Stat 414 - Day 5
Multiple Explanatory Variables

Last Time:

- Effect vs. Indicator coding
  - You can change the coding in R using contrasts, e.g.,
    lm(Heart ~ Region, data = PaceData, contrasts = list(Region="contr.sum"))
    lm(Heart ~ Region, data = PaceData, contrasts =list(Region="contr.treatment", base="South"))
  - You may also be able to change the reference group,e.g.,
    Region2 = relevel(factor(PaceDta$Region), ref="South")
  - The multilevel models we will run later in the course tend to use effect coding.
- How you code the categorical variable does not impact overall significance of the variable (just which comparisons are 'built into model')

Example 1: Reconsider the Pace of Life data.
Here is a better link to the original article [Levine (1990)]
(https://www.jstor.org/stable/29774181)

The response variable was a city-wide age-adjusted death rate from ischemic heart disease for the year. See RMarkdown file for more details on the variables.

(a) Which of these variables do you think is a better (or worse) indicator of pace of life?

(b) What do you learn from the scatterplot matrix?

```
pairs(PaceData[,2:6], pch = 19, lower.panel=NULL)
```

(c) There is another interesting feature to the pace variables that you might not have expected… what is going on/why do you think the data is set up this way?

(d) Does one of these variables seem more important than the others in predicting heart disease rate?

```
summary(model1 <- lm(data = PaceData, Heart ~ Walk + Talk + Bank + Watch))
```

One type of "feature scaling" that is often recommended when you want to compare the impact of different variables is to standardize the variables (i.e., subtract the mean and divide by the standard deviation)

(e) Why might standardizing variables be advantageous? Is that what was done here? How can you tell?

Another type of feature scaling is "min-max scaling" (i.e., subtract the min and divide by the range).

(f) What are the range of values after this rescaling? Why might that be advantageous? Is that what was done here? How can you tell? What do you think was done and why?

It certainly doesn't hurt to standardize all quantitative variables.

```
PaceDataZ <- PaceData
PaceDataZ[3 : 6] <- as.data.frame(scale(PaceData[3 : 6]))
summary(model2 <-lm(data = PaceDataZ, Heart ~ Walk + Talk + Bank + Watch))
```

(g) How has this changed the relative importance of the variables? How do we interpret the intercept?

The most common method to check for linear associations among the explanatory variables is variance inflation factors. The car package quickly gives us VIF values.

```
#install.packages("car")
car::vif(model1); car::vif(model2)
```

(h) Does standardizing impact the VIF values? What else might we do if we thought multicollinearity was a problem here?

Example 2 From Day 3 handout

Multicollinearity often arises with quadratic models. Centering or Standardizing can reduce multicollinearity between "product terms":

```
centeredyear = KYDerby23$Year - mean(KYDerby23$Year)
model2b = lm(speed ~ centeredyear + I(centeredyear^2))
car::vif(model2b)
```

(a) Did we improve multicollinearity?

Not sure why these $year$ and $year^2$ are no longer collinear?

```
plot(centeredyear ~ I(centeredyear^2))
```

All we have done is move the curve in the quadratic relationship to be in the middle of our x-space. This is fine, even beneficial, having strongly related x-variables just causes problems with they "line up."

(b) How do we interpret the intercept of this model with both variables centered?

Notes

- See Day 3 notes on benefits of centering

For next time Partial F-tests