Structural Estimation

Christopher Taber

University of Wisconsin

November 21, 2016

Structural Models

So far in this class we have been thinking about the “evaluation problem”

$$Y_i = \alpha T_i + \varepsilon_i$$

However, this always required data on T_i

What do we if we don’t have data on the treatment or policy we are interested in?

Most interesting reason is because it is a proposed policy-one that has never been previously implemented.

We need to estimate a “structural model” and use the structure to simulate the policy

Quote from Frank Knight

> *The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of a problem of knowledge depends on the future being like the past*

What does structural mean?

No obvious answer, it means different things to different people

3 Definitions:

- Parameters are policy invariant
- Estimation of preference and technology parameters in a maximizing model (perhaps combined with some specification of markets)
- The structural parameters a simultaneous equations model

For that matter what does reduced form mean

Now for many people it essentially means anything that is not structural

What I think of as the classic definition is that reduced form parameters are a known function of underlying structural parameters.

- fits classic Simultaneous Equation definition
- might not be invertible (say without an instrument)
- for something to be reduced form according to this definition you need to write down a structural model
- this actually has content-you can sometimes use reduced form models to simulate a policy that has never been implemented (as often reduced form parameters are structural in the sense that they are policy invariant)

Advantages and disadvantages of “structural" and “design-based”

Two caveats first

- To me the fact that there are advantages and disadvantages makes them complements rather than substitutes
- These are arguments that different people make, but obviously they don’t apply to all (or maybe even most) structural work or non-structural work-there are plenty of good and bad papers of any type

Differences between “structural" and “design-based” approaches

Structural	Design-Based
More emphasis on External Validity	More emphasis on Internal Validity
Tends to be more complicated involving many parameters	Focuses on estimation of a single (or small number of) parameters
Map from parameters to implications clearer	Map from data to parameters more transparent
Formalizes conditions for external validity	Requires fewer assumptions
Forces one to think about where data comes from	Might come from somewhere else

Steps for writing a structural paper for policy evaluation

1. Identify the policy question to be answered
2. Write down a model that can simulate policy
3. Think about identification/data (with the goal being the policy counterfactual)
4. Estimate the model
5. Simulate the policy counterfactual

Formalization

Lets use our notation from the identification notes

In the current state of the world the data is generated by

$$\begin{aligned} X_i \sim & H(X_i) \\ u_i \sim & F(u_i;\theta) \\ \Upsilon_{0i} = & y(X_i, u_i;\theta) \end{aligned}$$

Assume that under the policy regime π the data generation process is

$$\begin{aligned} X_{\pi i} \sim & H_\pi(X_{\pi i}) \\ u_{\pi i} \sim & F_\pi(u_{\pi i};\theta) \\ \Upsilon_{\pi i} = & y_\pi(X_{\pi i}, u_{\pi i};\theta) \end{aligned}$$

where H_π, F_π, and y_π are known up to θ

The counterfactual is often an expected difference in some outcome in the two regimes

$$\begin{aligned}\psi(\theta) =& E\left(\Gamma(\Upsilon_{\pi i}) - \Gamma(\Upsilon_i)\right)\\ =& \int\int \Gamma(y_{\pi}(X_{\pi i}, u_{\pi i}; \theta)) dF_{\pi}(u;\theta) dH_{\pi}(X)\\ &- \int\int \Gamma(y(X_i, u_i; \theta)) dF(u;\theta) dH(X)\end{aligned}$$

(there is nothing special about expected values, it could be some other function of the data but this covers most cases)

We can go beyond that and consider functions of things not observed in the data (most obvious example utility) where

$$\begin{aligned}
V_i =& v(X_i, u_i; \theta) \\
V_{\pi i} =& v_\pi(X_{\pi i}, u_{\pi i}; \theta) \\
\psi(\theta) =& E(V_{\pi i}) - E(V_i)
\end{aligned}$$

The most standard way to identify the policy effect is though the use of the full structural model.

If θ is identified, $\psi(\theta)$ is identified

This takes 2 main assumptions

1. H_π, F_π, and y_π are known up to θ
   - we require that either the data generating process is policy invariant, or we know precisely how it will change with the policy
   - this is in some sense the classic definition “structure," its generally not testable
2. θ is identified
   - That is we have point identified the data generating process
   - **and** the θ that determine F_0 and y_0 are the same θ that determine F_π and y_π

One can see how these relate to the Knight quote at the beginning

Sometimes you don’t always need to identify the full structural model but only part of it

That is you might only be able to partially identify θ but thats all you need

These cases are rare but important

I want to focus on estimation of the full structural model

Example 1

Lets consider to the classic simultaneous equations model

Model for gas:

Supply Curve

$$Q_t = \alpha_s P_t + X_{1t}'\beta + u_{1t}$$

Demand Curve

$$Q_t = \alpha_d P_t + X_{2t}'\gamma + u_{2t}$$

So $X_t = (X_{1t}, X_{2t})$, $u_t = (u_{1t}, u_{2t})$, $\Upsilon_t = (P_t, Q_t)$

We can solve for prices and quantities as

$$\begin{aligned}(P_t, Q_t) =& y(X_i, u_i; \theta) \\ =& \left[\begin{array}{c} \frac{Z_t'\gamma - X_{1t}'\beta + u_{2t} - u_{1t}}{\alpha_s - \alpha_d} \\ \alpha_s \frac{X_{2t}'\gamma + u_{2t}}{\alpha_s - \alpha_d} - \alpha_d \frac{X_{1t}'\beta + u_{1t}}{\alpha_s - \alpha_d} \end{array}\right]\end{aligned}$$

Now consider a gas tax imposed on consumers, so now

$$A_{\pi t} = \alpha_d \left(1+\pi\right) P_{\pi t} + Z_t' \gamma + v_t$$

The equilibrium effect is

$$\begin{aligned}
(P_{\pi t}, Q_{\pi t}) =& y_{\pi}(X_i, u_i; \theta) \\
=& \begin{bmatrix} \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d(1+\pi)} \\ \alpha_s \frac{Z_t'\gamma + v_t}{\alpha_s - \alpha_d(1+\pi)} - \alpha_d \left(1+\pi\right) \frac{X_t'\beta + u_t}{\alpha_s - \alpha_d(1+\pi)} \end{bmatrix}
\end{aligned}$$

Note that you are taking the model seriously here-all of the parameters are policy invariant

Example 2

Lets think about the Generalized Roy Model

For

$$\begin{aligned} U_{ji} =& Y_{ji} + \varphi_j(Z_i, X_{0i}) + \nu_{ji} \\ Y_{ji} =& g_j(X_{ji}, X_{0i}) + \varepsilon_{ji} \end{aligned}$$

But now let $j = hs$ represent high school and $j = col$ represent college

$$\begin{aligned} \Upsilon_i =& (j, y_{ji}) \\ =& y_0(X_i, u_i; \theta) \\ =& \begin{cases} (hs, Y_{hsi}) & U_{hsi} \geq U_{coli} \\ (col, Y_{coli}) & U_{hsi} < U_{coli} \end{cases} \end{aligned}$$

Now suppose we subsidize college by lowering the cost of college by π We are going to ignore equilibrium effects

$$\begin{aligned}\Upsilon_{\pi i} =& y_{\pi}(X_i, u_i; \theta) \\ =& \begin{cases} (hs_{\pi}, Y_{hsi}) & U_{hsi} \geq U_{coli} + \pi \\ (col_{\pi}, Y_{coli}) & U_{hsi} < U_{coli} + \pi \end{cases}\end{aligned}$$

What will this do to average wages?

$$\begin{aligned}\psi\left(\pi\right) = & E\left(1\left(U_{hsi} \geq U_{coli} + \pi\right) Y_{hsi} + 1\left(U_{hsi} < U_{coli} + \pi\right) Y_{coli}\right) \\ & - E\left(1\left(U_{hsi} \geq U_{coli}\right) Y_{hsi} + 1\left(U_{hsi} < U_{coli}\right) Y_{coli}\right)\end{aligned}$$

Other Examples

- Effects of Affordable Care Act on labor market outcomes (Aizawa and Fang, 2015)
- Tuition Subsidies on Health (Heckman, Humphries, and Veramundi, 2015)
- Effects of extending length of payment for college loan programs on college enrollment (Li, 2015)
- Peer effects of school vouchers on public school students (Altonji, Huang, and Taber, 2015)
- Tax credits versus income support (Blundell, Costa Dias, Meghir, and Shaw, 2015)
- Effects of immigration on short and long run wages of natives (Colas, 2016)
- Welfare effects of alternative designs of school choice programs (Calsamiglia, Fu, and Guell, 2016)

Other reasons to write structural models

While this is the classic use of a structural model it is not the only one.

Other motivations:

- Further evaluation of an established policy: we might want to know welfare effect (V_i)
- Basic Research-we want to understand the world better
  - Use data to help understand model
  - Use model to help understand data (use structural model as a lens)
  - A policy doesn't really have to be an actual policy
  - One methodological step in estimation of model

Estimation

So how do we estimate the counterfactual?

I want to focus on the following procedure:

1. Specify full data generating process (as above)
2. Estimate θ
3. Calculate $\psi(\widehat{\theta})$

There are special cases that don't do this exactly-but this is the typical case

The part of this I want to focus on for the rest of these lecture notes is estimating θ

Lets focus on a simple parametric version of the generalized Roy model

$$
\begin{aligned}
d_i =& 1\left(Z_{it}'\gamma + \alpha(Y_{i1} - Y_{i0}) + u_i\right) \\
Y_{i0} =& X_i'\beta_0 + \varepsilon_{i0} \\
Y_{i1} =& X_i'\beta_1 + \varepsilon_{i1}
\end{aligned}
$$

Background maximum likelihood

For some random vector Y, let $f(Y;\theta)$ be the density of Y if it is generated by a model with parameter θ

The likelihood function just writes the function the other way:

$$\ell(\theta;Y) = f(Y;\theta).$$

Let θ_0 represent the true parameter

We use Jensen’s inequality which implies that for any random variable X_i, the fact that log is concave implies that:

$$E(log(X_i)) \leq log(E(X_i))$$

We apply this with

$$X_i = \frac{\ell(\theta; Y_i)}{\ell(\theta_0; Y_i)}$$

The key result is that

$$\begin{aligned} E\left(\frac{\ell(\theta; Y_i)}{\ell(\theta_0; Y_i)}\right) &= \int \frac{\ell(\theta; Y_i)}{\ell(\theta_0; Y_i)} f(Y_i; \theta_0) dY_i \\ &= \int \frac{f(Y_i; \theta)}{f(Y_i; \theta_0)} f(Y_i; \theta_0) dY_i \\ &= \int f(Y_i; \theta) dY_i \\ &= 1 \end{aligned}$$

because $f(Y_i; \theta)$ is a density.

Thus

$$
\begin{aligned}
E\left(log\left(\frac{\ell(\theta;Y_i)}{\ell(\theta_0;Y_i)}\right)\right) =& E\left(log\left(\ell(\theta;Y_i)\right)\right) - E\left(log\left(\ell(\theta_0;Y_i)\right)\right) \\
\leq& log(1)
\end{aligned}
$$

or

$$E\left(log\left(\ell(\theta;Y_i)\right)\right) \leq E\left(log\left(\ell(\theta_0;Y_i)\right)\right)$$

thus we know that the true value of θ maximizes $E\left(log\left(\ell(\theta;Y_i)\right)\right)$

Maximum likelihood is just the sample analogue of this

Choose $\widehat{\theta}$ as the argument that maximizes

$$\frac{1}{N}\sum_{i=1}^{N}\log(\ell(\theta;Y_i))$$

With our model we take $Y_i = (X_i, \Upsilon_i)$ and being loose with notation

$$\begin{aligned}
\widehat{\theta} =& argmax \frac{1}{N}\sum_{i=1}^{N} \log(\ell(\theta; X_i, \Upsilon_i)) \\
=& argmax \frac{1}{N}\sum_{i=1}^{N} \log(\ell(\theta; \Upsilon_i \mid X_i) H(X_i)) \\
=& argmax \frac{1}{N}\sum_{i=1}^{N} \log(\ell(\theta; \Upsilon_i \mid X_i)) + \log(H(X_i)) \\
=& argmax \frac{1}{N}\sum_{i=1}^{N} \log(\ell(\theta; \Upsilon_i \mid X_i))
\end{aligned}$$

Example

Lets think of a binary variable so $Y_i = 1$ with probability p and 0 with probability $(1-p)$

The true value is p_0

Then

$$
\begin{aligned}
E\left(log\left(\ell(p;Y_i)\right)\right) =& p_0 log\left(\ell(p;1)\right) + (1-p_0) log\left(\ell(p;0)\right) \\
=& p_0 log(p) + (1-p_0) log(1-p)
\end{aligned}
$$

this is maximized at

$$\frac{p_0}{p} = \frac{1-p_0}{1-p}$$

or

$$p = p_0$$

The sample analogue will be

$$\frac{1}{N}\sum_{i=1}^{N} Y_i log(p) + (1 - Y_i) log(1 - p) = \bar{Y} log(p) + (1 - \bar{Y}) log(1 - p)$$

and the solution will be

$$\widehat{p} = \bar{Y}$$

Cramer-Rao Lower Bound

The most important result for MLE is that it is efficient

In particular no alternative estimator can have a lower asymptotic variance

The Likelihood function for the Generalized Roy Model

Its a real mess-I want to go through with it just to demonstrate as an example how you calculate the likelihood function

Its easier to break it down into two pieces.

If $d_i = 1$ we observe Y_{i1} but not Y_{i0}

The messy thing here is that we can't identify the joint distribution of ε_{i0} and ε_{i1}

First consider the case in which $d_i = 1$

$$\begin{aligned} d_i =& 1\left(Z_{it}'\gamma + \alpha(Y_{i1} - X_i'\beta_0) + u_i - \alpha\varepsilon_{i0}\right) \\ \equiv& 1\left(Z_{it}'\gamma + \alpha(Y_{i1} - X_i'\beta_0) + v_{i1}\right) \\ Y_{i1} =& X_i'\beta_1 + \varepsilon_{i1} \end{aligned}$$

Let Σ_1 be the variance/covariance matrix of $(v_{i1}, \varepsilon_{i1})$ so

$$\ell(\theta; Y_{i1}, 1) = \int_{-Z_{it}'\gamma - \alpha(Y_{i1} - X_i'\beta_0)}^{\infty} \phi(v, Y_{i1} - X_i'\beta_1; \Sigma_1) dv$$

(where θ is all of the parameters of the model) and analogously for $d_i = 0$

$$\begin{aligned} d_i =& 1\left(Z_{it}'\gamma + \alpha(X_i'\beta_1 - Y_{i0}) + u_i + \alpha\varepsilon_{i1}\right) \\ \equiv& 1\left(Z_{it}'\gamma + \alpha(X_i'\beta_1 - Y_{i0}) + v_{i0}\right) \\ Y_{i0} =& X_i'\beta_0 + \varepsilon_{i0} \end{aligned}$$

Let Σ_0 be the variance/covariance matrix of $(v_{i0}, \varepsilon_{i0})$ so

$$\ell(\theta; Y_{i0}, 0) = \int_{-\infty}^{-Z_{it}'\gamma - \alpha(X_i'\beta_1 - Y_{i0})} \phi(v, Y_{i0} - X_i'\beta_0; \Sigma_0) dv$$

Thus the log-likelihood function is

$$\frac{1}{N}\sum_{i=1}^{N}\Bigg[d_i\int_{-Z_{it}'\gamma-\alpha(Y_{i1}-X_i'\beta_0)}^{\infty}\phi(v,Y_{i1}-X_i'\beta_1;\Sigma_1)dv \\ +(1-d_i)\int_{-\infty}^{-Z_{it}'\gamma-\alpha(X_i'\beta_1-Y_{i0})}\phi(v,Y_{i0}-X_i'\beta_0;\Sigma_0)dv\Bigg]$$

Often in these models the integral is a big problem in calculating the model.

To see an interesting version of this, consider a panel data version of the treatment effect model

$$\begin{aligned} d_{it} =& 1\left(Z_{it}'\gamma + \mu_i^d + \varepsilon_{it}\right) \\ Y_{1it} =& X_{it}'\beta_1 + \mu_i^1 + v_{it}^1 \\ Y_{0it} =& X_{it}'\beta_0 + \mu_i^0 + v_{it}^0 \end{aligned}$$

and take a simple version of the model where the ε_{it}, v_{it}^1, and v_{it}^0 are all jointly independent and normal

Write the likelihood function as if μ_i were known:

$$\tilde{\ell}_i(\theta;\mu_i) \equiv \prod_{t=1}^{T} \Big[d_{it}\Phi(Z_{it}'\gamma + \mu_i^d)\phi(Y_{1it} - X_{it}'\beta_1 - \mu_i^1;\sigma_1^2) \\ +(1-d_{it})(1-\Phi(Z_{it}'\gamma + \mu_i^d))\phi(Y_{0it} - X_{it}'\beta_0 - \mu_i^0;\sigma_0^2) \Big]$$

Then one can write

$$\ell_i\left(\theta\right) = \int\int\int \tilde{\ell}_i(\theta;\mu)d\Phi(\mu;\Sigma_\theta)$$

Even though this is a relatively simple problem, approximating this three dimensional integral is very difficult-and we have to do it for every person in the data every time we evaluate the likelihood function

Lets think about three solutions

Gauss Hermite Quadrature

Lets focus on the one dimensional case in which the error term is normally distributed

We want to estimate something like $E(F(u_i))$ where $u_i \sim N(\mu, \sigma^2)$ Then write

$$u_i = \mu + \sigma\varepsilon_i$$

Then

$$E(F(u_i)) = \int F(\mu + \sigma\varepsilon)\frac{1}{\sqrt{2\pi}}e^{\frac{-\varepsilon_i^2}{2}}$$

It turns out that we can approximate this well as

$$\int F(\mu+\sigma\varepsilon)\frac{1}{\sqrt{2\pi}}e^{\frac{-\varepsilon_i^2}{2}} \approx \sum_{\ell=1}^{K} F(\mu+\sigma e_\ell)p_\ell$$

where one can look up the p_ℓ and the e_ℓ

This approximation is exact if F is a $(K-1)^{th}$ order polynomial

My experience is that this works great with $K \approx 7$

Higher dimensions are easy to incorporate since we can generate model with

$$
\begin{aligned}
u_{1i} &= \varepsilon_{1i} \\
u_{2i} &= a_1\varepsilon_{1i} + a_2\varepsilon_{2i} \\
u_{3i} &= b_1\varepsilon_{1i} + b_2\varepsilon_{2i} + b_3\varepsilon_{3i}
\end{aligned}
$$

where the εs are iid normals

Even with three dimensions this is still pretty big: $7^3 = 343$

“Heckman Singer” Heterogeneity

One way to make the problem simpler is to use a different assumption about the distribution of θ.

Rather than assuming it is Normal assume it takes on a finite number of values K each with probability p_j.

Index each of those values as $\mu_{(j)} = (\mu^d_{(j)}, \mu^1_{(j)}, \mu^0_{(j)})$

Then the likelihood function becomes a one dimensional integral

$$\ell_i\left(\theta\right) = \sum_{j=1}^{K} p_j \tilde{\ell}_i(\theta; \mu_{(j)})$$

When K is small this makes this pretty easy.

However, this often adds a lot of parameters and is not really that flexible.

With $K = 5$ it still seems pretty restrictive

However you have 19 parameters (15 values of $\mu_{(j)}$+ 4 values of p_j) to estimate

Also often a lot of local optima making it even harder to find maximum

If K increases with sample size it is semi-parametric (as in the original Heckman Singer paper)

Simulation

Another way to evaluate the likelihood function is to simulate.

In particular suppose we drew random variables $\mu_s \equiv (\mu_s^d, \mu_s^1, \mu_s^0)$ from a normal with mean zero and variance Σ_θ then as $S \to \infty$

$$\begin{aligned} \frac{1}{S}\sum_{s=1}^{S} \tilde{\ell}_i(\theta;\mu_s) \xrightarrow{p} & E\left[\tilde{\ell}_i(\theta;\mu_i)\right] \\ = & \int\int\int \tilde{\ell}_i(\theta;\mu) d\Phi(\mu;\Sigma_\theta) \\ = & \ell_i(\theta) \end{aligned}$$

This can be nice

One advantage is that we don’t need to waste time getting a good estimate of integral in places in which the density is small

However notice that this is a law of large numbers that has to hold for every single observation in our data every single time we do a function evaluation

Might be pretty tricky

Generalized Method of Moments

Another way to estimate such a model is by GMM, simulated method of moments, or indirect inference

I am not sure these terms mean the same thing to everyone, so I will say what I mean by them but recognize it might mean different things to different people.

Lets continue to use our data generation process above with (X_i, Υ_i) the observed data

The standard GMM model would come up with a set of moments

$$m(X_i, \Upsilon_i, \theta)$$

for which

$$E(m(X_i, \Upsilon_i, \theta_0)) = 0$$

the sample analogue comes from recognizing that

$$\frac{1}{N}\sum_{i=1}^{N} m(X_i, \Upsilon_i, \theta_0) \approx 0$$

But more generally we are overidentified so we choose $\widehat{\theta}$ to minimize

$$\left[\frac{1}{N}\sum_{i=1}^{N} m(X_i, \Upsilon_i, \theta)\right]' W \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i, \Upsilon_i, \theta)\right]$$

Relationship between GMM and MLE

Actually in one way you can think of MLE as a special case of GMM

We showed above that

$$\theta_0 = argmax \left[E \left(log \left(\ell(\theta; X_i, \Upsilon_i) \right) \right) \right]$$

but as long as everything is well behaved this means that

$$E \left(\frac{\partial log \left(\ell(\theta; X_i, \Upsilon_i) \right)}{\partial \theta} \right) = 0$$

We can use this as a moment condition

In fact this is kind of equivalent to MLE

We choose $\widehat{\theta}$ to maximize

$$\frac{1}{N}\sum_{i=1}^{N} log\left(\ell(\theta; X_i, \Upsilon_i)\right)$$

and at the maximum

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial log\left(\ell(\widehat{\theta}; X_i, \Upsilon_i)\right)}{\partial \theta} = 0$$

which is exactly what one would get out of GMM

The one very important caveat is that this is only true if the log likelihood function is strictly concave

Otherwise there might be multiple solutions to the first order conditions, but only one actual maximum to the likelihood function

In that case “locally” they are identical but not globally

Simulated Method of Moments

The classic reference is “A Method of Simulated Moments of Estimation of Discrete Response Models Without Numerical Integration,” McFadden, EMA, 1989

However, I will present it in a different way

To simplify first consider the case without X

Take any function of the data that you like say $g(\Upsilon_i)$ (where the dimension of g is K_g)

then notice that since y_0 and F represent the data generating process

$$E(g(\Upsilon_i)) = \int (g(y(u;\theta_0))dF(u;\theta_0)$$

So this means that we do GMM with

$$m(\Upsilon_i,\theta) = g(Y_i) - \int (g(y(u;\theta))dF(u;\theta)$$

So what?

Here is where things get pretty cool

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left[g(\Upsilon_i)-\int(g(y(u;\theta))dF(u;\theta)\right]\\
&=\left[\frac{1}{N}\sum_{i=1}^{N}g(\Upsilon_i)\right]-\int(g(y(u;\theta))dF(u;\theta)
\end{aligned}
$$

but we can approximate the thing on the right hand side by simulating from the distribution function $dF(u;\theta)$

Notice that if we simulate from the true value

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}g(\Upsilon_i)-\frac{1}{S}\sum_{s=1}^{S}(g(y(u_s;\theta_0)) \approx & E(g(\Upsilon_i))-\int(g(y(u;\theta_0))dF(u;\theta_0)\\
= & 0
\end{aligned}
$$

The nice thing about this is that we didn’t need S to be large for every N, we only needed S to be large for the one integral.

This makes this much much easier computationally

Adding X’s

Adding $X's$ back in is straight forward, just messier

Now let g be a function of X_i and Υ_i

$$\begin{aligned}
E(g(X_i,\Upsilon_i)) =& E\left[E(g(X_i,\Upsilon_i)\mid X_i)\right] \\
=& E\left[\int g(X_i, y_0(X_i,u;\theta_0))dF(u;\theta_0)\right]
\end{aligned}$$

Now when we do the simulation, we draw u_s from $F(\cdot;\theta)$ and X_s from the empirical distribution of X_i, then

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} g(X_i,\Upsilon_i) - \frac{1}{S}\sum_{s=1}^{S}(g(X_s,y(u_s;\theta_0)) \\
\approx & E(g(X_i,\Upsilon_i)) - E\left[\int g(X_i,y(X_i,u;\theta_0))dF(u;\theta)\right] \\
= & 0
\end{aligned}$$

In practice we minimize

$$\left[\frac{1}{N}\sum_{i=1}^{N} g(X_i,\Upsilon_i) - \frac{1}{S}\sum_{s=1}^{S}(g(X_s,y(u_s;\theta))\right]' W \left[\frac{1}{N}\sum_{i=1}^{N} g(X_i,\Upsilon_i) - \frac{1}{S}\sum_{s=1}^{S}(g(X_s,y(u_s;\theta))\right]$$

where u_s is simulated from $F(u;\theta)$

Indirect inference

The classic reference here is “Indirect Inference” Gourieroux, Monrort, and Renault, Journal of Applied Econometrics, 1993

Again I will think about this in a different way then them

Think about the intuition for the SMM estimator

$$\frac{1}{N}\sum_{i=1}^{N} g(X_i, \Upsilon_i) \approx \frac{1}{S}\sum_{s=1}^{S}(g(X_s, y(u_s;\theta_0))$$

If I have the right data generating model taking the mean of the simulated data should give me the same answer as taking the mean of the actual data

But we can generalize that idea

If I have the right data generating model the simulated data should look the same as the actual data

That means whatever the heck I do to the real data-if I do exactly the same thing to the simulated data I should get the same answer

So we define auxiliary parameters as

$$\widehat{\beta} = argmin_{\beta} F\left(\frac{1}{N}\sum_{i=1}^{N} g(X_i, \Upsilon_i, \beta), \beta\right).$$

Examples

- Moments

$$\widehat{\beta} = argmin_{\beta}\left(\frac{1}{N}\sum_{i=1}^{N} g(X_i, \Upsilon_i) - \beta\right)^2$$

- Regression models

$$\widehat{\beta} = argmin_{\beta}\frac{1}{N}\sum_{i=1}^{N}\left(\Upsilon - X_i'\beta\right)^2$$

- Misspecified MLE

$$\widehat{\beta} = argmin_{\beta} \frac{1}{N} \sum_{i=1}^{N} -\log(l(X_i, \Upsilon_i, \beta))$$

- Misspecified GMM

$$\widehat{\beta} = argmin_{\beta} \left[\frac{1}{N} \sum_{i=1}^{N} m(X_i, \Upsilon_i, \beta) \right]' W \left[\frac{1}{N} \sum_{i=1}^{N} m(X_i, \Upsilon_i, \beta) \right]$$

The most important thing: this can be misspecified, it doesn't have to estimate a true causal parameter

Creates a nice connection with reduced form stuff, we can use 2SLS or Diff in Diff as auxiliary parameters and it is clear where identification comes from

Define the population value of $\widehat{\beta}$ to be

$$\begin{aligned}\beta_0 \equiv & argmin_\beta F(E\left[g(X_i, \Upsilon_i, \beta)\right], \beta). \\ = & argmin_\beta F\left(\int g(X_i, y(X_i, u_i; \theta), \beta) dF(u; \theta), \beta\right)\end{aligned}$$

Then we define our simulated auxiliary model as

$$\widehat{B}(\theta) \equiv \frac{1}{H}\sum_{h=1}^{H} argmin_\beta F\left(\frac{1}{S}\sum_{s=1}^{S} g(X_{hs}, \Upsilon_{hs}(\theta); \beta), \beta\right)$$

H does not need to get large but often people think it is better to use $H > 1$

and we now get that

$$\widehat{B}(\theta) \approx \beta_0$$

as S gets large

We then do Indirect inference

$$\widehat{\theta} = argmin_{\theta} \left(\widehat{B}(\theta) - \widehat{\beta}\right)' \Omega \left(\widehat{B}(\theta) - \widehat{\beta}\right)$$

It turns out that it is optimal to choose Ω to be the inverse of the variance/covariance matrix for $\widehat{\beta}$

Often people just use the diagonal of this matrix

Panel Data Roy Model and Indirect Inference

So what parameters do we use as auxiliary parameters?

Important: you need to estimate the auxiliary model every time you do a function evaluation, so keep it simple

Some moments to use:

- Fixed Effect Linear Probability model for participation
- Fixed Effect Wage Regressions
- Variance Matrix of residuals from panel model

Maximum Likelihood versus Indirect Inference

- MLE is efficient
- Indirect inference you pick auxiliary model

Which is better is not obvious. Picking auxiliary model is somewhat arbitrary, but you can pick what you want the data to fit.

MLE essentially picks the moments that are most efficient-a statistical criterion

- Indirect inference is often computationally easier because of the simulation approximation of integrals
- With confidential data, Indirect Inference often is easier because only need to use the actual data to get $\widehat{\beta}$
- A drawback of simulation estimators is that they often lead to nonsmooth objective functions