Section 9 Regression with Stochastic Regressors

Meaning of random regressors

- Until now, we have assumed (against all reason) that the values of x have been controlled by the experimenter.
- Economists almost never actually control the regressors
- We should usually think of them as random variables that are determined jointly with y and e
- With a small adaptation of our assumptions, OLS still has the desirable properties it had before

OLS assumptions with random regressors

With fixed x	With random x
SR1: $y = \beta_1 + \beta_2 x + e$ with x fixed	A10.1: $y = \beta_1 + \beta_2 x + e$ with x, y, e random
SR2: $E(e) = 0$	A10.2: (x, y) obtained from IID sampling
SR3: $\text{var}(e) = \sigma^2$	A10.3: $E(e \mid x) = 0$
SR4: $\text{cov}(e_i, e_j) = 0$	A10.4: x takes on at least two values
SR5: x takes on at least two values	A10.5: $\text{var}(e \mid x) = \sigma^2$
SR6: e is normal	A10.6: e is normal

- Note that A10.2 implies SR4 (and A10.5?)
- Note that A10.3 implies both $\text{cov}(x, e) = 0$ and $E(e) = 0$
    - This assumption is a critical one.
    - Instead of assuming that x is a fixed value and e is random, we make the properties of e conditional on the particular outcome of x
    - This allows us to operate in very much the same way as if x is fixed, as long as A10.3 holds.
    - In the next section of the course, we will discuss how to deal with violations of A10.3.

OLS properties

- **Small-sample properties**
    - If A10.1–A10.6 hold, then
        - OLS is unbiased
        - OLS is BLUE
        - OLS standard errors are unbiased

        - OLS coefficient estimators (conditional on *x*) are normal
- **Asymptotic properties**
    - We can replace A10.3 by the weaker A10.3*:
        - $E(e) = 0$, $\text{cov}(x, e) = 0$.
        - OLS is biased in small samples if A10.3* is true but A10.3 is not
    - Under A10.1–A10.5, replacing with A10.3*:
        - OLS coefficient estimators are consistent
        - OLS coefficient estimators are asymptotically normal

If x *is correlated with* e

- If A10.3* is violated, then OLS is biased and inconsistent
    - Coefficient on *x* will pick up the effects of the parts of *e* that are correlated with it in addition to the direct effects of *x*
    - Direction of bias depends on sign of correlation between *x* and *e*
    - This is exactly analogous to omitted-variables bias
- **Measurement error** (discussed above under internal validity)
    - Suppose that the dependent variable is measured accurately but that we measure *x* with error: $\tilde{x}_i = x_i + \eta_i$.
    - The estimated model is $y_i = \beta_1 + \beta_2 \tilde{x}_i + (e_i - \beta_2 \eta_i)$.
    - Because η is part of $\tilde{x}$ and therefore correlated with it, the composite error term is now correlated with the actual regressor, meaning that b_2 is biased and inconsistent.
        - If *e* and η are independent and normal, then $\text{plim}\, b_2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2}\beta_2$.
        - The estimator is biased toward zero.
        - If most of the variation in $\tilde{x}$ comes from *x*, then the bias will be small.
        - As the variance of the measurement error grows in relation to the variation in the true variable, the magnitude of the bias increases.
        - As a worst-case limit, if the true *x* doesn't vary across our sample of observations and all of the variation in our measure $\tilde{x}$ is random noise, then the expected value of our coefficient is zero.
    - Best solution is getting a better measure.
    - Alternatives are instrumental variables or direct measurement of degree of measurement error.
        - For example, if an alternative, precise measure is available for some arguably random sub-sample of observations, then we can calculate the variance of the true variable and the variance of the measurement error and correct the estimate.

- **Omitted-variables bias**
  - We derived this result at the beginning of the multiple regression analysis
  - Omitted variable is included in error. If omitted variable is correlated with included variable, then OLS estimator of coefficient on included variable is biased and inconsistent.
- **Simultaneous-equations bias (simultaneity bias)**
  - Suppose that *y* and *x* are part of a larger theoretical system of equations:
    $y = \beta_1 + \beta_2 x + e$
    $x = \gamma_1 + \gamma_2 y + u$
  - The two variables are "jointly determined" and both are endogenous.
    - There is "feedback" from *y* to *x*, or "reverse causality" (actually, bidirectional)
  - $e \Rightarrow y \Rightarrow x$, so *e* and *x* are correlated
  - Supply and demand curves are difficult to estimate because both *q* and *p* are endogenous

Instrumental variables

- Recall the method of moments analysis by which we derived the OLS estimators
  - We used the assumed population moment conditions $E(e) = 0,\ \mathrm{cov}(x,e) = 0$ to derive the OLS normal equations as sample moment conditions: $\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2 = 0,\ \frac{1}{N}\sum_{i=1}^{N}x_i\hat{e}_i = 0$
  - If $\mathrm{cov}(x,e) \neq 0$, then the population moment conditions are invalid and we will get biased and inconsistent estimators from the OLS sample moment conditions.
- The **instrumental-variables estimator** can be derived from the method of moments.
- As usual, suppose that $y = \beta_1 + \beta_2 x + e,$ but suppose that $\mathrm{cov}(x,e) \neq 0.$
- Let *z* be a variable with the following properties:
  - *z* does not have a direct effect on *y*. It does not belong in the equation alongside *x*. (*z* affects *y* <u>only</u> through *x*, not independently.)
  - *z* is exogenous. It is not correlated with *e*.
  - *z* is strongly correlated with *x*, the endogenous regressor.
- This makes *z* a valid instrumental variable.
- We can exploit $\mathrm{cov}(z,e) = 0$ as our second moment condition in place of $\mathrm{cov}(x,e) = 0$, which is not true for this model.

- The sample moment conditions are

$$\sum_{i=1}^{N}\hat{e}_i = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0$$

$$\sum_{i=1}^{N}\hat{e}_i z_i = \sum_{i=1}^{N} z_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0.$$

- Solving the normal equations yields $\hat{\beta}_2 = \dfrac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})}$.

- Compare this to the standard OLS slope estimator $b_2 = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})}$.

- In matrix terms, $\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ vs. $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- Properties of IV estimator:
    - Consistent as long as z is exogenous
    - Asymptotically normal
    - $\hat{\beta}_2 \sim N\left(\beta_2, \dfrac{\sigma^2}{r_{xz}^2\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$, $r_{xz} \equiv \widehat{\text{corr}}(x, z)$
    - As usual, we estimate σ^2 by $\sigma_{IV}^2 = \dfrac{\sum_{i=1}^{N}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2}{N-2}$
- Weak instruments: If r_{xz} is near zero, then the variance of $\hat{\beta}_2$ is large and the IV estimator is unreliable.

Two-stage least squares

- What if we have multiple strong instruments and/or multiple endogenous regressors in a multiple regression?
- With more instruments than endogenous regressors, we have an "overidentified" system with alternative choices of instruments.
    - Suppose that x_K is endogenous but the first $K-1$ regressors are exogenous
    - Suppose that z_1 through z_L are L valid instruments
    - Any linear combination of the instruments is admissible
    - Let's choose the one that is more correlated with x_K
        - To get that, we regress $x_K = \gamma_1 + \gamma_2 x_2 + \ldots + \gamma_{K-1}x_{K-1} + \theta_1 z_1 + \ldots + \theta_L z_L + \nu_K$ and use the fitted values $\hat{x}_K$ as the instrument for x_K

- This amounts to doing two separate regressions, the **first-stage** regression of x_K on the exogenous x variables and the instruments z, then a **second-stage** regression of $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_{K-1} x_{K-1} + \beta_K \hat{x}_K + e^*$
- The estimators of β from the second-stage regression are called 2SLS estimators.
- But it's not exactly like doing two separate regressions because our estimator of the error variance uses the actual values of x_K rather than the fitted values:

$$\hat{\sigma}^2_{IV} = \frac{\sum_{i=1}^{N} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i,2} - \ldots - \hat{\beta}_K x_{i,K} \right)^2}{N - K}$$

  - (If you do the second regression manually substituting in the fitted values, Stata will use the fitted values to calculate the residuals rather than the actual.)
- 2SLS easily extends to multiple endogenous regressors, as long as there are more independent instruments than endogenous regressors.
  - Suppose there are G "good" exogenous regressors, $B = K - G$ "bad" endogenous regressors, and L "lucky" instrumental variables.
  - $L > B$ means overidentified, $L = B$ is just identified, $L < B$ means underidentified (and can't be estimated by IV)
  - $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_G x_G + \beta_{G+1} x_{G+1} + \ldots + \beta_K x_K + e$
  - First-stage regressions:
    $x_{G+j} = \gamma_{1j} + \gamma_{2j} x_2 + \ldots + \gamma_{Gj} x_G + \theta_{1j} z_1 + \ldots + \theta_{Lj} z_L + v_j, \; j = 1, \ldots, B$
  - Get fitted values:
    $\hat{x}_{G+j} = \hat{\gamma}_{1j} + \hat{\gamma}_{2j} x_2 + \ldots + \hat{\gamma}_{Gj} x_G + \hat{\theta}_{1j} z_1 + \ldots + \hat{\theta}_{Lj} z_L, \; j = 1, \ldots, B$
  - Regress original equation replacing endogenous regressors with fitted values
    $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_G x_G + \beta_{G+1} \hat{x}_{G+1} + \ldots + \beta_K \hat{x}_K + e^*$
- To implement 2SLS in Stata, use ivregress 2sls depvar exvars (endvars = instvars) , options

Overidentification and generalized method of moments

- If we have additional instruments beyond the minimum (i.e., an overidentified system), then we have more information than we need to estimate the model.
- Suppose that z_1 and z_2 are both valid instruments for endogenous x

- All three moment conditions:

$$\sum_{i=1}^{N}\left(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i\right)=\hat{m}_1=0$$

$$\sum_{i=1}^{N} z_{i,1}\left(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i\right)=\hat{m}_2=0$$

$$\sum_{i=1}^{N} z_{i,2}\left(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i\right)=\hat{m}_3=0$$

  are theoretically true.
- This opens the door for two possibilities:
  - We can determine the degree to which we cannot satisfy all three of these conditions simultaneously and use that as evidence of whether the model's assumptions are valid. (If the model is perfect, then all three should be zero except for sampling error.)
    - These are specification tests discussed below
  - We can think about alternative estimators (called GMM estimators) that would minimize a weighted average of the squares of the *m* moments.
  - 2SLS is a GMM estimator with a particular weighting of the moment conditions.

Instrument strength

- A strong instrument must provide correlation with part of the endogenous regressor that is *not* explained by the other (exogenous) regressors.
- Regression of $x_K=\gamma_1+\gamma_2 x_2+\ldots+\gamma_{K-1}x_{K-1}+\theta_1 z_1+v_K$ allows us to test $\theta_1 = 0$ with a standard $F = t^2$ test.
  - However, conventional wisdom says that the instrument is weak unless $F > 10$ rather than the standard critical values for testing this hypothesis.
  - This test can be applied with multiple instruments and one endogenous regressor, with 10 still being the traditional threshold for weak instruments.
  - (See HGL's Appendix 10E for a really confusing exposition of the general test with multiple endogenous regressors)

Specification tests

- If the model is overidentified, then we can do two kinds of tests:
  - A **Hausman test** of whether the *x* variables that we are treating as endogenous truly are endogenous
  - A test of the overidentifying restrictions, which can be interpreted as a test of instrument validity
- **Hausman test**
  - $H_0:\operatorname{cov}(x,e)=0,\ \ H_1:\operatorname{cov}(x,e)\neq 0$

- Under null hypothesis, OLS is consistent and efficient, IV is consistent but inefficient. Since both are consistent, $q \equiv b - \hat{\beta} \to 0$ in large samples
- Under alternative hypothesis, OLS is inconsistent but IV is consistent, so $q = b - \hat{\beta} \to c \neq 0$ in large samples.
- Stata command hausman implements the procedure
- HGL gives alternative implementation adding residuals from first-stage regression to OLS of original equation and testing whether they are significant

- **Tests for instrument validity**
  - Is z correlated with e?
    - Can't do direct test because we can't get consistent estimators for e without valid instruments, and we can't know whether instruments are valid without consistent estimator of e.
    - With extra instruments (overidentified model), we can use some to test the others.
  - **LM test:** Do 2SLS/IV, get residuals, regress $\hat{e}$ on all z instruments and exogenous regressors, under null hypothesis that all instruments are valid, NR^2 from this regression ~ χ^2 with $L - B$ degrees of freedom.
  - The ***J* statistic** is another common test of overidentifying restrictions:
    - As above, regress the 2SLS/IV residuals on the exogenous variables in the equation and all the instruments.
    - Compute the F statistic for the null hypothesis that the coefficients on the instruments are zero.
    - The test statistic LF (where L is the number of instruments) is asymptotically distributed as a χ^2 with $L - B$ degrees of freedom (number of instruments – number of endogenous regressors = number of overidentifying restrictions to be tested).
    - Why does the J test or the LM test work?
      - If the instruments are exogenous, then they should not be correlated with y except through their effects on x.
      - The 2SLS residuals are the part of y that is orthogonal to the part of z that works through x.
      - If that is the only correlation that z has with y (there is no direct effect either direction), then the residuals should be uncorrelated with z, conditional on the other x variables, the included exogenous variables.
  - Rejection of the null hypothesis tells us that at least one of the overidentifying restrictions does not hold, which may mean that one or more of the instruments is invalid.