Section 10 Endogenous Regressors and Simultaneous Equations

The most crucial of our OLS assumptions (which carry over to most of the other estimators that we have studied) is that the regressors be *exogenous*—uncorrelated with the error term. This assumption is violated if we have "reverse causality" in which $u \uparrow \rightarrow Y \uparrow \rightarrow X \uparrow\downarrow$.

System estimation vs. single-equation

- The first essential question to ask in a situation where the regressor may be endogenous is "What is the model that determines the endogenous regressor?"
    - This question, which must be answered at least partially to use any of the techniques in this section, suggests that our single econometric equation should be thought of as part of a *system of simultaneous equations* that jointly determine both our *Y* and our endogenous *X* variables.
    - For example, one of the most common applications in economics is attempting to estimate a demand curve: quantity is a function of price.
        - However, shocks to demand (*u*) affect price, so price cannot generally be taken as exogenous.
        - The demand curve is part of a system of simultaneous equations along with the supply curve that jointly determine quantity and price.
    - Thinking of the joint determination of *Y* and (at least some) *X* focuses our attention on a crucial set of variables: the exogenous variables that are in the "other" equation that determines *X* but that are not in the equation as separate determinants of *Y*.
        - Whether we end up modeling the second equation explicitly or not, these variables are crucial to *identifying* the effects of *X* on *Y*.
- The two main approaches to endogeneity revolve around our degree of interest in the determination of the endogenous regressors:
    - **System estimation** involves estimating a full set of equations with two or more dependent variables that are on the left-hand side of one equation and the right-hand side of others. (Example: both the supply and demand equations.)
    - **Single-equation estimation** involves estimating only the one equation of interest, but we still need to consider the variables that are in the other equation(s). (Example: estimate only the demand equation, but the exogenous variables in the supply equation are used as instruments.)

Instrumental-variables regression and 2-stage LS

- Suppose that *X* is endogenous in the regression equation $Y_i = \beta_0 + \beta_1 X_i + u_i$. Because it is endogenous (affected by *Y*), it is correlated with the error term and OLS is biased and inconsistent.
- We need an **instrumental variable** *Z* that has two properties:
    - o The instrument must be **relevant** in that it is correlated with *X*.
    - o The instrument must be **exogenous** in that it is not correlated with *u* (not affected by *Y*).
- Consider the matrix version of the model $Y = X\beta + u$. In the present case, *X* is correlated with *u*, so $\text{plim } X'u/n = \text{cov}(X, u) \neq 0$. Thus,

$$\begin{aligned}\text{plim}\,\hat{\beta}_{OLS} &= \text{plim}\left[\left(\frac{X'X}{n}\right)^{-1}\frac{X'Y}{n}\right]\\ &= \text{plim}\left[\left(\frac{X'X}{n}\right)^{-1}\frac{X'X}{n}\beta\right] + \text{plim}\left[\left(\frac{X'X}{n}\right)^{-1}\frac{X'u}{n}\right]\\ &= \beta + \text{plim}\left[\left(\frac{X'X}{n}\right)^{-1}\right]\text{plim}\left[\frac{X'u}{n}\right].\end{aligned}$$

    - o Because this last expression does not equal zero, the OLS estimator is inconsistent.
    - o This expression suggests the possibility of replacing *X* in the offending expression by something uncorrelated with *u*: our instrument(s) *Z*.
        - ▪ The matrix *Z* would consist of the constant vector and the column vector of the *Z* variable itself.
        - ▪ More generally, *Z* includes all exogenous regressors in *X*, plus sufficient exogenous instruments to replace all endogenous regressors.
        - ▪ Assume for the moment that there are the same number of instruments as endogenous regressors, so that *Z* and *X* are both *n* × (*k* + 1).
- Consider the instrumental-variables estimator $\hat{\beta}_{IV} = \left(Z'X\right)^{-1} Z'Y$.

$$\begin{aligned}\text{plim}\,\hat{\beta}_{IV} &= \text{plim}\left[\left(\frac{Z'X}{n}\right)^{-1}\frac{Z'Y}{n}\right]\\ &= \text{plim}\left[\left(\frac{Z'X}{n}\right)^{-1}\frac{Z'X}{n}\beta\right] + \text{plim}\left[\left(\frac{Z'X}{n}\right)^{-1}\frac{Z'u}{n}\right]\\ &= \beta + \text{plim}\left[\left(\frac{Z'X}{n}\right)^{-1}\right]\text{plim}\left[\frac{Z'u}{n}\right] = \beta.\end{aligned}$$

    - o The last equality is assured because $\text{plim}\left[\frac{Z'u}{n}\right] = \begin{pmatrix} E(u) \\ \text{cov}(Z,u)\end{pmatrix} = \vec{0}.$

    - This shows that it is crucial that the instrument *Z* be uncorrelated with the error term.
  - In order to calculate the IV estimator, we need to compute the inverse of *Z′X*. This inverse will not exist if *Z* is uncorrelated with *X*. (It also requires that *X* and *Z* have the same dimensions.)
- Instrumental-variables estimator via method of moments
  - We can use the moment condition $E(Z'u)=0$ to motivate the instrumental-variables estimator via methods of moments.
  - This estimator sets $Z'\left(Y - X\hat{\beta}\right)=\vec{0}$, which yields two equations for the two elements of $\hat{\beta}$.
  - We have assumed that $k = 1$, so that there is one (endogenous) regressor and one instrument.
    - This equation directly generalizes to multiple regressors, both endogenous and exogenous.
    - All exogenous regressors plus the instruments are in *Z*; all endogenous and exogenous regressors are in *X*.
    - If the number of instruments equals the number of endogenous regressors, then the method of moments matrix equation has $k + 1$ individual linear equations to estimate the $k + 1$ elements of $\hat{\beta}$.
  - Note that if you had an extra instrument, so that *Z* was $n \times (k + 2)$, you would have $k + 2$ linear equations to determine the $k + 1$ elements of $\hat{\beta}$, so the system would be over-determined (overidentified).
    - We will return to this case when we discuss the identification problem in detail.
- **Two-stage least squares**
  - We usually implement a variant of IV regression using a technique called two-stage LS.
  - This can be implemented as two OLS regressions:
    - First, regress *X* on *Z* to get a "predicted" $\hat{X}$ that is independent of the error term *u*.
    - Second, regress *Y* on $\hat{X}$ to estimate β.
      - Standard errors for the second stage must be adjusted for the fact that you are using $\hat{X}$ rather than *X*.
      - SER is calculated using sum of squares of $\hat{u} = Y - X\hat{\beta},$ not $Y - \hat{X}\hat{\beta}$.
  - Stata implements 2SLS as ivregress 2sls depvar exvars (endvars = instvars) , options

Simultaneous equations and the identification problem

- In the simple case above, we had one endogenous variable on the right-hand side and one exogenous variable available to act as an instrument.
  - In the more general case, there may be multiple endogenous variables and multiple instruments.
  - This forces us to think about the problem of whether there is sufficient exogenous variation to *identify* the coefficients we want to estimate: the **identification problem**.
- We will examine an extended example of a set of supply and demand curves to explore the identification problem.
  - **Model I:**

    Demand curve: $Q = \alpha_0 + \alpha_P P + u$

    Supply curve: $Q = \beta_0 + \beta_P P + v$

    - Solving for the **reduced form:**

      $$\beta_0 + \beta_P P + v = \alpha_0 + \alpha_P P + u$$

      $$(\beta_P - \alpha_P) P = (\alpha_0 - \beta_0) + (u - v)$$

      $$P = \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{u - v}{\beta_P - \alpha_P} \equiv \pi_{P0} + \varepsilon_P,$$

      $$Q = \alpha_0 + \alpha_P \left[\frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{u - v}{\beta_P - \alpha_P} \right] + u$$

      $$Q = \frac{\alpha_0 (\beta_P - \alpha_P) + \alpha_P (\alpha_0 - \beta_0)}{\beta_P - \alpha_P} + \frac{u(\beta_P - \alpha_P) + \alpha_P (u - v)}{\beta_P - \alpha_P}$$

      $$Q = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} + \frac{\beta_P u - \alpha_P v}{\beta_P - \alpha_P} \equiv \pi_{Q0} + \varepsilon_Q.$$

    - The equations

      $P = \pi_{P,0} + \varepsilon_P$

      $Q = \pi_{Q,0} + \varepsilon_Q$

      are called the reduced-form equations. We have solved the system of simultaneous linear equations for separate linear equations each of which has an endogenous variable on the left and none on the right.
    - The π coefficients are the **reduced-form coefficients**: they are nonlinear combinations of the **structural coefficients** α and β.
    - We can estimate the reduced-form coefficients by OLS because there are no endogenous variables on the right-hand side.
    - In this case, there are no variables at all on the RHS! We can estimate $\pi_{P,0}$ and $\pi_{Q,0}$ as the means of *P* and *Q*.

- Does this give us enough information to **identify** the α and β parameters?
- No. There are four structural coefficients (two α and two β) and only two reduced-form coefficients (π). There is no way to construct a unique estimator of and of the α or β coefficients from the estimate of π.
- Thus, in Model I *neither of the equations is identified.*

▪ Show graph: all variation in *P* and *Q* are due to unobserved error terms.

o **Model II:**

Demand curve: $Q = \alpha_0 + \alpha_P P + \alpha_M M + u$, where *M* is income and is exogenous

Supply curve: $Q = \beta_0 + \beta_P P + v$

▪ Solving for the reduced form:

$$\beta_0 + \beta_P P + v = \alpha_0 + \alpha_P P + \alpha_M M + u$$

$$P = \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M}{\beta_P - \alpha_P} M + \frac{u - v}{\beta_P - \alpha_P} \equiv \pi_{P0} + \pi_{PM} M + \varepsilon_P,$$

$$Q = \beta_0 + \beta_P \left[\frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M}{\beta_P - \alpha_P} M + \frac{u - v}{\beta_P - \alpha_P} \right] + v$$

$$Q = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M \beta_P}{\beta_P - \alpha_P} M + \frac{\beta_P u - \alpha_P v}{\beta_P - \alpha_P} \equiv \pi_{Q0} + \pi_{QM} M + \varepsilon_Q.$$

▪ Suppose we estimate the four reduced-form coefficients π_{P0}, π_{PM}, π_{Q0}, π_{QM} by OLS. Can we identify the five structural coefficients?

- Obviously not: can't identify five coefficients uniquely from four.
- However, we can identify *some of them*:

$$\frac{\pi_{QM}}{\pi_{PM}} = \frac{\dfrac{\alpha_M \beta_P}{\beta_P - \alpha_P}}{\dfrac{\alpha_M}{\beta_P - \alpha_P}} = \beta_P$$

$$\pi_{Q0} - \beta_P \pi_{P0} = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} - \beta_P \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} = \beta_0.$$

  o This is called **indirect least squares** and is an antiquated method for estimating these models.

- The presence of the income term in the demand equation *identifies* the slope and intercept of the supply equation. Changes in income affect demand but not supply, so we can use these changes to trace out the slope of the supply curve. How much does an increase income affect *P* and how much does it affect *Q*?

            - The demand equation is **just identified** because there is only one way of extracting the structural parameters from the reduced-form parameters.
            - 2SLS of the supply equation using income as an instrument gives us the same estimator as ILS in the just-identified case.
        - The demand equation is not identified: the only variation in the supply curve is the unobserved random shock.
        - What would happen if income also affected supply?
- **Model III:**

  Demand curve: $Q = \alpha_0 + \alpha_P P + \alpha_M M + u$

  Supply curve: $Q = \beta_0 + \beta_P P + \beta_M M + v$

    - Solving for the reduced form:

      $$\beta_0 + \beta_P P + \beta_M M + v = \alpha_0 + \alpha_P P + \alpha_M M + u$$

      $$P = \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M - \beta_M}{\beta_P - \alpha_P} M + \frac{u - v}{\beta_P - \alpha_P} \equiv \pi_{P0} + \pi_{PM} M + \varepsilon_P,$$

      $$Q = \beta_0 + \beta_P \left[\frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M - \beta_M}{\beta_P - \alpha_P} M + \frac{u - v}{\beta_P - \alpha_P} \right] + v$$

      $$Q = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M \beta_P - \beta_M \alpha_P}{\beta_P - \alpha_P} M + \frac{\beta_P u - \alpha_P v}{\beta_P - \alpha_P} \equiv \pi_{Q0} + \pi_{QM} M + \varepsilon_Q.$$

    - It's no longer possible to identify either equation. None of the six structural coefficients can be identified from estimates of the four reduced-form coefficients.
    - We can no longer use changes in *M* to trace out either curve because it affects both curves.
    - Note that nothing in the data has changed: we have merely changed our *assumption* (lens analogy) about how the data were generated.
        - If the assumption in Model II that income does not affect supply is incorrect, our estimates of the supply curve would be nonsense.
- **Model IV:**

  Demand curve: $Q = \alpha_0 + \alpha_P P + \alpha_M M + u$

  Supply curve: $Q = \beta_0 + \beta_P P + \beta_R R + v$, where *R* is rainfall (exogenous)

    - Solving for the reduced form:

      $$\beta_0 + \beta_P P + \beta_R R + v = \alpha_0 + \alpha_P P + \alpha_M M + u$$

      $$P = \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M}{\beta_P - \alpha_P} M - \frac{\beta_R}{\beta_P - \alpha_P} R + \frac{u - v}{\beta_P - \alpha_P} \equiv \pi_{P0} + \pi_{PM} M + \pi_{PR} R + \varepsilon_P,$$

$$Q = \beta_0 + \beta_P \left[\frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M}{\beta_P - \alpha_P} M - \frac{\beta_R}{\beta_P - \alpha_P} R + \frac{u - v}{\beta_P - \alpha_P} \right] + \beta_R R + v$$

$$Q = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} + \frac{\alpha_M \beta_P}{\beta_P - \alpha_P} M - \frac{\beta_R \alpha_P}{\beta_P - \alpha_P} R + \frac{\beta_P u - \alpha_P v}{\beta_P - \alpha_P} \equiv \pi_{Q0} + \pi_{QM} M + \pi_{QR} R + \varepsilon_Q .$$

- There are now six estimable coefficients and six structural coefficients we would like to estimate. Just identification of all coefficients is possible based on the numbers.
- In fact, as before,

$$\frac{\pi_{QM}}{\pi_{PM}} = \frac{\frac{\alpha_M \beta_P}{\beta_P - \alpha_P}}{\frac{\alpha_M}{\beta_P - \alpha_P}} = \beta_P$$

$$\pi_{Q0} - \beta_P \pi_{P0} = \frac{\beta_P \alpha_0 - \alpha_P \beta_0}{\beta_P - \alpha_P} - \beta_P \frac{\alpha_0 - \beta_0}{\beta_P - \alpha_P} = \beta_0 .$$

- Now, we can do the same thing with the rainfall coefficients:

$$\frac{\pi_{QR}}{\pi_{PR}} = \frac{\frac{\alpha_P \beta_R}{\beta_P - \alpha_P}}{\frac{\beta_R}{\beta_P - \alpha_P}} = \alpha_P$$

$$-\pi_{P0}\left(\beta_P - \alpha_P\right) + \beta_0 = \alpha_0 .$$

$$\pi_{PM}\left(\beta_P - \alpha_P\right) = \alpha_M$$

$$-\pi_{RM}\left(\beta_P - \alpha_P\right) = \beta_R$$

- Both equations are just identified:
  - Rainfall identifies the demand equation because it is exogenous, affects the endogenous variable price, and is not in the demand equation on its own.
  - Income identifies the supply equation because it is exogenous, affects the endogenous variable price, and is not in the supply equation on its own.
- Again, 2SLS gives us the same estimators as ILS in the just-identified case:
  - ivregress 2sls q m (p = r) to estimate the demand equation
  - ivregress 2sls q r (p = m) to estimate the supply equation

o **Model V:**

Demand curve: $Q = \alpha_0 + \alpha_P P + \alpha_M M + u$

Supply curve: $Q = \beta_0 + \beta_P P + \beta_R R + \beta_W W + v$, where W is wages (exogenous)

  - We now have *two* exogenous variables in the supply equation that are not in the demand equation. Two alternative ways of identifying the demand curve.
  - Solving for the reduced form:

$$\beta_0+\beta_P P+\beta_R R+\beta_W W+v=\alpha_0+\alpha_P P+\alpha_M M+u$$

$$P=\frac{\alpha_0-\beta_0}{\beta_P-\alpha_P}+\frac{\alpha_M}{\beta_P-\alpha_P}M-\frac{\beta_R}{\beta_P-\alpha_P}R-\frac{\beta_W}{\beta_P-\alpha_P}+\frac{u-v}{\beta_P-\alpha_P}$$

$$P\equiv\pi_{P0}+\pi_{PM}M+\pi_{PR}R+\pi_{PW}W+\varepsilon_P,$$

$$Q=\beta_0+\beta_P\left[\frac{\alpha_0-\beta_0}{\beta_P-\alpha_P}+\frac{\alpha_M}{\beta_P-\alpha_P}M-\frac{\beta_R}{\beta_P-\alpha_P}R-\frac{\beta_W}{\beta_P-\alpha_P}W+\frac{u-v}{\beta_P-\alpha_P}\right]+\beta_R R+\beta_W W+v$$

$$Q=\frac{\beta_P\alpha_0-\alpha_P\beta_0}{\beta_P-\alpha_P}+\frac{\alpha_M\beta_P}{\beta_P-\alpha_P}M-\frac{\beta_R\alpha_P}{\beta_P-\alpha_P}R-\frac{\beta_W\alpha_P}{\beta_P-\alpha_P}W+\frac{\beta_P u-\alpha_P v}{\beta_P-\alpha_P}$$

$$Q\equiv\pi_{Q0}+\pi_{QM}M+\pi_{QR}R+\pi_{QW}W+\varepsilon_Q.$$

  - There are now eight estimable reduced-form coefficients and seven structural coefficients.
  - All six of the equations that we used in Model IV to get the those six coefficients still work.
  - Now we can estimate α_P *either* as $\alpha_P=\dfrac{\pi_{QR}}{\pi_{PR}}$ or as $\alpha_P=\dfrac{\pi_{QW}}{\pi_{PW}}$.
    - The demand equation is now **overidentified** because there are two exogenous variables that affect the single endogenous variable *P* that are not separately in the demand equation.
    - Will they be the same? Will $\dfrac{\pi_{QR}}{\pi_{PR}}=\dfrac{\pi_{QW}}{\pi_{PW}}$?
    - Generally they won't be identical even if the model is correct because of sampling error. Is there more inequality than would be expected randomly?
    - We can test this nonlinear null hypothesis.
      - If the model is valid, we should not be able to reject this null hypothesis.
      - Rejecting these **overidentifying restrictions** suggests that the model is not valid.
  - There are two different ILS estimates for the coefficients of the demand equation. 2SLS will be a combination of them.
    - Estimate demand equation by ivregress 2sls q m (p = r w)
    - The instrument used is the prediction of *Q* based on *R* and *W*.
- Note several properties of identification

  - Identification is usually by equation/coefficient, not necessarily of the whole system.
    - It's possible to have one equation that is identified with others not.
    - If there are multiple endogenous regressors it is possible to have one identified and others not.
  - Identification depends crucially on three assumptions:
    - That the instrument is exogenous
    - That the instrument does not itself appear in the equation
    - That the instrument does appear in another equation that influences the endogenous regressor
    - If any of these assumptions is violated, then the 2SLS estimator is biased and inconsistent.
  - In general, there needs to be one omitted exogenous variable for each included endogenous variable. (Order condition for identification)
    - However, if you have two instruments that are correlated with one endogenous variable but neither is correlated with the other, then identification of the second endogenous regressor fails.
    - Order is not enough; the rank condition applies as well.
- Good instrumental variables
  - Instrument **relevance**
    - Weak instruments explain little of the variation in X, the endogenous regressor.
    - Rule of thumb test for weak instruments: F test of zero restriction on all instruments in the first-stage regression is less than 10.
  - Instrument **exogeneity**
    - With just-identification, you cannot test for exogeneity.
    - When there are enough instruments that the model is overidentified, you can test the overidentifying restrictions as discussed above.
    - The J statistic is a common test of overidentifying restrictions:
      - Regress the 2SLS 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 mF (where m is the number of instruments) is asymptotically distributed as a χ^2 with $m - k$ degrees of freedom (number of instruments – number of endogenous regressors = number of overidentifying restrictions to be tested).
      - Why does the J 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 *W*, the included exogenous variables.

- Matrix notation of the 2SLS estimator
    - Consider our 2-equation system of Model V.
        - Let $Y = [Q \;\; P]$ be an $n \times 2$ matrix of the two endogenous variables.
        - Let $Z = [1 \;\; M \;\; R \;\; W]$ be an $n \times 4$ matrix of the four exogenous variables (which are instruments for one equation and included exogenous variables for the other).
        - Let $e = [u \;\; v]$ be an $n \times 2$ matrix of error terms, which are probably correlated within a single observation.
        - Let Γ be the 2 × 2 matrix of coefficients applied to the endogenous variables (which will often be 1 or –1 by normalization)
        - Let Β be the 4 × 2 matrix of coefficients applied to the exogenous variables, which must have some elements that are known to be zero in order for identification to be achieved.
        - The two equations of the model can be written as $Y\Gamma + Z\mathrm{B} = e,$ where

$$Y = \begin{pmatrix} Q_1 & P_1 \\ Q_2 & P_2 \\ \vdots & \vdots \\ Q_n & P_n \end{pmatrix}, \quad \Gamma = \begin{pmatrix} -1 & -1 \\ \alpha_P & \beta_P \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & M_1 & R_1 & W_1 \\ 1 & M_2 & R_2 & W_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & M_n & R_n & W_m \end{pmatrix},$$

$$\mathrm{B} = \begin{pmatrix} \alpha_0 & \beta_0 \\ \alpha_M & 0 \\ 0 & \beta_R \\ 0 & \beta_W \end{pmatrix}, \quad e = \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \\ \vdots & \vdots \\ u_n & v_n \end{pmatrix}.$$

        - The reduced-form equations are obtained by post-multiplying the equation by the inverse of Γ (which must exist for the model to be solvable): $Y\Gamma\Gamma^{-1} + Z\mathrm{B}\Gamma^{-1} = e\Gamma^{-1},$ or $Y = -Z\mathrm{B}\Gamma^{-1} + e\Gamma^{-1} \equiv Z\Pi + \varepsilon,$ where

$$\Pi \equiv -\mathrm{B}\Gamma^{-1} \text{ and } \varepsilon \equiv e\Gamma^{-1} = \begin{pmatrix} \varepsilon_{Q1} & \varepsilon_{P1} \\ \varepsilon_{Q2} & \varepsilon_{P2} \\ \vdots & \vdots \\ \varepsilon_{Qn} & \varepsilon_{Pn} \end{pmatrix}.$$

        - If the system is identified, then there are enough restrictions on the Γ and Β matrices (five in the model above—two –1s and three 0s) to assure that the remaining elements can be obtained uniquely from the Π matrix.

- Estimating overidentified systems via instrumental variables
  - Suppose that the matrix *Z* of exogenous variables and instruments is *n* × (*m* + 1) and that there are only *k* < *m* total endogenous and exogenous regressors in the equation. This is an overidentified model.
  - The *X* matrix is only *n* × (*k* + 1), so *Z′X* in the IV estimator equation is *m* + 1 × *k* + 1: it is not square and doesn't have an inverse.
  - The moment condition $Z'\left(Y - X\hat{\beta}\right) = \vec{0}$ is *m* + 1 equations in *k* + 1 elements of $\hat{\beta}$.
    - As noted above, in the overidentified case the coefficients are over-determined.
    - If there were no sampling variation and the model was perfectly specified, then the *m* + 1 equations would all hold exactly—they would be completely consistent with one another.
  - **Generalized method of moments**
    - Since we cannot generally achieve $Z'\left(Y - X\hat{\beta}\right) = \vec{0}$, it makes sense to make $Z'\left(Y - X\hat{\beta}\right)$ as close to zero as possible.
    - But recall that $Z'\left(Y - X\hat{\beta}\right)$ contains *m* + 1 elements whose units vary with the units of the columns of *Z*.
      - It's not obvious that we should just square these elements and "add them up".
      - Instead, we might want to weight the elements of $Z'\left(Y - X\hat{\beta}\right)$ to assure that they are commensurate.
    - **Quadratic forms**
      - If *x* is a vector with *m* + 1 elements and $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} & a_{1,m+1} \\ a_{12} & a_{22} & \cdots & a_{2m} & a_{2,m+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{mm} & a_{m,m+1} \\ a_{1,m+1} & a_{2,m+1} & \cdots & a_{m,m+1} & a_{m+1,m+1} \end{pmatrix}$ is a positive semi-definite matrix, then $x'Ax = \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} a_{ij}x_i x_j$. In other words, the quadratic form $x'Ax$ sums the squares and cross-products (all the second-order terms) of the elements of *x*, with each square and cross-product bearing a coefficient of a_{ij}. The results of a quadratic form is a scalar.

        - GMM estimators minimize a quadratic form involving the squares and cross-products of $Z'\left(Y - X\hat{\beta}\right)$. Denoting the weighting matrix as *A*, we choose $\hat{\beta}_A^{GMM}$ to minimize $\left(Y - X\hat{\beta}_A^{GMM}\right)' ZAZ'\left(Y - X\hat{\beta}_A^{GMM}\right)$.
        - The 2SLS estimator is a GMM estimator with $A = \left(Z'Z\right)^{-1}$.

Estimation of systems of equations

- The method of instrumental variables offers us a means to estimate a single equation from a larger system of simultaneous equations. Sometimes we want or need to estimate the entire system.
    - Estimates are generally more efficient if all equations are estimated together.
        - Taking account of the correlation between the error terms is beneficial
        - Suppose that we know that the error terms are positively correlated and that equation 2 seems to have a large positive error for observation *i*
        - Joint estimation allows us to take account of the likely positive error in equation 1 and not attempt to fit the outlying observation too closely
        - Adds information and thus improves efficiency
    - We may want to impose and/or test coefficient restrictions across the equations of a system.
        - Demand equations derived from a common utility function (or factor demands from a common cost function) have cross-equation "symmetry" restrictions. (The Slutsky condition for demand says that the income-compensated cross-price elasticity of demand for *x* with respect to the price of *y* equals the elasticity of demand for *y* with respect to the price of *x*.)
        - Might want to test whether the income elasticity of demand for apples exceeds that of bananas.
        - Might want to test whether all of the coefficients of the demand for apples are the same as those of bananas so that we can aggregate them together
- Two kinds of joint-system estimation
    - **Seemingly unrelated regressions (SUR)** (also called Zellner-efficient regression)
        - System of equations with no endogenous variables on right-hand side
        - Efficiency gains from taking account of correlation of error
        - Possibility of testing/imposing cross-equation coefficient restrictions
    - **Three-state least squares (3SLS)**
        - System of equations with endogenous regressors
        - Example would be estimating both demand and supply equations together

        - Adds efficiency gains (or cross-equation tests) to 2SLS/IV consistent estimator of equation(s) with endogenous regressors

- Estimation by seemingly unrelated regressions
    - Here we have a set of equations that have no endogenous regressors, but we want to estimate the equations jointly.
    - We can do this by "stacking" the regressions:
        - Suppose that there are 3 equations to be jointly estimated with dependent variables Y_1, Y_2, and Y_3, sets of regressors (which might overlap) X_1, X_2, and X_3, and error terms u_1, u_2, and u_3.
        - Separately, the equations can be written
          $$Y_1 = X_1\beta_1 + u_1,$$
          $$Y_2 = X_2\beta_2 + u_2,$$
          $$Y_3 = X_3\beta_3 + u_3.$$
        - Let Y be the $3n \times 1$ element column vector that stacks the 3 Y vectors:
          $$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}.$$
        - Let X be the $3n \times (k_1 + k_2 + k_3 + 3)$ matrix $X = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{pmatrix}$.
        - Let u be the $3n \times 1$ element column vector $u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$.
        - Let β be the $(k_1 + k_2 + k_3 + 3) \times 1$ element vector $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.
        - We can then write the combined system of equations as $Y = X\beta + u$.
    - Can we estimate this system by OLS?
        - Yes, except this not efficient because of probably correlation between the *i*th observation's error term across equations.
    - Specification of error term
        - If observations are IID, then, $\text{cov}(u_{mi}, u_{lj}) = 0$ if $i \neq j$. (First subscript is equation; second is the observation.)
        - However, it is likely that within each observation, $\text{cov}(u_{mi}, u_{li}) = \sigma_{ml} \neq 0$.
        - Heteroskedasticity is also almost certain since we have different dependent variables for each third of the stacked regression.

  - Let $\Sigma_u \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}$
  - Assume that there is no correlation across observations either within any of the equations or between them.
  - Then the covariance matrix of the stacked error term is the $3n \times 3n$ matrix $\Omega = \Sigma \otimes I_n = \begin{pmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \sigma_{13}I_n \\ \sigma_{12}I_n & \sigma_{22}I_n & \sigma_{23}I_n \\ \sigma_{13}I_n & \sigma_{23}I_n & \sigma_{33}I_n \end{pmatrix}$.
- Given the non-scalar covariance matrix, this is another potential application of generalized least square: $\hat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$.
  - (This general formula specializes to weighted least-squares when there is heteroskedasticity but no autocorrelation. We will also see a GLS application for serial correlation of the error.)
- Of course, we don't know σ_{ml}, so we must estimate it.
  - We can do so based on OLS residuals because OLS is consistent (if not efficient).
- SUR is a two-step procedure
  - First estimate the three regressions by OLS and calculate the residual vectors $\hat{u}_1, \hat{u}_2$, and $\hat{u}_3$.
  - Next estimate $\hat{\sigma}_{ml} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_{mi}\hat{u}_{li}$, $m = 1, 2, 3$; $l = 1, 2, 3$, and assemble these estimators into $\hat{\Sigma}_u \equiv \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} & \hat{\sigma}_{23} \\ \hat{\sigma}_{13} & \hat{\sigma}_{23} & \hat{\sigma}_{33} \end{pmatrix}$ and $\hat{\Omega} \equiv \hat{\Sigma}_u \otimes I_n$.
  - Finally, use "feasible" GLS to estimate β as $\hat{\beta}_{FGLS} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}Y$.
- This procedure can be iterated:
  - Because $\hat{\beta}_{FGLS}$ is a more efficient estimator than the OLS estimator, we should get "better" residuals be calculating them based on $\hat{\beta}_{FGLS}$ rather than on OLS.
  - Iterated seemingly unrelated regressions (ISUR) repeatedly re-estimates σ_{ml} based on the FGLS coefficient estimator, then recalculates $\hat{\Omega}$ and re-estimates β by FGLS.
  - This can be repeated over and over until the elements of $\hat{\Omega}$ do not change from iteration to iteration.

  - SUR is more efficient than separate OLS except in two situations (in which they are identical):
    - First, there is no correlation between error terms across equations. In other words, $\sigma_{ml} = 0$ for all $m \neq l$.
    - Second, the same regressors appear in all equations: $X_m = X_l$ for all m, l.
  - In Stata, we use sureg (dvar1 indvars1) (dvar2 indvars2) (dvar3 indvars3)
    - The option isure iterates to convergence.
    - The option constraints ([dvar1]indvar1j = [dvar2]indvar2j) imposes the constraint that the indvar1j coefficient in the equation for dvar1 equals the indvar2j coefficient in the equation for dvar2.
    - If constraints are complex, can also use
      constraint 1 [dvar1]indvar1j = [dvar2]indvar2j
      constraint 2 [dvar2]indvar2j = [dvar3]indvar3j
      sureg (dvar1 indvars1) (dvar2 indvars2) (dvar3 indvars3), constraint(1 2)
- Estimation by three-stage least squares
  - If endogenous variables appear on the RHS of equations, then we must combine the system estimation of SUR with the instrumental variables method of 2SLS.
  - The resulting estimator is 3SLS:
    - Estimate the reduced-form equations by OLS.
      - Don't need SUR because all exogenous variables in the system appear in each equation, so the X_m matrices are identical and OLS is equivalent to SUR.
      - Calculate fitted values of the endogenous variables based in the reduced-form regressions on the exogenous variables as in 2SLS.
    - Estimate the individual equations by 2SLS, using their fitted values in place of the endogenous regressors.
      - Calculate the residuals of each equation from the 2SLS regressions.
      - Calculate estimates of σ_{ml} and assemble them into $\hat{\Omega}$.
    - Estimate the system of equations jointly by FGSL using the estimated $\hat{\Omega}$.
      - As with SUR, this can be iterated.
  - 3SLS has the same advantages relative to 2SLS that SUR has relative to OLS:
    - Efficiency gain by taking account of cross-equation correlation of error (if it exists)
    - Possibility of imposing or testing cross-equation coefficient restrictions
  - Stata will do 3SLS using the reg3 command, which combines the forms of the ivregress (with endregrs = instvars in the variable list of each regression) and multiple equations enclosed in parentheses.
- **Maximum-likelihood estimators for simultaneous equations**
  - Unlike OLS, 2SLS is *not* an MLE, nor is 3SLS.

- There are MLEs that apply to these models under (usually) the normal distribution.
- **Limited-information maximum likelihood** is a single-equation MLE that is analogous to 2SLS.
- **Full-information maximum likelihood** is the multiple-equation MLE analog of 3SLS.
- Stata will do LIML (and GMM) by changing the 2SLS option in the ivregress command to liml or gmm.
    - Each of these has different options that will need to be set.
- I can't find any FIML procedure in Stata, but it may be possible to program it in the general-purpose ml command.

Experimental and quasi-experimental data

- Random, controlled experiments
    - Individuals for treatment and control groups are selected randomly
    - Often use "double-blind" technique in medical treatments where neither the patient nor the doctor knows which group the patient is in: avoiding the **Hawthorne effect**.
    - If selection is truly random, then it is uncorrelated with other variables and we have no omitted variable bias from leaving out other variables:
        - Can just do a *t* test of the means: $\mu_{treatment} = \mu_{control}$.
        - This is equivalent to a regression on a treatment dummy and a *t* test of the coefficient.
        - This is the "differences" estimator
- Problems with experiments:
    - Lack of randomization can lead to correlation between group selection and other variables.
        - Can control for this by controlling for these variables by including them in a regression.
    - Partial compliance
        - Did the treatment and control groups actually do what they were supposed to do?
        - Did the job-training selectees actually attend training?
        - Did the patient take the drug?
        - Is this behavior correlated with *u*?
    - Attrition
        - Some drop out of both groups during the experiment.
        - Were they random or did people with high (or low) values of *u* drop out?
    - Hawthorne effect
        - Double-blind is not possible in many experiments.

        - Experimenter bias may result from incentives to make results look significant.
- External validity issues
    - Is sample representative?
    - Is the treatment representative of real-world applications it is designed to simulate/test?
    - Will large-scale application of the test program affect other markets in ways that can't be predicted from small-scale experiment?
    - Is the effect the result of treatment or eligibility for treatment?
- Can add regressors to differences estimator in order to control for variables that may be correlated with group selection.
    - Allows for "conditional random assignment" based on control variables *W*
    - Need $E\left(u_i \mid X_i, W_i\right) = W_i \gamma$ independent of X_i.
        - This assumes that the coefficient on *X* is consistent, though coefficients on *W* will not be.
- Differences-in-differences estimator
    - We encountered a non-experimental version of this estimator in our panel-data discussion.
        - It had dummies for both time and unit, so that only the differences across units in the changes over time were used to estimate the coefficients.
    - In an experimental setting, the d-in-d estimator is $\left(\overline{Y}^{treatment,\, after} - \overline{Y}^{treatment,\, before}\right) - \left(\overline{Y}^{control,\, after} - \overline{Y}^{control,\, before}\right)$
    - This estimator controls for changes in *Y* over time (during the experiment) that would be common to both groups.
        - For example, incomes generally grow over time, so just seeing that the treatment group had higher incomes would not be enough, we would have to see that the *difference* in their income over time is greater than the *difference* in incomes of the control group.
    - You need to add regressors (*W*) if the general *change* in *Y* from before to after is affected by these variables.
- **Quasi-experiments**
    - Natural experiments are frequently a source of economic data
    - May be effectively random selection if unrelated to variables of interest
    - Selection may be partially random and partially related to variables we can observe, which must then be used as instruments or controls.
    - Example: assignment to Hum 110 sections
        - Largely random (fully so at Whitman)
        - Class conflicts affect assignment, but can be controlled for with time-of-day dummies.