Applied Econometrics
Prof. Leo Feler
Quiz 5: Instrumental Variables

Name: Key

For notational simplicity, y is your dependent variable, s is your endogenous variable, x is your exogenous variable, and z is your instrumental variable.

1. What is an instrumental variables procedure? Why do we use it? What two conditions must an instrument satisfy?

An instrumental variables procedure is a way to estimate a causal effect in the presence of endogeneity, for example, omitted variable bias. We essentially "clean up" the estimate of our explanatory variable by using an instrumental variable to filter out the bias.

A good instrument must satisfy two conditions:

1) Strong – highly correlated with the endogenous variable
2) Valid – does not affect dependent variable other than through the endogenous variable

$\equiv$

independent of the error term $Cov(\varepsilon_i, z_i) = 0$

2. What is a structural, reduced form, first-stage, and second-stage equation? Show the structural, reduced form, first-stage, and second stage equations, and then show how you can estimate the causal effect of *s* on *y*. Assume only one endogenous variable, s, and one instrument, z. Show the calculation of indirect least squares (think about the Wald estimator).

Structural equation $Y_i = \alpha + \rho S_i + \eta_i$

The relationship of interest

First Stage

Relationship between instrument and endogenous variable $S_i = \pi_{10} + \pi_{11} Z_i + u_i$

Reduced form

Relationship between instrument and dependent variable (substitute in 1st stage)

$$\begin{aligned} Y_i &= \alpha + \rho\left[\pi_{10} + \pi_{11} Z_i + u_i\right] + \eta_i \\ &= (\alpha + \rho\pi_{10}) + \rho\pi_{11} Z_i + (\rho u_i + \eta_i) \\ &= \pi_{20} + \pi_{21} Z_i + \varepsilon_i \end{aligned}$$

To estimate the causal effect of S on Y using indirect least squares, divide the reduced form coefficient on z_i by the first stage coefficient on z_i.

$$\frac{\pi_{21}}{\pi_{11}} = \frac{\frac{\text{cov}(Y_i, Z_i)}{\text{var}(Z_i)}}{\frac{\text{cov}(S_i, Z_i)}{\text{var}(Z_i)}} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(S_i, Z_i)} = \frac{\rho\pi_{11}}{\pi_{11}} = \rho$$

3. When you run 2SLS in Stata, with two instruments z_1 and z_2, for one endogenous variable, s, what is Stata doing? Show in terms of equations. What do you need to check on the Stata output to make sure you satisfied (or at least didn't violate) the two conditions you mentioned in question 1?

With multiple instruments:

First stage $\quad s_i = \pi_{10} + \pi_{11} z_{i1} + \pi_{12} z_{i2} + u_i$

Reduced form $\quad Y_i = \alpha + \rho[\pi_{10} + \pi_{11} z_{i1} + \pi_{12} z_{i2} + u_i] + \zeta_i$

$= (\alpha + \rho\pi_{10}) + \rho(\pi_{11} z_{i1} + \pi_{12} z_{i2}) + (\rho u_i + \zeta_i)$

$= \pi_{20} + \rho$ (weighted avg. of z_{i1} & z_{i2}) $+ \varepsilon_i$

In this case we cannot use indirect least squares. Instead, Stata estimates $\hat{s}_i$ and uses that estimate in the structural equation:

$Y_i = \tilde{\alpha} + \tilde{\rho}\hat{s}_i + \zeta_i$

When you run 2SLS in stata with more instruments than you have endogenous variables, you need to check the F-stat (strength condition) and the Hansen/Sargan test (validity condition) → $F \geq 10$

→ fail to reject H_0 ($\approx p > .10$)

4. What is the Hansen/Sargan test for overidentifying restrictions? What is the null and alternative hypothesis you are testing. How do you conduct this test?

The Hansen/Sargan test is a way to test the orthogonality assumption (validity) when using an instrumental variables procedure.

H_0: The instruments are not invalid

H_a: The instruments are invalid (correlated with the error term)

To conduct the test:

- estimate with 2SLS $Y_i = \alpha + \rho S_i + X_i + \varepsilon_i$
  $\Downarrow$
  $S_i = \pi_{10} + \pi_{11} Z_{i1} + \pi_{12} Z_{i2} + X_i + \zeta_i$
- Save $\hat{\varepsilon}_i$
- estimate with OLS $\hat{\varepsilon}_i = \gamma_{10} + \gamma_{11} z_i + \gamma_{12} z_{i2} + X_i + \xi_i$
- Test whether γ_{11} and $\gamma_{12} = 0$
  - $R^2 \approx 0$ if instruments are not correlated with the error term
  - Test stat $NR^2 \sim \chi^2(\#Z_i - \#S_i)$
    - N: # observations
    - χ^2: "chi-square"
    - $\#Z_i$: # instruments
    - $\#S_i$: # endogenous variables
  - obtain p-value

5. If you have multiple instruments for multiple endogenous variables, in addition to the two conditions that must be satisfied from question 3, what is a third condition that must be satisfied and why? Could you generate a bunch of random variables as instruments to make sure you have at least as many instruments as endogenous variables? Why or why not?

In addition to instruments being strong and valid, our equation cannot be underidentified. This means that each endogenous variable must have at least one instrument that explains it, and that is not already explaining some other endogenous variable. For example: $y_i = \alpha + \beta_1 s_{1i} + \beta_2 s_{2i} + \beta_3 x_i + \varepsilon_i$

Instruments z_{1i} and z_{2i}. If z_{1i} and z_{2i} are both strongly correlated (i.e., predict) s_{1i} but neither are strongly correlated with s_{2i} (i.e., F-stat < 10), then even though we have two instruments for two endogenous variables, our equation is underidentified. Alternatively if z_{1i} is strongly correlated with s_{1i} and s_{2i} but z_{2i} is strongly correlated with neither, then our equation is again underidentified. We effectively have only one instrument for

two endogenous variables, since z_{2i} is not doing anything in our regression.

We cannot generate a bunch of random variables as instruments. Since they are <u>randomly</u> generated, by construction, they are uncorrelated with any of our endogenous variables and so do not "effectively" count as having more instruments.

"Instruments have to be instrumental in order to count as instruments"