Regression Models Course Project

Sarah Thiesen

10 September 2017

Summary

In this report have taken a look at the mtcars dataset and analyzed whether automatic or manual transmission is better for MPG. After an exploratory data analysis, I performed a linear regression analysis in which "am" (transmission; automatic is coded as 0 and manual as 1) is the independent variable and "mpg" (miles per gallon) is the dependent variable. Other than transmission and MPG, the mtcars dataset includes 9 other variables that might also affect MPG. I selected a model that other than transmission also accounts for weight and gross horsepower. Even after controlling for these variables, transmission has an effect on MPG. Cars with manual transmission generally have higher MPG. However I hesitate to make a recommendation since I don't know if this effect is causal.

Exploratory Data Analysis

First I will perform some exploratory data analysis and create a boxplot to compare the MPG of cars with automatic (am = 0) and manual (am = 1) transmission.

```
data(mtcars)
library(ggplot2); library(broom); library(car); library(dplyr)
mtcars %>% group_by(am) %>% summarize(mean = mean(mpg), sd = sd(mpg)) %>% as.data.frame()
```

```
##   am     mean       sd
## 1  0 17.14737 3.833966
## 2  1 24.39231 6.166504
```

It would seem like cars that have manual transmission have higher mean MPG, but also a higher variation. A boxplot to illustrate this can be found in the appendix. Plotting the relationship between weight and MPG while differentiating between automatic and manual transmission shows that while higher weight is associated with less MPG, the relationship is a bit different for the two groups. This plot can be found in the appendix as well.

Model Selection

In a first regression model that only includes "am" as an independent variable, manual transmission seems to result in higher mpg.

```
options(scipen=999) #Turn off scientific notation
fit1 <- lm(mpg ~ factor(am), data = mtcars)
tidy(fit1)
```

```
##          term  estimate std.error statistic                 p.value
## 1 (Intercept) 17.147368  1.124603 15.247492 0.00000000000000001133983
## 2 factor(am)1  7.244939  1.764422  4.106127 0.00028502074393506755 52
```

Cars with automatic transmission have a mean MPG of 17.147. MPG of cars with manual transmission is higher by 7.245. The effect is highly significant with a p value of 0.0003 (although I'm not sure how

interpretable this is since I don't know if the sample is random). About one third of the variation in MPG can be explained by the variation in transmission ($R^2 = 0.3598$).

Of course, other variables other than transmission could also be affecting MPG. Now I don't know anything about cars, but of the 9 other variables in the mtcars dataset, I imagine that gross horsepower (hp), the weight of the car in 1000 lbs (wt) and the number of cylinders (cyl) could have an effect on MPG. I introduce the three variables gradually in a stepwise regression and create an analysis of variance (anova) table to test whether or not these variables can help explain variance and thus should be included in the model. After some trial and error, I've also decided to include an interaction term between transmission and weight.

```
fit2 <- lm(mpg ~ factor(am) + wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am) + wt + factor(am) * wt, data = mtcars)
fit4 <- lm(mpg ~ factor(am) + wt + factor(am) * wt + hp, data = mtcars)
fit5 <- lm(mpg ~ factor(am) + wt + factor(am) * wt + hp + cyl, data = mtcars)

anovas <- anova(fit1, fit2, fit3, fit4, fit5)
```

The anova table (which can be found in the appendix on page 4) shows that the inclusion of the variables horsepower (hp), weight (wt) and the interaction term results in a significant decrease in variance, but the inclusion of the number of cylinders (cyl) does not. Model 4, which controls for horsepower and weight as well as the interaction between transmission and weight, seems to be the most appropriate model.

Looking at the summary of model 3 (page 4), we can see that cars with manual transmission still have higher MPG. This effect remains significant even after controlling for the other variables ($p = 0.008$). An increase in horsepower ($p = 0.01$) or weight ($p = 0.006$) both result in lower MPG. The interaction term shows that the MPG of cars with manual transmission decreases more as weight increases than that of cars with automatic transmission. R^2 has increased from 0.3598 in the first model to 0.8503 in model 4, which means the new model is a lot better at explaining variation in MPG.

```
data.frame(intercept = c(coef(fit1)[1], coef(fit4)[1]), am = c(coef(fit1)[2],
  coef(fit4)[2]), wt = c(NA, coef(fit4)[3]), hp = c(NA, coef(fit4)[4]), am_wt = c(NA,
  coef(fit4)[5]), row.names = c("fit1", "fit4"))
```

```
##      intercept        am        wt          hp    am_wt
## fit1  17.14737  7.244939        NA          NA       NA
## fit4  30.94733 11.554813 -2.515586 -0.02694935 -3.57791
```

Regression Diagnostics

Diagnostic plots can be found in the appendix. In the residuals vs fitted plot, the residuals are about equally distributed around a vaguely straight line. It seems unlikely that there are non-linear patterns in the data. In the Q-Q plot, the residuals deviate slightly from the line. It might be that the residuals are not normally distributed. The scale-location plot does not show many abnormalities, which indicates that the variance is about equal (homoscedastic). The residuals vs leverage plot shows that the Maserati Bora, and to a lesser extent the Chrysler Imperial, are influential outliers with high leverage.

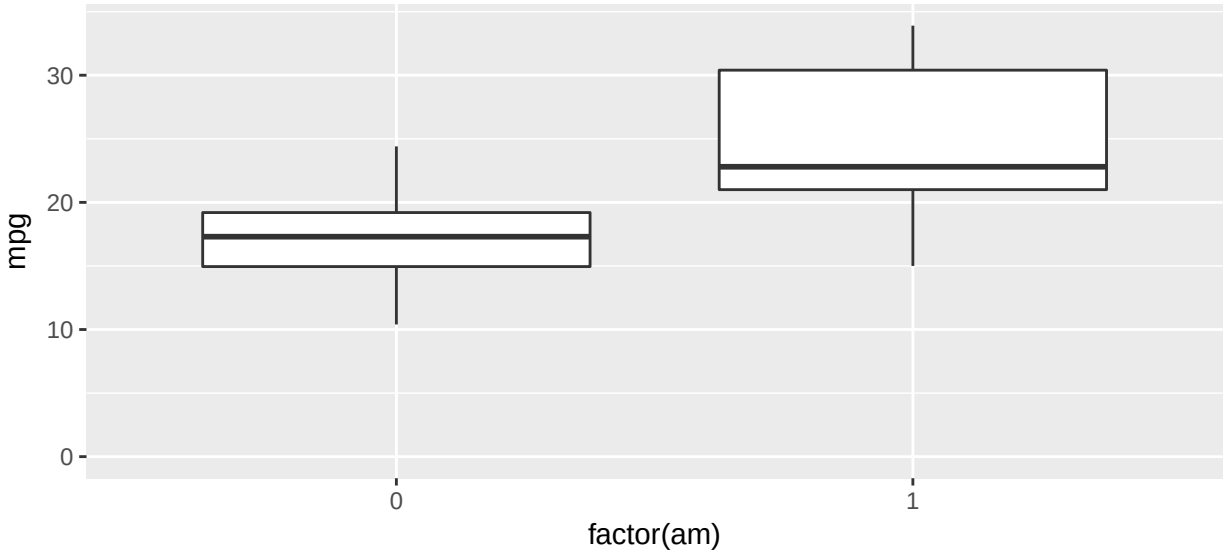
Conclusion

Cars with manual transmission generally have higher MPG, even when controlling for weight and gross horsepower. However, I don't know if this relationship is causaul, so I hesitate to give a general recommendation.

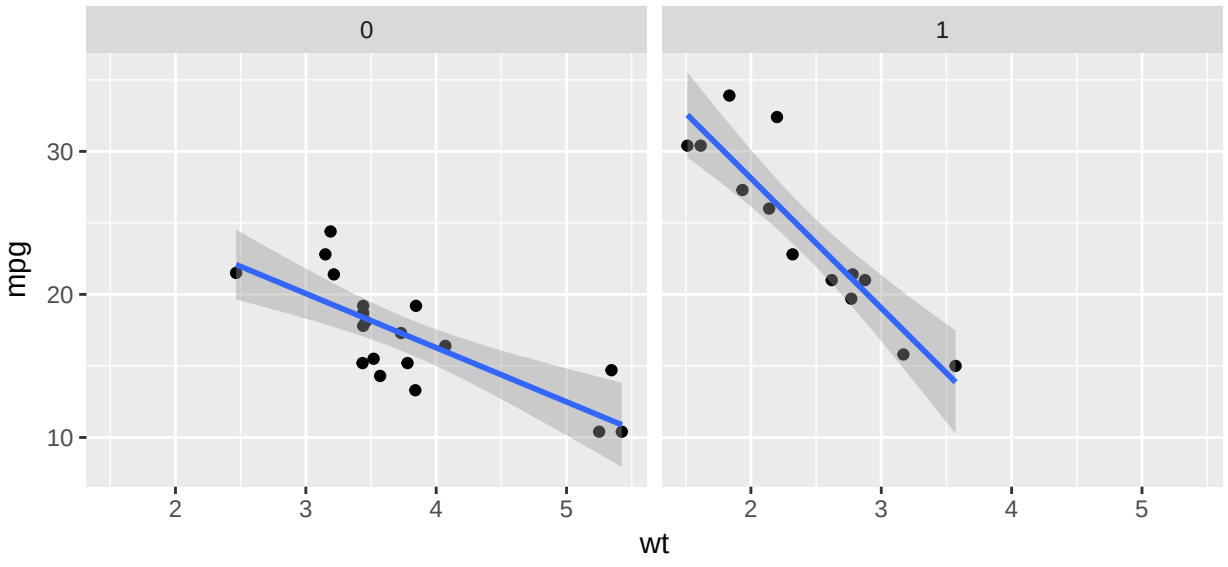
Appendix

Exploratory plots

```
m <- ggplot(mtcars, aes(factor(am), mpg))
m + geom_boxplot() + expand_limits(y = 0)
```




```
g <- ggplot(mtcars, aes(wt, mpg))
g + geom_point() + facet_grid(. ~ factor(am)) + geom_smooth(method = "lm")
```



Anova table

```
anovas
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + factor(am) * wt
## Model 4: mpg ~ factor(am) + wt + factor(am) * wt + hp
## Model 5: mpg ~ factor(am) + wt + factor(am) * wt + hp + cyl
##   Res.Df    RSS Df Sum of Sq       F         Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 85.9190 0.000000001005 ***
## 3     28 188.01  1     90.31 17.5326      0.0002865 ***
## 4     27 146.85  1     41.16  7.9910      0.0089223 **
## 5     26 133.93  1     12.92  2.5075      0.1253926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

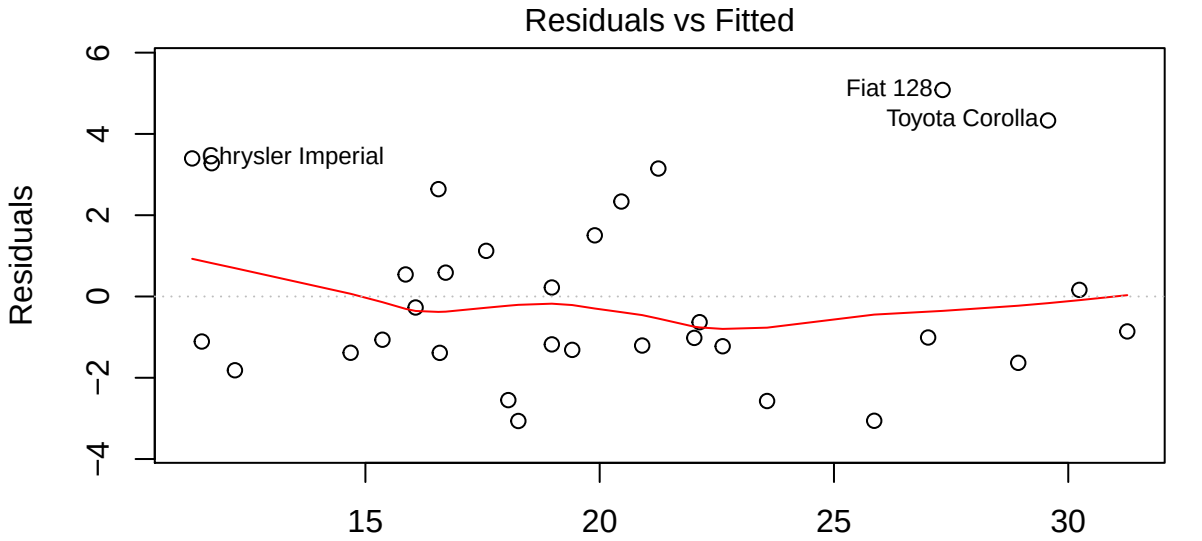
Summary of model 4

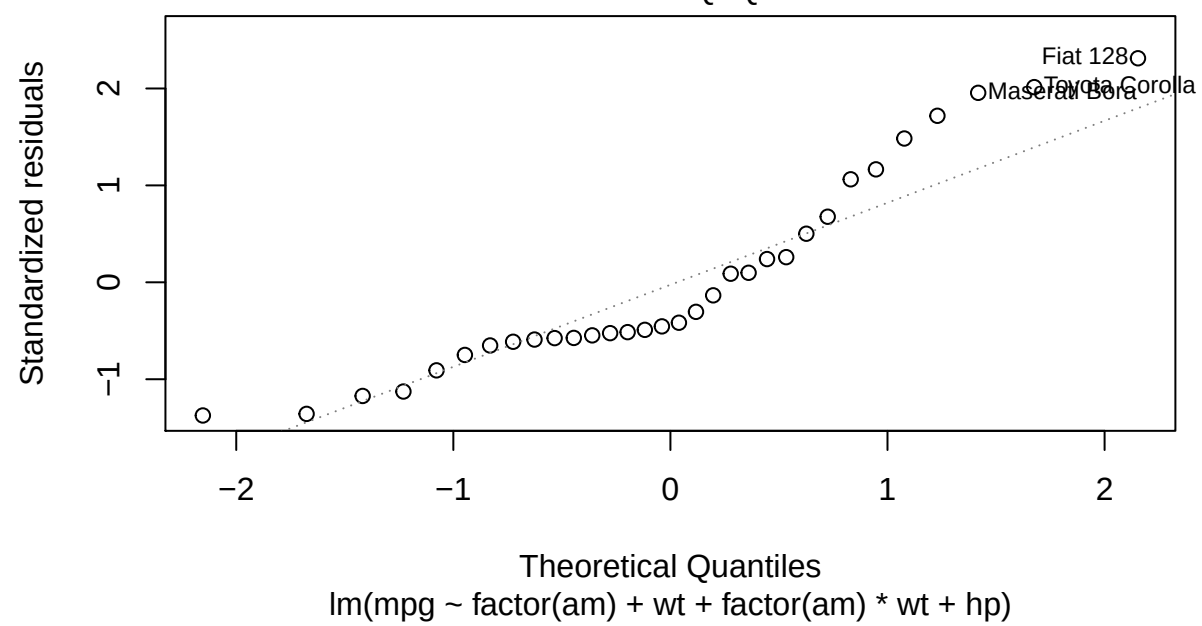
```
summary(fit4)
```

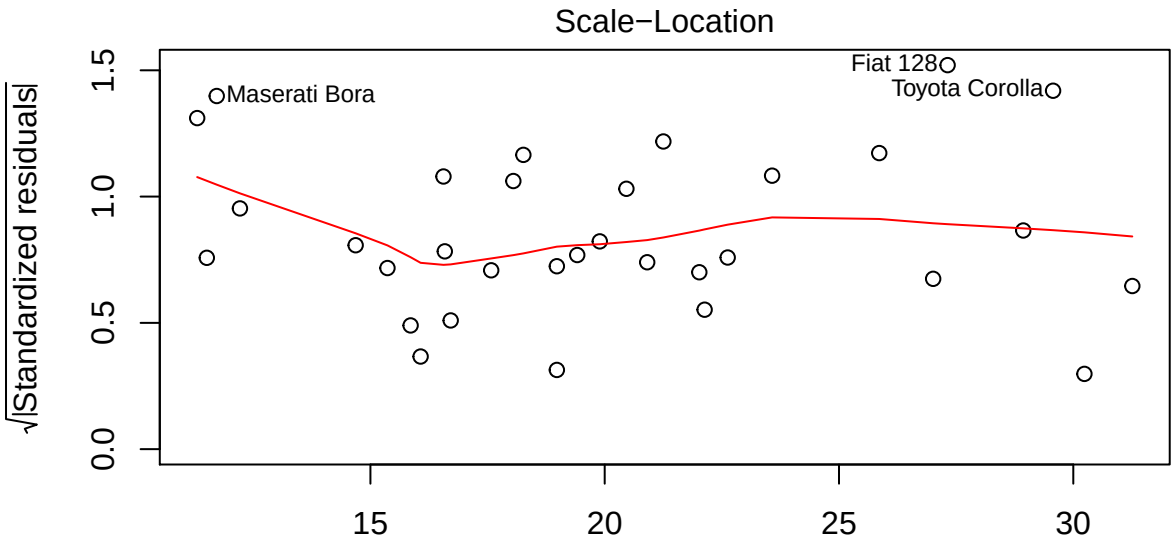
```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + factor(am) * wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0639 -1.3315 -0.9347  1.2180  5.0822
##
## Coefficients:
##                 Estimate Std. Error t value         Pr(>|t|)
## (Intercept)    30.947333   2.723411  11.363 0.00000000000855 ***
## factor(am)1    11.554813   4.023277   2.872          0.00784 **
## wt             -2.515586   0.844497  -2.979          0.00605 **
## hp             -0.026949   0.009796  -2.751          0.01048 *
## factor(am)1:wt -3.577910   1.442796  -2.480          0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.332 on 27 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8503
## F-statistic: 45.01 on 4 and 27 DF,  p-value: 0.00000000001451
```

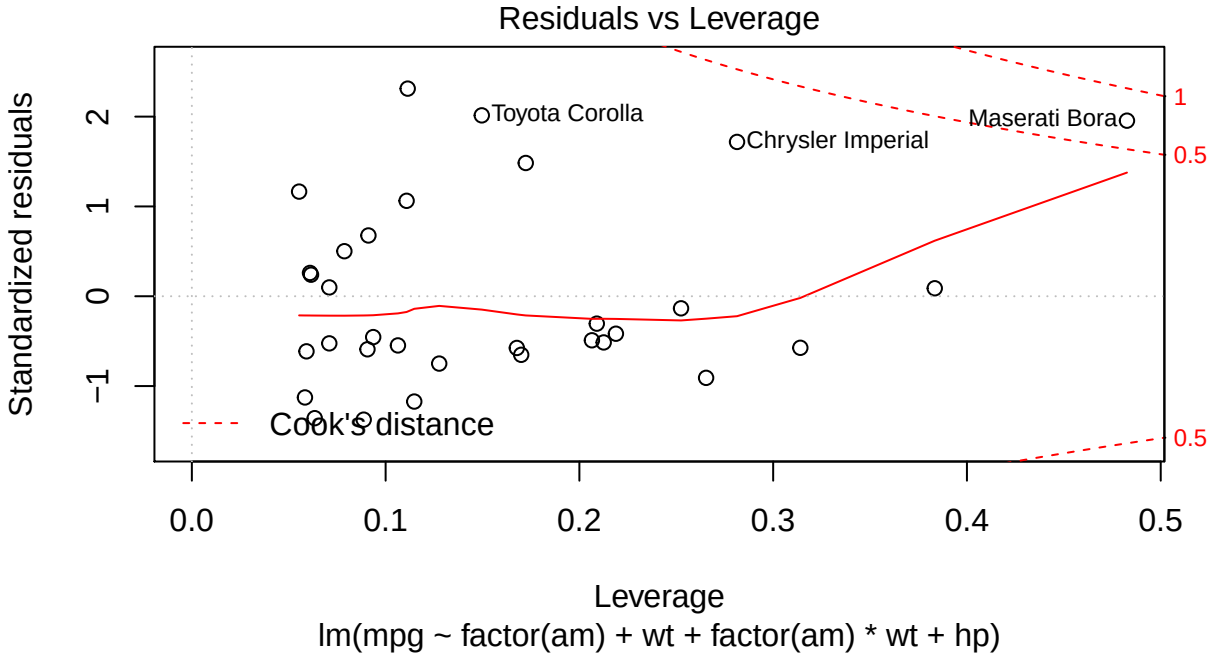
Regression diagnostics

```
plot(fit4)
```






Scale−Location
√|Standardized residuals|
Maserati Bora
Fiat 128
Toyota Corolla
0.0
0.5
1.0
1.5
15
20
25
30
Fitted values
lm(mpg ~ factor(am) + wt + factor(am) * wt + hp)


Residuals vs Leverage
Standardized residuals
Toyota Corolla
Chrysler Imperial
Maserati Bora
Cook's distance
1
0.5
0.5
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(mpg ~ factor(am) + wt + factor(am) * wt + hp)