Homework 5

Goals

- Practice interaction terms.
- Explore the impact of added predictors on R^2.

Directions & Getting Started

- Read the “General homework directions” which can be found on Moodle.
- Load the `ggplot2` and `dplyr` packages at the top of your homework.

```
library(ggplot2)
library(dplyr)
```

Part 1: hearts, numbers, and pets

Allen, Blascovich, Tomaka, and Kelsey (1991) asked a sample of 45 people to perform a stressful task: count backwards by 13’s and 17’s in front of the experimenter while either alone, in the presence of a friend, or in the presence of their pet. To assess stress levels, the researchers recorded the subjects’ mean and maximum heart rates during the experiment:

```
ps <- read.csv("https://www.macalester.edu/~ajohns24/data/PetStress.csv")
```

The data set includes the following variables:

Variable	Meaning
`MaxHeartRate`	maximum observed heartrate in beats per minute (bpm)
`MeanHeartRate`	average heartrate (bpm)
`Condition`	`C` =control (alone), `F` =friend, `P` =pet

1. Our primary goal throughout this section will be to explain the variability in maximum heart rates (response) from subject to subject. Let’s get to know this variable before modeling it!
    a. How many subjects participated in the experiment?
    b. Construct and interpret a univariate visualization of how the maximum heart rates varied among these subjects.

c. Provide a measure of the typical maximum heart rate. Be sure to include units.
d. Provide a measure of the variability among maximum heart rates. Be sure to include units.

2. Let's try to explain some of this variability. To this end, fit the following three models in RStudio.

```
#model MaxHeartRate by MeanHeartRate
mod1 <- lm(MaxHeartRate ~ MeanHeartRate, ps)

#model MaxHeartRate by MeanHeartRate & Conditions WITHOUT interaction
#i.e. assume the relationship between MaxHeartRate and MeanHeartRate is independent
 of Condition
mod2 <- lm(MaxHeartRate ~ MeanHeartRate + Condition, ps)

#model MaxHeartRate by MeanHeartRate & Conditions WITH an interaction
mod3 <- lm(MaxHeartRate ~ MeanHeartRate * Condition, ps)
```

a. Report AND interpret the R^2 value for `mod1`.
b. Report the R^2 values for `mod2` and `mod3`. (No need to interpret.)
c. Notice that `mod3` adds a variable to `mod2` which adds a variable to `mod1`. What pattern do you see in the R^2 values as we add more model terms?

3. Let's take a pause from our R^2 discoveries to explore interactions and `mod3`.
   a. Construct and interpret a visualization of the relationship between `MaxHeartRate` by `MeanHeartRate` and `Condition`. Include a representation of `mod3` on this visualization using `geom_smooth(method="lm")`.
   b. Report the *single* model formula for `mod3` (the one with the interaction term!).
   c. Predict the `MaxHeartRate` of a subject in the pet group who had a `MeanHeartRate` of 80 bpm. Show your work.
   d. From the model formula, construct equations that capture the relationship between each of the three treatment groups. Be sure to give three different equations, each of the form `MaxHeartRate = a + b MeanHeartRate` where *you* specify `a` and `b`.
   e. Interpret the coefficient on the `MeanHeartRate:ConditionP` interaction term in a contextually meaningful way.
   f. Given your visualization and model equations, do you think that there's a significant interaction between `MeanHeartRate` and `Condition`? Explain. (NOTE: We'll have a rigorous way to address such questions later in the semester!)

Part 2: R^2 Cautions!!

4. Consider another variable in the `ps` data set. `YearUnion` gives the year in which each subject's home state entered the union (United States).
   a. Construct and interpret a visualization of the relationship between `MaxHeartRate` and `YearUnion`. Be sure to comment on the strength of this relationship!
   b. Intuitively (and as evidenced by your plot above), I think we'd agree that there's no real relationship between `YearUnion` and `MaxHeartRate`. That said, let's add it to our model of maximum heart rates using the code below. How does the R^2 from `mod4` compare to `mod3`? Does it decrease or

stay the same or increase?

```
mod4 <- lm(MaxHeartRate ~ MeanHeartRate*Condition + YearUnion, ps)
```

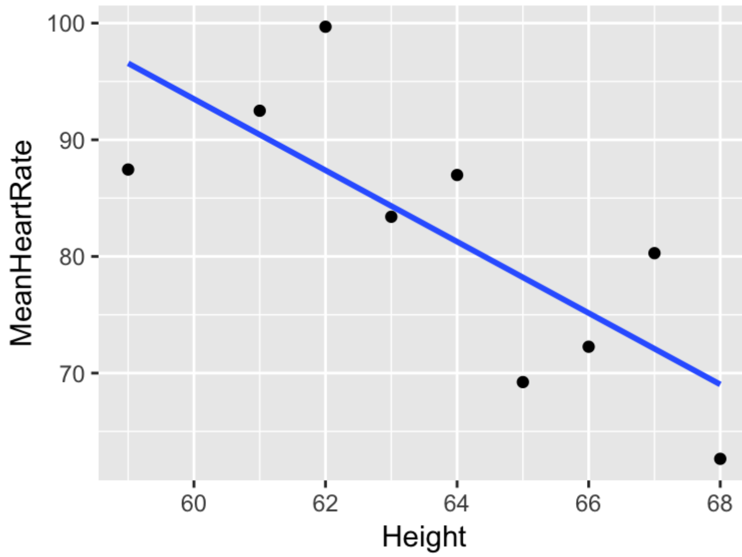
c. In light of this observation, explain what caution we should take in using R^2 to measure model quality.

5. Let's narrow our focus to just 9 of our subjects, each of a different height. Use the exact code below to select these subjects. (If you don't do this, you won't be able to complete the exercise below!)

```
#be sure to set the random number seed so that you get the same random jitter!
set.seed(2000)
new_ps <- ps %>%
    group_by(Height) %>%
    sample_n(1)
```

To begin, consider the model of `MaxHeartRate` and `Height` alone:

```
controlMod0 <- lm(MaxHeartRate ~ Height, new_ps)
summary(controlMod0)
##
## Call:
## lm(formula = MaxHeartRate ~ Height, data = new_ps)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.50 -12.05   3.60   6.85  18.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  447.750    109.662   4.083  0.00467 **
## Height        -5.450      1.715  -3.178  0.01553 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.23 on 7 degrees of freedom
## Multiple R-squared:  0.5907, Adjusted R-squared:  0.5322
## F-statistic:  10.1 on 1 and 7 DF,  p-value: 0.01553
ggplot(new_ps, aes(y=MeanHeartRate, x=Height)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```



Notice that the R^2 value for this relationship is 0.5907. Let's try to improve our model by adding some new predictors (polynomial terms) to our model!

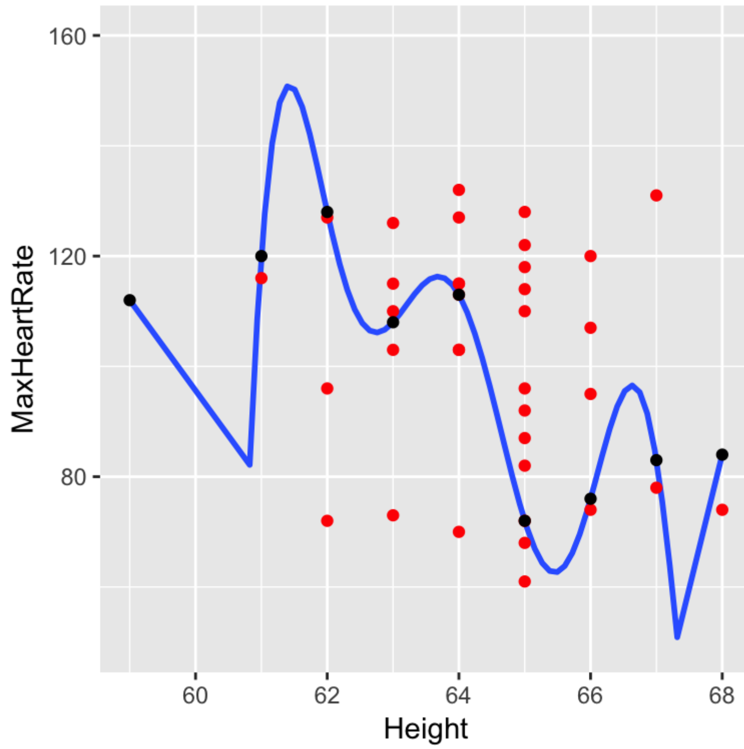
a. Consider `controlMod1` , the model of `MaxHeartRate` by `Height` AND `Height` ^{2}, a quadratic term for `Height` . Report the R^2 value for this model and visualize the model using the code below.

```
controlMod1 <- lm(MaxHeartRate ~ poly(Height,2), new_ps)
summary(controlMod1)
ggplot(new_ps, aes(y=MaxHeartRate, x=Height)) +
  geom_point() +
  geom_smooth(method="lm", formula=y~poly(x, 2), se=FALSE) +
  lims(y=c(50,160))
```

b. Consider `controlMod5` , the model of `MeanHeartRate` by `Height` AND `Height` ^{2} AND … AND `Height` ^{5}. Adapt the code above to fit this model, report its R^2, and construct a visualization of this model. HINT: use `poly(Height,5)` and `poly(x,5)` .

c. I want more!! I'm not satisfied with this R^2 value. Let's consider `controlMod8` , the model of `MaxHeartRate` by `Height` AND `Height` ^{2} AND … AND `Height` ^{8}. Adapt the code above to fit this model, report its R^2, and construct a visualization of this model.

d. OK. We've learned that it's always *possible* to get a perfect $R^2 = 1$. However, the model in part c demonstrates the drawbacks of **overfitting** a model to our sample data. Comment on the following:
   - How easy is it to interpret this model?
   - How well does this model capture the general trend of the relationship between `MaxHeartRate` and `Height` ?
   - How well does this model generalize to the subjects that were not included in the `new_ps` data (shown in red below):


160
120
80
MaxHeartRate
60
62
64
66
68
Height