SOC 3811/5811:
BASIC SOCIAL STATISTICS

Multiple Regression

What if X is a Discrete Variable?

Regression vs. ANOVA

The regression techniques we have used thus far require continuous predictor (X) variables

It would be wrong—technically and conceptually—to simply enter nominal or ordinal variables as predictor variables, since it is wrong to compute means and standard deviations for these variables

Regression vs. ANOVA

For example, the variable X = “favorite color” might be coded as:

X = 1 = Favorite color = blue

X = 2 = Favorite color = red

X = 3 = Favorite color = green

X = 4 = Favorite color = some other color

Because X is discrete, its mean and standard deviation are meaningless.

Regression vs. ANOVA

Since X ("favorite color") is discrete, it would also be meaningless to put it into a regression relating X to continuous variable Y ("level of happiness" between 0 and 100):

$$\hat{Y} = a + b_1 X$$

That's because a "one unit increase in X" is meaningless. As are the means, standard deviations, and correlations involving X that go into the computations.

So what do we do?

Regression vs. ANOVA

For a discrete variable X that has *j* categories, we can construct *j* "dummy" variables—each of which has possible values 0 and 1—and each of which indicates whether an individual falls into a particular category of X

Regression vs. ANOVA

For the “favorite color” example:

We can construct j=4 dummy variables, X_1 through X_4:

	X_1	X_2	X_3	X_4
Blue	1	0	0	0
Red	0	1	0	0
Green	0	0	1	0
Other	0	0	0	1

Notice that knowing the value of j—1 of the X_k values allows you to infer the value of the j^{th} value

Regression vs. ANOVA

If we then regress Y (happiness) on j—1 of these dummy variables—excluding one of them—we get:

$$\widehat{\mathrm{Y}} = \mathrm{a} + \mathrm{b}_2\mathrm{X}_2(\mathrm{Red}) + \mathrm{b}_3\mathrm{X}_3(\mathrm{Green}) + \mathrm{b}_4\mathrm{X}_4(\mathrm{Other})$$

We might observe:

$$\widehat{\mathrm{Y}} = 80 - 10\mathrm{X}_2(\mathrm{Red}) - 20\mathrm{X}_3(\mathrm{Green}) - 15\mathrm{X}_4(\mathrm{Other})$$

Given that X_1 through X_4 all equal 0 or 1, the predicted values from this regression exactly reproduce the means of Y for each (discrete) value of X!

Thus, we have simply done ANOVA. (We will, for example, get exactly the same F statistic)

Regression vs. ANOVA

So why not just do ANOVA to look at the association between a continuous variable Y and a discrete variable X?

Because in the regression framework we can statistically control for confounders. For example, what if Z="age" confounds the association between X and Y?

Model 1:

$$\widehat{\mathrm{Y}} = 80 - 10\mathrm{X}_2(\mathrm{Red}) - 20\mathrm{X}_3(\mathrm{Green}) - 15\mathrm{X}_4(\mathrm{Other})$$

Model 2:

$$\widehat{\mathrm{Y}} = 75 - 5\mathrm{X}_2(\mathrm{Red}) - 10\mathrm{X}_3(\mathrm{Green}) - 7.5\mathrm{X}_4(\mathrm{Other}) + 10\mathrm{X}_5(\mathrm{Age})$$

Model 1 doesn't control for age; Model 2 does.

WORKSHEET

The prediction equation below is from the regression of continuous variable Y (income) on discrete variable X (highest degree attained). Note that X_1=1 if people did not complete high school (and 0 otherwise).

$$\widehat{Y} = 20{,}000 + 15{,}000X_2(\text{High School Diploma Only}) + 25{,}000X_3(\text{Bachelors Degree Only}) + 45{,}000X_4(\text{Advanced Degree})$$

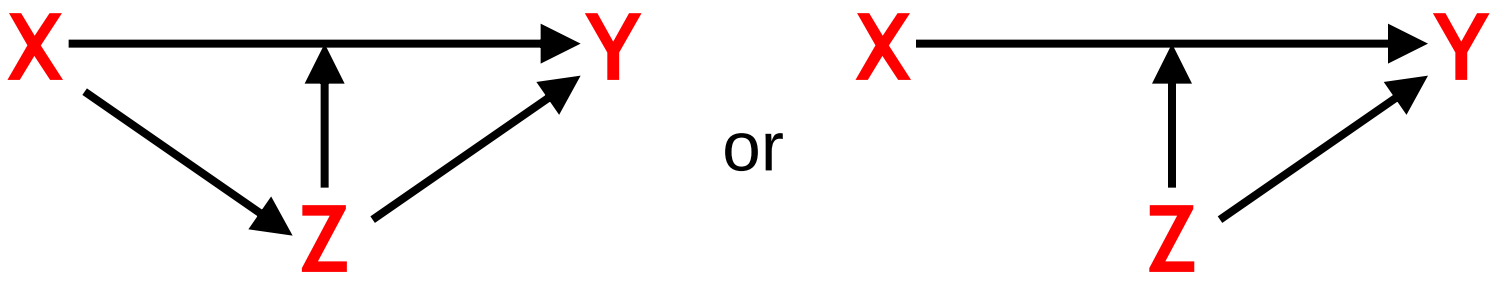
Report the mean value of Y for each of the 4 discrete values of X. That is, what is the mean level of income for people with different levels of education?

How Do We Allow For Moderation?

Interaction Effects

In some cases the association between X and Y may be different across levels of Z (or “conditioned by” Z)

In these cases we say that there is an interaction between X and Z … Z is known as a moderating variable

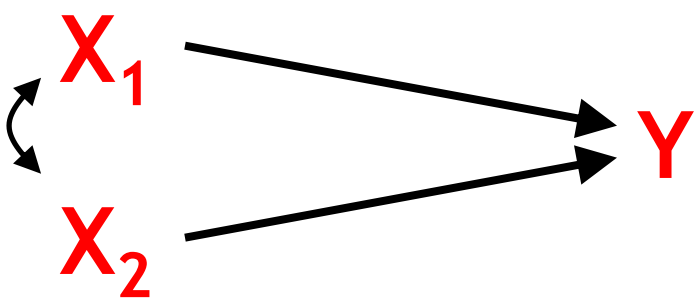


Interaction Effects

How do we model this in the regression context?

Imagine that we regress a continuous measure of income (Y) on a continuous measure of education (X_1) and a dummy variable for gender (X_2) that equals 0 for women and 1 for men

That model looks like this:

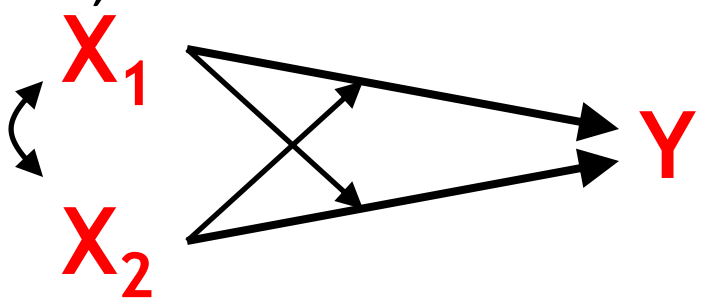


Interaction Effects

To allow for an "interaction effect" — such that the effect of X_1 on Y varies across levels of X_2 and the effect of X_2 on Y varies across levels of X_1 — we add an interaction term

An interaction term between X_1 and X_2 is simply a new variable creating by multiplying X_1 by X_2

If we add the interaction term to the model that includes X_1 and X_2 as predictors, we have:



Interaction Effects

Imagine we are interested in the effects of continuous variable X_1=education (in years) and discrete variable X_2=gender (0=female, 1=male) on continuous variable Y=income.

$$\widehat{Y}_i = a + b_1 X_1(\text{Education}) + b_2 X_2(\text{Male})$$

This model stipulates that there is only ONE effect of education … b_1 … and thus that this is the effect for both females and males.

What if we think there are different effects of education by gender? Maybe education pays off more for men?

Interaction Effects

To allow gender to moderate the effect of education (and vise versa), we create a new "interaction" variable:

$$X_3 = X_1(\text{Education}) \times X_2(\text{Male})$$

Then, we add the interaction term X_3 to the model:

$$\hat{Y}_i = a + b_1X_1(\text{Education}) + b_2X_2(\text{Male}) + b_3X_3$$

What is the effect of education for men (X_2=1)?

$b_1 + b_3$ because X_3=1

What is the effect for women (X_2=0)?

Just b_1 because X_3=0

Interaction Effects

How do we know whether the interaction term improves the power of the model to predict Y? We may theorize that an interaction effect exists … but how do we test the hypothesis that this is true in the data?

Look at the statistical significance of the coefficient for the interaction term.

WORKSHEET

The prediction equation below is from a regression of continuous variable Y ("happiness" where 100=maximum happy and 0=maximum unhappy) on continuous variable X_1 ("age in years"); discrete variable X_2 (0=Packers fan, 1=Vikings fan); and discrete interaction term X_3 which equals X_1 times X_2.

$$\widehat{Y}_i = 10 + 1.0X_1(\text{Age}) - 10.0X_2\,(\text{Fan}) - 0.5X_3$$

a. How do you interpret the coefficient (aka, slope) for X_1?
b. What is the effect of age for Vikings fans?
c. What is the effect of age for Packers fans?