Estimation of Policy Counterfactuals

Christopher Taber

University of Wisconsin

July 20, 2016

Outline

1. Structural and Reduced Form Models
   - Simultaneous Equations Models
   - The Roy Model
2. Identification
   - Definition of Identification
   - Identification of the Roy Model
3. Estimation

Estimation of Policy Impacts

Jeff talked about estimation of treatment effects

$$Y_1 - Y_0$$

where Y_1 is the outcome if the treatment was taken and Y_0 is the outcome if the treatment is not taken

I want to expand this to think about policy effects where now the treatment effect is

$$Y_{\pi} - Y_0$$

where Y_{π} is the outcome from an alternative policy environment π and Y_0 is the outcome under the status quo

The difference is that π represents a policy that has never been implemented

Estimation of this type of requires some structure

Heckman article on reading list has this 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*

Outline

1. Structural and Reduced Form Models
   - Simultaneous Equations Models
   - The Roy Model
2. Identification
   - Definition of Identification
   - Identification of the Roy Model
3. Estimation

Lets consider to the classic simultaneous equations model in a policy regime with no taxes

Supply Curve

$$Q_t = \alpha_s P_t + X_t' \beta + u_t$$

Demand Curve

$$Q_t = \alpha_d P_t + Z_t' \gamma + v_t$$

We can solve for prices and quantities as

$$\begin{aligned} P_t &= \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d} \\ Q_t &= \frac{\alpha_s(Z_t'\gamma + v_t) - \alpha_d(X_t'\beta + u_t)}{\alpha_s - \alpha_d} \end{aligned}$$

Now suppose we want to introduce a tax on this good imposed on consumers, so now

$$Q_t = \alpha_d \left(1+\tau\right) P_t + Z_t'\gamma + v_t$$

The equilibrium effect is

$$\begin{aligned} P_t =& \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d \left(1+\tau\right)} \\ Q_t =& \frac{\alpha_s(Z_t'\gamma + v_t) - \alpha_d(1+\tau)(X_t'\beta + u_t)}{\alpha_s - \alpha_d(1+\tau)} \end{aligned}$$

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

Outline

1. Structural and Reduced Form Models
   - Simultaneous Equations Models
   - The Roy Model
2. Identification
   - Definition of Identification
   - Identification of the Roy Model
3. Estimation

Example 2: The Roy Model

Labor Market is a Village

There are two occupations

- hunter
- fisherman

Fish and Rabbits are completely homogeneous

No uncertainty in number you catch

Let

- π_F be the price of fish
- π_R be the price of rabbits
- F number of fish caught
- R number of rabbits caught

Wages are thus

$$\begin{aligned} W_F &= \pi_F F \\ W_R &= \pi_R R \end{aligned}$$

Each individual chooses the occupation with the highest wage

Lets assume this village trades with the rest of the economy so prices fish and rabbits is taken as given.

Thats it, that is the model

Once we know the model we could think of several different policies

One is suppose we impose a minimum wage $\bar{w}$ in the fishing sector but not in the hunting sector?

What will this due to earnings inequality?

Anyone with $W_F < \bar{w}$ will no longer be employed in the fishing sector and must now hunt where they earn lower wages.

Inequality will likely rise and we can determine the magnitude from the model

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 border tightening on the U.S. government budget constraints (Nakajima, 2015)
- Welfare effects of alternative designs of school choice programs (Calsamiglia, Fu, and Guell, 2014)

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

(Possible) Steps for writing this type of paper

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

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
- 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)
- Methodological-this is a step in these directions, but we haven't gotten there yet

Outline

1 Structural and Reduced Form Models

- Simultaneous Equations Models
- The Roy Model

2 Identification

- Definition of Identification
- Identification of the Roy Model

3 Estimation

Why is thinking about nonparametric identification useful?

- Speaking for myself, I think it is. I always begin a research project by thinking about nonparametric identification.
- Literature on nonparametric identification not particularly highly cited
- At the same time this literature has had a huge impact on empirical work in practice. A Heckman two step model without an exclusion restriction is often viewed as highly problematic these days- because of nonparametric identification
- It is also useful for telling you what questions the data can possibly answer. If what you are interested is not nonparametrically identified, it is not obvious you should proceed with what you are doing

Outline

1. Structural and Reduced Form Models
   - Simultaneous Equations Models
   - The Roy Model
2. Identification
   - Definition of Identification
   - Identification of the Roy Model
3. Estimation

Definition of Identification

Another term that means different things to different people

I will base my discussion on Matzkin's (2007) formal definition of identification but use my own notation and be a bit less formal

This will all be about the Population in thinking about identification we will completely ignore sampling issues

We first need to define a data generation process

Data Generating Process

Let me define the data generating process in the following way

$$\begin{aligned} X_i &\sim H_0(X_i) \\ u_i &\sim F_0(u_i;\theta) \\ \Upsilon_i &= y_0(X_i,u_i;\theta) \end{aligned}$$

The data is (Υ_i, X_i) with u_i unobserved.

We know this model up to θ

To think of this as non-parametric we can think of θ as infinite dimensional

For example if F_0 is nonparametric we could write the model as $\theta = (\theta_1, F_0(\cdot))$

To relate this to our examples, for example 1 (being very loose with notation)

$$
\begin{aligned}
\Upsilon_t &= (P_t, Q_t) \\
X_t &= (X_t, Z_t) \\
u_t &= (u_t, v_t) \\
\theta &= (\gamma, \beta, \alpha_d, \alpha_s, G(u_t, v_t)) \\
y_0(X_i, u_i; \theta) &= \begin{bmatrix} \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d} \\ \frac{\alpha_s(Z_t'\gamma + v_t) - \alpha_d(X_t'\beta + u_t)}{\alpha_s - \alpha_d} \end{bmatrix}
\end{aligned}
$$

For the Roy Model we need to add some more structure to go from an economic model into an econometric model .

This means writing down the full data generation model.

First a normalization is in order.

We can redefine the units of F and R arbitrarily Lets normalize

$$\pi_F = \pi_R = 1$$

We consider the model

$$\begin{aligned} W_{fi} &= g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\ W_{hi} &= g_h(X_{hi}, X_{0i}) + \varepsilon_{hi} \end{aligned}$$

where the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ is G.

Let F_i be a dummy variable indicating whether the worker is a fisherman.

We can observe F_i and

$$W_i \equiv F_i W_{fi} + (1 - F_i) W_{hi}$$

Thus in this case

$$\begin{aligned}
\Upsilon_i &= (F_i, W_i) \\
X_i &= (X_{0i}, X_{fi}, X_{hi}) \\
u_i &= (\varepsilon_{fi}, \varepsilon_{hi}) \\
\theta &= (g_h, g_h, G) \\
y_0(X_i, u_i; \theta) &= \begin{bmatrix} 1\left(g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} > g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}\right) \\ \max\left\{g_f(X_{fi}, X_{0i}) + \varepsilon_{fi}, g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}\right\} \end{bmatrix}
\end{aligned}$$

You can see the selection problem-we only observe the wage in the occupation the worker chose, we don't observe the wage in the occupation they didn't

Point Identification of the Model

The model is identified if there is a unique θ that could have generated the population distribution of the observable data (X_i, Υ_i)

A bit more formally, let Θ be the parameter space of θ and let θ_0 be the true value

- If there is some other $\theta_1 \in \Theta$ with $\theta_1 \neq \theta_0$ for which the joint distribution of (X_i, Υ_i) when generated by θ_1 is identical to the joint distribution of (X_i, Υ_i) when generated by θ_0 then θ is not identified
- If there is no such $\theta_1 \in \Theta$ then θ is (point) identified

Set Identification of the Model

Define Θ_I as the identified set.

I still want to think of there as being one true θ_0

Θ_I is the set of $\theta_1 \in \Theta$ for which the joint distribution of (X_i, Υ_i) when generated by θ_1 is identical to the joint distribution of (X_i, Υ_i) when generated by θ_0.

So another way to think about point identification is the case in which

$$\Theta_I = \{\theta_0\}$$

Identification of a feature of a model

Suppose we are interested not in the full model but only a feature of the model: $\psi(\theta)$

We can identify

$$\Psi_I \equiv \{\psi(\theta) : \theta \in \Theta_I\}$$

Most interesting cases occur when Θ_I is a large set but Ψ_I is a singleton

In practice $\psi(\theta)$ could be something complicated like a policy counterfactual in which we typically need to first get θ and then simulate $\psi(\theta)$

However, often it is much simpler and we can just write it as a known function of the data.

Classic example is the reduced form parameters in the simultaneous equations model. They are a known function of the data

Lets define γ^*, β^*, and v_t^* implicitly as

$$\begin{aligned} P_t =& \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d} \\ \equiv& Z_t'\gamma^* + X_t'\beta^* + \nu_t^* \end{aligned}$$

Since $E(\nu_t^* \mid X_t, Z_t) = 0$, so one can identify $\psi \equiv (\gamma^*, \beta^*)$ and by regressing P_t on $W_t = (X_t, Z_t)$

That is

$$\psi = \left(E\left[W_t W_t'\right]\right)^{-1} E\left[W_t P_t\right]$$

Without exclusion restrictions we know we can't identify the structural parameters so we don't have point identification

However the reduced form parameters are still identified

To see how this relates to our definition of identification note that

- By definition of Θ_I, if $\theta \in \Theta_I$ then the joint distribution of w_t and P_t when generated by θ is the same as the joint distribution of w_t and P_t when generated by θ_0
- this also means that the value of $\left(E\left[W_t W_t'\right]\right)^{-1} E\left[W_t P_t\right]$ in a model generated by θ is the same as the value of $\left(E\left[W_t W_t'\right]\right)^{-1} E\left[W_t P_t\right]$ in a model generated by θ_0
- thus $\psi(\theta) = \psi(\theta_0)$ for every $\theta \in \Theta_I$
- thus Ψ_I is a singleton, so $\psi(\theta)$ is (point) identified

Identification of a policy counterfactual

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

$$\begin{aligned} X_i \sim & H_0(X_i) \\ u_i \sim & F_0(u_i;\theta) \\ \Upsilon_i = & y_0(X_i, u_i;\theta) \end{aligned}$$

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

$$\begin{aligned} X_i \sim & H_\pi(X_i) \\ u_i \sim & F_\pi(u_i;\theta) \\ \Upsilon_i = & y_\pi(X_i, u_i;\theta) \end{aligned}$$

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

$$\begin{aligned}\psi(\theta) =& E\left(Y_\pi - Y_0\right)\\ =& \int\int g(\Upsilon_i, u_i, X_i;\theta)dF_\pi(u_i;\theta)dH_\pi(X_i)\\ &- \int\int g(\Upsilon_i, u_i, X_i;\theta)dF_0(u_i;\theta)dH_0(X_i)\end{aligned}$$

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

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

One Example

The reduced form simultaneous equation model

$$\begin{aligned} P_t =& \frac{Z_t'\gamma - X_t'\beta + v_t - u_t}{\alpha_s - \alpha_d} \\ =& Z_t'\gamma^* + X_t'\beta^* + \nu_t^* \end{aligned}$$

Suppose we want to increase X_t to $X_t + \pi$ and look at effect on P_t

The reduced form is enough to answer this

$$\begin{aligned} P_t =& \frac{Z_t'\gamma - \left(X_t + \pi\right)' \beta + v_t - u_t}{\alpha_s - \alpha_d} \\ =& Z_t'\gamma^* + \left(X_t + \pi\right)' \beta^* + \nu_t^* \end{aligned}$$

Identifying the effects of a policy that has never been enacted is a difficult problem

To illustrate this clearly, consider a nonparametric version of an exercise we would teach in an introductory econometrics class

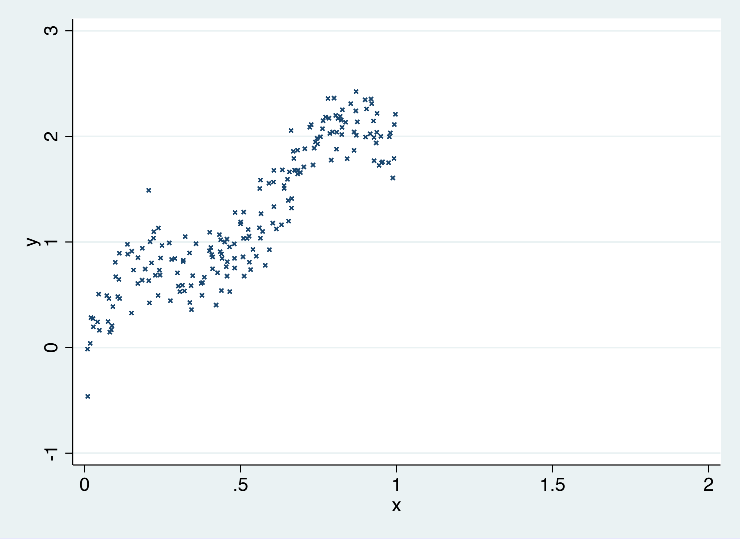
Suppose that

$$Y = g(X) + u$$

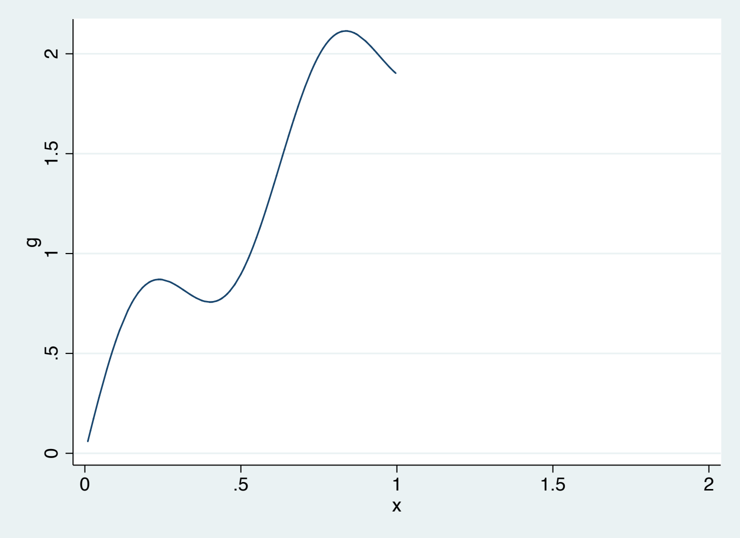
and

- Put aside the main focus of non-structural work by ignoring questions of endogeneity and assume that $E(u \mid X)$
- However suppose that the support of the data is $X \in [0, 1]$
- We want to estimate the effects of a policy π that sets $X = 1.5$

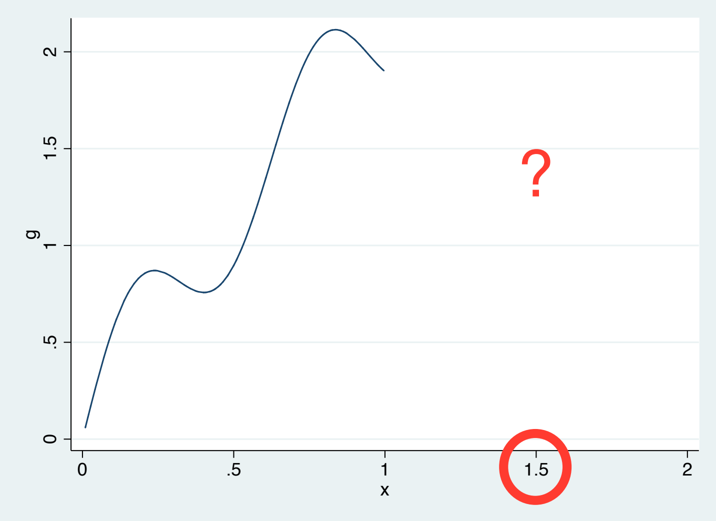
The raw data look like this



We can identify g



But we can't identify the effect of the policy without more assumptions



The obvious way to estimate such a model is to be parametric.

That is assume that

$$g(X) = g\left(X;\theta\right)$$

and estimate θ

Then we can predict the policy change as

$$g\left(1.5,\widehat{\theta}\right)$$

In some cases one can be completely non-parametric and use economic assumptions to solve this problem

For example suppose we have

- two variables X_1 and X_2
- both have support $[0,1]$ (and jointly they do as well)
- we want to predict Y when $X_1 = 0.75$ and $X_2 = 1.25$

Now suppose that

- X_1 is the tax rate
- X_2 is a wage subsidy rate

Economic theory tells us that all that matters is $X_2 - X_1$ so we can identify the policy effect as

$$g\left(0.75, 1.25\right) = E\left(Y \mid X_1 = 0.25, X_2 = 0.75\right)$$

You can see how both of these relate to the Knight quote

Outline

1. Structural and Reduced Form Models
    - Simultaneous Equations Models
    - The Roy Model
2. Identification
    - Definition of Identification
    - Identification of the Roy Model
3. Estimation

Identification of the Roy Model

Lets think about identifying this model

The reference is Heckman and Honore (EMA, 1990)

I follow the discussion in French and Taber, Handbook of Labor Economics, 2011

While the model is about the simplest in the world, identification is difficult

We consider the model above

$$\begin{aligned} W_{fi} &= g_f(X_{fi}, X_{0i}) + \varepsilon_{fi} \\ W_{hi} &= g_h(X_{hi}, X_{0i}) + \varepsilon_{hi}, \end{aligned}$$

where the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ is G.

In this case $\theta = (g_f, g_h, G)$

Assumptions

- $(\varepsilon_{fi}, \varepsilon_{hi})$ is independent of $X_i = (X_{0i}, X_{fi}, X_{hi})$
- Normalize $E(\varepsilon_{fi})$=0
  To see why this is a normalization we can always subtract $E(\varepsilon_{fi})$ from ε_{fi} and add it to $g_f(X_{fi}, X_{0i})$ making no difference in the model itself
- Normalize the median of $\varepsilon_{fi} - \varepsilon_{hi}$ to zero.
  A bit non-standard but we can always add the median of $\varepsilon_{fi} - \varepsilon_{hi}$ to ε_{hi} and subtract it from $g_h(X_{hi}, X_{0i})$
- $supp(g_f(X_{fi}, x_0), g_h(X_{hi}, x_0)) = \mathbb{R}^2$ for all $x_0 \in supp(X_{0i})$

Step 1: Identification of Reduced Form Choice Model

This part is well known in a number of papers (Manski and Matzkin being the main contributors) We can write the model as

$$\begin{aligned} Pr(F_i = 1 \mid X_i = x) &= Pr(g_h(x_h, x_0) + \varepsilon_{ih} \leq g_f(x_f, x_0) + \varepsilon_{if}) \\ &= Pr(\varepsilon_{ih} - \varepsilon_{if} \leq g_f(x_f, x_0) - g_h(x_h, x_0)) \\ &= G_{h-f}(g_f(x_f, x_0) - g_h(x_h, x_0)), \end{aligned}$$

where G_{h-f} is the distribution function for $\varepsilon_{ih} - \varepsilon_{if}$

We can not separate $g_f(x_f, x_0) - g_h(x_h, x_0)$ from G_{h-f}, but we can identify $Pr(F_i = 1 \mid X_i = x)$

This turns out to be very useful

It means that we know that for any two values x_1 and x_2, if

$$Pr(F_i = 1 \mid X_i = (x_0^a, x_h^a, x_f^a)) = Pr(F_i = 1 \mid X_i = (x_0^b, x_h^b, x_f^b))$$

then

$$g_f(x_f^a, x_0^a) - g_h(x_h^a, x_0^a) = g_f(x_f^b, x_0^b) - g_h(x_h^b, x_0^b)$$

Step 2: Identification of the Wage Equation g_f

Next consider identification of $g_{f.}$ This is basically the standard selection problem.

Notice that we can identify the distribution of W_{fi} conditional on $(X_i = x, F_i = 1.)$

In particular we can identify

$$\begin{aligned} E(W_i \mid X_i = x, F_i = 1) =& g_f(x_f, x_0) \\ &+ E(\varepsilon_{if} \mid \varepsilon_{ih} - \varepsilon_{if} < g_f(x_f, x_0) - g_h(x_h, x_0)). \end{aligned}$$

Lets think about identifying g_f up to location.

That is, for any $\left(x_f^a, x_0^a\right)$ and $\left(x_f^b, x_0^b\right)$ we want to identify

$$g_f(x_f^b, x_0^b) - g_f(x_f^a, x_0^a)$$

An exclusion restriction is key

Take x_h^b to be any number you want. From step 1 and from the support assumption we know that we can identify a x_h^a such that

$$Pr(F_i = 1 \mid X_i = (x_0^a, x_h^a, x_f^a)) = Pr(F_i = 1 \mid X_i = (x_0^b, x_h^b, x_f^b))$$

which means that

$$g_f(x_f^a, x_0^a) - g_h(x_h^a, x_0^a) = g_f(x_f^b, x_0^b) - g_h(x_h^b, x_0^b)$$

But then

$$
\begin{aligned}
&E(W_i \mid X_i = (x_0^a, x_h^a, x_f^a), F_i = 1) - E(W_i \mid X_i = (x_0^b, x_h^b, x_f^b), F_i = 1) \\
=&g_f(x_f^b, x_0^b) - g_f(x_f^a, x_0^a) \\
&+ E(\varepsilon_{if} \mid \varepsilon_{ih} - \varepsilon_{if} < g_f(x_f^a, x_0^a) - g_h(x_h^a, x_0^a))) \\
&- E(\varepsilon_{if} \mid \varepsilon_{ih} - \varepsilon_{if} < g_f(x_f^b, x_0^b) - g_h(x_h^b, x_0^b)) \\
=&g_f(x_f^b, x_0^b) - g_f(x_f^a, x_0^a)
\end{aligned}
$$

Identification at Infinity

What about the location?

Notice that

$$
\begin{aligned}
&\lim_{g_h(x_h,x_0)\to-\infty;\left(x_0,x_f\right)\text{ fixed}} E(W_i \mid X_i=(x_0,x_h,x_f),F_i=1)\\
&= g_f(x_f,x_0)\\
&+ \lim_{g_h(x_h,x_0)\to-\infty;\left(x_0,x_f\right)\text{ fixed}} E(\varepsilon_{fi} \mid \varepsilon_{ih}-\varepsilon_{if} < g_f(x_f,x_0)-g_h(x_h,x_0)))\\
&= g_f(x_f,x_0)+E(\varepsilon_{fi})\\
&= g_f(x_f,x_0)
\end{aligned}
$$

Thus we are done

Another important point is that the model is not identified without identification at infinity.

To see why suppose that $g_f(x_f, x_0) - g_h(x_h, x_0)$ is bounded from above at g^u then if $\varepsilon_{ih} - \varepsilon_{if} > g^u$, we know for sure that $F_i = 0$. Thus the data is completely uninformative about

$$E(\varepsilon_{fi} \mid \varepsilon_{ih} - \varepsilon_{if} > g^u)$$

so the model is not identified.

Parametric assumptions on the distribution of the error term is an alternative.

Who cares about Location?

Actually we do, a lot

- Without our intercept we know something about wage variation within fishing
- However we can not compare the wage level of fishing to the wage level of hunting
- If our policy involves moving people from one to the other we need the intercepts

Step 3: Identification of g_h

For any (x_h, x_0) we want to identify $g_h(x_h, x_0)$

What will be crucial is the other exclusion restriction (i.e. X_{fi}).

Again from step 1 and the other support condition, we know that can find an x_f so that

$$Pr(F_i = 1 \mid X_i = (x_0, x_h, x_f)) = 0.5.$$

This means that

$$0.5 = \Pr\left(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0)\right).$$

But the fact that $\varepsilon_{hi} - \varepsilon_{fi}$ has median zero implies that

$$g_h(x_h, x_0) = g_f(x_f, x_0).$$

Since g_f is identified, $g_f(x_f, x_0)$ is known, so $g_h(x_h, x_0)$ is identified from this expression.

To identify the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$ note that from the data one can observe

$$\begin{aligned} & \Pr(J_i = f, \log(W_i) < s \mid X_i = x) \\ = \; & \Pr(g_h(x_h, x_0) + \varepsilon_{hi} \leq g_h(x_h, x_0) + \varepsilon_{hi}, g_f(x_f, x_0) + \varepsilon_{fi} \leq s) \\ = \; & \Pr(\varepsilon_{hi} - \varepsilon_{fi} \leq g_f(x_f, x_0) - g_h(x_h, x_0), \varepsilon_{fi} \leq s - g_f(x_f, x_0)) \end{aligned}$$

which is the cumulative distribution function of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$ evaluated at the point $(g_f(x_f, x_0) - g_r(x_r, x_0), s - g_f(x_f, x_0))$

Thus by varying $(g_f(x_f, x_0) - g_h(x_h, x_0)$ and $(s - g_f(x_f, x_0))$ we can identify the joint distribution of $(\varepsilon_{hi} - \varepsilon_{fi}, \varepsilon_{fi})$

from this we can get the joint distribution of $(\varepsilon_{fi}, \varepsilon_{hi})$.

Outline

1. Structural and Reduced Form Models
   - Simultaneous Equations Models
   - The Roy Model
2. Identification
   - Definition of Identification
   - Identification of the Roy Model
3. Estimation

Estimation

So how do we estimate the model and do policy analysis? There are really 3 different approaches

1. Estimate full structural model (and thus data generating process) and simulate policy effect
2. Estimate reduced form of data generating process and simulate policy effect
3. Try to estimate policy directly without estimating full DGP

There are a few examples of the third. Some of them:

- Heckman and Vytlacil in a series of papers show how to use local instrumental variables to estimate policy relevant treatment effects (this is an empirical way to deal with the support problem we discussed above)
- Sufficient Statistics can be used to identify some policy effects. Associated with many Chetty papers, but Hendren talk gave the basic idea
- I have a paper with Hide Ichimura where we use “policy replicating variation” to show how to non-parametrically estimate the model

These are sufficiently different from each other and special that I want to focus on the first two approaches.

By the second approach I mean that we can often write the data generation model in terms of reduced forms as discussed above

$$P_t = Z_t'\gamma^* + X_t'\beta^* + \nu_t^*$$

We can just think of this as the data generating process

For some random variable 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

From Jensen's inequality we know

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

Maximum likelihood is just the sample analogue of this

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

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

The most important result for MLE is that it is efficient

In particular no alternative estimator can have a lower asymptotic variance

Often in these models the integral is a big problem in calculating the likelihood

For example in our data generating process if Υ_i were discrete the likelihood function would be

$$\ell(\theta; (X_i, \Upsilon_i)) = \int 1\left(y_0(X_i, u_i; \theta) = \Upsilon_i\right) dF(u_i; \theta)$$

(with something even more complicated for continuous variables)

The problem is this integral

Often we have a very large number of u_i to integrate over making this very difficult.

To make it harder, calculating y_0 can also be very complicated (if it involves solving a dynamic programming problem or an equilibrium)

Simulation

Another way to evaluate the likelihood function is to simulate.

Conditioning on X_i draw random variables u_s from the distribution $F(u_i;\theta)$ then as $S \to \infty$

$$\begin{aligned}\frac{1}{S}\sum_{s=1}^{S} 1\left(y_0(X_i,u_s;\theta)=\Upsilon_i\right) \xrightarrow{p} & Pr\left(y_0(X_i,u_s;\theta)=\Upsilon_i \mid X_i,\Upsilon_i;\theta\right)\\ =&\ell(\theta;(X_i,\Upsilon_i))\end{aligned}$$

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 (at least for every distinct value of X_i)

(the probability could also be zero for some observation if S is not sufficiently large)

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 assume that the econometrician observes (Υ_i, X_i) which are i.i.d. and both X_i and Υ_i are potentially large dimensional.

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

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

but as long as everything is well behaved this means that

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

We can use this as a moment condition

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

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

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

$$E[g(\Upsilon_i, X_i)] = \int \int (g(y_0(X, u; \theta_0), X_i) dF(u; \theta_0) dH(X)$$

So this means that we can do GMM with

$$m(\Upsilon_i, X_i, \theta) = g(\Upsilon_i, X_i) - \int \int (g(y_0(X, u; \theta), X_i) dF(u; \theta) dH(X)$$

So what?

Here is where things get pretty cool

Notice that if we simulate from the true value

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} g(\Upsilon_i, X_i) - \frac{1}{S}\sum_{s=1}^{S}(g(y_0(x_s, u_s; \theta_0))\\
\approx & E[g(\Upsilon_i, X_i)] - \int\int (g(y_0(X, u; \theta_0))dF(u; \theta_0)dH(X)\\
= & 0
\end{aligned}
$$

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 easier computationally

Also with confidential data can make things much easier to deal with

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

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

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, if I use the true parameter value, 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

Procedure

- Estimate auxiliary parameter $\widehat{\beta}$ using some estimation scheme in real data
- for any particular value of θ
    - Simulate data using data generation process: $y_0(x,u;\theta), H(X), G(u;\theta)$
    - Estimate $\widehat{B}(\theta)$ using exactly the same estimation scheme on simulated data
- Choose θ to minimize

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

This is consistent because

$$\widehat{B}(\theta_0)-\widehat{\beta}\xrightarrow{p}0$$

Examples:

- Moments
- Regression models
- Misspecified MLE
- Misspecified GMM
- IV
- Difference in Differences
- Regression Discontinuity
- Even Randomized Control Files

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

Can think of the analogue to the forecasting out of the sample-we use Indirect Inference to extend the convincing identification scheme into a structural framework

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
- Indirect inference preserves some of the advantages of non-structural estimation: Map from data to parameters is more transparent