18th Class **7/2/10**

"Out of the air a voice without a face
Proved by statistics that some cause was just
In tones as dry and level as the place."
--W.H. Auden

handouts:
--practice test (first two hours' worth)
--short answers for practice test
--essay questions

--test next Wed.
--project due by 6:00 next Fri. as an email attachment

final's format: open-book for the problems, closed-book for the essays
should allocate an hour or less to the essays, two hours to the problems

go over first problem on problem set #11 [see handout]

recall we were going through cases of departure from the assumptions regarding error term structure that lead to OLS being the preferred method

review Ch. 24 briefly: predicting a series from itself and from other lagged and contemporaneous series

if mean of error terms does not appear to be zero, may need to try a nonlinear fit instead, or use an instrumental variables method (Ch. 25)

Ch. 25: addresses the problem of simultaneous equation bias

often in economics, each single relationship/equation is part of a bigger model, so some of the X variables in any one equation are Y variables in other equations. Ignoring these interrelationships leads to biased estimates of parameters for single equations.

consider the macroeconomic model:

consumption $= C_0 + C_1$(national income)

national income = consumption + investment

we will write this in statistical form as:

$Y = \alpha + \beta X + e$
$X = Y + I$ (no error term because this is a definition)

Note that by assumption, I is exogenous and is therefore statistically independent of e. X and Y, however, are endogenous and are thus influenced by both I and e. This model is mathematically complete; i.e., the two equations in this system can be solved for the two unknowns/endogenous variables. However, in an equation in a simultaneous system, regressors that are not predetermined (X in this case) are not independent of the error term (e). In other words, the assumption underlying OLS that X is statistically independent of e is violated. In this model, X and e are positively correlated: when X is large, e tends to be positive; when X is small, e tends to be negative. Therefore, β is biased upward. In general, we can determine the direction of simultaneous equation bias on each affected coefficient.

you can see this by solving for Y and X in terms of α, β, I, and e:

for the definition equation:

$X = \alpha + \beta X + e + I$

so $X = \frac{\alpha}{1-\beta} + \left(\frac{1}{1-\beta}\right)I + \frac{e}{1-\beta}$

so since X is a positive function of e, the correlation between X and e must be positive

this means that if e is large, Y is larger and therefore so is X
and if e is small, Y is smaller and therefore so is X

this violates the assumption of OLS that X and e are uncorrelated
(which is the assumption that the mean of e is zero)

How to remedy the problem of simultaneous equation bias? Use the technique of instrumental variables (IV). The trick is to find a new variable V that has the following properties:

1) V should be uncorrelated with e
2) V should be highly correlated with X

We can calculate the covariance of V and Y:

$$s_{VY} = \frac{\sum (V - \overline{V})(Y - \overline{Y})}{n - 1}$$

consider our basic regression model:

$$Y = \alpha + \beta X + e$$

and express the variables in deviation form (relative to the mean), which causes the constant to drop out:

$$(Y - \overline{Y}) = \beta(X - \overline{X}) + (e - \overline{e})$$

note the mean $\overline{e}$ is not equal to zero if there is simultaneous equation bias, so the last term does not reduce to e.

Then multiply both sides by $(V - \overline{V})$:

$$(V - \overline{V})(Y - \overline{Y}) = \beta(V - \overline{V})(X - \overline{X}) + (V - \overline{V})(e - \overline{e})$$

sum both sides from 1 to n:

$$\sum (V - \overline{V})(Y - \overline{Y}) = \beta \sum (V - \overline{V})(X - \overline{X}) + \sum (V - \overline{V})(e - \overline{e})$$

and divide through by (n - 1):

$$s_{VY} = \beta s_{VX} + s_{Ve}$$

so we have transformed the equation relating variables into an equation relating covariances. Now we can rearrange this to show the formula for β:

$$\beta = \frac{s_{VY}}{s_{VX}} - \frac{s_{Ve}}{s_{VX}}$$

define the first term on the right as:

$$b_V = \frac{s_{VY}}{s_{VX}}$$

if the last term on the right is small, then b_V is a good estimator of β . b_V is the instrumental variable (IV) estimator of β. This term will in fact be small if the two desirable properties of V listed above are met (so s_{Ve} approaches 0, causing the estimator to be consistent; and s_{VX} is large if V and X are highly correlated, again reducing the last term through increasing the size of the denominator)

Note that if V = X, the IV estimator is the OLS estimator (in other words, you are using X as an instrumental variable) and it will be a good estimator in the single equation model where σ_{Ve} (σ_{Xe}) is 0 by assumption

However, in the simultaneous equation system, using X as an instrument causes the estimator to be inconsistent because X and e are correlated ($\sigma_{Xe} \neq 0$).

However, exogenous variables will yield consistent estimates because they are uncorrelated with e. They can still be better or worse instruments depending on their correlation with X. I in our model above passes this criterion well, so we would use the estimator for β:

$$b_I = \frac{s_{IY}}{s_{IX}}$$

This can be generalized to use a vector of instruments to construct the instrumental variable estimator for any given β. The technique of two stage least squares (2SLS) does exactly this; to estimate any single equation in a system which contains a mix of exogenous and endogenous regressors:

1) (first stage) regress endogenous regressors on all the exogenous variables in the sytem, using OLS. Then the predicted value of the endogenous regressor in each case will be independent of e by construction.

2) (second stage) use the predicted values of the endogenous variables contained in the equation and the exogenous variables as instruments to estimate the single equation. This yields a system of n equations in n unknowns, where n = the number of parameters to be estimated in the equation. This system can then be solved for estimates of the parameters.

IV is a much more general technique however.

Note that in the example on p. 732 in the book, all the X and Y variables could not enter into each equation.

Consider the identification problem in general: an equation will be unidentified (incapable of being estimated) if the number of slope parameters to be estimated in the equation exceeds the number of exogenous variables appearing anywhere in the system.

do problem 25-8: 3 exogenous variables, so the 2nd and 5th equations cannot be estimated

a graphical example: the two-equation model of a market: supply and demand; it is really a three-equation model, where the equilibrium condition reduces the number of variables to be determined to 2 (Q and P):

$$Q_S = b_0 + b_1P + b_2W$$
$$Q_D = c_0 + c_1P + c_2I$$

Here both the supply and demand curves can be identified because there are 2 exogenous variables which cause shifts in the curves when they are graphed in P-Q space.

But consider a simpler model:

$$Q_S = b_0 + b_1P + b_2W$$
$$Q_D = c_0 + c_1P$$

We can only identify the line that does not move; movements in the other line trace out the stationary line; in this case only the demand curve is identified [show graph]

Note that identification is different from mathematical completeness, where one needs as many equations as unknowns in order to solve for the unknowns. Therefore, in both of these market models, if one adds the equilibrium condition:

$$Q_S = Q_D = Q$$

one can solve for Q and P in terms of the parameters (as functions of the parameters). But if we have data on Q, P, I, and W, only in the first model can all the parameters be estimated individually.

so we see modeling has a number of issues: what is exog., what is endog., are there enough equations to solve for the endog, what variables should be in each equation, are all the equations identified?