M 374G/384G, Fall 04

FACTORS AND INTERACTIONS

Example: Testing a new teaching method.

We use a treatment and a control group. But some students will do better because of prior learning or because of aptitude. So we give a pretest to measure prior learning, a post-test to measure learning at the end of instruction, and an aptitude test. We ask: Do students do better under the experimental method, other things being equal?

Variables:

Response: y = (post-test score) - (pretest score)

Predictors:

x_1 = score on aptitude test (a *covariate*)

$$x_2 = \begin{cases} 0 & \text{control group} \\ 1 & \text{experimental group} \end{cases}$$

Possible models:

I. $E(y \mid x_1, x_2) = \eta_0 + \eta_1 x_1 + \eta_2 x_2$

This says:

For the control group, $E(y \mid x_1) =$

For the treatment group, $E(y \mid x_1) =$

Picture:

If the model is correct, then

$\eta_2 > 0$ says:

$\eta_2 = 0$ says:

$\eta_2 < 0$ says:

Will this model fit if, for example, the new method helps poorer students more than better students?

II. $E(y \mid x_1, x_2) = \eta_0 + \eta_1 x_1 + \eta_2 x_2 + \eta_3 x_1 x_2$

(Like the first model, but with the *interaction term* $x_1 x_2$ added.)

This says:

For the control group, $E(y \mid x_1) =$

For the treatment group, $E(y \mid x_1) =$

If $\eta_2 > 0$ and $\eta_3 < 0$, we have the picture:

This says:

If $\eta_2 < 0$ and $\eta_3 > 0$, we have the picture:

This says:

USING CATEGORICAL VARIABLES WITH MORE THAN ONE CATEGORY

Examples:

- Two treatments plus control (3 categories)
- Socioeconomic class (high, medium, low)
- Amount of water received by plants might be set at high, medium, low

Terminology:

Level = one category of a categorical variable.

Factor = a set of indicator variables used to describe a single categorical predictor.

Example: If a categorical variable has three levels, we could define three indicator variables:

$$v_1 = \begin{cases} 1 & level\ 1 \\ 0 & level\ 2 \\ 0 & level\ 3 \end{cases} \qquad v_2 = \begin{cases} 0 & level\ 1 \\ 1 & level\ 2 \\ 0 & level\ 3 \end{cases} \qquad v_3 = \begin{cases} 0 & level\ 1 \\ 0 & level\ 2 \\ 1 & level\ 3 \end{cases}$$

However,

$$v_1 + v_2 + v_3 = 1,$$

so using all three

- is not necessary
- introduces multicollinearity.

Which two we use will depend on which level we prefer to act as "baseline". If we choose v_2 and v_3, then level 1 is our baseline: It is the "default" case when the coefficients of the terms involving the indicator variables are all zero. In the case of treatment and control groups, typically the control is chosen as baseline.