2)[[3]]+coef(fit2)[[5]],

coef(fit2)[[1]]+coef(fit2)[[2]]+coef(fit2)[[4]]+coef(fit2)[[6]]))) +
        ylab(label = "Miles Per Gallon [mpg]") +
        xlab(label = "Transmission Type [0 = automatic; 1 = manual]")
```