General IV regression: TSLS with multiple endogenous regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

- Instruments: $Z_{1i},\ldots,Z_m$
- Now there are k first stage regressions:
  - Regress X_1 on $W_1,\ldots, W_r, Z_1,\ldots, Z_m$ by OLS
  - Compute predicted values $\hat{X}_{1i}$, $i = 1,\ldots,n$
  - Regress X_2 on $W_1,\ldots, W_r, Z_1,\ldots, Z_m$ by OLS

- Compute predicted values $\hat{X}_{2i}$, $i = 1,\ldots,n$
- Repeat for all X's, obtaining $\hat{X}_{1i}$, $\hat{X}_{2i},\ldots, \hat{X}_{ki}$

TSLS with multiple endogenous regressors, ctd.

- Second stage
  - Regress Y on $\hat{X}_{1i}$, $\hat{X}_{2i}$,…, $\hat{X}_{ki}$, W_1,…, W_r by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SE*s are wrong
- To get correct *SE*s, do this in a single step
- *What would happen in the second stage regression if the coefficients were underidentified (that is, if #instruments < #endogenous variables); for example, if k = 2, m = 1?*

Sampling distribution of the TSLS estimator in the general IV regression model

- Meaning of “valid” instruments in the general case
- The IV regression assumptions
- Implications: if the IV regression assumptions hold, then the TSLS estimator is normally distributed, and inference (testing, confidence intervals) proceeds as usual

A “valid” set of instruments in the general case

The set of instruments must be relevant and exogenous:

1. Instrument relevance: *Special case of one X*

   At least one instrument must enter the population counterpart of the first stage regression.

2. Instrument exogeneity

   All the instruments are uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_m,u_i) = 0$

“Valid” instruments in the general case, ctd.

General instrument relevance condition:

- *General case, multiple X’s*

  Suppose the second stage regression could be run using the predicted values from the *population* first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression

- *Special case of one X*

  At least one instrument must enter the population counterpart of the first stage regression.

The IV Regression Assumptions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \ldots + \beta_{k+r} W_{ri} + u_i$$

1. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ are i.i.d.
3. The X's, W's, Z's, and Y have nonzero, finite 4^{th} moments
4. The W's are not perfectly multicollinear
5. The instruments $(Z_{1i}, \ldots, Z_{mi})$ satisfy the conditions for a valid set of instruments.

- #1 says “the exogenous regressors are exogenous.”
- #2 – #4 are not new; we have discussed #5.

Implications: Sampling distribution of TSLS

- If the IV regression assumptions hold, then the TSLS estimator is normally distributed in large samples.
- Inference (hypothesis testing, confidence intervals) proceeds as usual.
- Two notes about standard errors:
  - The second stage *SE*s are incorrect because they don't take into account estimation in the first stage; to get correct *SE*s, run TSLS in a single command

    - Use heteroskedasticity-robust *SE*s, for the usual reason.
- *All this hinges on having valid instruments…*

Checking Instrument Validity (SW Section 10.3)

Recall the two requirements for valid instruments:

1. *Relevance* (special case of one X)

   At least one instrument must enter the population counterpart of the first stage regression.

2. *Exogeneity*

   All the instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_{mi},u_i) = 0$

What happens if one of these requirements isn't satisfied? How can you check? And what do you do?