1 Endogeneity

1. Formally, the problem is that, in a model

$$Y = g(X, \beta) + \varepsilon,$$

the disturbances are *endogenous*, or equivalently, correlated with the regressors, as in

$$E[X'\varepsilon] \neq 0$$

In the Venn Diagram (Ballentine) on page 167 of Kennedy, we get a picture of this. There is a variable X which does covary with Y (red+blue+purple), but sadly some of the covariance with Y is through covariance with the error term (red). This covariance leads to bias in the OLS estimator.

2. Why do you get bias?

   (a) Covariation of Y with the disturbance term, which is correlated with X, is attributed to X. Consider the case where

$$Y = X\beta + \varepsilon,\ \varepsilon = X\Gamma + \eta \Leftrightarrow Y = X(\beta + \Gamma) + \eta.$$

   (b) The regression *loads* the response of Y to X entirely on to X. But in reality, the response of Y to X has two channels: the direct channel through β, and the indirect channel through Γ. The indirect channel is through the disturbance term: it is the derivative of ε with respect to X.

   (c) Think of the simplest possible regression model, where X is a univariate continuous variable in $[-1, 1]$ and the true coefficient is 1, but where

$$E[X'\varepsilon] = 1.$$

   Here, X is positively correlated with the disturbance, which has expectation 1 when $X = 1$, and expectation -1 when $X = -1$. Draw the picture and you will see that the OLS estimator gives you the wrong slope. The reason is that it assigns variation that is due to the covariance of X and the disturbance to covariation between X and Y. Here, it will give you an estimate of 2 rather than 1.

   (d) The covariance between X and the disturbance pollutes the OLS estimator with variation that violates its identifying assumptions.

3. What causes endogeneity?

   (a) Nothing need cause endogeneity. It is just about covariances.

(b) *Simultaneity*: Equations are said to be simultaneous if stuff on the RHS in one equation shows up in the LHS in other equation(s). For example, if

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ X &= Y\alpha + Z\Gamma, \end{aligned}$$

then, substituting into X yields

$$\begin{aligned} X &= X\beta\alpha + Z\Gamma + \varepsilon\alpha \\ X - X\beta\alpha &= Z\Gamma + \varepsilon\alpha \\ X &= [I - \beta\alpha]^{-1}(Z\Gamma + \varepsilon\alpha), \end{aligned}$$

which is obviously correlated with ε. A classic reason that this could happen is if an underlying cause of the error term in your primary equation is also an underlying cause of variation in one of the X's.

(c) *Correlated Missing Variables*. If the true model is

$$\begin{aligned} Y &= X\beta + Z\Gamma + \varepsilon, \\ E\left[[X\ Z]'\varepsilon\right] &= 0 \end{aligned}$$

but we estimate a model based on

$$Y = X\beta + \varepsilon,$$

then, the expectation of the OLS estimator is

$$\begin{aligned} E\left[\hat{\beta}_{OLS}\right] &= E\left[(X'X)^{-1}X'(X\beta + Z\Gamma + \varepsilon)\right] \\ &= \beta + E\left[(X'X)^{-1}X'(Z\Gamma + \varepsilon)\right] \\ &= \beta + (X'X)^{-1}X'Z\Gamma + (X'X)^{-1}E[X'\varepsilon] \\ &= \beta + (X'X)^{-1}X'Z\Gamma \end{aligned}$$

Here, even though the ε are exogenous to the X's, there is a bias term depending on the empirical covariance between X and Z. If $X'Z = 0$, then $\hat{\beta}_{OLS}$ is unbiased. If Z is a random variable, then if $E[X'Z] = 0$, then $\hat{\beta}_{OLS}$ is unbiased. This is why only *correlated* missing variables are a problem: uncorrelated missing variables do not induce bias.

i. If X and Z are correlated, then the OLS estimator is comprised of two terms added together: (1) the true coefficient on X, and (2) the marginal effect of X on $Z\Gamma$. The latter effect may be interpreted as Γ times the regression coefficients on X in a regression of Z on X.

(d) Consider the model of earnings Y given characteristics X and schooling levels W. Let η be a measure of smartness that affects schooling choice. In a model like this, schooling will be correlated with the disturbance because smart people both get more schooling and get more money given schooling. Thus, the OLS estimate of the return to schooling is biased upwards:

$$Y = X\beta + W\delta + \varepsilon$$

where

$$W = Z\gamma + \eta$$

and the two disturbances are correlated, implying

$$E[\eta\varepsilon] \neq 0$$

Here W is correlated with the disturbance ε, which captures among other things the impact of smartness on income conditional on schooling.

(e) *Selection Bias.* Classic selection bias is when your sample selection has different characteristics from the population in a way that matters for the coefficient you are estimating. Eg, leaving out nonworkers leaves out those who get no return to schooling. Formally, the problem is that the same thing that causes selection into the sample$X$an observed wage$X$is correlated with having a higher value of the disturbance term.

4. Corrections for endogeneity bias are always going to lie in dealing with this pollution by either (1) removing it; (2) controlling for it; or (3) modelling it.

2 Dealing With Endogeneity

1. how do you solve an endogeneity problem?

   (a) Go back to the Venn Diagram. The problem is that we have red variation that OLS mistakenly attaches to X. Three general approaches are

      i. Clever Sample Selection. Drop the polluted observations of X that covary with the disturbance;

      ii. Instrumental Variables or Control Variables. In each observation, drop the polluted component of X or control for the polluted component of X.

      iii. Full Information Methods. Model the covariation of errors across the equations.

(b) Clever Sample Selection. Okay, so some of our observations have covariance of X and the disturbances. But, maybe some don't.

i. In the returns to schooling example, if some people, for example the very tall, were not allowed to choose their schooling level, but rather were just assigned an amount of time in school, then their W would be exogenous.

ii. Run a regression on the very tall only. You would have to assume that their returns to schooling are the same as everyone else's though.

iii. If you are interested, for example, in the wage loss due to layoff, you might worry that firms layoff big losers, so that layoff (on the RHS) is correlated with the disturbance in the pre-layoff wage.

iv. When firms close entire plants, they don't pick and choose who to layoff.

v. run a regression on people who were laid off due to plant closures only.

(c) Instrumental Variables.

i. Say you have a variable, called an instrument, usually labelled Z, that is correlated with the polluted X, but not correlated with the disturbance term. You could use that information instead of X. This is the purple area in the Venn Diagram.

ii. The purple area is not correlated with the disturbance term by construction. Thus, you can use it in an OLS regression.

iii. The only problem is a regression of Y on the instrument gives you the marginal impact of the instrument on Y, when you really wanted the marginal impact of X on Y. You can solve this by expressing the instrument in units of X. This is the spirit of two stage least squares.

iv. It is important to assess the strength of the instrument. After all, identification and precision is coming from the covariation of the instrument with X and Y. If there isn't much of this joint covariation, you won't get much precision. You will also get bias if the model is overidentified. These problems are called *Weak Instruments* problems.

2. Two Stage Least Squares

(a) find an *instrument* Z such that

$$E[X'Z] \neq 0, \ E[Z'\varepsilon] = 0,$$

and then use Z variation instead of X variation. The trick is to use Z variation which is correlated with X variation to learn about the marginal effect of X on Y.

(b) Denote $\widehat{X}$ as the predicted values from a regression of X on Z. Since Z is assume uncorrelated with the disturbances, $\widehat{X}$ must also be uncorrelated with the disturbances. Run a regression of Y on $\widehat{X}$. This regression will satisfy the conditions which make OLS unbiased. This regression uses only the purple area in the Venn Diagram to identify the marginal effect of X on Y. Here,

$$\hat{\beta}_{2SLS} = \left(\widehat{X}'\widehat{X}\right)^{-1}\widehat{X}'Y$$

where

$$\widehat{X} = Z\left(Z'Z\right)^{-1}Z'X.$$

(c) Since, $\widehat{X}$ is a constructed variable, the output from this OLS regression will give the wrong covariance matrix for the coefficients. The reason is that what you want is

$$\widehat{\sigma}^2\left(\widehat{X}'\widehat{X}\right)^{-1}$$

where

$$\widehat{\sigma}^2 = \left(Y - X\hat{\beta}_{2SLS}\right)'\left(Y - X\hat{\beta}_{2SLS}\right)/\left(N-K\right)$$

This is because the estimate of the disturbance is still

$$e = Y - X\hat{\beta}_{2SLS}$$

even though the coefficients are the IV coefficients. But what you'd get as output is

$$\widehat{s}^2\left(\widehat{X}'\widehat{X}\right)^{-1}$$

where

$$\widehat{s}^2 = \left(Y - \widehat{X}\hat{\beta}_{2SLS}\right)'\left(Y - \widehat{X}\hat{\beta}_{2SLS}\right)/\left(N-K\right)$$

which is not the estimated variance of the disturbance.

3. Exactly Identified (and homoskedastic) Case: Method of Moments Estimator:

   (a) Recall: a *method-of-moments* approach substitutes sample analogs into theoretical moments:

      i. a moment is an expectation of a power of a random variable.

      ii. sample analogs are observed variables that can be plugged in.

iii. For example, given the assumption that

$$E[X'\varepsilon] = 0,$$

the method-of-moments approach to estimating parameters is to substitute sample moments into the restriction:

$$X'e = 0 \Leftrightarrow X'(Y - X\beta) = 0.$$

The unique solution to this equation is the OLS estimate.

iv. The method-of-moments approach tells you that the covariance assumption on X and the disturbances is extremely closely linked to the solution of minimising vertical distances.
me

(b) Definition: *Projection Matrix*

i. For any variable W, denote the projection matrix

$$P_W \equiv W(W'W)^{-1}W',$$

and note that in OLS, we may express predicted values in terms of this projection matrix as

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = P_X Y$$

ii. projection matrices are idempotent: they equal their own square:

$$P_W P_W = P_W.$$

iii. projection matrice are symmetric, so that $P_W = P'_W$.

iv. projection matrices have rank T where T is the rank of W.

(c) You can think about Two Stage Least Squares in a method-of-moments way, too. The moment restrictions are

$$E[Z'\varepsilon] = 0$$

If Z is full rank with K columns, then we say that the model is exactly identified. Substituting in sample moments yields

$$Z'\left(Y - X\hat{\beta}_{MoM}\right) = 0$$

Solving this for the coefficients yields:

$$\widehat{\beta}_{MoM} = (Z'X)^{-1}Z'Y$$

This is sometimes referred to as *Indirect Least Squares*.

(d) Indirect Least Squares is equivalent to two-stage least squares when Z has the same rank as X:

$$\begin{aligned}
\hat{X} &= P_Z X \\
\hat{\beta}_{2SLS} &= \left(\widehat{X}'\widehat{X}\right)^{-1}\widehat{X}'Y = \\
&= \left(X'P_Z P_Z X\right)^{-1} P_Z Y = \left(X'P_Z X\right)^{-1} X'P_Z'Y \\
&= \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1}Z'Y \\
&= \left(Z'X\right)^{-1}\left(Z'Z\right)\left(X'Z\right)^{-1} X'Z\left(Z'Z\right)^{-1}Z'Y \\
&= \left(Z'X\right)^{-1}Z'Y,
\end{aligned}$$

because with square matrices, you can rearrange within the inverse.

(e) Intution: Consider a one-dimensional X and Z, and consider regressing Y on Z and X on Z:

$$\begin{aligned}
\widehat{\beta}_{OLS}^{YonZ} &= (Z'Z)^{-1}Z'Y \\
\widehat{\beta}_{OLS}^{XonZ} &= (Z'Z)^{-1}Z'X \\
\widehat{\beta}_{2SLS}^{YonX} = \widehat{\beta}_{MM}^{YonX} &= \frac{\widehat{\beta}_{OLS}^{YonZ}}{\widehat{\beta}_{OLS}^{XonZ}}
\end{aligned}$$

i. Here, we see that the IV estimator (2SLS, MM) is given by the ratio of two OLS estimators. It is the ratio of the derivatives of Y and X with respect to Z, and for that reason, we think of it is as indirectly illuminating the derivative of Y with respect to X.

ii. Thus, regressing Y on Z looks the same as 2SLS of Y on X in the Ballentines, but the Ballentine misses out the rescaling shown above.

(f) **Asymptotic** bias and variance of the indirect least squares estimator (small sample later):

i. the bias of the ILS estimator is formed by plugging Y into the estimator:

$$\begin{aligned}
E\left[\hat{\beta}_{ILS}\right] - \beta &= E\left[\left(Z'X\right)^{-1}Z'Y\right] - \beta \\
&= E\left[\left(Z'X\right)^{-1}Z'X\beta + \left(Z'X\right)^{-1}Z'\varepsilon\right] - \beta \\
&= E\left[\beta + \left(Z'X\right)^{-1}Z'\varepsilon\right] - \beta \\
&= E\left[\left(Z'X\right)^{-1}Z'\varepsilon\right] = \left(Z'X\right)^{-1}E\left[Z'\varepsilon\right] \\
&= \left(Z'X\right)^{-1}0_K = 0_K
\end{aligned}$$

ii. since the ILS estimator is asymptotically unbiased, aka, **consistent**, the variance is formed by squaring the term inside the expectation in the bias expression:

$$E\left[\left(\hat{\beta}_{ILS}-\beta\right)\left(\hat{\beta}_{ILS}-\beta\right)'\right]=E\left[\left(Z'X\right)^{-1}Z'\varepsilon\varepsilon'Z\left(Z'X\right)^{-1}\right].$$

Here, we need to assume something about $E[\varepsilon\varepsilon']$ to make progress (just as in the OLS case). So, use the usual homoskedasticity assumption:

$$E[\varepsilon\varepsilon']=\sigma^2 I_N,$$

so,

$$\begin{aligned}E\left[\left(\hat{\beta}_{ILS}-\beta\right)\left(\hat{\beta}_{ILS}-\beta\right)'\right] &= E\left[\left(Z'X\right)^{-1}Z'\varepsilon\varepsilon'Z\left(Z'X\right)^{-1}\right]\\ &= \left(Z'X\right)^{-1}Z'\sigma^2 I_N Z\left(Z'X\right)^{-1}\\ &= \sigma^2\left(Z'X\right)^{-1}Z'Z\left(Z'X\right)^{-1}.\end{aligned}$$

If $Z=X$, then we have the OLS variance matrix.

iii. This variance is similar to

$$\frac{V(\varepsilon)}{Cov(X,Z)}\frac{V(Z)}{Cov(X,Z)},$$

so, to get precise estimates, you keep the variance of the disturbance $V(\varepsilon)$ low, and the covariance between instruments and regressors $Cov(X,Z)$ high. The right-hand term is the reciprocal of the share of the variance of Z that shows up in X. If you get a big share of X in Z, the variance of the estimate is smaller.

4. Overidentified (and homoskedastic) Case: Generalised Method of Moments Estimator

   (a) If Z contains more columns than X, we say that the model is *overidentified*. In this case, you have too much information, and you can't get the moment restriction all the way to zero. Instead, you solve for the coefficients by minimising a quadratic form of the moment restriction. This is called *Generalised Method-of-Moments* (GMM). The GMM estimator is exactly equal to the two stage least squares estimator in linear homoskedastic models.

   (b) Assume Z has $J>K$ columns. GMM (Hansen 1982) proposes a criterion "get $Z'e$ as close to zero as you can by choice of the parameter vector β". Since β only has K elements and $Z'e$ has more than K elements, you can't generally get $Z'e$ all the way to zero. So, you minimize a quadratic form in $Z'e$ with some kind of weighting matrix in the middle. Consider

$$\min_{\beta} e'Z\Omega^{-1}Z'e$$

where Ω is a matrix that puts weight on the bits of $Z'e$ that you think are most important, or most informative about $Z'e$.

(c) Since $E[Z'\varepsilon] = 0$, it doesn't matter asymptotically how you weight the various bits of it. You could use $\Omega = I_J$ if you wanted, and the estimator would be consistent. Minimisation with $\Omega = I_J$ is often referred to as *Minimum Distance Estimation*. The point here is that **any choice of** Ω yields a consistent estimator. However, different choices of Ω yield estimators of different efficiency.

(d) A natural choice of Ω is the covariance matrix of $Z'\varepsilon$. By assumption, $E[Z'\varepsilon] = 0$, so its covariance $E[Z'\varepsilon\varepsilon'Z]$ is equal to its mean squared error. If an element of $Z'\varepsilon$ is always really close to zero, then you'd want to pay a lot of attention to keeping this element close to zero. In contrast, if an element of $Z'\varepsilon$ varied wildly, you wouldn't want to pay too much attention to it in choosing your β. Since the weighting matrix is Ω^{-1}, elements of $Z'\varepsilon$ that are wildly varying, and have big variance, get small weight in the quadratic form; elements which are always close to zero, and have small variance, get big weight in the quadratic form.

(e) Hansen (1982) shows that if Ω is the covariance matrix of $Z'\varepsilon$, then the GMM estimator is asymptotically efficient.

(f) *Given homoskedasticity*,

$$V[Z'\varepsilon] = E[Z'\varepsilon\varepsilon'Z] = E[Z'\sigma^2 I_N' Z] = \sigma^2 Z'Z$$

So, we minimize

$$\min_\beta (Y - X\beta)' Z \frac{1}{\sigma^2} (Z'Z)^{-1} Z' (Y - X\beta) =$$
$$\min_\beta (Y - X\beta)' Z (Z'Z)^{-1} Z' (Y - X\beta)$$
$$\min_\beta (Y - X\beta)' P_Z (Y - X\beta)$$
$$\min_\beta (Y - X\beta)' P_Z' P_Z (Y - X\beta)$$

yielding a first-order condition

$$\begin{aligned} 2X'P_Z'(Y - X\beta) &= 0 \\ &\Leftrightarrow \\ X^{1rststage\prime}(Y - X\beta) &= 0. \end{aligned}$$

That is, choose β to make e orthogonal to $\hat{X}$.

i. That is, given homoskedasticity, the GMM estimator is the 2SLS estimator:

$$\begin{aligned} X'P_Z'(Y - X\beta) &= 0 \\ X'P_Z'Y &= X'P_Z'X\beta \end{aligned}$$

$$\hat{\beta}_{GMM} = \hat{\beta}_{2SLS} = \left(X' P_Z' X\right)^{-1} X' P_Z' Y$$

(g) Since all exogenous variables correspond to columns of X that are also in Z, and defining as an instrument as an exogenous variable that is in Z but not in X, this implies that you need at least one instrument in Z for every endogenous variable in X. Consider

$$X = [X_1 \ X_2], \ Z = [X_1 \ Z_2],$$

a case where a subvector of X is exogenous (because it shows up in Z), and the rest is endogenous. Here, $X_2$is endogenous.

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

The two stage approach to IV is to estimate X on Z and use its predicted values on the RHS instead of X.
What are its predicted values? Since Z contains the exogenous pieces of X, this is equivalent to regressing

$$Y = X_1\beta_1 + \widehat{X}_2\beta_2 + \varepsilon$$

where

$$\widehat{X}_2 = Z\left(Z'Z\right)^{-1} Z' X_2$$

the predicted values from a regression of the endogenous X's on the entire set of Z.

i. What if X_2contains interactions or squared terms, for example, if schooling years are endogenous, then so, too, must be schooling years squared, and the interaction of schooling years with gender. So, even if the root is the endogeneity of schooling years, all these other variables must be endogenous, too.

ii. In this case, two stage least squares must treat the endogenous variables just as columns of X_2. Thus, the $\widehat{X}_2$that gets used in the final regession has the predicted value of schooling years and a separate predicted value of schooling years squared. This latter variable is not the square of the predicted value of schooling. You may wish to impose this restriction, but two stage least squares does not do it automatically. It is easy to see how to impose it when using full information methods (below), but can be cumbersome in limited information (IV and control variate) methods.

(h) The **Asymptotic** bias of this estimator is (small sample later)

$$\begin{aligned} E\left[\hat{\beta}_{2SLS}\right] - \beta &= E\left[\left(X' P_Z' X\right)^{-1} X' P_Z' Y\right] - \beta \\ &= \left(X' P_Z' X\right)^{-1} X' P_Z' X\beta + E\left[\left(X' P_Z' X\right)^{-1} X' P_Z' \varepsilon\right] - \beta \\ &= E\left[\left(X' P_Z' X\right)^{-1} X' P_Z' \varepsilon\right] \\ &= \left(X' P_Z' X\right)^{-1} X' Z' \left(Z'Z\right)^{-1} E\left[Z'\varepsilon\right]. \end{aligned}$$

If Z is orthogonal to ε, then this asymptotic bias is zero, so that GMM/2SLS is **consistent**.

(i) The **Asymptotic** variance of the estimator is the square of this object. If the ε are homoskedastic, then

$$\begin{aligned} V\left(\hat{\beta}_{2SLS}\right) &= \sigma^2 \left(X'P_Z'X\right)^{-1} X'P_Z'\varepsilon\varepsilon P_Z X \left(X'P_Z'X\right)^{-1} \\ &= \sigma^2 \left(X'P_Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1} Z'Z\left(Z'Z\right)^{-1} Z'X \left(X'P_Z'X\right)^{-1} \\ &= \sigma^2 \left(X'P_Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1} Z'X \left(X'P_Z'X\right)^{-1} \\ &= \sigma^2 \left(X'Z\left(Z'Z\right)^{-1} Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1} Z'X \left(X'P_Z'X\right)^{-1} \\ &= \sigma^2 \left(X'P_Z'X\right)^{-1}, \end{aligned}$$

which is a reweighted version of the variance of the OLS estimator.

5. 2SLS small-sample bias and Weak Instruments (see Angrist and Pisschke, pg 170ish).

(a) **2SLS is biased but consistent**.

(b) Consider the model

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E[X'\varepsilon] &\neq 0 \\ E[Z'\varepsilon] &= 0 \\ E[Z'X] &\neq 0 \\ E[\varepsilon\varepsilon'] &= \sigma^2 I_N \end{aligned}$$

where, X, Z, ε are all random variables. And, additionally, let

$$X = Z\Gamma + u$$

with

$$E\left(u'\varepsilon\right) \neq 0,$$

so that we have in mind that X has a part correlated with ε given by u, and a part uncorrelated with ε given by $Z\Gamma$.

(c) The estimated coefficient vector via GMM/2SLS is

$$\widehat{\beta} = \widehat{\beta}_{2SLS} = \widehat{\beta}_{GMM} = \left(X'P_Z X\right)^{-1} X'P_Z Y$$

(d) Its bias is

$$E\left[\hat{\beta}_{2SLS}\right] - \beta = E\left[\left(XP_Z X\right)^{-1} X'P_Z\varepsilon\right].$$

Notice that the expectation is now over X and over the inverse. It turns out that expectations of matrix inverses times other matrices are pretty close to the product of the two expectations, so that

$$E\left[\hat{\beta}_{2SLS}\right] - \beta \approx E\left[\left(XP_Z X\right)^{-1}\right] E\left[X'P_Z\varepsilon\right].$$

(e) Substituting in for $X = Z\Gamma + u$ in the 2nd chunk, we have

$$\begin{aligned} E\left[\hat{\beta}_{2SLS}\right] - \beta &\approx E\left[\left(XP_ZX\right)^{-1}\right] E\left[(Z\Gamma + u)'P_Z\varepsilon\right] \\ &= E\left[\left(XP_ZX\right)^{-1}\right] E\left[(Z\Gamma)'P_Z\varepsilon\right] + E\left[\left(XP_ZX\right)^{-1}\right] E\left[u'P_Z\varepsilon\right]. \end{aligned}$$

(f) The problem is the term $E\left[\left(XP_ZX\right)^{-1}\right] E\left[u'P_Z\varepsilon\right]$. If

$$E\left(u'\varepsilon\right) \neq 0,$$

then a quadratic form in u, ε will be nonzero for some values of the weighting matrix P_Z. If Z were fixed, we could restrict that that particular quadratic form had expectation zero. But, since Z can take on any value, we cannot be assured of this for all values of P_Z unless we restrict that $E\left(u'\varepsilon\right) = 0$, which would imply no endogeneity in the first place.

(g) If we had $Z\Gamma$ in our pockets, there would be no bias, because $Z\Gamma$ is uncorrelated with ε by assumption. Unfortunately, we don't have $Z\Gamma$, all we have is its estimator, the first-stage estimate P_ZX, which contains a little P_Zu, which **is** correlated with ε by assumption.

(h) Of course, asymptotically, the first stage estimate is exactly right, and we do have $Z\Gamma$. That is why 2SLS is consistent (asymptotically unbiased), even though it is biased for every finite sample.

6. Weak Instruments

(a) Here, we see that if Z is 'too good' at predicting X, then P_ZX will look a lot like X, which is not orthogonal to ε.

(b) In exactly identified models where Z is orthogonal to ε, P_ZX must also be orthogonal to ε. However, when the model is overidentified, in finite samples, $E\left[\left(P_ZX\right)'\varepsilon\right]$ is not zero. The basic reason is that if J (the rank of Z) is large relative to N, then Z will pick up some ε no matter how Z is correlated with ε. Consider the case where $J = N$. Then, $P_ZX = X$, because there is one column of Z for each observation. In this case, $E\left[\left(P_ZX\right)'\varepsilon\right] = E\left[X'\varepsilon\right]$, which is not zero by assumption.

(c) On the other hand, if $P_ZX = X$, then $\left(\hat{X}'\hat{X}\right)^{-1} = \left(X'X\right)^{-1}$, so that the bias does not get magnified a whole lot, and in the limit, it disappears as $\left(X'X\right)^{-1}$ goes to zero.

(d) A large number of weak instruments is the worse case: you get $E\left[P_Z'X\right]$ spruriously close to X, and in the small sample a nonzero $\left(\hat{X}'\hat{X}\right)^{-1}$.

(e) If you have lots of overidentification, then even with homoskedasticity, the term $P_Z'\varepsilon\varepsilon P_Z$ will not have expectation $\sigma^2 P_Z$ in the small sample. Thus, when you have overidentification and weak instruments, you get both bias and weird standard errors.

(f) What if your instruments are only weakly correlated with X. This is the problem of *weak instrument*s (see the suggested reading). Weak instruments are an issue in overidentified models where the instruments are only weakly correlated with the endogenous regressor(s). In this case, you have small-sample bias, and high variance at any sample size.

(g) This problem goes away asymptotically, but if you have crappy enough instruments with weak enough correlation with X, you can induce a huge small sample bias even if the instruments are truly exogenous.

(h) We can approximate the small-sample bias. Hahn and Hausman (2005), who explore the case with 1 endogenous regressor, write the (second-order) approximation of the small-sample bias of the 2SLS estimator for that endogenous regressor proportional to

$$E\left[\hat{\beta}_{2SLS}\right] - \beta \propto \frac{L\rho\left(1-\tilde{R}^2\right)}{N\tilde{R}^2},$$

where $\tilde{R}^2$ is the R^2 value in the first-stage regression, J is the number of instruments, ρ is the correlation between the endogenous regressor and the model disturbance term, and N is the sample size.

i. Obviously, this goes to zero as N goes to ∞, so the the 2SLS estimator is consistent.

ii. However, if the instruments are weak, so that $\tilde{R}^2$ is small, the bias gets large.

iii. Further, if the model is highly overidentified, so that J is very large, the bias gets large.

iv. the correlation ρ is the reason that there is an endogeneity problem in the first place. Indeed, the bias of the OLS regression coefficient is proportional to ρ.

v. We may rewrite this expression in terms of the first stage F statistic, too:

$$E\left[\hat{\beta}_{2SLS}\right] - \beta \approx \frac{\sigma_{u\varepsilon}}{\sigma^2}\frac{1}{F+1},$$

(i) one can express the bias of the 2SLS estimator in terms of the bias of the OLS estimator:

$$Bias\left[\hat{\beta}_{2SLS}\right] = \frac{J}{N\tilde{R}^2} Bias\left[\hat{\beta}_{OLS}\right],$$

so if $J > N\tilde{R}^2$, 2SLS is certainly worse than OLS: it would have bigger bias and, as always, it would have bigger variance. Of course, even if 2SLS has smaller bias, it may enough more variance that it is still undesirable in Mean Squared Error terms.

(j) A simple comparison of J to $N\tilde{R}^2$ gives you a sense of how big the 2SLS bias is. Since that fraction is strictly positive, 2SLS is biased in the same direction as OLS, so if you have theory for the direction of OLS bias, the same theory works for 2SLS bias.

7. Rules of Thumb

   (a) Report the first stage and think about whether it makes sense. Are the magnitude and sign as you would expect, or are the estimates too big or large but wrong-signed? If so, perhaps your hypothesized first-stage mechanism isn't really there, rather, you simply got lucky.

   (b) Report the F-statistic on the excluded instruments. The bigger this is, the better. Stock, Wright, and Yogo (2002) suggest that F-statistics above about 10 put you in the safe zone though obviously this cannot be a theorem.

   (c) Pick your best single instrument and report just-identifed estimates using this one only. Just-identifed IV is approximately unbiased and therefore unlikely to be subject to a weak-instruments critique.

   (d) Check over-identifed 2SLS estimates with LIML. LIML is less precise than 2SLS but also less biased. If the results come out similar, be happy. If not, worry, and try to find stronger instruments.

   (e) Look at the coefficients, t-statistics, and F-statistics for excluded instruments in the reduced-form regression of dependent variables on instruments. Remember that the reduced form is proportional to the causal effect of interest. Most importantly, the reduced-form estimates, since they are OLS, are unbiased. Angrist and Krueger (2001), and many others, believe that if you can't see the causal relation of interest in the reduced form, it's probably not there.

8. GMM in non-homoskedastic and nonlinear models.

   (a) If the disturbances are not homoskedastic *and* the model is overidentified (Z has higher rank than X), then GMM is not equivalent to 2SLS. GMM would give different numbers.

   (b) Here, the key is to get an estimate of Ω, given by $\hat{\Omega}$. Any consistent estimate of Ω will do the trick. How big is Ω? It is the covariance matrix of $Z'\varepsilon$, so it is a symmetric JxJ matrix, with $J(J+1)/2$ distinct elements. Note that this number does not grow with N.

(c) Recall that GMM is consistent regardless of which Ω is used. Thus, one can implement GMM with any old Ω (such as the 2SLS estimator), collect the residuals, e, and construct

$$\hat{\Omega} = \frac{1}{N} \sum \left(Z' e e' Z \right),$$

and then implement GMM via

$$\min_{\beta} e' Z \hat{\Omega}^{-1} Z' e$$

(d) GMM can be used in nonlinear models, too. Consider a model

$$Y = f(X, \beta) + \varepsilon$$

The moment restrictions are

$$E[Z' \varepsilon] = 0$$

Substituting in sample moments yields

$$Z' \left(Y - f(X, \beta) \right) = 0.$$

Thinking in a GMM way, it is clear that if the rank of Z is less than the length of β, you won't have enough information to identify β. This is equivalent to say (in full-rank linear models) that you need at least one column of Z for each column of X.

(e) If $J > K$, you can't get these moments all the way to zero, so one would minimize

$$\min_{\beta}{}' \left(Y - f(X, \beta) \right)' Z \hat{\Omega}^{-1} Z' \left(Y - f(X, \beta) \right)$$

where $\hat{\Omega}$ is an estimate of the variance of the moment conditions.

9. Are the Instruments Exogenous? Are the Regressors Endogeneous?

(a) If the regressors are endogeneous, then the OLS estimates should differ from the endogeneity-corrected estimates (as long as the instruments are exogenous).

(b) The test of this hypothesis is called a Hausman Test.

(c) If the instruments are exogenous and you have an overidentified model, then you can test the exogeneity of the instruments.

(d) Under the assumption that at least one instrument is exogenous, you can test whether or not all the instruments are exogenous.

(e) *Test of Overidentifying Restrictions*

(f) Under the assumption that all the instruments are exogenous, we have from the model that

$$\begin{aligned} E\left[Z'\varepsilon\right] &= 0 \\ E\left[Z'\varepsilon\varepsilon'Z\right] &= \Omega. \end{aligned}$$

(g) The GMM estimator, minimised a quadratic form of $Z'e$:

$$\min_{\widehat{\beta}} e'Z\Omega^{-1}Z'e,$$

where $\Omega = E\left[Z'\varepsilon\varepsilon'Z\right]$ is the (possibly estimated) covariance matrix of $Z'\varepsilon$.

(h) If we had observed ε, we might have considered the optimisation

$$\min_{\widehat{\beta}} \varepsilon'Z\Omega^{-1}Z'\varepsilon,$$

and, so, if we premultiplied everything by $\Omega^{-1/2}$, we'd have

$$\begin{aligned} E\left[\Omega^{-1/2}Z'\varepsilon\right] &= 0, \\ E\left[\varepsilon'Z\Omega^{-1}Z'\varepsilon\right] &= I_J, \end{aligned}$$

so the quadratic form would be the square of a bunch of mean-zero, unit variance, uncorrelated random variables.

(i) If the model is homoskedastic, then $\Omega^{-1} = \frac{1}{\sigma^2}\left(Z'Z\right)^{-1}$, so you don't even need to know Ω. You'd have

$$\min_{\widehat{\beta}} \frac{1}{\sigma^2}\varepsilon'Z\left(Z'Z\right)^{-1}Z'\varepsilon = \min_{\widehat{\beta}} \frac{1}{\sigma^2}\varepsilon'P_Z\varepsilon = \min_{\widehat{\beta}} \frac{1}{\sigma^2}\varepsilon'P_ZP_Z'\varepsilon$$

and, so, if we premultiplied everything by $\Omega^{-1/2}$, we'd have

$$\begin{aligned} E\left[P_Z'\varepsilon\right] &= 0, \\ E\left[\varepsilon'P_ZP_Z'\varepsilon\right] &= I_J. \end{aligned}$$

(j) But, e isn't ε. The residual is related to the model via:

$$\begin{aligned} e &= \left(Y - X\widehat{\beta}\right) \\ &= \left(Y - X'\left(X'P_Z'X\right)^{-1}X'P_Z'Y\right) \\ &= \left[I - X'\left(X'P_Z'X\right)^{-1}X'P_Z'\right]Y \\ &= \left[I - X'\left(X'P_Z'X\right)^{-1}X'P_Z'\right]\left(X\beta + \varepsilon\right) \\ &= \left[X\beta - X'\left(X'P_Z'X\right)^{-1}X'P_Z'X\beta\right] + \left[I - X'\left(X'P_Z'X\right)^{-1}X'P_Z'\right]\varepsilon \\ &= \left[I - X'\left(X'P_Z'X\right)^{-1}X'P_Z'\right]\varepsilon. \end{aligned}$$

(k) Sticking with the homoskedastic case, the GMM objective function is

$$\min_{\widehat{\beta}} e'Z\Omega^{-1}Z'e = \min_{\widehat{\beta}} \frac{1}{\sigma^2} e'P_Z P_Z' e,$$

so the thing we are summing and squaring is

$$\begin{aligned} P_Z' e &= \left[P_Z' - P_Z' X' \left(X' P_Z' X\right)^{-1} X' P_Z'\right] \varepsilon \\ &= \left[P_Z' - \widehat{X}' \left(\widehat{X}'\widehat{X}\right)^{-1} \widehat{X}\right] \varepsilon \\ &= \left[P_Z' - P_{\widehat{X}}'\right] \varepsilon \end{aligned}$$

i. The rank of P_Z' is J and the rank of $P_{\widehat{X}}$ is K, so the rank of $\left[P_Z' - P_{\widehat{X}}'\right]$ is $J - K$. Thus, we are summing and squaring $J - K$ independent linear combinations of ε, whose mean is 0 and whose variance matrix is I.

ii. Thus,

$$e' P_Z P_Z' e$$

is a sum of squares of $J - K$ things with mean zero and variance 1.

iii. Central Limit Theorem: if things have finite variances, then finite-order linear combinations of them will behave asymptotically and approximately like linear combinations of normals.

iv. So, each element of and $P_Z' e$ is approximately asymptotically $N(0, 1)$.

v. The sum-of-squares of it is approximately and asympotically a chi-square:

$$e' P_Z P_Z' e \overset{asy, approx}{\sim} \chi^2_{J-K}.$$

(l) Consider a homoskedastic case with 1 endogenous regressor, no exogenous regressors, and 2 instruments.

i. The quadratic form over

$$P_Z' e = \left[P_Z' - P_{\widehat{X}}'\right] \varepsilon$$

cannot be brought to zero, because there are 2 equations and only one coefficient to choose.

ii. However, if $E[Z'\varepsilon] = 0$, then in a given sample, $P_Z'\varepsilon$ should be 'close' to zero in both of its elements.

iii. Further, since each element of $\Omega^{-1/2} P_Z' \varepsilon$ is a mean-zero and unit-variance, each element should be between $[-2, 2]$ a lot of the time.

iv. Asymptotically and approximately, each element of $\Omega^{-1/2} P_Z' \varepsilon$ is a standard normal.

v. But, the 2 elements of $\Omega^{-1/2}P_Z'\varepsilon$ are linear in each other, so if you move one of them, you move the other by a fixed amount.

vi. Thus, the sum of squares of these is not the sum of 2 squared normals, but just a single squared normal.

vii. So, if the GMM objective function is 50 at its minimum, this number is too high to believe that you really had exogenous instruments—more likely, you had an endogenous instrument.

10. What are good instruments?

(a) Going back to the Ballentine/Venn Diagrams in Kennedy, you want an instrument Z that has no purple stuff in it (no correlation with the disturbance terms),

$$E\left[Z'\varepsilon\right] = 0,$$

and which has lots of correlation with X,

$$E\left[Z'X\right] \neq 0.$$

(b) What if you have instruments that violate this? Consider violation of exogeneity. In the linear exactly identified case:

$$\widehat{\beta}_{2SLS} = \widehat{\beta}_{(G)MM} = (Z'X)^{-1}Z'Y,$$

so, the bias term is

$$\begin{aligned} E\left[\widehat{\beta}_{2SLS} - \beta\right] &= E\left[(Z'X)^{-1}Z'Y - \beta\right] \\ &= E\left[(Z'X)^{-1}Z'X\beta + (Z'X)^{-1}Z'\varepsilon - \beta\right] \\ &= E\left[(Z'X)^{-1}Z'\varepsilon\right] = (Z'X)^{-1}E\left[Z'\varepsilon\right], \end{aligned}$$

so, the degree to which $E\left[Z'\varepsilon\right]$ is not zero gets reweighted by $(Z'X)^{-1}$ to form the bias. If $E\left[Z'\varepsilon\right]$ is big, the bias is big, unless $(Z'X)^{-1}$ is really small. The matrix $(Z'X)^{-1}$ is small if $Z'X$ is big, which is the case if Z and X covary a lot. Thus, even if your instrument is not perfectly exogenous, you still want lots of correlation with X.

11. Control Variables

(a) Denote $\widehat{\eta}$ as the residuals from a regression of X on Z. The residual $\widehat{\eta}$ contains the same information about the endogeneity as $\widehat{X}$. If $\widehat{X}$ is a part of X that is not polluted, then $\widehat{\eta}$ must contain *all* the pollution.

(b) That means that if we add $\widehat{\eta}$ to the RHS of the original specification, it will *control out* the endogeneity.

(c) estimate by OLS

$$Y = X\beta + \widehat{\eta}\gamma + \varepsilon$$

The intuition is that $\widehat{\eta}$ is a control for the pollution in X. So, if X is positively correlated with the disturbance, then

$$\widehat{\eta}$$

will be a big number when X is big, and small when X is small. This polluted covariation with Y will be assigned to $\widehat{\eta}$ (measured by γ), and not to X.

(d) In linear homoskedastic models, the estimated coefficients on X (and their variance) is the same whether you use GMM, 2SLS or control variables.

(e) The fact that the control variable comes in additively is important. It makes the whole system triangular, and therefore writeable in this fashion. In nonlinear, semiparametric and nonparametric models, sometimes the control variable approach is easier, but it has to come in additively (Newey & Powell Econometrica 1997, 2003).

(f) Most everyone uses IV and not control variables.

12. Model the endogeneity (Full Information)

(a) Consider the two equation model

$$\begin{aligned} Y_i &= X_i\beta + \varepsilon_i \\ X_i &= Z_i\Gamma + \eta_i \\ E[\varepsilon_i\eta_i] &\neq 0. \end{aligned}$$

Assume that Z has at least the rank of X, and that Z is exogenous. An example of this might be

$$X_i = [X_{1i}\ X_{2i}],\ Z_i = [X_{1i}\ Z_{2i}],$$

a case where a subvector of X is exogenous (because it shows up in Z), and the rest is endogenous.

(b) We could assume that

$$\begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \tilde{} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon\sigma_{\varepsilon\eta} \\ \rho\sigma_\varepsilon\sigma_\eta & \sigma_\eta^2 \end{pmatrix}\right) = N(0_2, \Sigma)$$

and estimate the parameters inside this normal, and the parameters of the two equations, by maximum likelihood. This would be called "full information maximum likelihood" (FIML).

(c) The FIML estimator would use the fact that

$$\Sigma^{-1/2}\begin{bmatrix}\varepsilon_i\\ \eta_i\end{bmatrix} \tilde{} N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right) = N(0_2, I_2)$$

where

$$\Sigma^{-1/2} = \begin{pmatrix}\sigma_\varepsilon^2 & \rho\sigma_\varepsilon\sigma_{\varepsilon\eta}\\ \rho\sigma_\varepsilon\sigma_\eta & \sigma_\eta^2\end{pmatrix}^{-1/2}.$$

(d) Let

$$\begin{bmatrix}\widetilde{\varepsilon}_i\\ \widetilde{\eta}_i\end{bmatrix} = \Sigma^{-1/2}\begin{bmatrix}\varepsilon_i\\ \eta_i\end{bmatrix}.$$

Then, the density of observation i is

$$\phi\left(\widetilde{\varepsilon}_i\right)\phi\left(\widetilde{\eta}_i\right)$$

and the likelihood to be maximized by choice of $\beta, \Gamma, \sigma_\varepsilon^2, \sigma_\eta^2, \rho$ is

$$\max_{\beta,\Gamma,\sigma_\varepsilon^2,\sigma_\eta^2,\rho} \ln L = \sum_{i=1}^{N} \ln\phi\left(\widetilde{\varepsilon}_i(Y_i, X_i; \beta, \Gamma, \sigma_\varepsilon^2, \sigma_\eta^2, \rho)\right) + \ln\phi\left(\widetilde{\eta}_i(Y_i, X_i; \beta, \Gamma, \sigma_\varepsilon^2, \sigma_\eta^2, \rho)\right)$$

(e) Note that for a mean-zero bivariate normal, the conditional distribution (http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions) of one given the other is

$$\varepsilon_i|_{\eta_i=\eta} \tilde{} N(\frac{\sigma_\varepsilon}{\sigma_\eta}\rho\eta, (1-\rho^2)\sigma_\varepsilon^2).$$

This conditional distribution is not mean-zero. Indeed, its mean is linear in η.

(f) If you just regress Y on X, the expectation of ε_i is $\frac{\sigma_\varepsilon}{\sigma_\eta}\rho\eta_i$. That is the bias term in the endogenous regression. It is zero if and only if $\rho = 0$.

(g) Consider adding η_i as a regressor. Its coefficient in the regression would be $\frac{\sigma_\varepsilon}{\sigma_\eta}\rho$ and it would soak up all the endogeneity.

(h) Thus, a control function approach to this ML problem would be to regress X on Z, collect residuals $n_i = X_i - Z_i\widehat{\Gamma}$, and then regress Y on X and n.

(i) This is called "limited information maximum likelihood" (LIML) because we don't take all the information about the distribution of ε_i given η_i. Instead, we just try to control out is mean. We leave all the other moments of that distribution unconstrained.

13. Selection-Corrections

(a) Consider the two equation model

$$Y_i = X_i\beta + \varepsilon_i$$
$$Y_i \text{ observed if } Y^* = Z_i\Gamma + \eta_i > 0,$$

$$\begin{bmatrix} \varepsilon_i \\ \eta_i \end{bmatrix} \tilde{} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon\sigma_{\varepsilon\eta} \\ \rho\sigma_\varepsilon\sigma_\eta & \sigma_\eta^2 \end{pmatrix}\right)$$

Assume that Z has at least the rank of X, and that Z is exogenous. An example of this might be

$$X_i = [X_{1i}\ X_{2i}],\ Z_i = [X_{1i}\ Z_{2i}],$$

a case where a subvector of X is exogenous (because it shows up in Z), and the rest is endogenous.

(b) Here, Y_i is observed only for some observations. For other observations, we observe Z (which includes X) but not Y.

(c) Y_i could be wages. Wages are only observed for workers, but labour force attachment is observed for all people (be they workers or non-workers).

(d) Suppose that an unobservable, like ability, is correlated with both labour force attachment (the probability of working) and with wages. If ability were correlated with, e.g., an observable like education, this would induce endogeneity in the first equation above if we were to just regress Y X. The reason is that the sample of workers with observed wages would be richer with high-ability people than the full population of people which included nonworkers.

(e) The challenge is to figure out the conditional expectation of ε_i for observations with values of η_i that put them into the category of having observed Y_i .

(f) Since the conditional distribution

$$\varepsilon_i|_{\eta_i=\eta} \tilde{} N(\frac{\sigma_\varepsilon}{\sigma_\eta}\rho\eta, (1-\rho^2)\sigma_\varepsilon^2),$$

the conditional mean is

$$E[\varepsilon_i|_{\eta_i=\eta}] = \frac{\sigma_\varepsilon}{\sigma_\eta}\rho\eta,$$

which is linear in η, so the mean of ε_i across all the values of η consistent with Y being observed is linear in the average value of those η's.

(g) Consider the values of η consistent with Y being observed. For any $Z_i\Gamma$, to get $Y^* > 0$ we need $\eta_i > -Z_i\Gamma$. So we need the expectation of η_i given that it lies above $-Z_i\Gamma$.

(h) This expectation is called the "truncated mean" and the truncated mean of a normal variate is easily looked up on Wikipedia http://en.wikipedia.org/wiki/Truncated_n In our context, it is

$$E\left[\eta_i|_{\eta_i>-Z_i\Gamma}\right] = \sigma_\eta \lambda(\frac{-Z_i\Gamma}{\sigma_\eta}),$$

where

$$\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$$

is the "Inverse Mills Ratio".

(i) Consequently,

$$E[\varepsilon_i|_{Y_i \text{ observed}}] = \frac{\sigma_\varepsilon}{\sigma_\eta}\rho E\left[\eta_i|_{\eta_i>-Z_i\Gamma}\right] = \sigma_\varepsilon \rho \lambda(\frac{-Z_i\Gamma}{\sigma_\eta}).$$

This is the bias term in the regression of Y on X.

(j) So, a FIML approach would be analogous to that in 12 above. A LIML approach is very easy to implement via the 2-step "Heckman Regression":

i. probit (Y_i observed) on Z_i
ii. predict λ_i, invmills
iii. regress Y_i X_i λ_i
iv. The coefficient on λ_i goes to $\sigma_\varepsilon\rho$ and corrects for bias induced by unobservables correlated with selection into the labour force.