Introduction: structural econometrics

Jean-Marc Robin

Abstract

1. Descriptive vs structural models
2. Correlation is not causality
    a. Simultaneity
    b. Heterogeneity
    c. Selectivity

Descriptive models

Consider a sample of N observations of a couple of variables $S = \{(y_i, x_i), i = 1, ..., N\}$.

A **descriptive model** is a **statistical** restriction (a list of assumptions) on the distribution of S.

It describes the statistical link between x_i and y_i, not the **causal** relationship between these variables.

Example: the linear model.

Assume $x_i \in \mathbb{R}^K$ and $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}$ is full column rank (the columns of X are linearly indep.).

Then, the symmetric matrix $X^T X = \sum_{i=1}^N x_i x_i^T$ is invertible and the **Ordinary Least Squares** (OLS) estimator exists:

$$\widehat{b} = \left[\sum_{i=1}^N x_i x_i^T\right]^{-1} \sum_{i=1}^N x_i y_i = \left(X^T X\right)^{-1} X^T \mathbf{y} \quad \text{where} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

Assuming moreover iid observations and $\mathbb{E}(x_i x_i^T)$ non singular, the Law of Large Numbers implies that

$$\operatorname*{plim}_{N\to\infty} \widehat{b} = \left[\mathbb{E}(x_i x_i^T)\right]^{-1} \mathbb{E}(x_i y_i) \equiv b.$$

Parameter b is the coefficient of the theoretical regression of y_i on x_i. It describes the correlations between y_i and x_i in the **whole** population whereas $\widehat{b}$ describes the correlations between y_i and x_i in the sample.

$\widehat{b}$ and b always exist under these assumptions...

...even if the true **Data Generating Process** is not a linear model.

Examples: it may very well be that the true DGP is

- $y_i \sim \mathcal{N}\left(cx_i^2, \sigma^2\right)$, $y_i, x_i \in \mathbb{R}$.
- $y_i \in \{0, 1\}$ and $\Pr\{y_i = 1 | x_i\} = \Phi\left(x_i^T \gamma\right)$ (Probit model).

Correlation is not causality

In general, however, correlations can be pointing in the wrong direction of causality.

This happens in particular

- when x_i is endogenous: either because of **simultaneity biases**, or because of **unobserved heterogeneity**.
- when the sample is endogenous (**selectivity biases**).

Example of simultaneity bias: the supply-demand model

Aggregate demand: $D_i = a - bp_i + u_i$.

Aggregate supply: $S_i = \alpha + \beta p_i + v_i$.

Assume demand shock u_i and supply shock v_i uncorrelated.

At the equilibrium, supply equals demand equals exchanged quantity:

$$\begin{cases} y_i = a - bp_i + u_i \\ y_i = \alpha + \beta p_i + v_i \end{cases} \Rightarrow \begin{cases} y_i = \dfrac{a\beta + b\alpha}{b+\beta} + \dfrac{\beta u_i + b v_i}{b+\beta} \\ p_i = \dfrac{a-\alpha}{b+\beta} + \dfrac{u_i - v_i}{b+\beta} \end{cases}.$$

Regressing y_i on p_i yields

$$\gamma = \frac{\mathrm{Cov}\,(y_i, p_i)}{\mathrm{Var}\,(p_i)} = \frac{\beta\sigma_u^2 - b\sigma_v^2}{\sigma_u^2 + \sigma_v^2}$$

which is a weighted average of b and β.

Instrumental variables

Demand curves and supply curves are identifiable if there exist observed supply and demand shocks:

$$\begin{cases} u_i = x_i^T c + \varepsilon_i \\ v_i = z_i^T \gamma + \eta_i \end{cases}$$

Under the assumptions that

$$\begin{aligned} \mathrm{Cov}\left(x_i, \eta_i\right) &= 0 \\ \mathrm{Cov}\left(z_i, \varepsilon_i\right) &= 0 \end{aligned}$$

- regressing y_i on p_i by Two Stage Least Squares (2SLS) using z_i (supply shocks) to instrument p_i yields consistent estimates of a and b (demand curve);
- regressing y_i on p_i by Two Stage Least Squares (2SLS) using x_i (demand shocks) to instrument p_i yields consistent estimates of α and β (supply curve).

Example of heterogeneity bias: convergence and growth

Do LDCs grow faster than developed countries so that their wealths will converge?

Idea: regress $q_{it_1} - q_{it_0}$ on $q_{it_0} - \overline{q}_{t_0}$, for a sample $\{(q_{it_0}, q_{it_1})\,; i = 1, ..., N\}$ where q_{it} is per capita GDP (in log) of country i measured at two different times t_0 and t_1 (for example $t_0 = 1960$ and $t_1 = 1990$) and $\overline{q}_t = \frac{1}{N}\sum_{i=1}^{N} q_{it}$.

The OLS estimate of b, $\widehat{b}$, is found significantly negative, which seems to imply that the countries starting from a high GDP value relative to the mean ($q_{it_0} - \overline{q}_{t_0} > 0$) have a lower growth rate than the others.

Structural model.

Assume that the GDP of each country fluctuates around the same international trend but with different levels:

$$q_{it} = \overline{q}_t + \alpha_i + v_{it}.$$

where $\mathbb{E}\alpha_i = 0$ and $\operatorname{Var}\alpha_i = \sigma_\alpha^2$, and v_{it} is an iid shock, independent of α_i and $\overline{q}_t$, with mean 0 and $\operatorname{Var} v_{it} = \sigma_v^2$ (white noise).

Regression to the mean 1

OLS estimator of the regression of $q_{it_1} - \overline{q}_{t_1}$ on $q_{it_0} - \overline{q}_{t_0}$:

$$\widehat{\beta}_{10} = \frac{\sum_i \left(q_{it_1} - \overline{q}_{t_1}\right)\left(q_{it_0} - \overline{q}_{t_0}\right)}{\sum_i \left(q_{it_0} - \overline{q}_{t_0}\right)^2} = \frac{\sum_i \left(\alpha_i + v_{it_1}\right)\left(\alpha_i + v_{it_0}\right)}{\sum_i \left(\alpha_i + v_{it_0}\right)^2}$$
$$\xrightarrow{P} \frac{\mathrm{Cov}\left(\alpha_i + v_{it_1}, \alpha_i + v_{it_0}\right)}{\mathrm{Var}\left(\alpha_i + v_{it_0}\right)} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_v^2} = \beta < 1.$$

OLS estimator of the regression of $q_{it_0} - \overline{q}_{t_0}$ on $q_{it_1} - \overline{q}_{t_1}$:

$$\widehat{\beta}_{01} = \frac{\sum_i \left(q_{it_0} - \overline{q}_{t_0}\right)\left(q_{it_1} - \overline{q}_{t_1}\right)}{\sum_i \left(q_{it_1} - \overline{q}_{t_1}\right)^2} = \frac{\sum_i \left(\alpha_i + v_{it_1}\right)\left(\alpha_i + v_{it_0}\right)}{\sum_i \left(\alpha_i + v_{it_1}\right)^2}$$
$$\xrightarrow{P} \frac{\mathrm{Cov}\left(\alpha_i + v_{it_1}, \alpha_i + v_{it_0}\right)}{\mathrm{Var}\left(\alpha_i + v_{it_1}\right)} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_v^2} = \beta < 1.$$

Hence the regressing $q_{it_1} - \overline{q}_{t_1}$ on $q_{it_0} - \overline{q}_{t_0}$ or $q_{it_0} - \overline{q}_{t_0}$ on $q_{it_1} - \overline{q}_{t_1}$ yields two estimators $\widehat{\beta}_{10}$ and $\widehat{\beta}_{01}$ which converge to the same value $\beta < 1$!!!!

Regression to the mean 2

Lastly,

$$\begin{aligned} q_{it_1} - \overline{q}_{t_1} &= \widehat{\beta}_{10}\left(q_{it_0} - \overline{q}_{t_0}\right) + \widehat{u}_i \\ &\Leftrightarrow\ q_{it_1} - q_{it_0} = \overline{q}_{t_1} - \overline{q}_{t_0} + \underbrace{\left(\widehat{\beta}_{10} - 1\right)}_{=b<0}\left(q_{it_0} - \overline{q}_{t_0}\right) + \widehat{u}_i. \end{aligned}$$

So, one obtains $\widehat{b} < 0$ although all countries follow parallel GDP trajectories, which therefore cannot converge!

This result is known since Galton (1822-1911) who observed that regressing the sizes of sons on the sizes of fathers or the sizes of fathers on the sizes of sons produced a coefficient less than one. This is the "regression to the mean" phenomenon.

Alternative model

Construct a dynamic model of each country's GDP with heterogeneous levels:

$$q_{it} = \alpha_i + \beta q_{i,t-1} + v_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T.$$

If $\beta < 1$ and (v_{it}) is stationary, each country tends to fluctuate around a specific target $\frac{\alpha_i}{1-\beta}$ as

$$q_{it} - q_{i,t-1} = -(1-\beta)\left(q_{i,t-1} - \frac{\alpha_i}{1-\beta}\right) + v_{it}.$$

"Clubs" of convergence: countries with similar values of α_i.

Example of selectivity bias: productivity and wages

Based on Heckman, *Ecta*, 1974.

I will develop the different steps from the construction of an economic model, to the construction of the econometric model.

Individuals determine labour supply by trading off consumption for leisure:

$$\max_{c,\ell} U(c,\ell) \text{ s.t. } \begin{cases} c = \pi h + \mu \\ 0 < c, \quad 0 < \ell \leq T \end{cases} \tag{1}$$

where utility function $U(c,\ell)$ is increasing in consumption c and leisure ℓ, π is the wage (equal to labour productivity), μ is nonlabour income.

Let

$$\ell^*(\underset{+/-}{\pi}, \mu \underset{-}{+} \pi T) = \arg\max_{c,\ell} \{U(c,\ell) | c + \pi\ell = \pi T + \mu, 0 < c, 0 < \ell\}$$

be the Marshallian demand for leisure (interior solution to optimization programme, i.e. without $\ell \leq T$). The solution to (1) is

$$\ell\left(\pi, \mu + \pi T\right) = \left| \begin{array}{ll} \ell^* & \text{if } \ell^* < T \\ T & \text{if } \ell^* \geq T \end{array} \right. .$$

Proof

The Lagrangian is

$$L(c, \ell, \psi_1, \psi_2) = U(c, \ell) + \psi_1(\pi T + \mu - c - \pi\ell) + \psi_2(T - \ell),$$

where ψ_1 and ψ_2 are the Lagrange multipliers. A solution (c, ℓ) is such that

$$\begin{aligned} \frac{\partial L(c, \ell, \psi_1, \psi_2)}{\partial c} &= \frac{\partial U(c, \ell)}{\partial c} - \psi_1 = 0 \\ \frac{\partial L(c, \ell, \psi_1, \psi_2)}{\partial \ell} &= \frac{\partial U(c, \ell)}{\partial \ell} - \psi_1\pi - \psi_2 = 0 \end{aligned}$$

with

$$\begin{aligned} \psi_2 &\geq 0, \\ \psi_2(T - \ell) &= 0, \\ c + \pi\ell &= \pi T + \mu, \\ c &> 0, T \geq \ell > 0. \end{aligned}$$

Interior solution. An interior solution, such that $c > 0, T > \ell > 0$, has $\psi_2 = 0$, and

$$\begin{aligned} \frac{\partial U(c,\ell)}{\partial \ell} \Big/ \frac{\partial U(c,\ell)}{\partial c} &= \pi, \\ c + \pi\ell &= \pi T + \mu. \end{aligned}$$

Note that duality theory implies that the solution to this system is

$$\ell = \ell^* (\pi, \pi T + \mu) = -\frac{\partial V(\pi, \pi T + \mu)/\partial \pi}{\partial V(\pi, \pi T + \mu)/\partial y}$$

where

$$V(\pi, y) = \max_{c>0, \ell>0} U(c,\ell) \text{ s.t. } c + \pi\ell = y,$$

is the indirect utility function associated to U.

Corner solution. A corner solution has $\ell = T$, $c = \mu$, and

$$\frac{\partial U(\mu, T)}{\partial \ell} \Big/ \frac{\partial U(\mu, T)}{\partial c} = \frac{\psi_1 \pi + \psi_2}{\psi_1} = \pi + \frac{\psi_2}{\psi_1} \geq \pi.$$

as $\psi_1 > 0$ as $U(c, \ell)$ is strictly increasing wrt c. However, ψ_2 can be equal to 0.

Lastly, by definition, $\ell^*(\pi, \pi T + \mu)$, satisfies

$$\begin{aligned} \frac{\partial U(c^*, \ell^*)}{\partial \ell} \Big/ \frac{\partial U(c^*, \ell^*)}{\partial c} &= \pi, \\ c^* + \pi \ell^* &= \pi T + \mu. \end{aligned}$$

The TMS $\frac{\partial U(\pi(T-\ell)+\mu, \ell)}{\partial \ell} \Big/ \frac{\partial U(\pi(T-\ell)+\mu, \ell)}{\partial c}$ being a decreasing function (of ℓ) when U is concave, the inequality

$$\frac{\partial U(\mu, T)}{\partial \ell} \Big/ \frac{\partial U(\mu, T)}{\partial c} \geq \pi$$

holds iff $\ell^*(\pi, \pi T + \mu) \geq T$. (Draw a picture that shows that you want to be at the corner when $\ell^*(\pi, \pi T + \mu) \geq T$.)

Specification of labour supply function

The next step is to specify the labour supply functions.

First way. Choose a specification for U or V and determine ℓ.

Example:

$$V(\pi, y) = \pi^{-\delta}\left[y + \frac{\alpha}{\delta} - \frac{\beta\pi}{1-\delta} - \frac{\gamma\pi^2}{2-\delta}\right],$$

where $y = \pi T + \mu$. By Roy's identity,

$$\begin{aligned}
\ell^*(\pi, y) &= -\frac{\partial V(\pi, y)}{\partial \pi} \Big/ \frac{\partial V(\pi, y)}{\partial y} \\
&= -\frac{-\delta\pi^{-\delta-1}\left[y + \frac{\alpha}{\delta} - \frac{\beta\pi}{1-\delta} - \frac{\gamma\pi^2}{2-\delta}\right] + \pi^{-\delta}\left[-\frac{\beta}{1-\delta} - \frac{2\gamma\pi}{2-\delta}\right]}{\pi^{-\delta}} \\
&= \beta + \gamma\pi + \frac{\alpha}{\pi} + \delta\frac{y}{\pi}. \qquad \text{(LS1)}
\end{aligned}$$

Second way. Determine U or V such as ℓ has the desired form.

A good parametric specification is linear:

$$\ell^*(\pi, y) = \alpha - \beta \ln \pi + \gamma y, \tag{LS2}$$

which is parametrically simpler as (LS1) (less parameters, effects of π and y additively separable).

Is is possible to find a utility function which rationalizes this function?

Yes, one can show that

$$V(\pi, y) = \left(\frac{\alpha - \beta \ln \pi}{\gamma} + y \right) e^{-\gamma\pi} - \frac{\beta}{\gamma} E_1(\gamma\pi)$$

where $E_1(t) = \int_t^\infty \frac{e^{-x}}{x} dx = -\operatorname{Ei}(-t)$, works.

Proof

We search for a cost function $y(\pi, v)$ such that $V(\pi, y) = v$ and

$$-\frac{\partial V(\pi, y)}{\partial \pi} \Big/ \frac{\partial V(\pi, y)}{\partial y} = \alpha - \beta \ln \pi + \gamma y.$$

Fix v and differentiate $V(\pi, y) = v$ wrt π:

$$\frac{\partial V(\pi, y)}{\partial \pi} + \frac{\partial V(\pi, y)}{\partial y} \frac{\partial y}{\partial \pi} = 0$$

or, assuming that $\frac{\partial V(\pi, y)}{\partial y} \neq 0$,

$$\frac{\partial y(\pi, v)}{\partial \pi} = \alpha - \beta \ln \pi + \gamma y.$$

This is a linear ODE, the solution of which is easily found to be

$$y(\pi, v) = -\frac{\alpha - \beta \ln \pi}{\gamma} + \left[\frac{\beta}{\gamma} E_1(\gamma \pi) + c(v)\right] e^{\gamma \pi}$$

where $E_1(t) = \int_t^\infty \frac{e^{-x}}{x} dx = -\operatorname{Ei}(-t)$ is the E1-exponential-integral function and $c(v)$ is an arbitrary function of v, that has to be increasing for $y(\pi, v)$ to be a cost function. Function $c(v)$ is arbitrary, so one can take $c(v) = v$. By inverting $y(\pi, v) = y$ wrt v, one proves the announced result.

Theoretical predictions

A model is useful if/as it allows to make predictions.

Consumer theory tells us that the cost function

$$y(\pi, v) = -\frac{\alpha - \beta \ln \pi}{\gamma} + \left[v + \frac{\beta}{\gamma} E_1(\gamma\pi)\right] e^{\gamma\pi}$$

has to be increasing in v (true) and increasing and concave in π:

$$\frac{\partial y(\pi, v)}{\partial \pi} = \alpha - \beta \ln \pi + \gamma y(\pi, v) > 0, \quad \frac{\partial^2 y(\pi, v)}{\partial \pi^2} = -\frac{\beta}{\pi} + \gamma \frac{\partial y(\pi, v)}{\partial \pi} < 0.$$

Duality theory implies that $\frac{\partial y(\pi,v)}{\partial \pi} = \ell^*(\pi, y(\pi, v))$. Hence

$$\ell^*(\pi, y) > 0 \text{ and } -\frac{\beta}{\pi} + \gamma \ell^*(\pi, y) < 0.$$

These conditions cannot be imposed ex ante but should be verified over the whole support of the distribution of (π, y) after estimation.

Lastly, if leisure is a normal good,

$$\frac{\partial \ell^*(\pi, y)}{\partial y} = \gamma > 0.$$

Moreover, we see that $-\frac{\beta}{\pi} + \gamma \ell^* < 0$ and $\ell^* > 0$ imply that $\beta > 0$. Parameter α can be positive or negative.

Econometric model

Now, we want to use the previous model to analyze a sample of individual data on female participation. There is an iid sample of N observations of weekly hours worked ($h_i = T - \ell_i$), individual wage (w_i)–if i does not work, then $w_i = \cdot$, the number of years of education (Ed_i), the individual's age (Age_i) and the husband's wage (μ_i).

The theory yields the following model for (h_i, w_i):

$$\begin{pmatrix} h_i \\ w_i \end{pmatrix} = \begin{cases} \begin{pmatrix} h_i^* \\ \pi_i \end{pmatrix} & \text{if } h_i^* > 0, \\ \begin{pmatrix} 0 \\ \cdot \end{pmatrix} & \text{if } h_i^* \leq 0, \end{cases} \qquad \text{(selection rule)}$$

where

$$h^*(\pi, y) = T - \ell^*(\pi, y) = T - \alpha + \beta \ln \pi - \gamma y.$$

An econometric model requires to consider the modelling of observed and unobserved heterogeneity.

Modelling heterogeneity

We could assume that

$$\begin{cases} h_i^* = a^T x_i + \beta \ln \pi_i + u_i, \\ \ln \pi_i = b^T z_i + v_i, \end{cases} \qquad \text{(latent)}$$

where

$$\begin{aligned} x_i &= \left(1, Ed_i, Exp_i, (Exp_i)^2, \mu_i\right)^T, \quad a = (a_1, a_2, a_3, a_4, a_5), \\ z_i &= \left(1, Ed_i, Exp_i, (Exp_i)^2, Ed_i * Exp_i\right)^T, \quad b = (b_1, b_2, b_3, b_4, b_5), \end{aligned}$$

with $Exp_i = Age_i - Ed_i - 6$, and

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}\right).$$

The correlation between u_i and v_i reflects omitted variables determining both individual preferences and productivity.

Note that education and experience/age are likely to determine both preferences and productivity.

Note also that we have put some restrictions: μ_i only conditions preferences and the interaction $Ed_i * Exp_i$ only conditions productivity. Always useful to have these sorts of restrictions as they provide instruments for potential endogeneity or selectivity problem.

Identification

The next step is to discuss identification.

Given that one has specified a fully parametric model, parametric identification is to show that two sets of parameters yielding the same likelihood value are necessarily equal.

Here, one can show that the selection model makes no difference. The model is identified if the latent model (latent) is identified. This requires the existence of variables determining $\ln \pi_i$ but not h_i^*.

Nonparametric identification holds when the distribution of observables picks only one set of parameters irrespective of the stochastic assumption on the distribution of u_i and v_i.

Much more difficult to prove. We shall come back to this point later.

Limited information models

In general it is useful to search for model parts which are simpler to estimate.

a. Participation model. Eliminate $\ln \pi_i$ out of h_i^*:

$$\begin{aligned} h_i^* &= a^T x_i + \beta \ln \pi_i + u_i \\ &= a^T x_i + \beta b^T z_i + u_i + \beta v_i \\ &= c^T q_i + r_i, \end{aligned}$$

where

$$\begin{aligned} q_i &= \left(1, Ed_i, Exp_i, (Exp_i)^2, Ed_i * Exp_i, \mu_i\right)^T, \\ c &= (a_1 + \beta b_1, a_2 + \beta b_2, a_3 + \beta b_3, a_4 + \beta b_4, \beta b_5, a_5)^T, \\ r_i &= u_i + \beta v_i, \end{aligned}$$

and

$$\begin{pmatrix} v_i \\ r_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_v^2 & \sigma_{rv} \\ \sigma_{rv} & \sigma_r^2 \end{pmatrix}\right),$$

where $\sigma_r^2 = \sigma_u^2 + \beta^2 \sigma_v^2 + 2\beta\rho\sigma_u\sigma_v$ *and* $\sigma_{rv} = \beta\sigma_v^2 + \rho\sigma_u\sigma_v$.

Let

$$\delta_i = \begin{cases} 1 & \text{if } c^T q_i + r_i > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Probit)}$$

Then

$$\begin{aligned} \Pr(\delta_i = 1 | x_i, z_i) &= \Pr(h_i^* > 0 | x_i, z_i) \\ &= \Pr(r_i > -c^T q_i | q_i) \\ &= \Pr\left(\frac{r_i}{\sigma_r} > -\frac{c^T q_i}{\sigma_r} \Big| q_i \right) \\ &= 1 - \Phi\left(-\frac{c^T q_i}{\sigma_r} \right) \\ &= \Phi\left(\frac{c^T q_i}{\sigma_r} \right), \end{aligned}$$

where $\Phi(\cdot)$ is a standard normal distribution function. The Probit model of participation identifies c/σ_r.

b. Reduced form labour supply model. Consider regressing h_i on x_i and z_i using a TOBIT model:

$$h_i = \begin{cases} c^T q_i + r_i & \text{if } c^T q_i + r_i > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Tobit)}$$

This will separately identify c and σ_r.

c. Productivity equation. One has the following selection model for productivity:

$$\ln w_i = \begin{cases} b^T z_i + v_i & \text{if } c^T q_i + r_i > 0 \\ . & \text{otherwise} \end{cases} \qquad \text{(Selection)}$$

One can (should) use ML to estimate this selection model and consistently estimate $b, c/\sigma_r, \sigma_v$ and ρ.

Alternatively, one can use Heckman's two-step estimator:

$$\begin{aligned} \mathbb{E}(\ln \pi_i | r_i, x_i, z_i) &= b^T z_i + \mathbb{E}\left(v_i | r_i, x_i, z_i\right) \\ &= b^T z_i + \mathbb{E}\left(v_i | r_i\right) \\ &= b^T z_i + \frac{\sigma_{rv}}{\sigma_r^2} r_i. \end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}(\ln w_i | h_i^* > 0, x_i, z_i) &= b^T z_i + \frac{\sigma_{rv}}{\sigma_r^2}\mathbb{E}\left(r_i | r_i > -c^T q_i\right) \\
&= b^T z_i + \frac{\sigma_{rv}}{\sigma_r}\mathbb{E}\left(\frac{r_i}{\sigma_r}\Big|\frac{r_i}{\sigma_r} > -\frac{c^T q_i}{\sigma_r}\right) \\
&= b^T z_i + \frac{\sigma_{rv}}{\sigma_r}\frac{\varphi\left(c^T q_i/\sigma_r\right)}{\Phi\left(c^T q_i/\sigma_r\right)} \\
&= b^T z_i + \frac{\sigma_{rv}}{\sigma_r}\lambda\left(\frac{c^T q_i}{\sigma_r}\right).
\end{aligned}$$

Note that $\lambda\left(\frac{c^T q_i}{\sigma_r}\right)$ effectively corrects for selection only if there does not already exist some function of regressors in z_i which looks like it. This is why, and also not to rely too much on the stochastic assumptions, in order to strengthen identification, a good practice is to require the existence of variables in q_i (that is x_i) which are not in z_i (here μ_i).

(If $\lambda(t) = \mathbb{E}\left(\frac{r_i}{\sigma_r}\Big|\frac{r_i}{\sigma_r} > t\right)$ is unknown, the regression is called a semiparametric regression. See Robinson (ECMA, 1988).)

Semiparametric regression

Consider the regression:

$$y_i = x_i^T \beta + \lambda(z_i) + u_i$$

Then

$$y_i - \mathbb{E}(y_i|x_i) = \lambda(z_i) - \mathbb{E}(\lambda(z_i)|x_i) + u_i \equiv m(x_i, z_i) + u_i$$

Function $\mathbb{E}(y_i|x_i)$ is identified, and so is $m(x_i, z_i) = \mathbb{E}[y_i - \mathbb{E}(y_i|x_i)|x_i, z_i]$. Hence:

$$\frac{\partial m(x,z)}{\partial z} = \frac{d\lambda(z)}{dz}$$

is identified. So $\lambda(z)$ is identified up to an additive constant.

For estimation, replace conditional expectations by conditional mean estimators, like the Nadaraya-Watson kernel estimator:

$$x \mapsto \widehat{\mathbb{E}}(y_i|x) = \sum_{i=1}^{N} y_i \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right)}$$

where $K\left(\frac{t}{h}\right)$ is a function that puts a lot of weight at $t = 0$ et less and less when $|t|$ gets away from 0 (like the density of a continuous distribution centered at 0).

Identification from reduced forms

Reduced forms a, b and c suffice to identify all parameters.

In particular, β is identified from $c_5 = \beta b_5$; and σ_r, σ_v and $\frac{\sigma_{rv}}{\sigma_r}$ identify σ_u and ρ as

$$\begin{cases} \sigma_r^2 = \sigma_u^2 + \beta^2 \sigma_v^2 + 2\beta\rho\sigma_u\sigma_v \\ \sigma_{rv} = \beta\sigma_v^2 + \rho\sigma_u\sigma_v \end{cases}$$

Minimum distance can be used to recover structural parameters from the reduced form parameters.

In general, suppose that a vector m can be estimated by a root-N consistent estimator $\widehat{m}$. Suppose that $\exists\theta \in \Theta$ such that $m = \widetilde{m}(\theta)$, where $\widetilde{m}(\cdot)$ is injective. Parameter θ can be estimated by minimum distance (GMM):

$$\min_{\theta\in\Theta} \left[\widetilde{m}(\theta) - \widehat{m}\right]^T W^{-1} \left[\widetilde{m}(\theta) - \widehat{m}\right],$$

where W is the positive semi-definite matrix. The Minimum Distance estimator is consistent $\widehat{\theta} \xrightarrow{P} \theta_0$ along the lines of consistency of M-estimators. The optimal choice of the weighting matrix is $W_0 = \text{AVar}\sqrt{N}(\widehat{m} - m)$.

The asymptotic variance of the GMM estimator with the optimal weighting matrix is

$$\text{AVar}(\widehat{\theta}) = \frac{\left[M_0 W_0^{-1} M_0^T\right]^{-1}}{N},$$

where $M_0 = \frac{\partial \widetilde{m}(\theta_0)}{\partial \theta^T}$.

In this example,

$$m = \left(b^T, c^T, \frac{\sigma_{rv}}{\sigma_r}, \sigma_v, \sigma_r\right)^T$$

is related to

$$\theta = \left(a^T, \beta, b^T, \sigma_u, \sigma_v, \rho\right)^T$$

by the set of nonlinear equations sketched above.

Minimum distance

The MD estimator solves:

$$H\left(\widehat{m},\widehat{\theta}\right) \equiv \frac{\partial \widetilde{m}(\widehat{\theta})}{\partial \theta^T} W^{-1}\left[\widetilde{m}(\widehat{\theta}) - \widehat{m}\right] = 0.$$

Taylor expansion in the neighboorhood of $(m_0 \equiv \widetilde{m}(\theta_0), \theta_0)$:

$$\underbrace{H\left(\widehat{m},\widehat{\theta}\right)}_{=0} \simeq \underbrace{H\left(m_0,\theta_0\right)}_{=0} + \frac{\partial H\left(m_0,\theta_0\right)}{\partial m^T}\left(\widehat{m} - m_0\right) + \frac{\partial H\left(m_0,\theta_0\right)}{\partial \theta^T}\left(\widehat{\theta} - \theta_0\right)$$

$$\Leftrightarrow \quad \frac{\partial H\left(m_0,\theta_0\right)}{\partial \theta^T}\left(\widehat{\theta} - \theta_0\right) \simeq -\frac{\partial H\left(m_0,\theta_0\right)}{\partial m^T}\left(\widehat{m} - m_0\right)$$

where

$$\frac{\partial H\left(m_0,\theta_0\right)}{\partial \theta^T} = \frac{\partial \widetilde{m}(\theta_0)}{\partial \theta^T} W^{-1}\left[\frac{\partial \widetilde{m}(\theta_0)}{\partial \theta^T}\right]^T \equiv M_0 W^{-1} M_0^T$$

$$\frac{\partial H\left(m_0,\theta_0\right)}{\partial m^T} = -\frac{\partial \widetilde{m}(\theta_0)}{\partial \theta^T} W^{-1} = -M_0 W^{-1}$$

If matrix $M_0 W^{-1} M_0^T$ (i.e. $\widetilde{m}(\cdot)$ injective) is invertible, the consistency and asymptotic normality of $\widehat{m}$ implies that of $\widehat{\theta}$:

$$\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{L} \mathcal{N}\left(0, \Omega_0\left(W\right)\right)$$

with

$$\Omega_0(W) = \left(M_0 W^{-1} M_0^T\right)^{-1} M_0 W^{-1} W_0 W^{-1} M_0^T \left(M_0 W^{-1} M_0^T\right)^{-1}$$

which is minimal when $W = W_0$:

$$\begin{aligned}
\Omega_0(W) &\gg \Omega_0(W_0) \Leftrightarrow \Omega_0(W_0)^{-1} \gg \Omega_0(W)^{-1} \\
&\Leftrightarrow M_0 W_0^{-1} M_0^T \gg M_0 W^{-1} M_0^T \left(M_0 W^{-1} W_0 W^{-1} M_0^T\right)^{-1} M_0 W^{-1} M_0^T \\
&\Leftrightarrow W_0^{-1} \gg W^{-1} M_0^T \left(M_0 W^{-1} W_0 W^{-1} M_0^T\right)^{-1} M_0 W^{-1} \\
&\Leftrightarrow I \gg \underbrace{W_0^{1/2} W^{-1} M_0^T}_{=A} \underbrace{\left(M_0 W^{-1} W_0 W^{-1} M_0^T\right)^{-1}}_{=\left(A^T A\right)^{-1}} \underbrace{M_0 W^{-1} W_0^{T/2}}_{=A^T}
\end{aligned}$$

which is true as matrix $A\left(A^T A\right)^{-1} A^T$, for $A = W_0^{1/2} W^{-1} M_0^T$, is an orthogonal projector whose eigenvalues are zeroes and ones.

Full Maximum Likelihood

Full maximum likelihood yields efficiency.

Usually, involves numerical optimization techniques. Reduced forms are useful to provide starting values for the algorithms and control the results (easy to make programming mistakes; Monte Carlo simulations recommended).

The likelihood is given by

$$
\begin{aligned}
L(\theta) &= \prod_{h_i>0} \operatorname{pdf}(\ln \pi_i = \ln w_i, h_i^* = h_i) \prod_{h_i=0} \Pr(h_i^* \leq 0) \\
&= \prod_{h_i>0} \Pr(h_i^* = h_i | \ln w_i) \operatorname{pdf}(\ln \pi_i = \ln w_i) \prod_{h_i=0} \Pr(h_i^* \leq 0) \\
&= \prod_{h_i>0} \frac{1}{\sigma_u \sqrt{1-\rho^2}} \varphi \left(\frac{a^T x_i + \beta \ln \pi_i + \frac{\rho \sigma_u \sigma_v}{\sigma_v^2} \left(\ln w_i - b^T z_i \right)}{\sigma_u \sqrt{1-\rho^2}} \right) \frac{1}{\sigma_v} \phi \left(\frac{\ln \pi_i - b^T z_i}{\sigma_v} \right) \\
&\quad \prod_{h_i=0} \left[1 - \Phi \left(\frac{c^T q_i}{\sigma_r} \right) \right],
\end{aligned}
$$

Notice that $h_i^* = h_i$ when $h_i > 0$ implies that $h_i^* > 0$. So $\Pr(h_i^* > 0)$ is "embedded" in $\operatorname{pdf}(\ln \pi_i = \ln w_i, h_i^* = h_i)$.

Do not maximize $L(\theta)$ directly. Operate first the change in variables:

$$a \to \frac{a}{\sigma_u}, \beta \to \frac{\beta}{\sigma_u}, b \to \frac{b}{\sigma_v}, \sigma_u \to \frac{1}{\sigma_u}, \sigma_v \to \frac{1}{\sigma_v}.$$

The likelihood becomes more linear in the parameters, which makes numerical convergence easier to achieve. Moreover, parameter ρ is usually difficult to recover. It is advisable to use a grid search for ρ, i.e. maximize the likelihood wrt to $\frac{a}{\sigma_u}, \frac{\beta}{\sigma_u}, \frac{b}{\sigma_v}, \frac{1}{\sigma_u}$ and $\frac{1}{\sigma_v}$ for different values of ρ in $[-1, 1]$. This provides a first approximation of $\widehat{\rho}$ and the other estimates. Then maximize the likelihood wrt $\frac{a}{\sigma_u}, \frac{\beta}{\sigma_u}, \frac{b}{\sigma_v}, \frac{1}{\sigma_u}, \frac{1}{\sigma_v}$ and ρ starting from this point. This allows to find the right local maximum.

The MLE provides an efficient estimator under the joint normality assumption. However, the joint normality may be a strong assumption. Multi-step methods are often more robust as they are consistent under less restrictive conditions. Moreover, the MLE is sometimes compositionally cumbersome.

Conclusion

It is very difficult to properly analyze these issues of unobserved heterogeneity, simultaneity and selectivity without a proper economic model formatting the econometric equations.

Ideally, the econometric model should be *the* economic model. That is, heterogeneity and shocks should be proper variables of the economic model, and the stochastic assumptions about their distributions should be embedded in the economic model itself.

I will call this empirical approach, **structural econometrics**.

Econometrics would then be only about the statistical techniques of inference on the model parameters, not about modelling.

Unfortunately, this is not always the case. Papers start with a question. Then there is an economic interpretation. Sometimes, a formal economic model is developed to analyze this questions, but not always. Lastly, very often, a reduced form econometric study is provided which establishes a few correlations which are supposed to support the economic theory.

The aim of the remaining lectures is to study exemplary structural econometric papers.