Regression Models Final Project

Shelby Bachman

2019-01-27 10:08:39

Summary of results

In this report, I analyzed the relationship between transmission type and MPG in the `mtcars` dataset. I found that, other variables aside, cars with manual transmission had a significantly MPG value than did cars with automatic transmission. A simple linear model with transmission type as a predictor and MPG as the outcome variable confirmed this relationship between transmission type and MPG, but the model did not account for a majority of variance in MPG. I subsequently chose to include additional variables from the `mtcars` dataset (horsepower, weight, displacement, number of cyclinders) as predictors in an extended linear model in an attempt to account for more of the variance in MPG. When taking these variables into account, transmission type no longer had a significant effect on MPG; gross horsepower and weight were the variables related to MPG, with each variable having an inverse relationship with MPG.

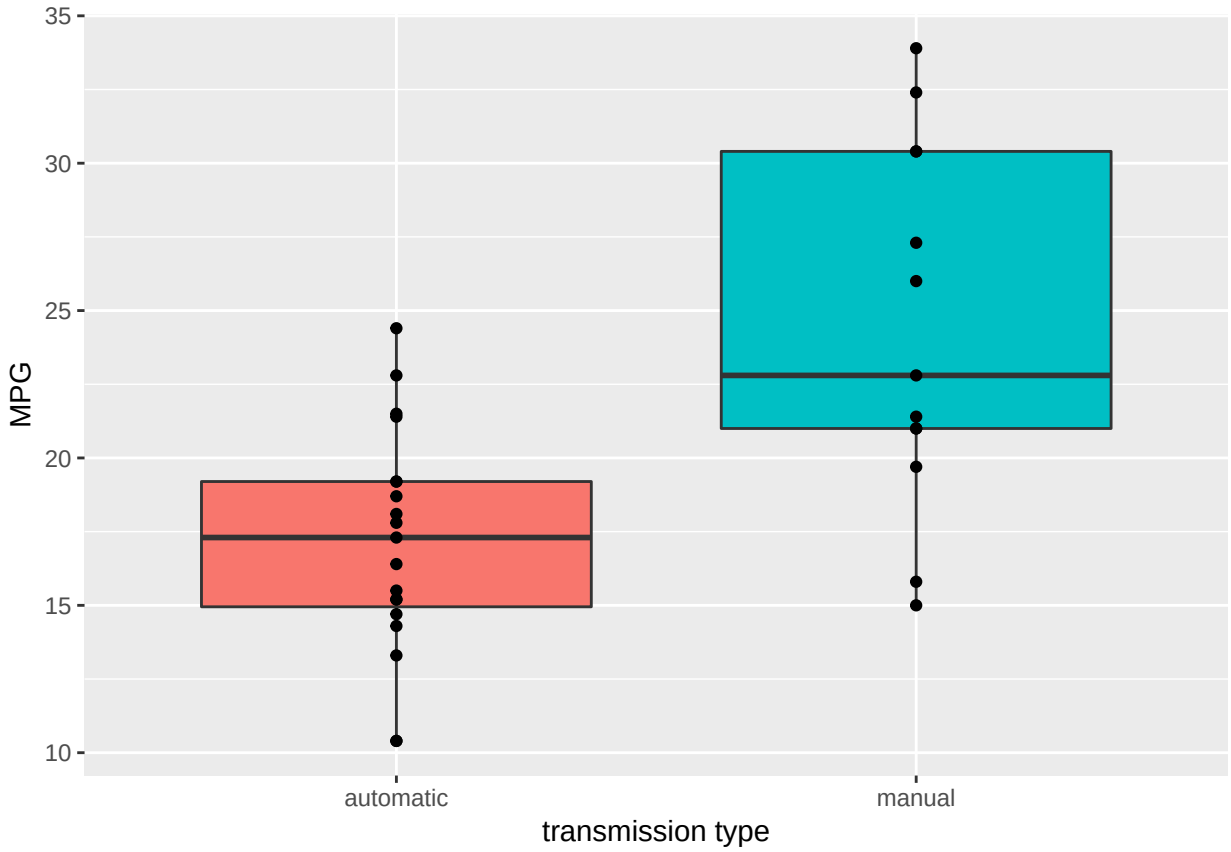
Is an automatic or manual transmission better for MPG?

```
data(mtcars)
```

Before doing any modeling, I will get an overview of the data by plotting. Below, I include a boxplot depicting mpg for each automatic (red) and manual (blue) cars. Dots indicate actual data points, and the horizontal lines through the boxes reflect the median for each group.

```
mtcars <- mtcars %>%
  mutate(am_name = ifelse(am == 0, 'automatic',
                          ifelse(am == 1, 'manual', NA)))
p1 <- ggplot(aes(x = am_name, y = mpg, fill = factor(am_name)), data = mtcars) +
  geom_boxplot(show.legend = FALSE) +
  geom_point(show.legend = FALSE) +
  labs(x = 'transmission type', y = 'MPG')

p1
```



This plot shows that MPG does differ by transmission: cars with manual transmission had a higher 1st quantile, median, and 3rd quantile MPG than did cars with an automatic transmission. We can test whether the MPG of the transmission types is significantly different using a t-test:

```
t.test(mtcars$mpg[mtcars$am==0], mtcars$mpg[mtcars$am==1], paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The results of the t-test indicate that the group difference is statistically signficant; we can reject the null hypothesis and conclude that cars with manual transmission have significantly higher MPG.

Quantify the MPG difference between automatic and manual transmissions

Below, I will quantify the difference in MPG between automatic and manual transmissions using linear regression. I will first fit a linear model using MPG as the output and transmission (factor variable) as the

predictor and interpret the regression coefficients to determine the impact of transmission type on MPG. This initial model will not take into account other variables in the `mtcars` dataset.

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

Inspecting the summary of the model fit, several points are apparent. First, there is a significant effect of transmission type on MPG ($Pr(>|t|) = 0.000285$), which supports the analyses above. Secondly, the intercept estimate tells us the mean MPG value for automatic transmissions (am = 0) according to the model, which was *17.147*. In addition, the slope estimate for transmission type tells us the estimated increase in MPG for manual transmissions (am = 1) relative to automatic transmissions according to the model; this value was *7.245*. Finally, the overall adjusted R-squared value for the model was *0.3385*, indicating that 33.85% of the variance in MPG could be explained by the model.

Using the output above, I will now calculate a 95% confidence interval for the intercept and slope estimates:

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(0.975, df = fit$df) * sumCoef[1, 2]
```

```
## [1] 14.85062 19.44411
```

```
sumCoef[2,1] + c(-1, 1) * qt(0.975, df = fit$df) * sumCoef[2, 2]
```

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

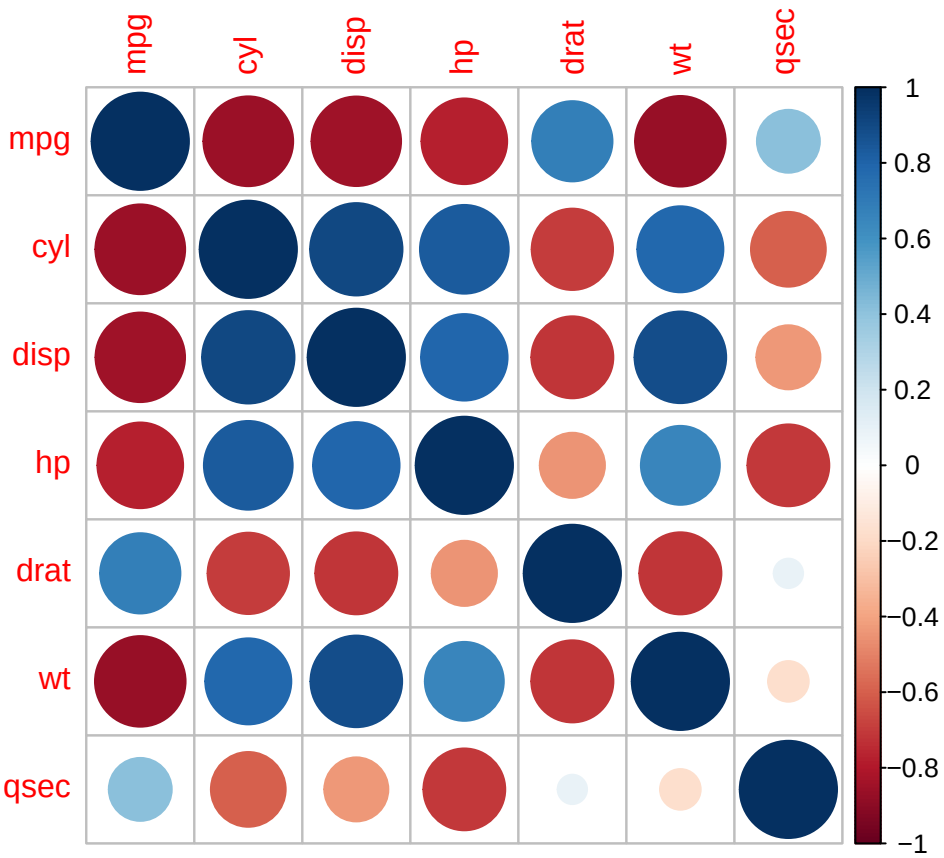
The latter confidence interval (for the slope) indicates that, with 95%, we estimate that moving from automatic to manual transmission will result in an MPG increase of *3.64 to 10.85*.

Appendix: Extended model

Although we found a significant effect of transmission type on MPG, the adjusted R-squared value for the above model suggests that other variables may be useful for accounting for more of the variance in MPG. Below I include a correlation matrix with the areas of the circles reflecting absolute values of the corresponding correlation coefficients. Positive correlations are depicted in blue and negative in red.

```
mtcars_forplot <- mtcars %>%  # choose only numeric variables
  select(-am_name, -vs, -am, -gear, -carb)
```

```
mtcars_cor <- cor(mtcars_forplot)
corrplot(mtcars_cor, method = 'circle')
```



The plot shows that quite a few of the variables in the matrix are strongly correlated with MPG. Next, I will include four of the variables that appear strongly correlated with MPG (cyl, disp, hp, and wt) and include them as predictors in a linear model, along with the already-used transmission type variable.

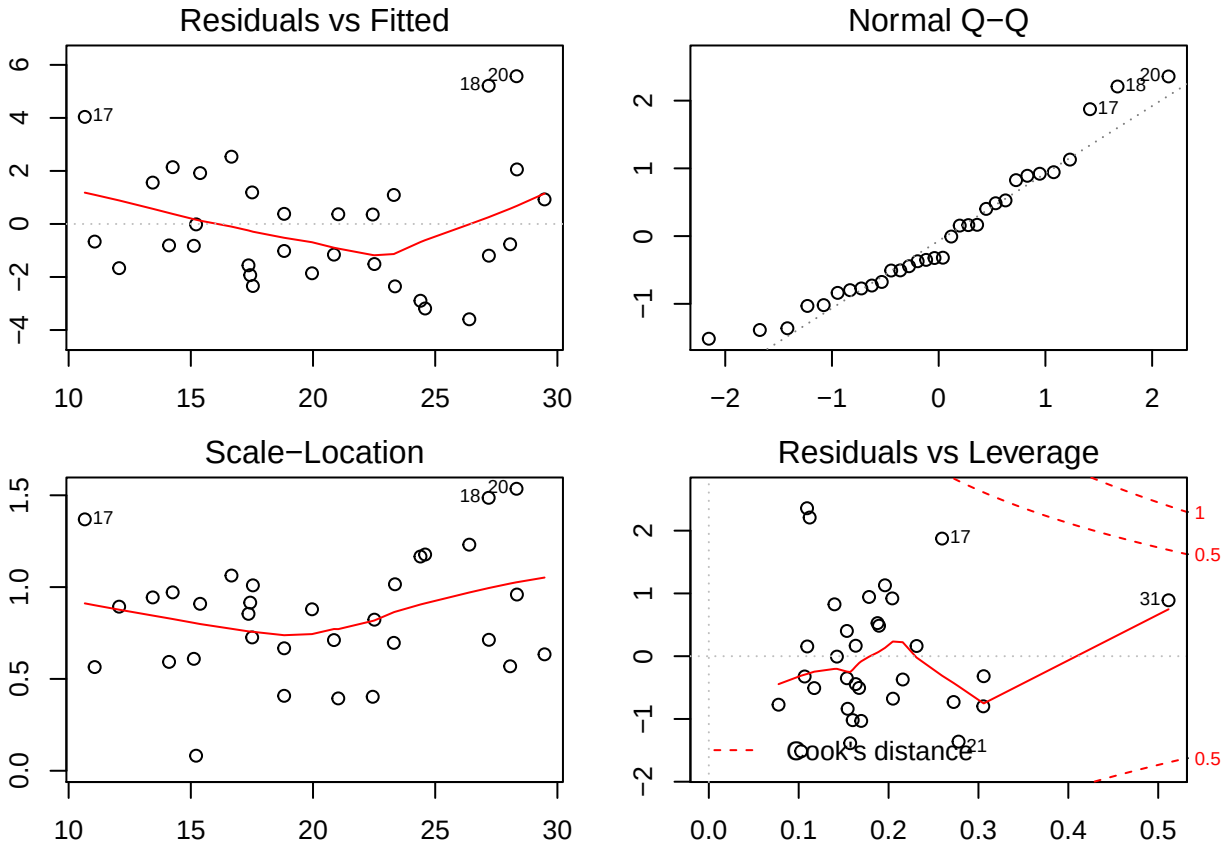
```
fit2 <- lm(mpg ~ factor(am) + cyl + disp + hp + wt, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## factor(am)1  1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

The model summary shows that this model accounts for much more variance in MPG (*Adjusted R-squared = 0.8273*). Critically, in this model, there is no longer a significant effect of transmission type on MPG. Instead, there is a significant effect of both gross horsepower (hp) and weight (wt) on MPG. More specifically, the slope terms for each variable indicate that as both gross horsepower and vehicle weight increase, MPG decreases. Number of cylinders (cyl) and displacement (disp) did not have a significant effect on MPG in this model. Next, I include some diagnostic plots to further inspect the fit of this extended model.

```
par(mfrow=c(2,2))
par(mar = rep(2, 4))
plot(fit2)
```



In the first plot, the residuals are approximately distributed along the horizontal line, but there are several values at each end of the x-axis with higher residual values. As such, since the red line is slightly curved at each end, including a quadratic term in the model may improve the model fit. The second plot indicates that the residuals are approximately normally distributed, with the exception of observations #17, 18, and 20. The third plot suggests that the residuals have approximately uniform variance across the predictor range. The fourth plot of residuals versus leverage shows that observation #31 has high leverage, so in subsequent analyses this observation could be excluded and the effect on the model fit examined.