Causality

Devesh Raval

1 Introduction

So far in this class I have introduced two major statistical concepts to estimate:

1. The Conditional Expectation Function, $E(Y|X)$
2. The Best Linear Predictor-$(X'X)^{-1}X'Y$

Both of these have strictly statistical interpretations- the CEF tells us the average value of Y for any given X. The BLP provides the best linear approximation to the CEF-so it tells us how much Y increases for any given change in X in the data, restricted to a linear functional form. The BLP is really just telling us the statistical correlation between Y and X in the data. With more than one X it is just telling you the statistical correlation of each variable after taking out the variation due to other variables in the regression.

In an experimental setting, where we set all of the Xs in advance, the CEF and BLP will also be telling us the "effect" of X on Y- if we change X, what will happen to Y? Thus, in a strict experiment, we can learn about the causal effect of X.

Almost all of economics is in nonexperimental settings, however, where we do not set X. So we do not know how X varies with any errors in our model. This makes it hard to assess causality. Take the basic model:

$$Y = \alpha + \beta X + \epsilon$$

Imagine we estimate this model by OLS and find b_N is positive. There are many casual stories that can result in this outcome:

1. True Causal Effect- when X increases, Y really increases by $\beta > 0$.
2. Reverse Causality- when Y increases, X increases. X is really a function of Y.
3. Third Variable- Some variable T exists- when T increases, so does X and Y.
4. Simultaneity- X and Y are determined simulatenously (i.e. price and quantity).

How did we solve this before? By making an assumption, as always- the first assumption of the CRM- that:

$$\begin{aligned} E(\epsilon|X) &= 0 \\ E(Y|X) &= X\beta \end{aligned}$$

This assumption states that the average value of ϵ doesnt change as we change X, so changes in X in the data are only reflected in the avg value of Y through a causal effect of X. All other mechanisms are ruled out since ϵ is not changing on avg.

Thus, we need to know what ϵ is- what is the interpretation of the error?

Some examples of interpretations:

1. Measurement error in y

2. Measurement error in X

3. Exogenous shocks in the economy- cost shocks or technology shocks or demand shocks

4. Omitted Variables

5. "Mistakes" or taste shocks of the agent

6. Nonlinearities in the CEF

For example, in the Third Variable example, T is an omitted variable that is in the error. In the Reverse Casuality example, ϵ partly determines X (and so they are not uncorrelated!)

Nice exs for the others: production function, consumption income model

Imagine we have reason to believe that $E(\epsilon|X) \neq 0$. What do we do to still try to estimate causal effects?

Two different approaches in economics today:

1. Structural approach: Write down the full model of the problem coming from some economic theory and use the model to get at causal effects. Here all errors should be "named". The example we will do in class is with a supply model and demand model- each equation has an economic interpretation, as do all the coefficients.

2. Reduced form approach (or Identification Strategy): Find a strategy to get the effect of X on Y- where we have a way to move X without moving the error. The most prominent of these is Instrumental Variables which we will cover now.

2 Instrumental Variables

2.1 A Simple Example

Imagine our economic model is as follows:

$$Health = f(visits)$$

where we do not know the f function. The regression model imposes linearity as below:

$$health = \alpha + \beta * visits + \epsilon$$

Lets look at the bivariate slope coefficient:

$$\begin{aligned}
Cov(health, visits) &= Cov(\beta * visits + \epsilon, visits) \\
&= Cov(\beta * visits, visits) + Cov(\epsilon, visits) \\
&= \beta Var(visits) + Cov(\epsilon, visits)
\end{aligned}$$

But imagine that the true model includes heartiness- which determines visits and health itself. Thus we have:

$$\begin{aligned}
\epsilon &= \gamma heartiness + \eta \\
visits &= \delta heartiness + \mu
\end{aligned}$$

This is just the Omitted Variable bias problem from earlier, where the second regression is the auxiliary regression. Clearly in this case:

$$\begin{aligned}
Cov(\epsilon, visits) &\neq 0 \\
&= Cov(\delta heartiness + \mu, \gamma heartiness + \eta) \\
&= \gamma\delta Var(heartiness)
\end{aligned}$$

This will lead to a negative bias if $\gamma > 0, \delta < 0$- so since more hearty people go to the doctor less, more visits will have lower impact on health than it should.

So, the true model is:

$$health = \alpha + \beta * visits + \gamma heartiness + \eta$$

If *heartiness* is not observed, can we do anything? What if we know that the distance to the doctor also matters?

Thus,

$$\begin{aligned} visits &= g(heartiness, dist) \\ &= \delta heart + \rho distance + \mu \end{aligned}$$

Imagine we could change people's distance from the doctor- this would increase the number of visits without changing the person's health except through the visits channel. The idea of IV is to use only the variance in visits through distance to doctor to get at the causal effect- ignoring the variance through other causes (like heartiness). We can formalize this intution as follows:

$$\begin{aligned} Cov(visits, dist) &= \rho * var(dist) \\ Cov(dist, health) &= Cov(dist, \beta * visits + \gamma * heartiness + \eta) \\ &= \beta Cov(dist, visits) = \beta\rho * var(dist) \end{aligned}$$

The IV estimator in this case is:

$$\begin{aligned} b_{IV} &= \frac{Cov(health, dist)}{Cov(visits, dist)} \\ &= \frac{\beta Cov(visits, dist)}{Cov(visit, dist)} = \beta \end{aligned}$$

So IV is consistent in this case. What is it doing? Its assuming that all the covariance between health and distance is coming from visits. If we divide by $var(visits)$ we can see this clearly:

$$\frac{Cov(health, dist)}{Cov(visits, dist)} = \frac{Cov(health, dist)/Var(dist)}{Cov(visits, dist)/Var(dist)}$$

The numerator is the effect of distance on health, the denominator the effect of distance on visits. This is really doing the following:

$$\frac{dhealth}{ddist} = \frac{\partial health}{\partial visits} * \frac{dvisits}{ddist}$$

We want $\frac{\partial health}{\partial visits}$ which we get by the division. Note that implicit in this is that distance has no partial effect on health itself- if it did this method would not work. I.e. we dont have:

$$\frac{dhealth}{ddist} = \frac{\partial health}{\partial visits} * \frac{dvisits}{ddist} + \frac{\partial health}{\partial dist}$$

This assumption is called an Exclusion Restriction- distance is in the visit equation but not in the health equation.

The problem with this is thats its difficult to find variables that truly satisfy an exclusion restriction- that have no partial effect on the dependent variable, only through the X variable.

For example, in the health case very sick people might choose to live near a hospital- in which case distance would have a negative effect on health. Or it could be the case that people that live far from a hospital are living in rural areas and are less healthy than urban/suburban people for other reasons (more alcohol/tobacco use, less education, etc.)

A final way to motivate this estimator is as follows. First, regress *visits* on *distance* so we get a fitted value of *visits*, $\hat{visits}$, which tells us the predicted value of the number of visits given the distance from the doctor. Then regress *health* on the fitted value $\hat{visits}$. The resulting coefficient in the population will be:

$$\begin{aligned}
\frac{Cov(\hat{visits}, health)}{Var(\hat{visits})} &= \frac{Cov(\rho dist, health)}{Var(\rho dist)} \\
&= \frac{Cov(\frac{Cov(dist,visits)}{Var(dist)} dist, health)}{Var(\frac{Cov(dist,visits)}{Var(dist)} dist)} = \frac{\frac{Cov(dist,visits)}{Var(dist)} Cov(dist, health)}{(\frac{Cov(dist,visits)}{Var(dist)})^2 Var(dist)} \\
&= \frac{Cov(dist, health)}{Cov(dist, visits)}
\end{aligned}$$

One thing we definitely do not want to do is include *dist* in the main regression and run OLS- b/c if we do, through the partitioned regression logic we get:

$$b_{OLS} = \frac{Cov(visits^*, health^*)}{Var(visits^*)}$$

where the * indicates that we have regressed out the dist variable. The endogeneity bias is still there, but now all our variation in visits is coming from heartiness, so the bias should be worse. The “good” variation from distance has all been stripped out by putting distance in the OLS regression.

What estimator should we use here? So far everything has been in the population. The analogy principle here suggests three different estimators, all of which are the same in the just identified case (as many instruments as endogenous variables):

For notation purposes lets call $dist = Z, health = Y, visits = X, heartiness = Q$. Then our underlying model is:

$$Y = \beta X + \gamma Q + \eta$$

$$X = \rho Z + \delta Q + \mu$$

These are the structural equations of the model.

Indirect Least Squares:

Regress Y on Z and X on Z and take the ratio, so the estimator is:

$$b_{ILS} = \frac{(Z'Z)^{-1}Z'Y}{(Z'Z)^{-1}Z'X}$$

IV:

Use the following moment condition: $E(Z'(\gamma Q+\eta)) = 0$ This is the exclusion restriction where $(\gamma Q + \eta)$ is the full error. We then have:

$$\begin{aligned} E(Z'(Y - \beta X)) &= 0 \\ E(Z'Y) &= \beta E(Z'X) \\ \beta &= E(Z'X)^{-1}E(Z'Y) \end{aligned}$$

whose sample analog is:

$$b_{IV} = (Z'X)^{-1}(Z'Y)$$

Thus, OLS is the IV estimator when X is uncorrelated with the residuals!

2SLS:

Regress X on Z: we then get the estimator:

$$(Z'Z)^{-1}Z'X$$

with fitted values P_ZX. Then regress Y on the fitted values P_ZX: the resulting 2SLS estimator is:

$$(X'P_ZX)^{-1}(X'P_ZY)$$

All three of these estimators should be consistent, and also be the same in sample (as I showed in the population case).

What happens if we have other X variables, that are exogenous? For example, a constant? Then divide up $X = [X_1\ X_2]$, $Z = [Z_1\ X_2]$

Here X_1 are the endogenous variables, X_2 are the exogenous variables, Z_1 are the instruments. The same formulas as above apply- we are just replacing the endogenous variables with their fitted values, the exogenous variables are remaining the same in the second stage of 2SLS. In general you want to keep the same exogenous variables in the first stage and second stage. Why?

The second stage equation is:

$$Y_i = \beta_2 X_2 + \beta_1 \hat{X}_1 + (\gamma Q + \eta + \beta_1(X_1 - \hat{X_1})$$

If X_2 has been put in the first stage it should be uncorrelated with $X_1 - \hat{X}_1$ by definition (the residual of the 1st stage is uncorrelated with all variables) but if it hasnt it could be correlated with this residual, which is in the error, which will cause inconsistency.

An example here- imagine X_2 is age- age is related to health but also to how often you visit the doctor. If not in the 1st stage, age will be correlated with the residual of the predicted visits and so cause inconsistency.

2.2 Identification

What happens if we have more than one instrument for the same variable? For example, in the health-visits to the doctor example- have a second instrument- some people in the sample were randomly given vouchers for a free visit to the doctor. Then Z_2 is 1 if the individual received a voucher and 0 o/w.

What do we do?

First- let me do a detour to explain identification.

Identification- A parameter is identified if, given an infinite amount of data, we could obtain the true value of the parameter.

The coefficient on visits in the health-visits model is not identified with just data on health and visits. I.e. the conditional distribution of health given visits or joint dist of both not enough! But it is with data on distance as well- thus the exclusion restriction gives us identification!

With one instrument we have the just identified case- one equation and one unknown- so all of the three estimators given earlier (ILS, 2SLS, IV) are the same. This should remind you how many OLS analogies produced the same estimator.

However, what happens when we have more than 1 instrument?

The model is still identified- we can use either instrument as before to identify the parameter. But the model is called over-identified as we have some extra restrictions beyond what are strictly necessary for identification.

Lets see the ILS case: we have an infinite number of possible estimators- can use the first iv, second iv, or a convex combination of both- all will be consistent!

What estimator should we use then? Or rather what is optimal?

Lets look at the IV moments:

$$\begin{aligned} E(Z_1'(Y - \beta X)) &= 0 \\ E(Z_2'(Y - \beta X)) &= 0 \end{aligned}$$

Here we have a system of two equations for 1 unknown (making things simple). In the population both of these will solve by inversion for the same parameter, but in sample estimation error will mean each moment condition will give different estimates.

The logic of GMM says put a weighting matrix on the moments- so weight the moments and get a consistent estimate. But what weighting matrix?

Since we cant just invert the moments, instead minimize their (weighted) square- in the population we know if:

$$E(Z'(Y - X\beta)) \quad = \quad 0$$

at the true value β_0

then

$$\arg\min_{\beta} E(Z'(Y - X\beta))'\Delta E(Z'(Y - X\beta)) \quad = \quad \beta_0$$

At β_0 the criterion function is zero, at other values it will be positive so the arg min is β_0.

Take the sample analog of this estimator:

$$b_N \quad = \quad \arg\min_c (\frac{1}{N}\sum_i Z_i'(Y_i - X_i c))'\Delta(\frac{1}{N}\sum_i Z_i'(Y_i - X_i c))$$

Hansen (1979) shows that the optimal Δ is Σ^{-1}- so this difference becomes like the Mahalanobis distance formula again!

Here Σ is the variance of the moments, or $\Sigma = V(Z'U) = E(Z'UUZ)$.

Taking the first order conditions:

$$\begin{aligned}
(\frac{1}{N}\sum_i Z_i'(X_i))'\Sigma^{-1}(\frac{1}{N}\sum_i Z_i'(Y_i - X_i c)) &= 0 \\
(\frac{1}{N}\sum_i Z_i'(X_i))'\Sigma^{-1}(\frac{1}{N}\sum_i Z_i'(Y_i)) &= (\frac{1}{N}\sum_i Z_i'(X_i))'\Sigma^{-1}(\frac{1}{N}\sum_i Z_i'X_i))c \\
c &= [(\frac{1}{N}\sum_i Z_i'(X_i))'\Sigma^{-1}(\frac{1}{N}\sum_i Z_i'X_i)]^{-1} \\
&\quad (\frac{1}{N}\sum_i Z_i'(X_i))'\Sigma^{-1}(\frac{1}{N}\sum_i Z_i'(Y_i)) \\
c &= ((Z'X)'\Sigma^{-1}(Z'X))^{-1}((Z'X)'\Sigma^{-1}(Z'Y))
\end{aligned}$$

This is just another type of GLS estimator- we can apply our GLS/FGLS ideas from before. Given this is a GMM estimator- it will be consistent, asymptotically normal, and obtain an asymptotic bound given that we use these moments.

The 2SLS estimator will remain the same- the first stage will just include more instruments than one.

The 2SLS estimator is then:

$$(X'P_Z X)^{-1}(X'P_Z Y)$$

If we plug in for $P_Z = Z(Z'Z)^{-1}Z'$ w e have:

$$((Z'X)'(Z'Z)^{-1}Z'X)^{-1}((Z'X)'(Z'Z)^{-1}Z'Y)$$

This is the same as the IV estimator if Σ is proportional to $E(Z'Z)$ i.e. we have conditional homoskedaticity or $E(U^2|Z) = \sigma^2$ so $\Sigma = \sigma^2 E(Z'Z)$.

Thus 2SLS is optimal given conditional homoskedasticity- if this assumption is violated, we can use IV to give a more efficient estimator using info on the variance-covariance matrix or attempt some kind of robust std. errors.

2.3 Problems with IV

2.3.1 Weak Instruments

We have shown that IV estimators are consistent. But in general they are biased-look at the ILS estimates- just b/c numerator and denominator are unbiased by themselves doesnt mean the ratio will be unbiased-b/c expectations are linear and can't be divided like can do with asymptotics.

It turns out this bias is a severe problem when the instruments are weak-so they dont have much predictive power on X. Think about the distance instrument- maybe it changes visits to the doctor, but not by very much.

The IV estimates end up biased in the direction of the probability limit of the OLS estimator.

First- show why IV in general has large std errors (at least with weak instruments)- b/c dividing by zero problem with denominator (1st stage estimates).

Why? Intuition as follows:

First stage estimates are random- randomness comes from the endogenous variable, as estimates are just fitted values from regression of endogenous var on instruments. This randomness will be correlated with second stage errors-so have a bias problem- and its the same bias as OLS.

Lets start with the following system:

$$\begin{aligned} Y &= X\beta + \eta \\ X &= Z\rho + \mu \end{aligned}$$

Here I am just folding the Q omitted variable into each error for simplicity. The omitted variable problem is that η and μ are correlated (for ex, through Q).

The 2SLS estimator is:

$$(X'P_Z X)^{-1}(X'P_Z Y)$$

Plugging in what Y is we have:

$$\begin{aligned}
b_{2SLS} &= (X'P_ZX)^{-1}(X'P_ZY) \\
&= (X'P_ZX)^{-1}(X'P_ZX\beta) + (X'P_ZX)^{-1}(X'P_Z\eta) \\
&= \beta + (X'P_ZX)^{-1}(X'P_Z\eta)
\end{aligned}$$

Then replacing for X:

$$\begin{aligned}
b_{2SLS} &= \beta + (X'P_ZX)^{-1}(X'P_Z\eta) \\
&= \beta + (X'P_ZX)^{-1}(Z\rho + \mu)'P_Z\eta \\
&= \beta + (X'P_ZX)^{-1}(Z\rho)'P_Z\eta + (X'P_ZX)^{-1}\mu P_Z\eta
\end{aligned}$$

Taking expectations and approximating, so expectations can go through divisions:

$$E(b_{2SLS} - \beta) \approx (E(X'P_ZX))^{-1}E(\rho'Z'\eta) + (E(X'P_ZX))^{-1}E(\mu P_Z\eta)$$

For this to be a valid instrument, $E(\rho'Z'\eta) = 0$ so we can simplify this:

$$E(b_{2SLS} - \beta) \approx (E(X'P_ZX))^{-1}E(\mu P_Z\eta)$$

In asymptotics $\mu P_Z\eta$ will converge to 0 (why?) but in finite samples correlation between μ and η will keep this nonzero- and depending on the correlation of the two errors. Thus the estimation error in the fitted values is correlated w/ the second stage error- leading to bias.

We can derive a formula to show what this bias will look like:

$$E(b_{2SLS} - \beta) \approx \frac{\sigma_{\mu\eta}}{\sigma_\mu^2}\frac{1}{F+1}$$

Thus the bias is in the direction of the OLS bias- coming from the correlation between 1st stage and 2nd stage errors. F is the F stat from the 1st stage- a strong instrument has a high F stat and weak instrument a low F stat. As the F stat gets smaller the bias gets bigger- towards the OLS bias- so IV has not solved anything!

There are alternative asymptotics possible here, or using LIML (max lik with normal errors), but really need better instruments!

2.3.2 Heterogenous Treatment Effects

So far we have assumed that the X variable has the same effect for everyone in the population, β. Even in the initial simple example this does not quite make sense- some people (those who are unhealthy) would have a much bigger gain from visiting the doctor than healthy people. Also, these are not pure random coefficients- people will likely respond to the gains from visiting the doctor, so sick people will be more likely to go to the doctor than healthy people. What does IV identify in this case?

To begin to think about this I will introduce the potential outcomes framework. Here:

- D_i is an indicator variable, 1 for person i if she takes the treatment
- Y_{1i} is the potential outcome for person i if she takes the treatment
- Y_{0i} is the potential outcome for person i if she does not take the treatment

In cross section data only it is impossible to observe the individual specific treatment effect: that is, $Y_{1i} - Y_{0i}$. This is because one only observes Y_{1i} if i takes the treatment. D_i is in general going to depend on both Y_{1i} and Y_{0i}- for example, one simple decision rule for D_i is:

$$D_i = 1 \quad if \quad Y_{1i} - Y_{0i} > 0$$

In our earlier example, D_i is 1 if the patient visits a doctor, Y_{1i} is health for the patient if she doesnt go to a doctor, and Y_{0i} is health for the patient if she goes to a doctor.

If we can never get the individual specific treatment effect, what can we say? Well, there are a number of potential interesting treatment effects averaging over the population:

- Average Treatment Effect: $E(Y_{1i} - Y_{0i})$- this is the average effect across everyone in the population
- Treatment on the Treated: $E(Y_{1i} - Y_{0i}|D_i = 1)$- this is the average effect for those who take the treatment
- Treatment on the Untreated:$E(Y_{1i} - Y_{0i}|D_i = 0)$- this is the average effect for those who dont take the treatment
- Policy Relevant Treatment effect: $E(Y|P = p) - E(Y|P = p')$−this is the effect on Y of changing a policy from p' to p

In our health-visit ex. the ATE is the effect on health on avg if everyone went to the doctor, TT just for those who actually went to the doctor, and TUT just for those who didnt actually go the doctor. Here we can see that TT can be positive and high (people who are really sick improve a lot after going to the doctor) and TUT can be negative (if you are healthy, going to the doctor might expose to sickness so you are worse than before).

A “perfect” randomized trial would give you ATE. A perfect randomized trial would give you $E(Y_{1i})$ for those randomized to get the treatment and $E(Y_{0i})$ for those randomized to not get the treatment. Combining these we get:

$$E(Y_{1i}) - E(Y_{0i}) = E(Y_{1i} - Y_{0i})$$

In general those it is hard to set up such a perfect experiment. This is b/c people who have very high values from treatment (for ex, are very sick) will still take it and those with very low values from treatment will not take it. In this case we dont get the expectation over the entire distribution for some parts of the potential outcome distn. This is the “imperfect compliance” problem.

Lets add some more content to the model:

$$\begin{aligned} Y_{1i} &= X\beta_1 + U_{1i} \\ Y_{0i} &= X\beta_0 + U_{0i} \end{aligned}$$

Here we have added some explanatory power for X variables that can affect the different potential outcomes in a separate way. These could be race, sex, age, smoking status, etc. In this case, ATE is:

$$E(Y_{1i} - Y_{0i}) = X\beta_1 - X\beta_0$$

and TT is:

$$E(Y_{1i} - Y_{0i}|D = 1) = X\beta_1 - X\beta_0 + E(U_{1i} - U_{0i}|D = 1)$$

We can then write, using switching regression notation, $Y_i = D_iY_{1i} + (1 - D_i)Y_{0i}$ - why?- so then we have:

$$Y_i = X\beta_0 + D_i(X\beta_1 + U_{1i} - X\beta_0 - U_{0i}) + U_{0i}$$

In this “standard” regression, the coefficient on D_i will be random in the population. What will OLS estimate?

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= X\beta_1 - X\beta_0 \\ &\quad + E(U_{1i} - U_{0i}|D = 1) \\ &\quad + E(U_{0i}|D = 1) - E(U_{0i}|D = 0) \end{aligned}$$

The first part is ATE, the first two TT, but OLS identifies neither parameter. The third part is the standard problem- endogenous var correlated with residual- would be true if $U_{1i} = U_{0i}$. The second part comes from the fact that the gain from D is correlated with the decision to undertake treatment D itself.

What can we do then?

One approach is matching- we assume that $D \perp \{U_1, U_0\}|X$. Under this assumption, people who have the same Xs are the same- so if we compare those with $D = 1$ and $D = 0$, we will get ATE for someone of those Xs. To get the ATE for the population we will have to integrate out over X. However, this is a very strong assumption- need extremely good data to do things nonparametrically, plus why is D changing across people?

Another approach is IV. What is the IV estimator in this simple case, with a discrete instrument as well? IV will be:

(implicitly conditioning on X everywhere here)

$$\frac{E[ZY] - E[Z]E[Y]}{E[ZD] - E[Z]E[D]} =$$

$$\frac{E(Y|Z = 1)Pr(Z = 1) - (Pr(Z = 1) * (E(Y|Z = 1) * Pr(Z = 1) + E(Y|Z = 0)Pr(Z = 0))}{E(D|Z = 1)Pr(Z = 1) - (Pr(Z = 1) * (E(D|Z = 1) * Pr(Z = 1) + E(D|Z = 0)Pr(Z = 0))}$$

$$\begin{aligned} &= \frac{(E(Y|Z = 1) - E(Y|Z = 0))Pr(Z = 1)Pr(Z = 0)}{(E(D|Z = 1) - E(D|Z = 0))Pr(Z = 1)Pr(Z = 0)} \\ &= \frac{E(Y|Z = 1) - E(Y|Z = 0)}{Pr(D = 1|Z = 1) - Pr(D = 1|Z = 0)} \end{aligned}$$

This estimator is sometimes called the Wald estimator- the numerator is the average change in Y through the instrument, the denominator is the change in the probability of the treatment D through the instrument.

Does this parameter identify any treatment effect? It does under some strict conditions, as shown by Imbens and Angrist (1994). They show that under some assumptions IV gives the Local Average Treatment Effect (LATE):

$$E(Y_1 - Y_0|D(1) - D(0) = 1)$$

Here $D(1)$ is the value of D given $Z = 1$ and $D(0)$ is the value of D given $Z = 0$. Thus LATE is the treatment effect for those induced into treatment via the instrument- i.e. if the instrument is zero they would not take the treatment and if the instrument is one they would take the treatment.

Assumption 1. Independence

$$Z_i \perp \{Y_{1i}, Y_{0i}, D_i(z)\}$$

This assumption is the exclusion restriction- Z does not affect Y_1 or Y_0, and the random variable Z does not affect the potential treatment assignments D_{1i}, D_{0i} (only which one is picked!)

Assumption 2. Rank

$$Pr(D = 1|Z)\text{depends on}Z.$$

This assumption just says that the instrument affects D.

The first two assumptions are really just the definition of an instrumental variable. The third assumption is new:

Assumption 3. Monotonicity or Uniformity

$$D_i(1) \geq D_i(0) \forall i \text{or} D_i(1) \leq D_i(0) \forall i$$

This assumption states that the instrument moves everyone in the same way- what is ruled out is an instrument that pushes some people into treatment and others out of treatment.

Under these conditions, we can derive that the IV estimator is equal to LATE:

$$\begin{aligned} E(Y_i|Z_i=1) - E(Y_i|Z_i=0) &= E(Y_{1i}D_i + (1-D_i)Y_{0i}|Z_i=1) - E(Y_{1i}D_i + (1-D_i)Y_{0i}|Z_i=0) \\ &= E(Y_{1i}D_i(1) + (1-D_i(1))Y_{0i}) - E(Y_{1i}D_i(0) + (1-D_i(0))Y_{0i}) \\ &= E((Y_{1i} - Y_{0i})(D_i(1) - D_i(0)) \\ &= E((Y_{1i} - Y_{0i})(D_i(1) - D_i(0)|D_i(1) - D_i(0) = 1)Pr(D_i(1) - D_i(0) = 1) \\ &+ E((Y_{1i} - Y_{0i})(D_i(1) - D_i(0)|D_i(1) - D_i(0) = 0)Pr(D_i(1) - D_i(0) = 0) \\ &+ E((Y_{1i} - Y_{0i})(D_i(1) - D_i(0)|D_i(1) - D_i(0) = -1)Pr(D_i(1) - D_i(0) = -1) \end{aligned}$$

The first line is via the exclusion restriction the independence of Z_i from the first line, the second line that Z will only affect D (hence $D_i(1)$ and $D_i(0)$). The next lines are the LIE over $D_i(1) - D_i(0)$- separated into three expectations times their probabilities.

We can divide up the population of people into four groups based on how the instrument affects treatment, i.e. $D_i(1)$ and $D_i(0)$

1. Always takers- $D_i(1) = D_i(0) = 1$ so changes in the instrument do not affect their treatment status.

2. Never takers- $D_i(1) = D_i(0) = 0$ so changes in the instrument do not affect their treatment status.

3. Compliers- $D_i(1) = 1 > 0 = D_i(0)$ so when the instrument changes to 1 compliers take the treatment where before they did not.

4. Defiers- $D_i(1) = 0 < 1 = D_i(0)$ so when the instrument changes to 1 defiers do not take the treatment where before they did.

The always takers and never takers are not changed by the instrument- so those terms (2nd line) are set to zero as $D_i(1) - D_i(0) = 0$. We then have:

$$\begin{aligned} E(Y_i|Z_i=1) - E(Y_i|Z_i=0) &= E((Y_{1i} - Y_{0i})|D_i(1) - D_i(0) = 1)Pr(D_i(1) - D_i(0) = 1) + \\ &- E((Y_{1i} - Y_{0i})|D_i(1) - D_i(0) = -1)Pr(D_i(1) - D_i(0) = -1) \end{aligned}$$

The monotonicity condition assumes that there are no two way flows- in this case (WLOG) there are no defiers so the last term is assumed zero (i.e. $Pr(D_i(1) - D_i(0) = -1) = 0$). This assumption is critically important- if there are defiers then even if all causal effects are positive (so $Y_{1i} - Y_{0i} > 0 \forall i$) we can have $E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) < 0$ so the IV gives a negative result and the LATE is negative.

So, we then have:

$$
E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \quad = \quad E((Y_{1i} - Y_{0i})|(D_i(1) - D_i(0) = 1)Pr(D_i(1) - D_i(0) = 1)
$$

If we take the denominator:

$$
\begin{aligned}
Pr(D = 1|Z = 1) - Pr(D = 1|Z = 0) &= Pr(D_i(1) = 1) - Pr(D_i(0) = 1) \\
&= Pr(D_i(1) - D_i(0) = 1)
\end{aligned}
$$

Combining these two expressions we have:

$$
\begin{aligned}
\frac{E(Y|Z = 1) - E(Y|Z = 0)}{Pr(D = 1|Z = 1) - Pr(D = 1|Z = 0)} &= \frac{E((Y_{1i} - Y_{0i})|D_i(1) - D_i(0) = 1)Pr(D_i(1) - D_i(0) = 1)}{Pr(D_i(1) - D_i(0) = 1)} \\
&= E((Y_{1i} - Y_{0i})|D_i(1) - D_i(0) = 1)
\end{aligned}
$$

What does LATE tell us (assuming all its assumptions hold)? It gives the treatment effect for the section of the population induced to switch due to the instrument- thus it is an instrument dependent effect. Different instruments will give different LATEs- LATE is really a slope parameter based on the instrument (change in Y over change in D).

One example- compulsory schooling laws as an instrument- meant had to go one extra year or one extra grade (so from 8th grade to 9th grade)- outcome var is health, endogenous var schooling.

The IV estimator here (or LATE) tells us the effect on health of an extra year of schooling from 8th grade to 9th grade for those people who would drop out at 8th grade otherwise- this is very different from the effect for those going from HS dropout to HS complete, or HS complete to college.

Does LATE correspond to any of the other treatment parameters? In general no- it will not correspond to treatment on the treated (as its getting a subset of the pop who would switch given the instrument, ignoring the always takers who are in the TT).

We can see this below:

$$
\begin{aligned}
E(Y_{1i} - Y_{0i}|D_i = 1) &= E(Y_{1i} - Y_{0i}|D_i(0) = 1)Pr(D_i(0) = 1|D_i = 1) \\
&+ E(Y_{1i} - Y_{0i}|D_i(1) > D_i(0))Pr(D_i(1) > D_i(0), Z_i = 1|D_i = 1)
\end{aligned}
$$

Thus the TT is a weighted average of the LATE effect (effect on compliers) and the effect on always takers. TuT will proceed in a similar way- avging with never takers.

LATE only provides ATE if there are values of the instrument where everyone enters the treatment and values of the instrument where no one enters- this is called “identification at infinity”. Otherwise ATE is unidentified.

What happens if there are multiple values of the instrument (but discrete for this example- can always make a continuous variable discrete)? The IV estimator is a weighted average of the LATE parameters for each discrete change (1 to 2, 2 to 3, etc.)- with nonnegative weights that sum to 1, as long as monotonicity holds.

Similarly, if the treatment D is discrete (think schooling- 11 grades to 12, 12 grades to 13, or doctor visits- 0 to1, 1 to 2, 2 to 3) the IV estimator is a weighted average weighting over the LATE for each change.

More than 1 instrument or X covariates are handled the same way.

LATE only holds when monotonicity or uniformity holds- when would this assumption be violated?

Some examples:

The instrument affects different people in different ways. For ex, imagine a housing subsidy program funded by taxed as an instrument for migration- this would induce poor people to enter the state and rich people to leave the state- so we have 2 way flows.

Another- two officials determine who enters a program. One is much stricter than the other- so on average its harder to get in- but the stricter one may still let some people in that the lax one does not.

Or- only use one instrument when they are many- if instrument used changes values of others when you change it (through say an interaction, or correlated between vars)- can have two way flows.

Ex: 1st instrument is distance to doctor, 2nd instrument is primary care voucher, only available in inner city- as change 1st instrument, value of 2nd changing which can cause two way flows.

Last comment: Can also look at this problem through an explicit structural model of the problem- look at what is identified given assumptions on errors:

$$\begin{aligned} Y_{1i} &= X\beta_1 + U_{1i} \\ Y_{0i} &= X\beta_0 + U_{0i} \\ D_i &= 1(Z\gamma > U_c) \end{aligned}$$

For example- if put joint normal errors- what is identified in this model? Selection equation determines which outcome you choose. Unfortunately not enought time to cover this!

3 Simulatenous Equations models

One of the most basic economic models is a model of Supply and Demand. In the standard model, there is a demand function and supply function- demand is decreasing in price and supply is increasing in price. In equilibrium quantity and price are set where the supply function intersects the demand function.

Thus, as long as we are in equilibrium, we will observe quantity and price pairs (q^*, p^*) from markets in equilibrium across time or space.

Now imagine we want to estimate the demand function or supply function (b/c we are interested in the elasticity of demand or elasticity of supply, say). What should we do?

Will simple OLS give us the demand function or supply function? It turns out that OLS will give neither. To see this we will need to set up the structural model:

Setup $\{S(), D(), p, q, X\}$

The Definition of Equilibrium in this model is that Supply=Demand, or:

$$\begin{aligned} q &= D(p) \\ q &= S(p) \end{aligned}$$

The Structural Model of Supply and Demand is:

$$\begin{aligned} S(t) &= \beta_1 t + X\alpha_1 + U_1 \\ D(t) &= \beta_2 t + X\alpha_2 + U_2 \end{aligned}$$

Here X is some vector of covariates potentially affecting both the supply and demand functions. β_1 and β_2 are related to the elasticity of supply and elasticity of demand:

$$\begin{aligned} \frac{d\log S(p)}{d\log p} &= \frac{\beta_1 p}{S(p)} \\ \frac{d\log D(p)}{d\log p} &= \frac{\beta_2 p}{D(p)} \end{aligned}$$

Here the elasticities will depend on the levels of price and quantity as well as the estimated effects from the supply and demand functions.

U_1 and U_2 are unobserved demand and supply shocks. By assumption they are mean zero conditional on covariates, and have a particular variance matrix:

$$\begin{aligned} E(U_1|X) &= 0 \\ E(U_2|X) &= 0 \\ V(\begin{matrix} U_1 \\ U_2 \end{matrix} |X) &= \Sigma^* \end{aligned}$$

Imagine we were to do OLS on the system. What would we get?

The system of equations is:

$$\begin{aligned} q &= \beta_1 p + X\alpha_1 + U_1 \\ q &= \beta_2 p + X\alpha_2 + U_2 \end{aligned}$$

where we have assumed markets are in equilibrium so demand equals supply. Putting endogenous variables on the LHS and going to matrix notation I have:

$$\begin{bmatrix} 1 & -\beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} X + \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$$

The inverse of this matrix is:

$$\begin{bmatrix} 1 & -\beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} = \frac{1}{\beta_1 - \beta_2} \begin{bmatrix} -\beta_2 & \beta_1 \\ -1 & 1 \end{bmatrix}$$

If we invert the matrix we will get the Reduced Form Equations:

$$\begin{aligned} q &= \Pi_1 X + V_1 \\ p &= \Pi_2 X + V_2 \end{aligned}$$

where the parameters are:

$$\begin{aligned} \Pi_1 &= \frac{\beta_1 \alpha_2 - \beta_2 \alpha_1}{\beta_1 - \beta_2} \\ \Pi_2 &= \frac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2} \\ V_1 &= \frac{\beta_1 U_2 - \beta_2 U_1}{\beta_1 - \beta_2} \\ V_2 &= \frac{U_2 - U_1}{\beta_1 - \beta_2} \end{aligned}$$

First of all, lets look at the model with no Xs- what happens if we regress quantity on price (the simple OLS estimator). Notice that here both quantity and price are just functions of the random shocks in equilibrium:

$$\begin{aligned} q &= V_1 \\ p &= V_2 \end{aligned}$$

The OLS estimator of the quantity-price relationship will then be:

$$\begin{aligned} \frac{Cov(q,p)}{Var(p)} &= \frac{Cov(V_1, V_2)}{Var(V_2)} \\ &= \frac{Cov(\beta_1 U_2 - \beta_2 U_1, U_2 - U_1)}{Var(U_2 - U_1)} \\ &= \frac{\beta_1 Var(U_2) + \beta_2 Var(U_1) - 2(\beta_1 + \beta_2) Cov(U_1, U_2)}{Var(U_2) + Var(U_1) - 2Cov(U_1, U_2)} \end{aligned}$$

If we have variances of 1 and a covariance of 0, this simplifies to:

$$\frac{\beta_1 + \beta_2}{2}$$

We can see that in either case, OLS does not deliver either the demand equation parameters or the supply equation parameters- just some mix of the two. It is not just biased in some direction (as in the Omitted Variable Bias case) but really just uninformative- a mix of the slopes of demand and supply (though I suppose it bounds the demand slope and supply slope)- show a graph with demand and supply moving around.

Another way to see this is that price is a function of both U_1 and U_2. If I estimate the model:

$$q = \beta_1 p + U_1$$

p will be correlated with the error as its a function of U_1!

What can we do then?

One can use OLS to estimate both of these reduced form equations and obtain Π_1, Π_2 since X is uncorrelated with U_1, U_2 and thus with V_1, V_2.

That is, we estimate:

$$\begin{aligned} E(q|X) &= X\Pi_1 \\ E(p|X) &= X\Pi_2 \end{aligned}$$

For simplicity let X be made up of X_1 and X_2 as scalars where X_1 is the average income in the market and X_2 is the oil price. Then we have:

$$\begin{aligned} E(q|X) &= X_1\Pi_{11} + X_2\Pi_{12} \\ E(p|X) &= X_1\Pi_{21} + X_2\Pi_{22} \end{aligned}$$

What do these reduced form regressions tell us? They tell us what happens to equilibrium price and quantity as we change X covariates. For example, as income changes, both supply and demand are potentially shifting until we reach a new equilibrium value of quantity and price.

What is identified then? Can we obtain any of the structural parameters from the reduced form parameters we just estimated?

Plug in the reduced form equations into the structural equations, we then have:

$$\begin{aligned} X\Pi_1 + V_1 - \beta_1 X\Pi_2 - \beta_1 V_2 &= X\alpha_1 + U_1 \\ X\Pi_1 + V_1 - \beta_2 X\Pi_2 - \beta_1 V_2 &= X\alpha_2 + U_2 \end{aligned}$$

For the X variable parameters to be the same, we must have:

$$\begin{aligned}\Pi_{1j} - \beta_1 \Pi_{2j} &= \alpha_{1j} \\ \Pi_{1j} - \beta_2 \Pi_{2j} &= \alpha_{2j}\end{aligned}$$

Writing these in derivatives, these mean:

$$\begin{aligned}\frac{\partial q}{\partial X_1} - \frac{\partial q_s}{\partial p}\frac{\partial p}{\partial X_1} &= \frac{\partial q_s}{\partial X_1} \\ \frac{\partial q}{\partial X_1} - \frac{\partial q_d}{\partial p}\frac{\partial p}{\partial X_1} &= \frac{\partial q_d}{\partial X_1}\end{aligned}$$

Thus, the changes in equilibrium quantity and price as a function of a covariate (X_1) depend on how the supply function and demand function change with price and with the covariate. Imagine that the covariate only affects demand- can think of tracing out supply curve- or moving to higher demand function with shock and then price adjusting to hit equilibrium.

Equating $\frac{\partial q}{\partial X_1}$ between equations we have:

$$\frac{\partial q_d}{\partial X_1} + \frac{\partial q_d}{\partial p}\frac{\partial p}{\partial X_1} = \frac{\partial q_s}{\partial X_1} + \frac{\partial q_s}{\partial p}\frac{\partial p}{\partial X_1}$$

If $\frac{\partial q_s}{\partial X_1} = 0$ then can see as two changes on demand side (to new demand curve, then along demand curve) or one change on supply curve from price- show this in graph.

If all of the αs are non zero, there is nothing we can do- none of the structural parameters are identified. The math reason is more parameters than variables- the economic intuition is that both supply and demand functions are moving around with each covariate that is changing- so any change in a covariate can be rationalized as the demand curve shifting or supply curve shifting. I could assume that X_1 is tracing out the demand curve, supply curve, or some mixture- these cant be separated in the data.

What if one of the αs is equal to zero? Then all of the structural parameters (of one equation) are identified.

This is an exclusion restriction again- some variable is in the demand eqn but not the supply eqn, or vice versa.

Lets assume $\alpha_{11} = 0$ but $\alpha_{21} \neq 0$. This factor X_1 , income, does not affect supply but does affect demand.

Then:

$$\begin{aligned}\frac{\partial q}{\partial X_1} - \frac{\partial q_s}{\partial p}\frac{\partial p}{\partial X_1} &= 0 \\ \frac{\Pi_{11}}{\Pi_{21}} &= \beta_1 = \frac{\partial q_s}{\partial p} \\ \Pi_{ij} - \beta_1 \Pi_{2j} &= \alpha_{ij} \forall j \neq 1\end{aligned}$$

Once we know β_1 or the price effect on demand we can get all the other demand parameters.

Suppose $\alpha_{12} = 0$ but $\alpha_{22} \neq 0$- then

$$\beta_1 = \frac{\Pi_{12}}{\Pi_{22}}$$

Then we have an over identified system- 2 ways of learning β_1.

In general, what kind of variables will satisfy this problem?

Instruments for demand- supply factors like weather or rainfall, oil prices, input costs, etc.

Instruments for supply- demand factors like "exogenous" advertising, demographics or size of mkt, govt spending or output sector spending, etc.

Notice we can write this problem as follows:

$$\begin{aligned} q &= \beta_1 p + \alpha_1 X + U_1 \\ p &= \Pi_1 X + V_1 \end{aligned}$$

and it looks just like the IV systems we were thinking about before- so all of the estimation theory will be the same. The estimators above were the ILS estimators of the IV system.