13th Class **6/24/10**

"Get the facts first, and then you can distort them as much as you please."--Mark Twain

view of this part of course: understanding regression techniques

Ch 13: 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)

general linear model (GLM):

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

$$E(Y) = b_0 + b_1X_1 + b_2X_2$$

using OLS, we estimate the parameters:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

where b_1 and b_2 are given by the equations:

$$\sum x_1 y = b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

$$\sum x_2 y = b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

(x and y are deviations from the means) (solve 2 equations for 2 unknowns)

and then substitute b_1 and b_2 into the following equation to get the intercept b_0:

$$b_0 = \overline{Y} - b_1 \overline{X_1} - b_2 \overline{X_2}$$

can still do this by hand, but obviously the more data you have and the more independent variables, the harder it gets; it is trivial for a computer to do though

can't show it geometrically any more after two independent variables either--goes to n dimensions where there are n-1 independent variables
1858: early example of use of Least Squares to solve a large multivariable problem: (Stigler, p. 158): "The 1858 Ordnance Survey of the British Isles required the reduction of an immense mass of data through the use of least squares. The main triangulation was cast as a system of 1554 equations involving 920 unknowns. Even though they broke the system into 21 pieces of no more than 77 unknowns each before attempting a solution, the calculations took two teams of human "computers" working independently and in duplicate, two and a half years to complete."

can still calculate standard errors for each coefficient separately and create confidence intervals and t ratios for each coefficient separately (d.f. now n - k - 1) where there are (k + 1) coefficients to estimate and n observations, k being the number of slope variables and 1 for the intercept)

[use example of predicting Olympic medal counts for different nations: goal of seeing what matters and how much; contrast to simple prediction based on last time's outcome]

Rest of the course is basically going to be extensions of the regression model, some of which will be directly relevant to your project

omitted variable bias: [come up with an example]
what are important omitted variables, important in the sense that leaving them out causes the coefficient on some included variables to be too large (in absolute value)?

underlying problem is that observational data are generally nonrandom; systematic variation occurs and pulls the regression line towards it

interpretation of each coefficient: the effect on the dependent variable if all other variables are held constant

--it is the direct effect only

finish Ch. 13 next class (path analysis), start Ch. 14