Biostatistics 201a - Lecture 19 - 11/7/11

Contents:

- Correlation / causation current event
- Interaction basics

pages = 11

11/7/11
Lecture 19

Interactions:

- When we have multicollinearity we look at the joint effect of our X's, and they may share some of the <u>same</u> <u>information</u> about Y.
- An <u>interaction</u> occurs when the relationship between Y and a predictor, X_1, <u>depends</u> on the value of another predictor, X_2, and vice versa - so X_1 and X_2 have a combined effect on Y.
* You do <u>not</u> need to have X_1 and X_2 be correlated with each other to have an interaction - in fact the more correlated they are the less likely there is to be an interaction.

Mathematically an interaction is just multiplication. If I want to know if there's an interaction between X_1 and X_2 in their effect on Y I create a new variable

$$X_3 = X_{1,2} = X_1 \cdot X_2$$

Example: Simplified version of jogging problem

Y = exercising heart rate

X_1 = time run in minutes

$X_2 = \begin{cases} 1 & \text{if exercises regularly} \\ 0 & \text{if don't} \end{cases}$

Interaction

$$X_{1,2} = X_1 \cdot X_2 = \begin{cases} X_1 & \text{if exercises} \\ 0 & \text{if not} \end{cases}$$

Let's suppose that we fit this model and we get the following estimated regression equation

$$\hat{Y} = 80 + 1.2X_1 - 15X_2 - .2X_{1,2}$$

We're thinking, intuitively, that the effect of running on $\dot{V}O_2$ rate depends on how fit you are.

Note: we can't interpret the coefficient of the interaction in the usual way since it's impossible to change $X_{1,2}$ while holding both X_1 and X_2 fixed. So what we have to do is interpret X_1 and X_2 as a pair.

Let's start with $X_2 = 0$ - people who don't exercise regularly:

$$\hat{y} = 80 + 1.2X_1 - 15(0) - .2(X_1)(0)$$

$$= 80 + 1.2X_1$$

This gives relationship between time run on $\dot{V}O_2$ rate in non-ex. group.

$b_0 = 80$ = resting $\dot{V}O_2$ rate in the non-exercise group

$b_1 = 1.2$ = change in $\dot{V}O_2$ rate per extra minute run in the non-exercise group.

for people who do exercise $X_2 = 1$

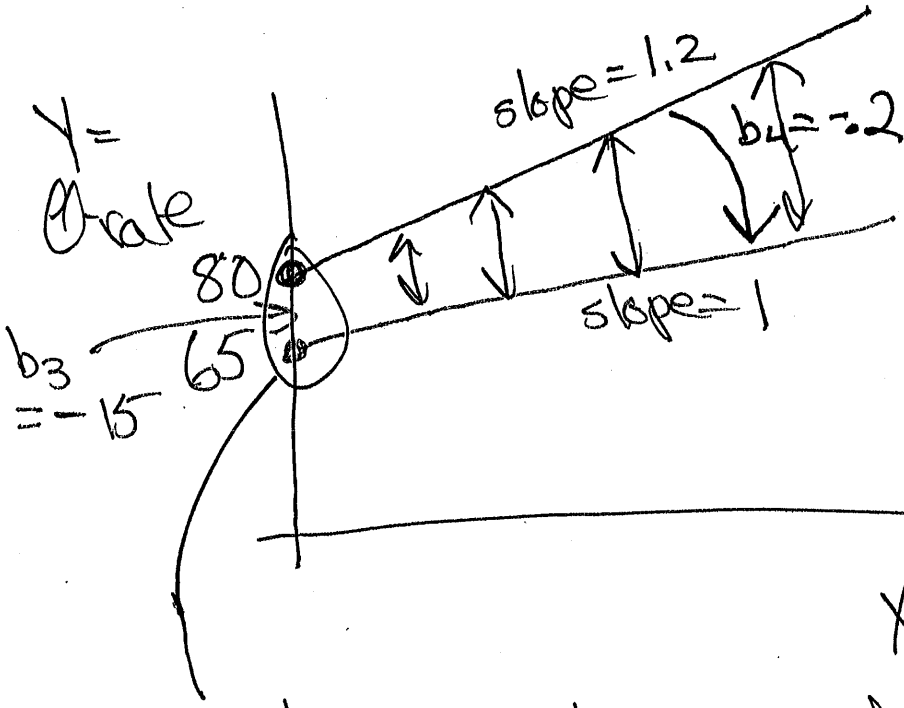
$$\hat{Y} = 80 + 1.2X_1 - 15(1) - .2(X_1)(1)$$

$$= 65 + 1 \cdot X_1$$

This gives the relationship between time run and rate in fit group

$b_2 = -15$ = difference in resting rate between people who exercise regularly and people who don't

$b_3 = -.2$ = difference in slope/rate of rate change, between people who exercise regularly and those who don't

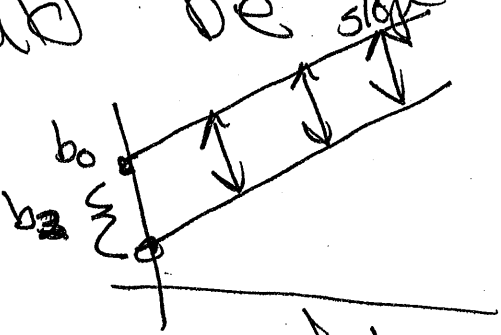


don't exercise
$\hat{Y} = 80 + 1.2X_1$

exercise regularly
$\hat{Y} = 65 + X_1$

X_1 = time run

We have both different intercepts and different slopes. If we didn't have the interaction term then the lines would be parallel



Why not just split the data set up and fit seperate models for each group?

reason to keep together — lowers power - to use each model - you only get half the points in stability of estimates - similarly reduces

reason to split. but we're making assumptions about errors in models being same after adjusting for $X_1, X_2, X_{1,2}$

- Having both variables and the interaction in the same model we can directly test whether the interaction is significant.

$H_0: \beta_3 = 0$ - no interaction between X_1 and X_2; i.e. relationship between time run and O_2 rate doesn't depend on fitness

$H_A: \beta_3 \neq 0$ - there is an interaction; the relationship between time run and O_2 rate is different for people who are fit and people who aren't.

H_0: two parallel lines for the groups with possibly different intercepts

H_A: different slopes as well as different intercepts.

We could also look at this interaction in terms of how the difference between fit and not fit people varies as a function of time run - we expect the gap to get bigger the longer they run!

Now we fix time run, X_1^*, and look at fit - unfit (instead of fixing X_2, fitness level, and looking at the slope).

Before we saw that for fit people

$$\hat{Y}_{fit} = 65 + X_1^*$$

for ~~couch potatoes~~ people

$$\hat{Y}_{don't\ ex} = 80 + 1.2X_1^*$$

$$\hat{Y}_{don't\ ex} - \hat{Y}_{do\ ex} = (80 + 1.2X_1^*) - (65 + X_1^*)$$

$$= 15 + .2X_1^*$$

$-b_2$ $-b_3$

If no time is spent running there's a difference of 15 beats per minute (diff. in resting HR rate; non-exercisers are higher)

AND for every extra minute run that gap grows by .2 bpm

or for extra 5 minutes run gap grows by 1 bpm

time run	gap
$X_i^* = 0$	$\longleftrightarrow$ 15 bpm
$X_i^* = 5$	$\longleftrightarrow$ 16 bpm
$X_i^* = 10$	$\longleftrightarrow$ 17 bpm
⋮	

Note. We have two interpretations of our interaction:

(1) Exercising HR rate goes up faster w/ length of run if you don't exercise regularly

(2) The difference in HR-rate between people who do and do not exercise increases as the length of run increases.

Recap:

Coefficients in model:

$b_0 = 80$ - resting $\dot{V}O_2$ rate of non-exercisers

$b_1 = 1.2$ - $\dot{V}O_2$ rate vs time slope for non-exercisers

$b_2 = -15$ = difference in resting $\dot{V}O_2$ rate between exercisers and non exercisers

$b_3 = -.2$ = difference in $\dot{V}O_2$ rate vs time slope between groups.

Note: You can have interactions that consist of 2 indicators or 2 continuous variables also. We'll look at that next time

Sequence of events is often used as an argument for causality - but this doesn't have to be correct!