Regression Models Course Project

Rachel Smith

Executive summary

In this report, we examine the mtcars data set to explore how miles per gallon (mpg) is affected by transmission. The questions addressed by this investigation are as follows:

(1) Is automatic of manual transmission better for mpg?
(2) Can we quantify the mpg difference between automatic and manual transmissions?

```
# import
data("mtcars")

# structure
df_mtcars <- as_tibble(mtcars) %>%
       mutate(am = as.factor(am)) %>%
       mutate(cyl = as.factor(cyl))
```

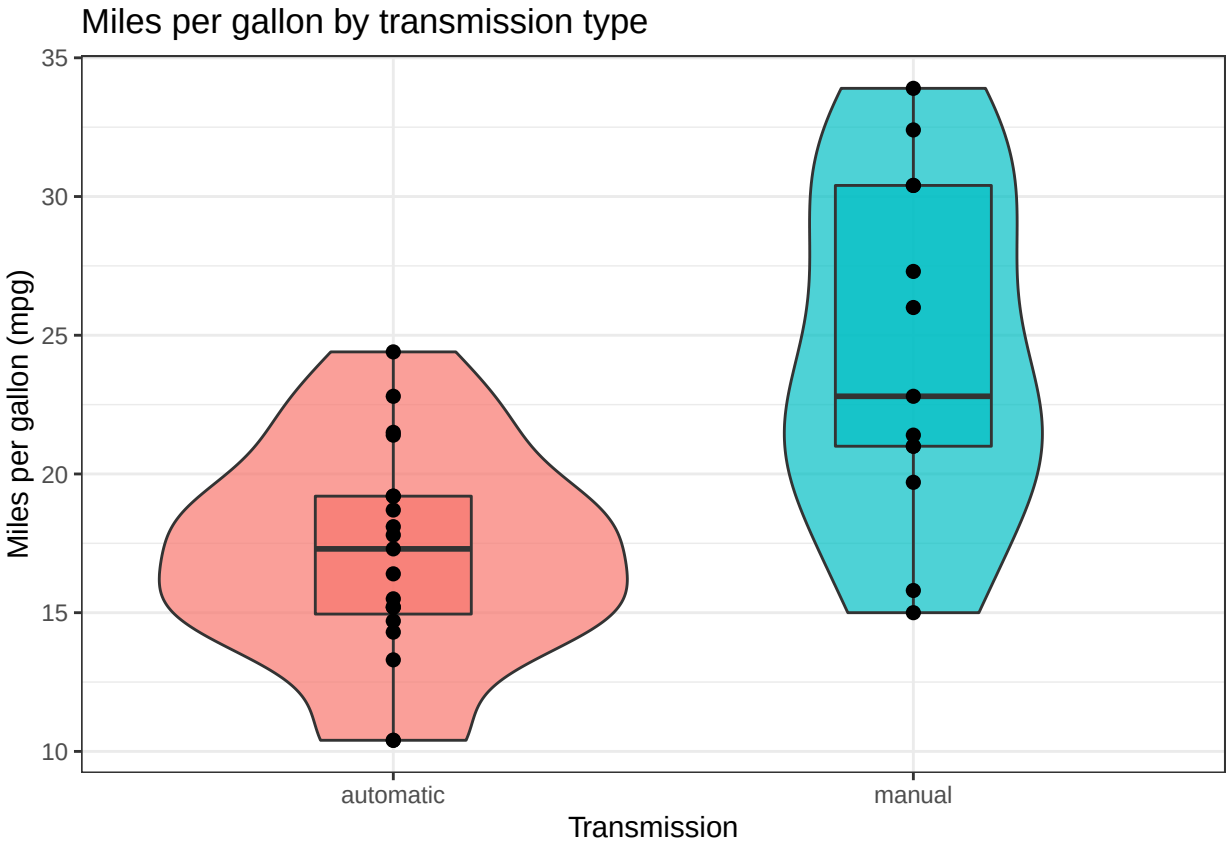
Exploratory analysis

```
df_mtcars %>% ggplot(aes(x = am, y = mpg, fill = am)) +

       geom_violin(alpha = 0.7) +
       geom_boxplot(width = 0.3, alpha = 0.7) +
       geom_point(size = 2) +

       xlab("Transmission") +
       ylab("Miles per gallon (mpg)") +
       ggtitle("Miles per gallon by transmission type") +
       scale_x_discrete(labels = c("0" = "automatic", "1" = "manual")) +

       theme_bw() +
       theme(legend.position = "none")
```



Based on this plot, it looks like manual vehicles have higher miles per gallon, on average. Let's fit some models to quantify that relationship.

Simple linear regression model

```
fit_lm <- lm(mpg ~ am, data = df_mtcars)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df_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 ***
## am1            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
```

```
# automatic
coef(fit_lm)[[1]]
```

```
## [1] 17.14737
```

```
# manual
coef(fit_lm)[[1]] + coef(fit_lm)[[2]]
```

```
## [1] 24.39231
```

Automatic cars (a0) have on average 17.147 mpg, while manual cars have on average 24.392 mpg. This aligns with what we saw in the exploratory analysis.

```
# P-values
summary(fit_lm)$coefficients[,4]
```

```
##  (Intercept)          am1
## 1.133983e-15 2.850207e-04
```

```
# R-squared
summary(fit_lm)$r.squared
```

```
## [1] 0.3597989
```

The p-values are statistically significant ($p < 0.05$), however, the r-squared value for the model is only ~0.36. This indicates that only about a third of the variance in mpg can be attributed to transmission type alone, suggesting that there are other confounding variables we need to include in the model.

Multivariable linear regression model

Our R-squared in the simple linear regression model indicated there are other confounding variables contributing to variance in mpg. Let's run an ANOVA to see what these variables are:

```
aov(mpg ~ ., data = df_mtcars) %>% summary()
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  61.863 2.72e-09 ***
## disp         1   57.6    57.6   8.647  0.00809 **
## hp           1   18.5    18.5   2.776  0.11130
## drat         1   11.9    11.9   1.787  0.19626
## wt           1   55.8    55.8   8.369  0.00900 **
## qsec         1    1.5     1.5   0.229  0.63768
## vs           1    0.3     0.3   0.045  0.83358
## am           1   16.6    16.6   2.485  0.13061
## gear         1    3.8     3.8   0.563  0.46183
## carb         1    1.9     1.9   0.292  0.59485
## Residuals   20  133.3     6.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

cyl, disp, and wt all have p-values < 0.05, suggesting they also contribute to the variance. hp is also very low (less than am), so we can explore including that in the model.

```
fit_multi <- lm(mpg ~ am + cyl + disp + wt + hp, data = df_mtcars)

summary(fit_multi)
```

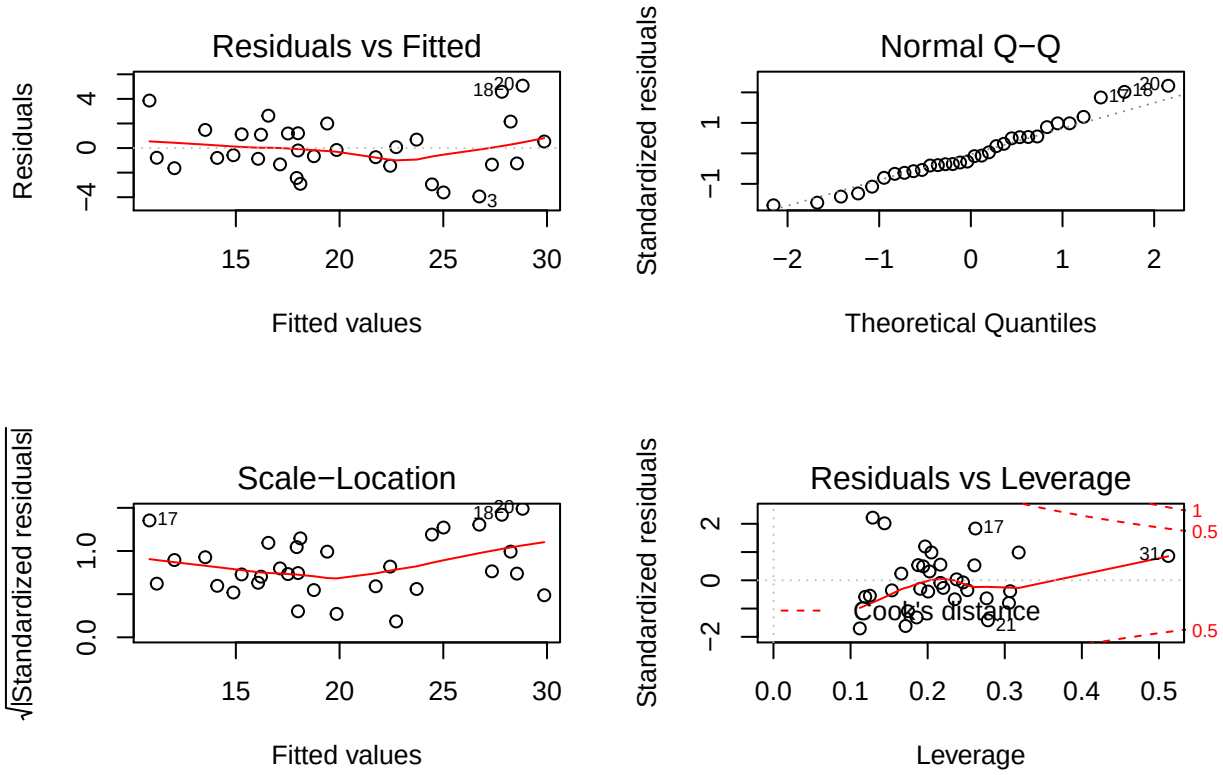
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt + hp, data = df_mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## am1          1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## wt          -2.738695   1.175978  -2.329   0.0282 *
## hp          -0.032480   0.013983  -2.323   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

```
summary(fit_multi)$r.squared
```

```
## [1] 0.8664276
```

The r-squared for this model is ~0.84 - much higher than the previous models. This suggests this model better represents these data.

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



The residuals are homoscedastic (have approximately the same variance) and are approximately normally distributed as per the quantile plot.

Conclusion

Thus, wt, hp, and 6 cyl are confounding variables in the relationship between am and mpg ($p < 0.05$). Holding these confounding variables constant, manual has on average 1.8 mpg higher than automatic