Regression Models Assignment: MTCARS

Seth Dimick

12/27/2017

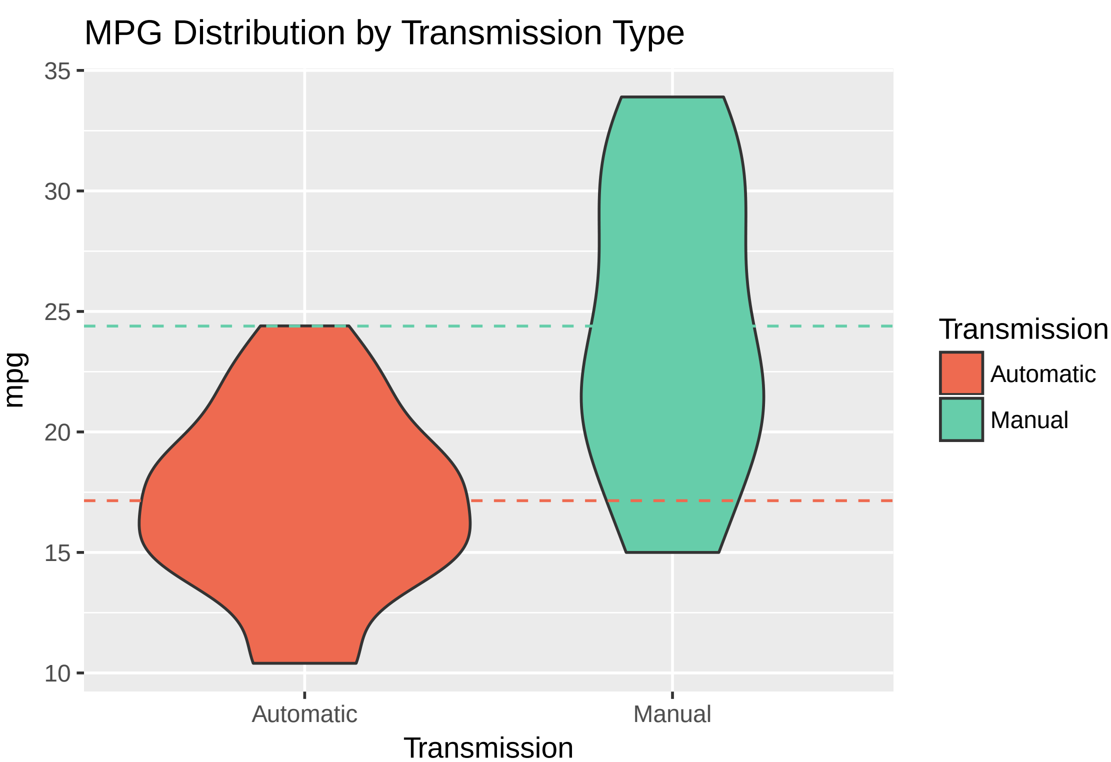
MPG: Automatic or Manual?

If you're looking for a vehicle with good fuel economy will you be buying an automatic or a manual transmission? Here at *Motor Trend* we took a look at the numbers and will halfheartedly tell you to look for a manual. What we found was that on average, cars with manual transmissions get more miles per gallon (MPG), but it's not due to the transmission type. If you were to hold other features of the vehicle constant, the transmission wouldn't make the difference in terms of MPG. However, automatics tend to have other features that make them guzzle more gas, like more cylinders and greater weight.

The Figures

The average MPG for automatics are 17.15 and the average MGP for manuals is 24.39. Take a look at the distribution of MPG for our sample of 32 different makes and models.

```
mtcars2 <- mtcars
mtcars2$Transmission <- ifelse(mtcars2$am==0,"Automatic","Manual")
ggplot(mtcars2, aes(x=Transmission, y=mpg, fill=Transmission)) +
  geom_violin() +
  scale_fill_manual(values = c("coral2","aquamarine3")) +
  geom_hline(yintercept = mean(mtcars[mtcars$am==0,]$mpg), color="coral2", linetype=2) +
  geom_hline(yintercept = mean(mtcars[mtcars$am==1,]$mpg), color="aquamarine3", linetype=2) +
  ggtitle("MPG Distribution by Transmission Type")
```



Without taking other attributes of the cars into consideration, we can see that transmission type is a significant factor in estimating MPG if that is all we know about the vehicle. See that output of a linear model will MPG as the dependent variable, and Transmission as the sole independent variable:

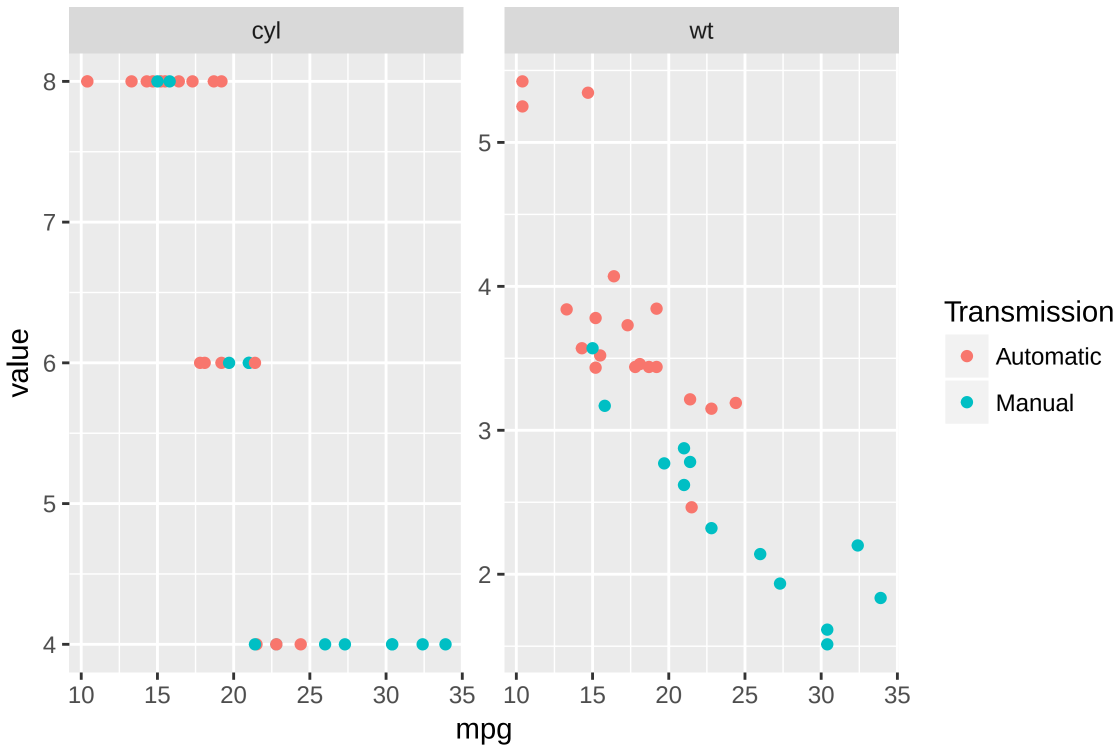
```r
fit1 <- lm(mpg ~ Transmission, data = mtcars2)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ Transmission, data = mtcars2)
##
## 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 ***
## TransmissionManual    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
```

We see our intercept is the mean of MPG for an automatic, and our coefficient for 'TransmissionManual' is the difference between our averages, and returns a significant p-value (less than 0.05). However, we see that the transmission type does a poor job of explaining variation of MPG alone since we have an Adjusted R-Squared of less than 0.35.

Looking at our available data points, we can spot a few attributes that intuitively give us a causal relationship with MPG, and also show a more linear relationship. See the relationship of the number of cylinders (which more cylinders allow more gas to combust) and the weight of the car (greater weight requires more energy to move) with MPG, and also notice how the automatics in our data set fall on the higher ends of these spectrums.

```r
mtcars2 %>%
  select(mpg, cyl, wt, Transmission) %>%
  gather(stat, value, -mpg, -Transmission) %>%
  ggplot(aes(x=mpg, y=value, color=Transmission)) +
  geom_point() +
  facet_wrap(~stat, scales = "free_y")
```



Creating a model with the number of cylinders and weight of the car as our determinants of MPG creates a much better model than the transmission type alone.

```r
fit2 <- lm(mpg ~ cyl + wt, data = mtcars2)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

We see all our terms significant, and a much higher Adjusted R-squared of 0.82. From this model we see that holding weight constant, each additional cylinder decreases MPG by 1.5, and holding the number of cylinders constant, every 1,000 additional pounds decreases the MPG by 3.2.

Once we have these key drivers of cylinders and weight in our model, we find that transmission type does not make a difference in our MPG, holding all else constant. With an ANOVA test of nested models, be see that adding an additional term for Transmission does not produce a significant impact on our model.

```
fit3 <- lm(mpg ~ cyl + wt + Transmission, data = mtcars2)
anova(fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + wt
## Model 2: mpg ~ cyl + wt + Transmission
##   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
```

Apendix

Ridisual Plot for Best Model:

```
plot(predict(fit2),resid(fit2))
```