Regression Models MPG project

Chris Sirico

2/5/2018

Objective

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

Executive Summary

The means of the data show a bias towards higher fuel economies in vehicles with manual transmissions. Those correlations, however, can be explained away by adjusting for cofounders such as weight, number of cylinders and displacement in linear regression. Transmission type does not show a statistically significant predictive effect for fuel economy.

Exploratory Analysis

Using the mtcars dataset, we find automatic vehicles have lower average gas mileage than manuals, 17 mpg to 24 mpg.

Note that 0 = automatic; 1 = manual.

```
cars_manual_mpg <- mtcars %>%
  group_by(am) %>%
  summarize(mpg = mean(mpg))
cars_manual_mpg
```

```
## # A tibble: 2 x 2
##      am      mpg
##   <dbl>    <dbl>
## 1     0 17.14737
## 2     1 24.39231
```

This finding is unscientific, however, because we haven't adjusted for factors like vehicle weight, number of cylinders or number of gears. We also aren't sure whether the finding is significant.

Let's look at average number of cylinders, number of gears, and weight to see if those vary between our automatic and manual data points.

```
cars_manual_cyl <- mtcars %>%
  group_by(am) %>%
  summarize(cyl = mean(cyl))
cars_manual_cyl
```

```
## # A tibble: 2 x 2
##      am      cyl
##   <dbl>    <dbl>
## 1     0 6.947368
## 2     1 5.076923
```

A quick look confirms that automatics tend to have more cylinders. Similar exploration reveals they also have fewer speeds and more weight than manuals. We need to adjust for these factors before drawing a confident conclusion.

Linear Regression and Adjustment

A quick linear regression tells us the same things as our tranmission-typed means.

```
mpglm <- lm(mpg ~ am, data = mtcars)
summary(mpglm)
```

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

Let's also try linear regression adjusting for other factors. Note that as soon as weight is added as a co-predictor, trans type loses statistical significance.

```
mpglm2 <- lm(mpg ~ am + wt, data = mtcars) # trans type quickly loses significance
summary(mpglm2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## am          -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```r
mpglm3 <- lm(mpg ~ am + wt + disp, data = mtcars)

mpglm4 <- lm(mpg ~ am + wt + disp + cyl, data = mtcars)

mpglm5 <- lm(mpg ~ am + wt + disp + cyl + gear, data = mtcars)

anova(mpglm, mpglm2, mpglm3, mpglm4, mpglm5) # adding gear count isn't significant
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + disp
## Model 4: mpg ~ am + wt + disp + cyl
## Model 5: mpg ~ am + wt + disp + cyl + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0849 2.024e-08 ***
## 3     28 246.56  1     31.76  4.5276  0.042999 *
## 4     27 188.43  1     58.13  8.2859  0.007888 **
## 5     26 182.41  1      6.02  0.8582  0.362765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Transmission type doesn't end up being significant in the presence of other mpg predictors for this data set. Fuel economy depends more heavily on weight and engine size / number of cylinders.

Analysis of variance shows significance in adding weight, displacement and number of cylinders as predictors (though not gear count).

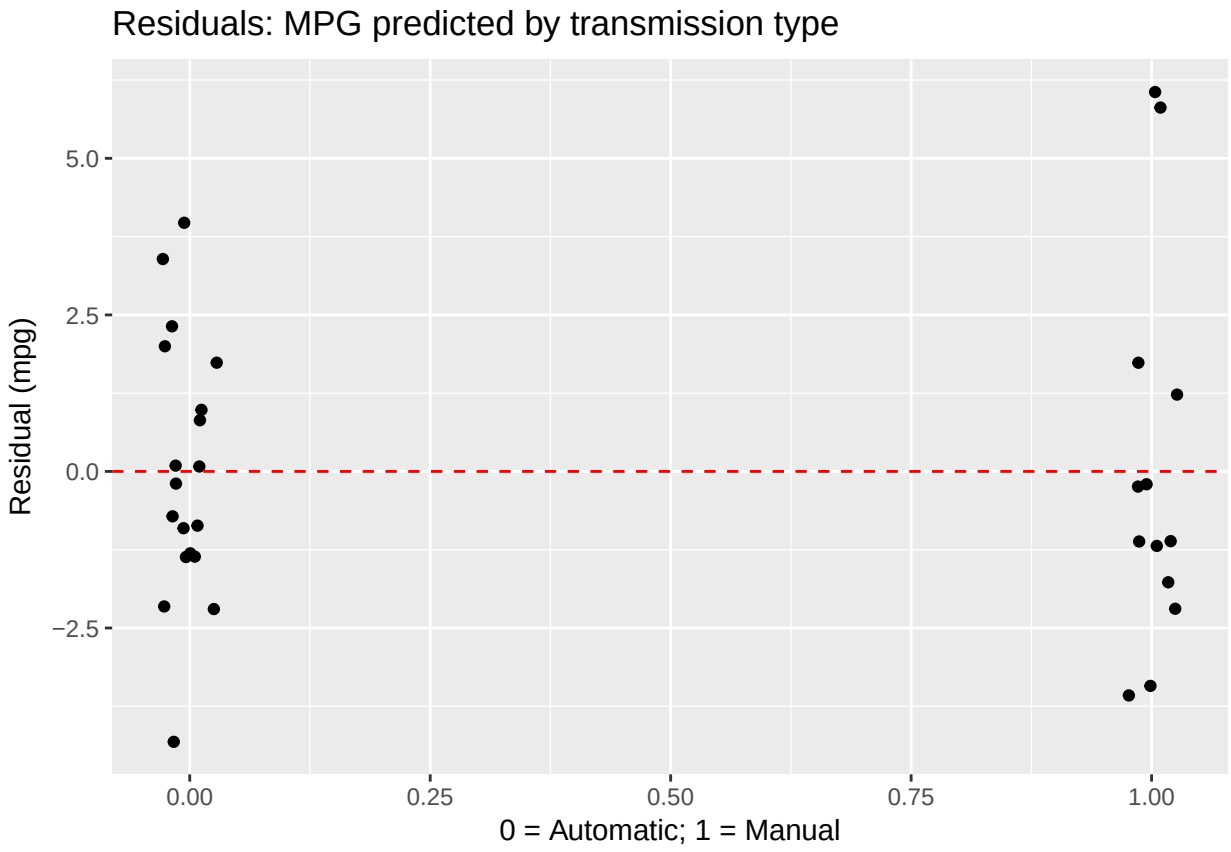
Appendix

Here's a residual plot of MPG by trans type.

```r
data <- tibble(x = mtcars$am, y = summary(mpglm4)$residuals)

# adapted from https://rpubs.com/therimalaya/43190

 ggplot(data = data, aes(x, y)) +
 geom_jitter(width = .03) +
 geom_hline(yintercept=0, col="red", linetype="dashed") +
 ggtitle("Residuals: MPG predicted by transmission type") +
 xlab("0 = Automatic; 1 = Manual") +
 ylab("Residual (mpg)")
```


Residuals: MPG predicted by transmission type
Residual (mpg)
5.0
2.5
0.0
−2.5
0.00
0.25
0.50
0.75
1.00
0 = Automatic; 1 = Manual