

# Analysis of mtcars: Automatic or Manual Transmission Better for MPG?

## Summary

When purchasing a vehicle, one of the most important factors for a prospective buyer is the car's mpg. This analysis looks at how the mpg is affected by the type of transmission on a vehicle, utilizing data from the 1974 Motor Trends US magazine for a set of cars from 1973-1974 using 11 characteristics. The results of this analysis determined that a vehicle's transmission type alone is not the best indicator of the fuel efficiency of the automobile.

## Data Processing/Exploration

The standard mtcars data sets was loaded. This data set is comprised of 32 observations of 11 variables. No cleansing of data was conducted.

First, a scatterplot of each pair of characteristics was generated with the "pairs"" command to get an initial visualization of potential data relationships (see Appendix Figure 1). Then correlations were checked:

```
sort(abs(LGcor[,1]))
      qsec      gear      carb      am      vs      drat      hp      disp      cyl
wt      mp
0.4186840 0.4802848 0.5509251 0.5998324 0.6640389 0.6811719 0.7761684 0.8475514 0.8521620
0.8676594 1.0000000
```

Initial check of the correlation between variables indicates a strong correlation between mpg and weight (.86766), number of cylinders (.85216), and displacement (.84755). However, further review of all variable correlation shows a strong correlation between displacement and both cylinders (.902) and weight (.88798) so displacement will be removed from further consideration in order to ensure we don't have collinear predictors that could make it difficult to separate the effects of these variables on the response. Figure 2 graphically shows the relationship between mpg and 1) weight and 2) weight and cylinders, respectively.

A simple linear regression, modelling mpg as a function of am (transmission type), results in an R-squared value of .3598, meaning that only 35.98% of the variation of the mpg can be explained by the variable am. Looking back to the results of the correlation, the next most likely values to model would be weight and cylinders. Displacement will not be included due to the previously determined likelihood of confounding the results.

```
lm(formula = mpg ~ am, data = mtcars)
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 ***
Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385      F-statistic: 16.86 on 1 and
30 DF, p-value: 0.000285
```

```
lm(formula = mpg ~ wt + cyl, data = mtcars)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.6863      1.7150   23.141 < 2e-16 ***
wt          -3.1910      0.7569   -4.216 0.000222 ***
cyl           0.4122      0.0170    24.210 < 2e-16 ***
```

```
cyl          -1.5078      0.4147  -3.636 0.001064 **
Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185      F-statistic: 70.91 on 2 and
29 DF, p-value: 6.809e-12
```

Since we want to evaluate the transmission type, a final linear model is generated that include this characteristic with the weight and number of cylinders.

```
lm(formula = mpg ~ am + wt + cyl, data = mtcars)
Multiple R-squared:  0.8303,    Adjusted R-squared:  0.8122
```

The inclusion of the transmission type into the model only increased the coefficient of determination (R-squared) by .0001109. Though this is far superior than the original model which only include the mpg characteristic, adding mpg to the model with weight and number of cylinders doesn't add much to explain the variation in mpg.

## Results

Performing an analysis of variance (ANOVA) on the model that includes only transmission type (am) to the model utilizing "wt" and "cyl" compared to a model with "wt" supports the original conclusion derived by looking at the respective R-squared values for the models, namely that there is a statistically significant difference in the models (since P-value easily exceeds .05).

Further, analyzing the models containing only "wt" and "cyl" to the model with these and "am" does NOT show a statistically significant difference, as demonstrated by the P value of .8933.

```
anova(fitT, fitWC)
Model 1: mpg ~ am
Model 2: mpg ~ wt + cyl
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      30 720.90
2       29 191.17  1    529.72 80.357 7.421e-10 ***
```

```
anova(fitWC, fitTWC)
Model 1: mpg ~ wt + cyl
Model 2: mpg ~ am + wt + cyl
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1       29 191.17
2       28 191.05  1    0.12491 0.0183 0.8933
```

Plotting the residuals for the model containing weight and cylinders shows no discernable or predictable pattern (see Figure 3); therefore we can conclude that our model is sound, indicating that vehicle weight combined with the number of cylinders is a good predictor of its fuel consumption (mpg).

## Conclusion

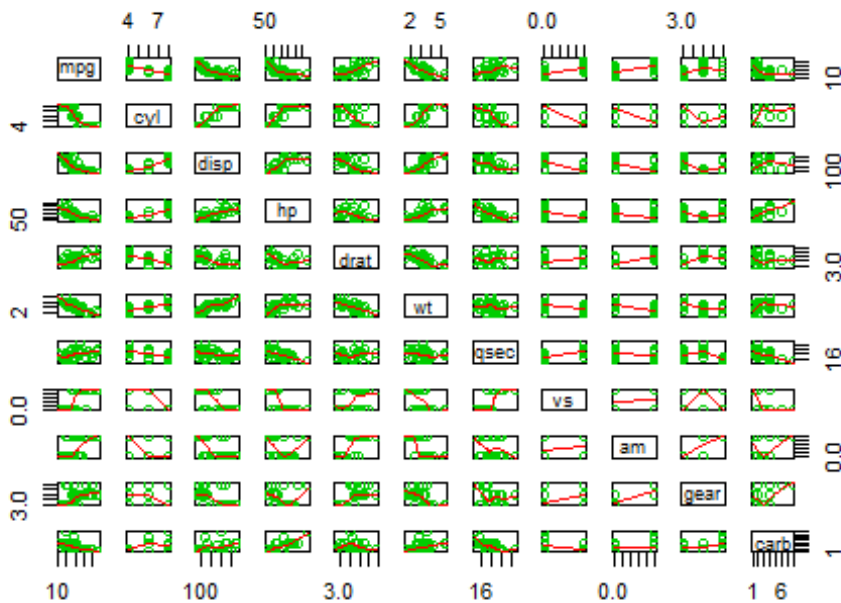
Attempting to predict the mpg based solely on the transmission type is not a solid predictor since it can only explain 35% of the variance. Predicting mpg based upon a vehicles weight and number of cylinders is a good predictor of mpg, accounting for 83.02% of mpg variance. Including the transmission type in the model only improved the accuracy of the model by .01% and therefore is not needed.

---

## Appendix

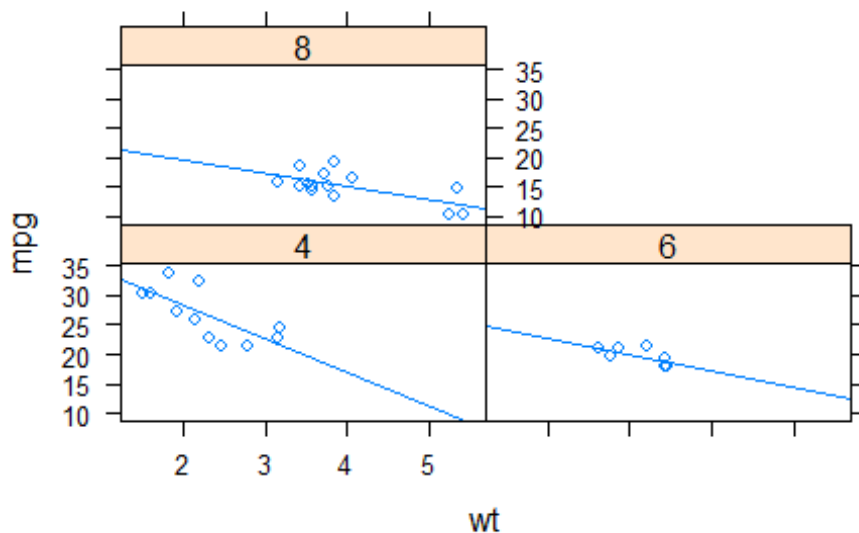
```
pairs(mtcars, panel=panel.smooth, main="Figure 1 \n Plot of pairs - mtcars data", col=3)
```

**Figure 1**  
**Plot of pairs - mtcars data**



```
library(lattice)
xyplot(mpg~wt|factor(cyl), data=mtcars, type=c("p", "r"), main="Figure 2\n Relationship B\n between MPG and Weight \n (By Number of Cylinders)")
```

**Figure 2**  
**Relationship Between MPG and Weight**  
**(By Number of Cylinders)**



```
plot(fitted(lm(mpg~wt+cyl+disp, data=mtcars)), resid(lm(mpg~wt+cyl+disp, data=mtcars)), m
ain="Figure 3\n Residual Plot (MPG~Weight+Cylinders)", xlab="Fitted", ylab="Residuals")
abline(h=0, lty=2)
```

