Biostatistics 201a - Lecture 20 - 11/9/11

Contents:

- More on Interactions
- Power transformations and the Hierarchical Principle

pages = 10

11/8/11
Lecture 21

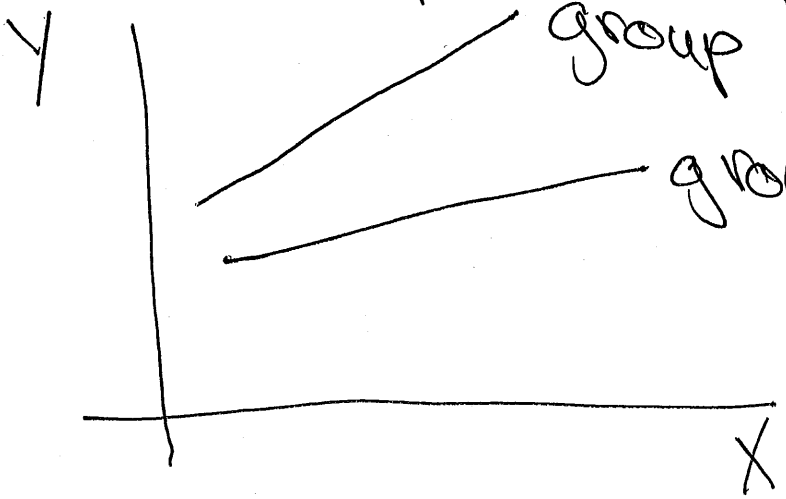
Announcements:

office hours: today 1-2
Tommorrow: cancelled but will be reachable by e-mail or on friday.
HW 5 Due Monday the 14th

More on Interactions:

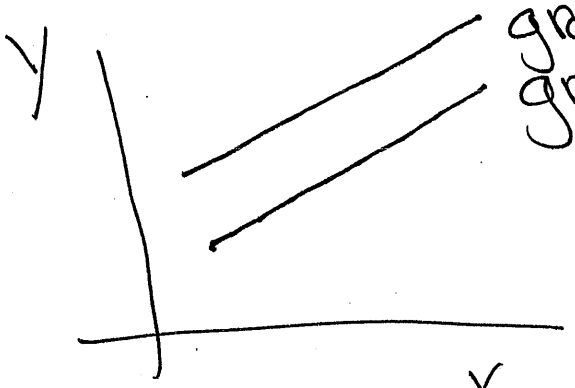
Last time we looked at the case for one continuous variable and one categorical variable

- Interaction in this cases allows you fit separate lines (i.e. possibly different slopes and intercepts) for each group



rate of change of Y with X_1 depends on group
OR difference in Y between groups depends on X_1

- If there is no interaction then the lines are parallel so there's just a shift or different intercept.



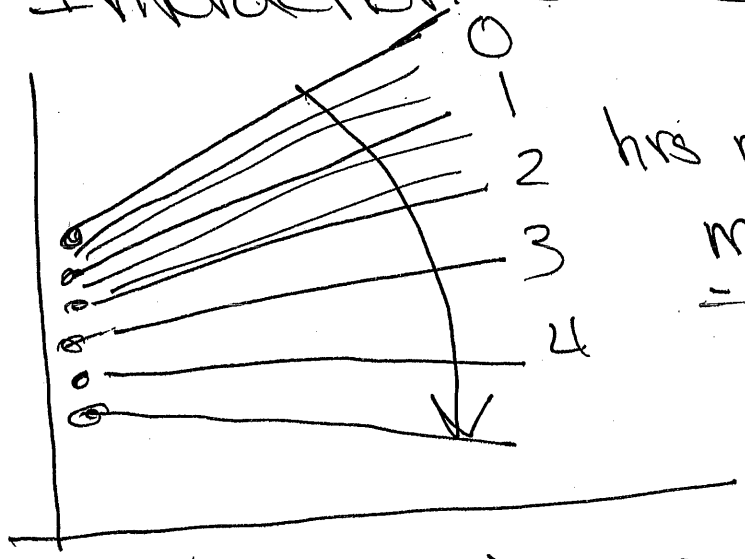
There's a special term for this - we say that the effects of X_1 and X_2 on Y are additive. (i.e. can be considered separately.)

What happens if we have either two continuous predictors, a categorical predictors, or a continuous predictor and >2 categories?

2 continuous variables:

Y = exercising @ rate
X_1 = time run (in minutes)
X_2 = # of hours per week exercises

Interaction as before is $X_1 * X_2$



hrs per week

more exercise = lower intercept and less steep slope

X_1 = time run

Way this is usually displayed since there are an infinite # of possible lines is to pick a few "useful" values of X_2 to draw lines for — e.g. no exercise, low, moderate, high.

What if you have other variables in model? Interpretations for those variables are as usual. If you want to plot the interaction, plug average values for other variables. (changes intercepts)

2 indicator variables

Let $X_1 = \begin{cases} 1 & \text{if person ran} \\ 0 & \text{if not} \end{cases}$

Let $X_2 = \begin{cases} 1 & \text{exercise regularly} \\ 0 & \text{if not} \end{cases}$

(4 possible combinations - like an ANOVA)

Let $X_1 \cdot X_2 = \text{interaction} = \begin{cases} 1 & \text{if ran \& fit} \\ 0 & \text{else} \end{cases}$

What the interaction does in this case is to allow each of the 4 combinations to have an arbitrary mean. With no interaction, the ran & fit mean has to be based on the sum of the ran and fitness individual effects

$$\hat{Y} = 80 + 36X_1 - 15X_2 - 6X_1 \cdot X_2$$

non-exerciser, no run: $\hat{Y} = b_0 = 80$

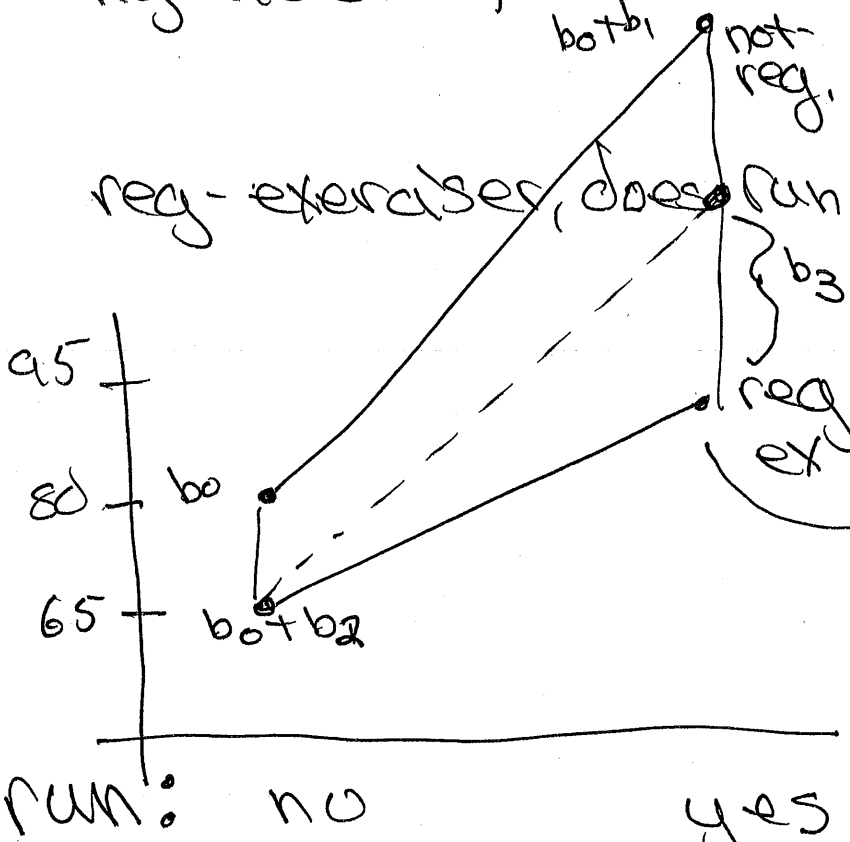
non-exerciser, does run: $\hat{Y} = b_0 + b_1 = 80 + 36 = 116$

reg-exerciser, no run: $\hat{Y} = b_0 + b_2 = 80 - 15 = 65$

reg-exerciser, does run: $\hat{Y} = b_0 + b_1 + b_2 + b_3$

$= 80 + 36 - 15 - 6$

$= 95$



One final case - what if you have a categorical variable with more than two categories? That just adds more interaction terms - one term for each indicator you use to describe your categorial variable.

One last comment: You'll often see the terms "moderator" or "effect modifier." If you say Z moderates or modifies the effect of X on Y you just mean that the relationship between X and Y depends on Z - i.e. there's a significant interaction between X and Z.

One really last comment: We looked at "two-way" interactions (between 2 predictors X_1 and X_2). You can have multiple 2-way interactions in a model. You can also have "higher order" interactions involving the product of 3 or more variables. You can get a lot of interaction terms ⇒ leads to overfitting unless you carefully select them.

Curvilinear Models + Transformations

- Like an indicator or interaction we create a special variable based on existing ones – our interpretation will be about the shape of the relationship between the original X and Y
- You can transform either the X or the Y variable depending on circumstances

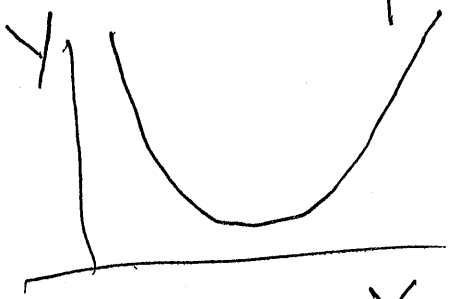
Most common choice – power transformations

On X → include in model terms like $X, X^2, X^3, \ldots$

On Y → raise Y to any power, Y^m to change shape.

e.g. A "quadratic" model has the form

$$Y = \underbrace{\beta_0 + \beta_1 X + \beta_2 X^2}_{\text{shape of a parabola}} + \varepsilon$$



This is treating X and X^2 as two separate variables in the model but of course they are linked and the interpretation if β_2 is significant is that a curved shape describes the relationship between X and Y better than a line does. Naturally we can't talk about the change in X^2 while holding X fixed!

Two key points

- No matter how many terms you have it's really all about the relationship between X and Y
- Hierarchical principal - you

can get an arbitrarily good approximation to any shape using powers of X (Taylor expansion) but you should stop when higher powers are not significant (i.e. making the fit much better) and you should keep all the terms below your highest significant one whether they are significant or not.

e.g. if you want X^3 you also should keep X and X^2