Regression Models: Impact of Transmission on MPG

by: 335emily

Executive Summary

In this report we aim to determine if automatic or manual transmission is better for miles per gallon (MPG), and quantify the impact automatic and manual transmissions have on MPG.

Our findings indicate that manual transmission is associated with higher MPG. Weight and quarter-mile time are also statistically significant predictors of MPG.

Weight and quarter-mile being equal, a manual car will have 2.9 more MPG than an automatic car.

Strategy for Model Selection

Given the coefficient of interest (MPG), we will assess:

1. A multi-variate linear model that considers all other variables as predictors
2. A single variable linear model that only considers transmission as a predictor
3. All variables regressed against each other
4. A multi-variate linear model created with the Stepwise algorithm to determine which variables are statistically significant

Then, a residual plot will be used to confirm the residuals are "balanced" among the data points.

Understanding the Data and Interpreting the Coefficients

The mtcars dataset has 32 observations of 11 variables. **See appendix [0] for variable definitions.** cyl, vs, am and gear should be classified as factor variables:

```
mtcarsClean <- mtcars
mtcarsClean$cyl <- as.factor(mtcarsClean$cyl)
mtcarsClean$vs <- as.factor(mtcarsClean$vs)
levels(mtcarsClean$vs) <- c("V", "S")
mtcarsClean$am <- as.factor(mtcarsClean$am)
levels(mtcarsClean$am) <- c("automatic", "manual")
mtcarsClean$gear <- as.factor(mtcarsClean$gear)
mtcarsClean$carb <- as.factor(mtcarsClean$carb)
```

Exploratory Data Analysis

1. Including All Variables in MPG Estimation Linear Model

First, consider the linear model for MPG with all variables included:

```
fitAll <- lm(mpg ~ . , data = mtcarsClean)
```

See appendix [1] for summary of fitAll. This model suggests that all things being equal, a manual car has an MPG 1.212 higher than automatic cars.

The Adjusted R-squared value is 0.779, which means that the model can explain about 77.9% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significance level.

2. Using Only Transmission to Estimate MPG

Consider the unadjusted comparison between automatic and manual cars; first a box and whisker plot. A box and whisker plot shows that cars with manual transmissions have a higher MPG than cars with automatic transmissions. However, other variables may impact MPG. **See appendix [2] for the box plot.**

Consider the unadjusted linear regression:

```
fitAM <- lm(mpg ~ am, data = mtcarsClean)
```

See appendix [3] for the summary of fitAM. This model suggests that manual cars have an MPG 7.245 higher than automatic cars.

The Adjusted R-squared value is 0.338, which means that the model can explain about 33.8% of the variance of the MPG variable. The relatively low R-squared indicates other variables may also explain model variance.

3. Comparing All Variables to Each Other

See appendix [4] for the graphs of the relationship between all variables. This shows that there are no notable outliers in the data for MPG.

4. Selecting a Model with the Stepwise Algorithm

In order to determine which variables to include, we will use the a Stepwise algorithm, with the step function, by B. D. Ripley. See more information at this link (click).

```
fitStep <- step(fitAll, k=log(nrow(mtcarsClean)),trace=0)
```

See appendix [5] for the summary of the fitStep model. From this, we see that weight (wt), 1/4 mile time (qsec) and transmission (am) are all significant at the 0.05 level. This model suggests that manual cars have an MPG 2.936 higher than automatic cars.

The Adjusted R-squared value is 0.834, which means that the model can explain about 83.4% of the variance of the MPG variable. This is the highest Adjusted R-squared of the assessed models.

Residual Plot

See appendix [6] for the residual plot. The residual plot confirm the residuals are "balanced" among the data points. Residuals must be uncorrelated with predictors, because if the residuals and predictors were correlated, we could make a better prediction to reduce the residuals.

Conclusions

Using the 3 variables shown as statistically significant in the Stepwise algorithm gets us close to a suitable prediction However, the relationships are not known to be causal, and correlation may be coincidental or not specific to these 3 predictors.

Next steps would be to test the model against additional data that was not included in the original analysis to determine if the model remains similarily accurate. Additionally, other variables not included in this data set may further reduce the correlation of residuals.

Appendices

[0] Variable Definitions

Variable Name	Variable Definition
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

[1] Summary of fitAll Model (all variables regressed against mpg)

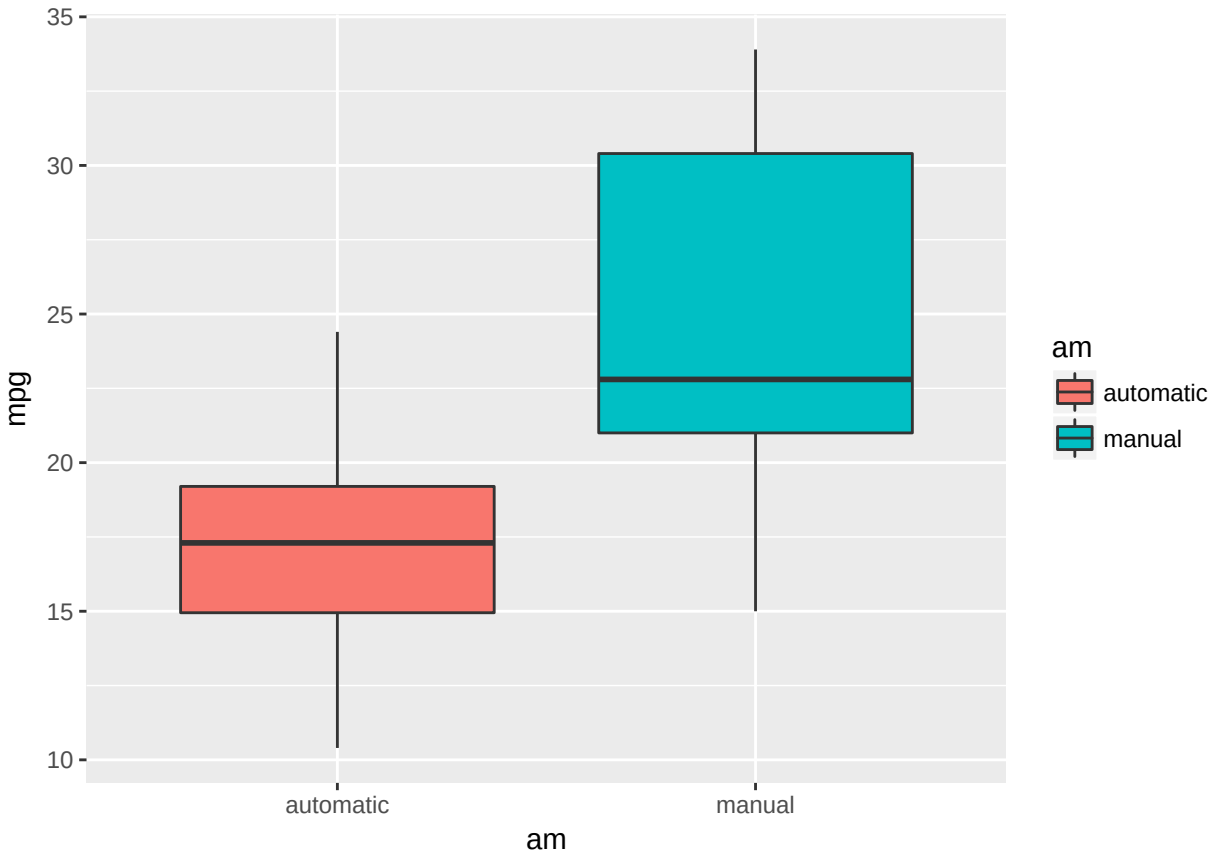
```
summary(fitAll)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcarsClean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vsS          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
```

```
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

[2] Box plot of Transmission to MPG

```
p <- ggplot(mtcarsClean, aes(am, mpg))
p + geom_boxplot(aes(fill = am))
```



[3] Summary of fitAM Model (transmission variable regressed against mpg)

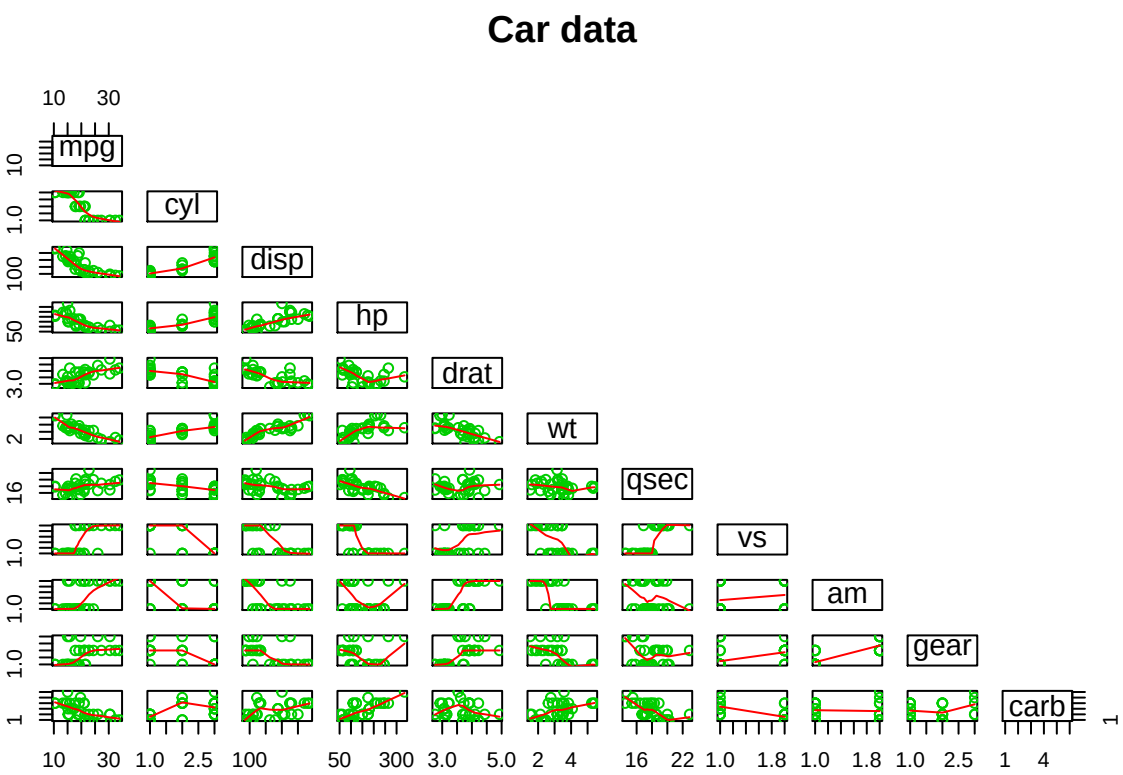
```
summary(fitAM)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcarsClean)
##
## 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 ***
## ammanual       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
```

[4] Relationships between All Variables

```
require(stats)
require(graphics)
pairs(mtcarsClean, lower.panel = panel.smooth, upper.panel=NULL, main = "Car data", col = 3 )
```



[5] Summary of fitStep Model (variables determined by the Stepwise Algorithm)

```
summary(fitStep)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcarsClean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
```

```
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

[6] Residual plot for fitStep Model

```
par(mfrow = c(2, 2))
plot(fitStep)
```