Regression Models Course Project

Eugene Fallon

July 26, 2015

Automatic or Manual Transmission: Which Has Better MPG?

Executive Summary

Context

The project assignment was to assume you work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1) Is an automatic or manual transmission better for MPG?
2) Quantify the MPG difference between automatic and manual transmissions.

The Analysis section of this document focuses on inference with a simple linear regression model and a multiple regression model. Both models support the conclusion that the cars in this study with manual transmissions have on average significantly higher MPG's than cars with automatic transmissions.

This conclusion holds whether we consider the relationship between MPG and transmission type alone or transmission type together with another predictors: wt / weight - the weight of the vehicle.

In the simple model, the mean MPG difference is 7.245 MPG; the average MPG for cars with automatic transmissions is 17.147 MPG, and the average MPG for cars with manual transmissions is 24.392 MPG. In the multiple regression model, an increase in weight resulted in a loss of .0236 MPG per unit.

Exploratory analysis and visualizations are located in the Appendix to this document.

Analysis

Simple Linear Regression - lm(mpg ~ am, data = mtcars)

```
data(mtcars)
n <- length(mtcars$mpg)
alpha <- 0.05
fit <- lm(mpg ~ am, data = mtcars)
coef(summary(fit))
```

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

The beta0 / intercept coefficient is mean MPG for cars with automatic transmissions; the beta1 / am coefficient is the mean increase in MPG for cars with manual transmissions (am = 1). The sum beta0 + beta1 is our mean MPG for cars with manual transmissions.

Using the output above, we can calculate a 95% confidence interval for beta1 (mean MPG difference) as follows:

```r
pe <- coef(summary(fit))["am", "Estimate"]
se <- coef(summary(fit))["am", "Std. Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1]  3.64151 10.84837
```

The p-value of 2.850207410^{-4} for beta1 is small and the CI does not include zero, so we can reject null in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at alpha = 0.05.

Multiple Regression - lm(mpg ~ wt + am, data=mtcars)

The predictor wt (weight) was added to the model based on initial analyses of the data. The predictor am (transmission type) is of course a requirement of the project. This set of predictors yields the highest adjusted R-squared.

```r
# fit a model using the regressors suggested by bestglm residual plot is in
# Appendix
bestfit <- lm(mpg ~ wt + am, data = mtcars)
coef(summary(bestfit))
```

```
##                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
## am          -0.02361522  1.5456453 -0.01527855 9.879146e-01
```

Using the output above, we can calculate a 95% confidence interval for beta3 / am as follows:

```r
pe <- coef(summary(bestfit))["am", "Estimate"]
se <- coef(summary(bestfit))["am", "Std. Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1] -3.180244  3.133014
```

Nested Model

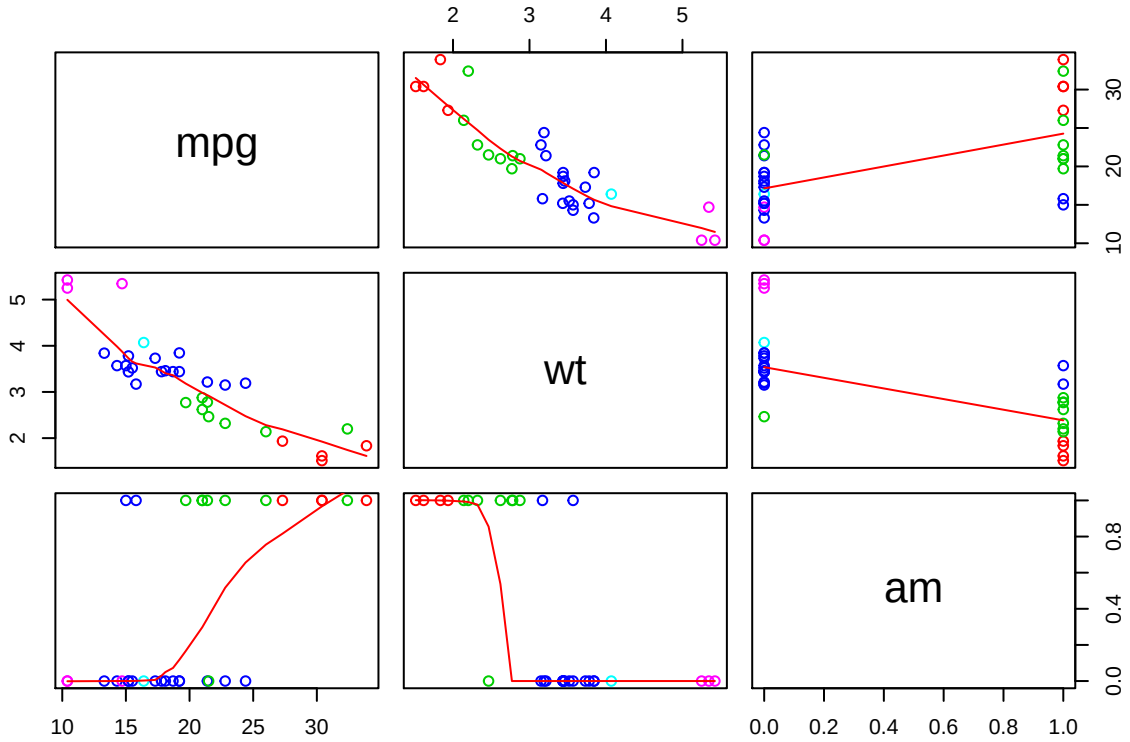
```r
# nested model testing of the model selected by bestglm
fit1 <- lm(mpg ~ wt, data = mtcars)
fit2 <- update(fit1, mpg ~ wt + am)
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + am
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     30 278.32
## 2     29 278.32  1 0.0022403 2e-04 0.9879
```

Appendix - Exploratory Analysis and Visualizations

Correlations

```
mtcars_vars <- mtcars[, c(1, 6, 9)]
mar.orig <- par()$mar  # save the original values
par(mar = c(1, 1, 1, 1))  # set your new values
pairs(mtcars_vars, panel = panel.smooth, col = 9 + mtcars$wt)
```




```
par(mar = mar.orig)  # put the original values back
cor(mtcars_vars)
```

```
##            mpg         wt         am
## mpg  1.0000000 -0.8676594  0.5998324
## wt  -0.8676594  1.0000000 -0.6924953
## am   0.5998324 -0.6924953  1.0000000
```

Histograms

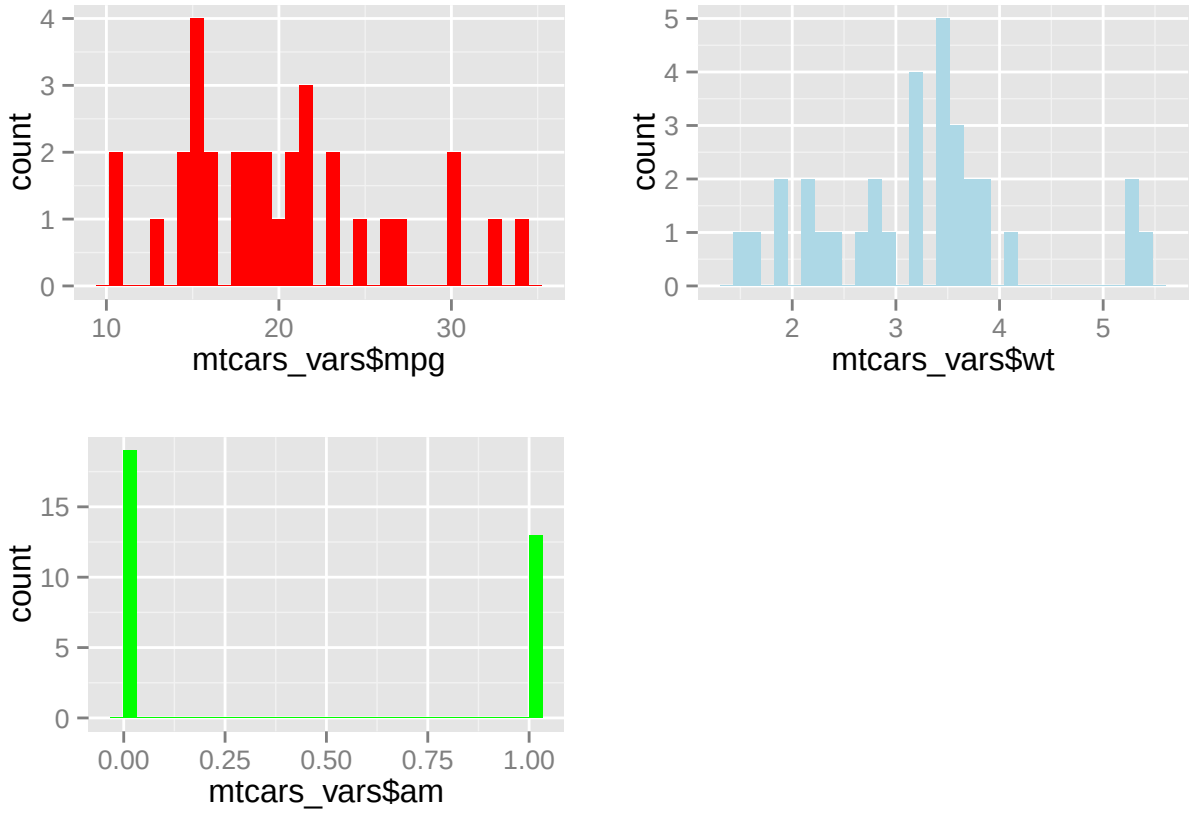
Nothing remarkable here except perhaps in the weight / wt histogram. The Cadillac Fleetwood, Lincoln Continental and Chrysler Imperial are quite a bit heavier than other cars in the dataset.

```
library(ggplot2)
library(gridExtra)
```

```
## Loading required package: grid
```

```
mpg_dist <- qplot(mtcars_vars$mpg, fill = I("red"))
wt_dist <- qplot(mtcars_vars$wt, fill = I("lightblue"))
am_dist <- qplot(mtcars_vars$am, fill = I("green"))
grid.arrange(mpg_dist, wt_dist, am_dist, ncol = 2)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Standard Deviation of MPG by Transmission Type

```
by(mtcars_vars$mpg, mtcars_vars$am, sd)
```

```
## mtcars_vars$am: 0
## [1] 3.833966
## ------------------------------------------------------------
## mtcars_vars$am: 1
## [1] 6.166504
```

Residual Plot

There is a bit of a curve to the residual plot, so that it departs slightly from normality. The residuals for the Chrysler Imperial, Fiat 128, and Toyota Corolla are called out because they exert some influence on the shape of the curve.

```r
mar.orig <- par()$mar  # save the original values
par(mar = c(2, 2, 2, 2))  # set your new values
plot(bestfit, which = c(1:1))
```