# Regression Models Course Project

## Executive Summary

Is having an automatic or manual transmission better for fuel efficiency (mpg), and by how much? The simple answer is that the average mpg is 17.14 for automatic transmissions, and 42% more (24.4mpg) for manual, so manual transmission does seem to be better for fuel efficiency. If we wanted to use this model alone for predicting fuel efficiency, we would expect 95% of the time the actual mpg would be within the range of 14.9 - 19.4 mpg for automatic and 18.5 - 30.3 mpg for manual transmissions. However, if we look at other factors such as the weight of the car, the answer is not that straight forward. First, car weight is a much better predictor of fuel efficiency than transmission type is; and second, there is a tendency for low-weight cars to have manual transmissions. With this in mind, we used a model where weight, transmission type and their interaction was considered, and this model explained more of the variation in fuel efficiency than our original model. For automatic transmission, each 1000 lb increase in car weight was predicted to result in a 3.8 miles per gallon reduction in fuel efficiency, while for manual transmission the reduction was much larger, close to 9.1 mpg. We can be 95% confident that our predictions for mpg fuel efficiency reduction for each 1000lb of car weight is within the range of 2.2 and 5.4 mpg for automatic transmissions, and within the range of 4.5 to 13.7 mpg for manual transmissions. In conclusion, it seems that automatic transmission is better for fuel efficiency, but other factors (such as weight) have even more effect on mpg than the transmission type.

## Methodology

The analysis was performed using the 1974 Motor Trend US magazine automobile data *mtcars*, consisting of 11 variables and 32 observations. First check to see if the proposed relationship between fuel efficiency (*mpg*, miles per gallon) and transmission type (*am*, 0 if automatic, 1 if manual) exists; **Figure 1A** (see Appendix) suggests that there is some relationship. Therefore use a simple linear regression model, where mpg is the outcome and am is the predictor, to fit the data.

```
fit_am <-lm(mpg~factor(am),mtcars); summary(fit_am)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

According to the results, cars with automatic transmission (*am*=0) average 17.147 miles per gallon, while cars with manual transmission average 17.147+7.245=24.392 mpg; the p values are small enough that we can say with 95% confidence that the averages are not 0. Unfortunately, the variation in mpg explained by this model ($R^2$) is about 36%, which suggests that there is more to the story. In addition, plotting the model residuals (see **Figure 1B** in Appendix) suggests that the model may be heteroskedastic, i.e. the prediction accuracy for manual transmission is lower than the prediction accuracy for automatic transmission.

Therefore let's examine the other variables in the **mtcars** data set to see if including one or more ofthem could improve our model. **Figure 2** in the Appendix shows the plots and correlations between each of the variable pairs in the **mtcars** data set.

The first thing to notice is that the variable *am* (automatic/manual) is not as highly correlated to *mpg* (.6), as some of the other variables in the data set. If other variables can explain the variation in *mpg* better, then that variation should be removed before we consider the effects of "am". The top four variables having the highest absolute correlation with mpg (miles per gallon) are: *wt* (weight) -> -0.868; *cyl* (number of cylinders) -> -0.852; *disp* (displacement cu. in.) -> -0.848; and *hp* (horsepower) -> -0.776. Visually examinig their pair plots with *mpg* confirms that these four variables are all similarly negatively correlated with *mpg*. The last three variables are indicators of engine size and power, and therefore it is not surprising that they are also highly positively correlated with each other (in the .8-.9 range). It also makes sense that the weight of the car is correlated with engine power, as the heavier the car the more engine power required to move it – and indeed, correlations are in the range of .65-.9. That given, the variable *wt* (weight), having the highest correlation with *mpg* and also highly correlated engine power indicator variables, was chosen to be examined in models with and without *am*.

```
fit_wt <- lm(mpg~wt,mtcars)
```

More than 75% of the mpg variation is explained by the car weight alone ($R^2$=.7528 in mpg~wt model regression). **Figure 3A** shows the data points and the fitted line of the model. Plotting the residuals (see **Figure 3B**) reveals that there is a slight pattern: the residuals at the low and high end of weight are consistently positive. This means that predictions using this model our predictions would be underestimated at the low and high end of car weight spectrum. Could this pattern be accounted for by the variable we are interested it, the transmission type?

To answer this question, we have fitten a multivariate model of *mpg* vs *wt* plus *am*. The result of this model should be two parallel lines, one intercept for predicting mpg for automatic, and another one for manual transmission.

```
fit2 <- lm(mpg~wt+factor(am),mtcars)
summary(fit2)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
## factor(am)1 -0.02361522  1.5456453 -0.01527855 9.879146e-01
```

There is barely any difference in the two intercepts (-0.02362), and even that is not significant (p-value is a huge .988). This can be seen in **Figure 4A**, where the fitted lines overlap. The residual plot (**Figure 4B**) is practically identical to that of the weight-only model. However, the actual data points in the graph, coloured red for manual and black for automatic, visually reveal an association of manual transmission with low-weight cars. In order to account for that association, a model that includes and interaction term between *wt* and *am* is fitted.

```
fitx <- lm(mpg ~ wt + factor(am) + wt*factor(am),mtcars); summary(fitx)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(am) + wt * factor(am), 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 ***
## wt              -3.7859     0.7856  -4.819 4.55e-05 ***
## factor(am)1     14.8784     4.2640   3.489  0.00162 **
## wt:factor(am)1  -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
```

This model has two predictor lines with different intercepts and slopes, one for automatic and one for manual. **Figure 5A** shows these, the manual one in red. The residuals plotted in **Figure 5B** show that this model no longer has the positive residual bias on the low and high end of the car weight spectrum. The $R^2$ of .833 suggests that this model explains 83% of the variability in *mpg*, which is better than any of the other models. The high F-statistic and the associated p-value of 0 confirm that at least one of the coefficient estimates is non-zero, so we can be confident that this is a valid model. Furthermore, each of the coefficients p-value is significant to at least 95%,i.e. we can be 95% certain that each of the coefficients is non-zero in this model.

Just to confirm, let's compare the three (nested) models involving weight:

```r
anova(fit_wt,fit2,fitx)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + factor(am)
## Model 3: mpg ~ wt + factor(am) + wt * factor(am)
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 278.32  1     0.002  0.0003 0.985556
## 3     28 188.01  1    90.312 13.4502 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, there is no value in adding the *am* factor alone as the residial sum of squares (RSS) remains constant. However there is a big improvement (32% RSS decrease) when the interaction is added, and its F-statistic, here testing the hypothesis that this additional interaction coefficient is non-zero, is significant to 99%.

The model implies that for automatic transmission, each 1000 lb increase in car weight is predicted to result in a 3.8 miles per gallon reduction in fuel efficiency, while for manual transmission the reduction is close to 9.1 mpg.

```r
ci<-confint(fitx,level=.95);c(ci[2,1],ci[2,2]);c(ci[2,1]+ci[4,1],ci[2,2]+ci[4,2])
```

```
## [1] -5.395234 -2.176581
```

```
## [1] -13.652927  -4.515609
```

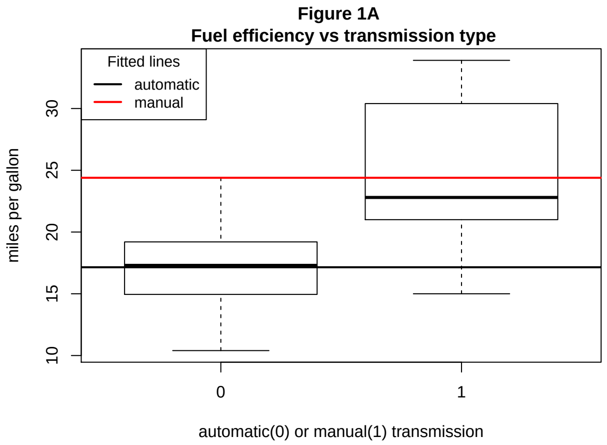
We are 95% confident that mpg fuel efficiency reduction for each 1000lb of car weight is within the range of 2.2 and 5.4 mpg for automatic transmissions, and within the range of 4.5 to 13.7 mpg for manual transmissions.

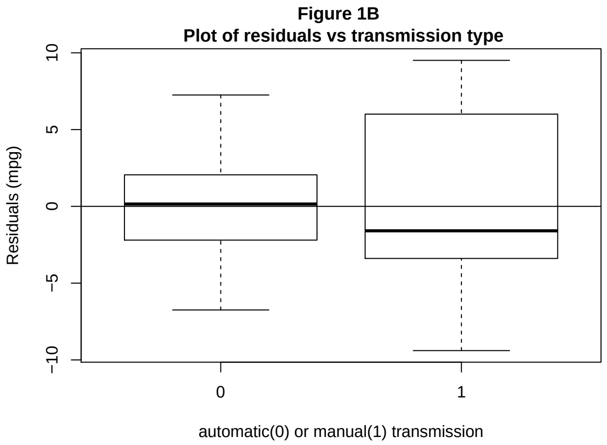
## Note

This document was produced by R markdown. The corresponding .Rmd file is available at .

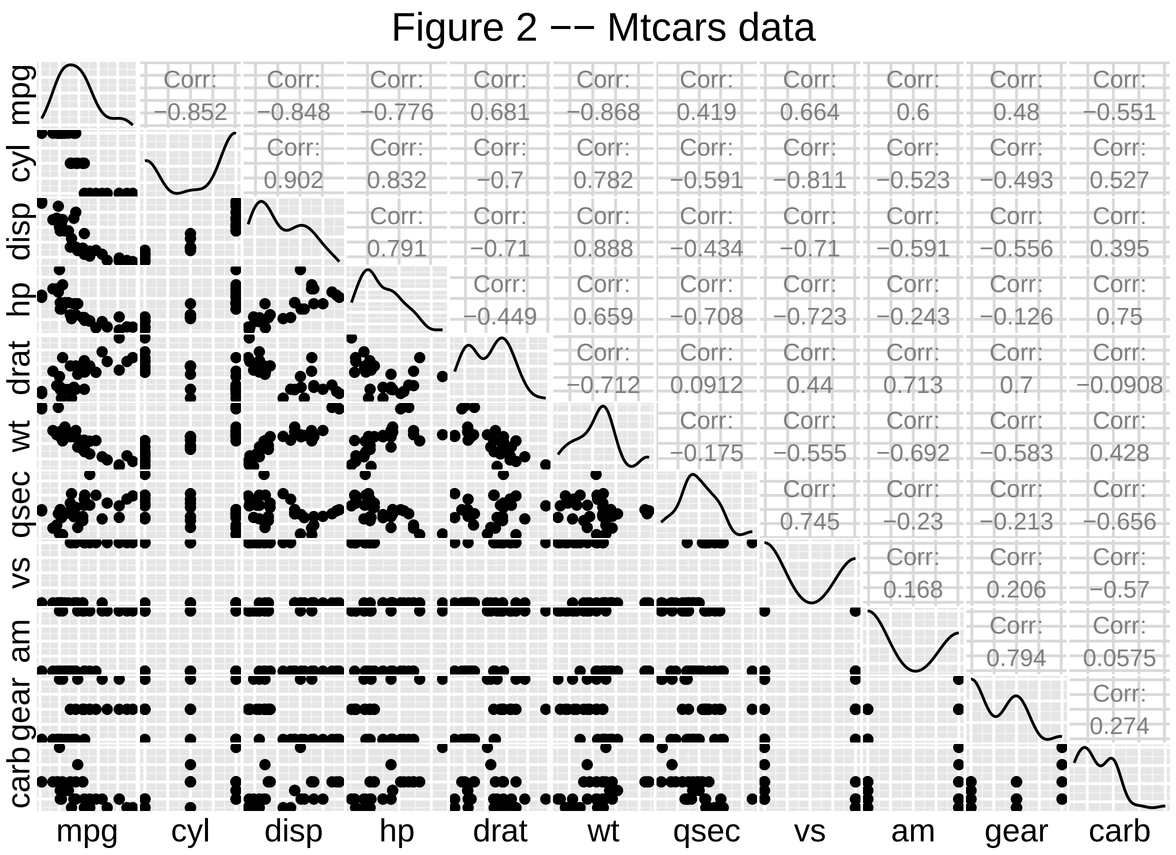
# Appendix

## Simple linear model: mpg vs am





## Dataset mtcars paired plots and correlations



## Simple linear model: mpg vs wt

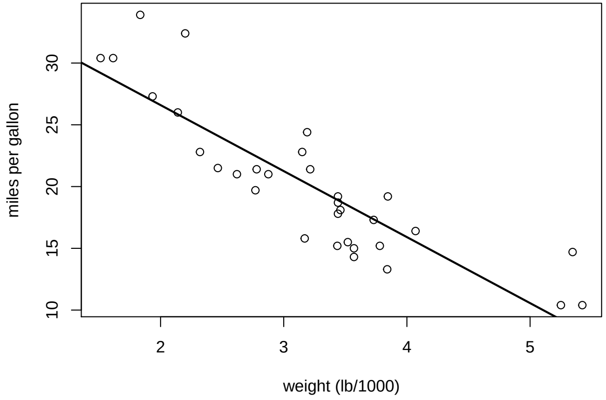


Figure 3A
Fuel efficiency vs car weight

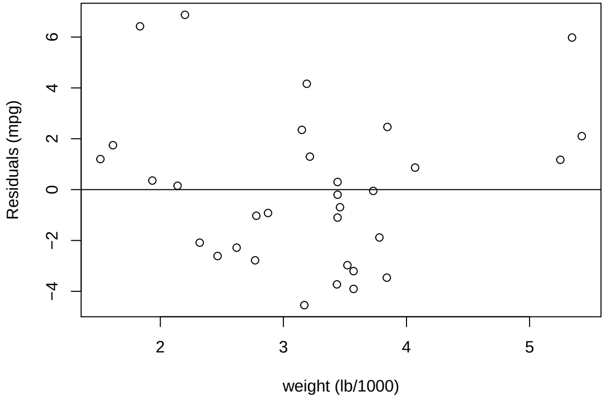


Figure 3B
Residuals vs car weight

## Multivariate model: mpg vs wt and am

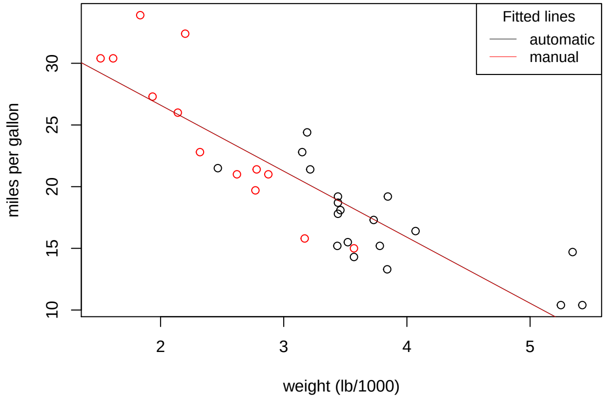


Figure 4A
Fuel efficiency vs car weight and transmission type

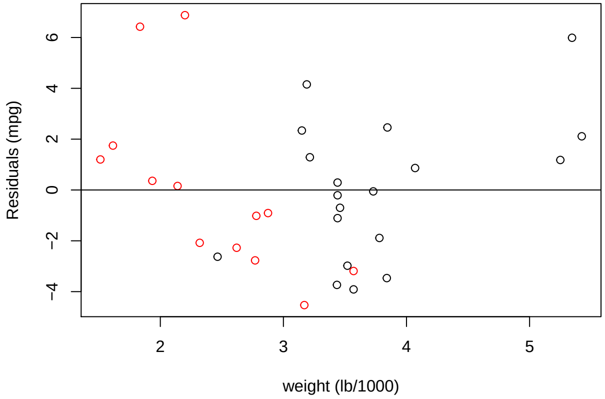


Figure 4B
Residuals vs car weight and transmission type

## Multivariate model: mpg vs wt and am with interaction

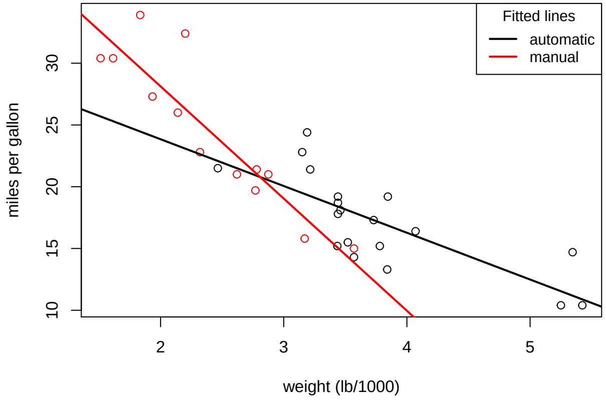


Figure 5A
Fuel efficiency vs car weight and transmission type and interaction

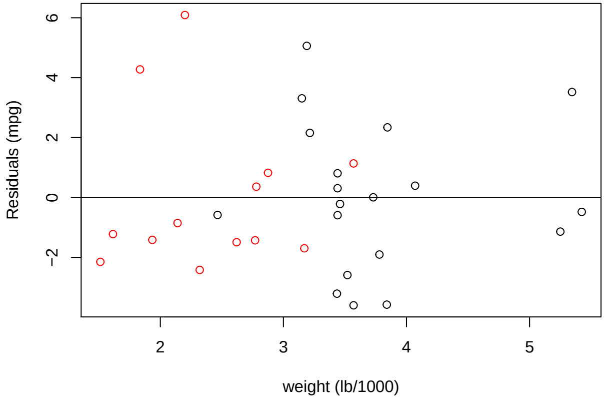


Figure 5B
Residuals vs car weight and transmission type and interaction