Econ 300/QAC 201: Quantitative Methods in Economics/Applied Data Analysis

14th Class **6/28/10**

"An appropriate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."--John Tukey

[as follow-up to question on the 8th problem set and on the midterm, show chicken cartoon and link to discussion of published research and science news cycle cartoon]

[review material on multiple regression:]
go to a multiple regression model for several reasons:

1) are simultaneously interested in the effects of several independent variables on a dependent variable (often because you are testing a theory which involves several variables, or fitting a relationship that involves several variables, e.g. a production function)
2) want to explain as much of the variation in Y as possible (subject to Occam's Razor)
3) are only interested in the effect of one variable, but want to measure its effect accurately
   a. need to eliminate bias on estimating the direct effect of this variable by including confounding factors
   b. want to distinguish between direct and indirect effects of a variable (path analysis)

discuss path analysis: allows us to decompose the effects of an independent variable on the dependent variable in a regression into indirect and direct effects

example of path analysis, calculating total effects as the sum of the direct effect and the indirect effects: problem 13-23:

a)

X_1 X_2 X_3 X_4

Y

b) X_3 effects on Y: direct (.40) + indirect (.21*.43 = .09) = .49

X_2 effects on Y: direct (.12)
+ indirect via X_4 (.21*.22 = .0462)
+ indirect via X_3 (.40*.28 = .112)
+ indirect via X_3 & X_4 (.21*.43*.28 = .025284)
= .303

X_1 effects on Y: direct (-.01)
+ indirect via X_4 (.21*.02 = .0042)
+ indirect via X_3 (.40*.31 = .124)
+ indirect via X_2 (.12*.52 = .0624)
+ indirect via X_3 & X_4 (.21*.43*.31 = .027993)
+ indirect via X_2 & X_4 (.21*.22*.52 = .024024)
+ indirect via X_2 & X_3 (.40*.28*.52 = .05824)
+ indirect via X_2 & X_3 & X_4 (.21*.43*.28*.52 = .0131477)
= .304

c) i. 3*(-.01) = -.03
ii. 3*(.304) = .912

econ: tend to either just put in the “truly exogenous” variables (in this case just X_1) or ignore all the indirect effects (assume all the Xs are exogenous)

what about cases where a causes b and c?
[show link to article about creativity]

next, Ch. 14: ways to extend the usefulness of the OLS method

Ch. 14:

expands usefulness of multiple regression analysis through use of two techniques: dummy variables and nonlinear specifications

dummy variables

--intercept can vary across groups

e.g. men and women; consider the relationship between earnings and experience. Perhaps men make more than women by a constant amount at any level of experience.

test this by setting up a dummy D = 1 if woman, 0 if man

then run the multiple regression:

earnings $= b_0 + b_1$*experience $+ b_2$*D

then you can derive two regression lines, one for men and one for women, which differ only in their intercept:

men: $b_0 + b_1$*experience
women: $(b_0 + b_2) + b_1$*experience

we can run a simple t-test on the coefficient on the dummy variable to see if it is different from zero; if it fails the test, cannot reject the hypothesis that the intercepts are the same for men and women

can also create what is called a "fully interacted" model, where slope is allowed to vary as well as intercept between groups (can also allow only slope to vary while holding intercept constant, but this is rarely done in practise)

Suppose that men make more than women at any level of experience, but that the gap widens at higher levels of experience, so that the slope of the men's regression line is steeper.

test this by setting up a dummy D = 1 if woman, 0 if man

now create a new variable Dex = D*experience

then run the multiple regression:

earnings $= b_0 + b_1$*experience $+ b_2$*D $+ b_3$Dex

then you can derive two regression lines, one for men and one for women, which differ in both their intercept and slope:

men: $b_0 + b_1$*experience
women: $(b_0 + b_2) + (b_1 + b_3)$*experience

now can use t-test on the coefficient b_3 to test if the slopes are the same or not (null hypothesis is that $b_3 = 0$, i.e., the slopes are the same)

can also set up dummy variables for several different indicators in the same equation, and can have both dummy and nondummy variables in the same equation

e.g. Mike Lovell ran a regression to try to calculate how much different factors about a car affected the car's mileage per gallon in the early 1980s:

MPG = $b_0 + b_1$*weight + b_2*transmission + b_3*engine

where transmission = 0 if standard, 1 if automatic
and engine = 0 if gasoline, 1 if diesel

and found MPG = 43.6 - 0.006*weight - 3.75*transmission + 6.26*engine

he was able to explain 74% of the variance in MPG across cartypes with this equation

in these two cases, there were only two groups for each indicator, but with dummy variables, if there are n groups related to an indicator, need n-1 dummies

e.g., three employment groups, red, blue, and green

set up D_1 = 1 if blue, 0 otherwise
D_2 = 1 if green, 0 otherwise

so red is the implicit reference group

then estimate:

earnings = $b_0 + b_1$*experience + b_2*D_1 + b_3*D_2

and can derive the three parallel lines:

red: $b_0 + b_1$*experience
blue: $(b_0 + b_2) + b_1$*experience
green: $(b_0 + b_3) + b_1$*experience

it is irrelevant which group is set up as the reference group; the coefficients on the dummy variables will adjust to yield the same intercepts no matter what

note that you can model group differences in an additive or separate way

e.g. consider differences in intercepts for whites vs. nonwhites, and Hispanics vs. nonHispanics; two ways of modeling differences, depending on whether you think racial and Hispanic status are interactive or not in terms of their effects on earnings; in both cases control group is white Hispanic:

i) a noninteractive (additive) model:

D_1 = 1 if nonwhite, 0 otherwise
D_2 = 1 if nonHispanic, 0 otherwise

estimate earnings $= b_0 + b_1$*experience $+ b_2{*}D_1 + b_3{*}D_2$

then can derive four regression lines:

w, H: $b_0 + b_1$*experience
nw, H: $(b_0 + b_2) + b_1$*experience
w, nH: $(b_0 + b_3) + b_1$*experience
nw, nH: $(b_0 + b_2 + b_3) + b_1$*experience

ii) an interactive model where each of the four possible groups is allowed to have a uniquely determined intercept

D_1 = 1 if nonwhite and Hispanic, 0 otherwise
D_2 = 1 if white and nonHispanic, 0 otherwise

D_3 = 1 if nonwhite and nonHispanic, 0 otherwise

estimate earnings $= b_0 + b_1$*experience $+ b_2{*}D_1 + b_3{*}D_2 + b_4{*}D_3$

then can derive four regression lines:

w, H: $b_0 + b_1$*experience
nw, H: $(b_0 + b_2) + b_1$*experience
w, nH: $(b_0 + b_3) + b_1$*experience
nw, nH: $(b_0 + b_4) + b_1$*experience

nonlinear specifications--can create new variables which "trick" the regression package into estimating Y as a nonlinear function of X

e.g. suppose we think returns to experience start to decrease after a certain point, so the earnings function flattens out

may prefer to estimate:

earnings $= b_0 + b_1$*experience $+ b_2$*(experience)^2

create a new variable expsq = (experience)^2

and run the equation:

earnings $= b_0 + b_1$*experience $+ b_2$*expsq

can use this technique to estimate any degree of polynomial; then can use t-tests to see if the coefficients on the higher-order terms are significantly different from zero to see if they should be included or not.

also allows us to avoid having to make linear projections forever

next class: finish Ch. 14, start Ch. 15