サンプルセレクションバイアスとその補正（統計）

1. Sample Selection Bias と対処法

(1) 設定 ・ ただし観測されない

$$(\text{賃金 } W_i^*) = \begin{cases} \beta_0 + \beta_1(\text{教育年数 } x_i) + \varepsilon_i & \text{if } s_i = 1 \\ (\beta_0 + \beta_1(\text{教育年数 } x_i) + \varepsilon_i) & \text{if } s_i = 0 \end{cases} \quad \text{———} (*)$$

・$$(\text{働いているか } s_i) = \begin{cases} 1 & \text{if } (\text{賃金 } W_i^*) - (\text{留保賃金 } \underline{W}_i) = \gamma_0 + \gamma_1(\text{貯金 } z_i) + u_i \\ 0 & \text{if } (\text{賃金 } W_i^*) - (\text{留賃 } \underline{W}_i) = \gamma_0 + \gamma_1(\text{貯金 } z_i) + u_i < 0 \end{cases}$$

ただし、$s_i = 1$ のときのみ W_i^* を観測し、

$s_i = 0$ ： W_i^* を観測できない。

また、$\begin{bmatrix} \varepsilon_i \\ u_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\varepsilon^2 & \rho_{\varepsilon u} \\ \rho_{\varepsilon u} & 1 \end{bmatrix}\right)$ とする。

※もちろん $X_i = Z_i$ であっても問題ない。

(2) 操作： 観測された、データ、今回は、$i = 1, 3$ に限定して β を推定すると、

i	W^*	S	X	Z
1	9	1	16	2
2		0	12	1
3	10	1	20	0
4		0	14	1

推定値 $\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 5 \\ 0.25 \end{pmatrix}$ を得るが……。

(3) 問題点： Selection bias（バイアス形成過程に注目）

推定量 $\hat{\beta}$（$s_i = 1$ に制限した data）は bias がある。$\mathbb{E}[\hat{\beta}(\text{observed data})] \neq \beta$。

どのようなメカニズムでどれだけの（大きさの）バイアスが発生しているのか？

$$\underbrace{\mathbb{E}[W_i \mid s_i = 1, X_i]}_{\text{$s_i=1$ に制限したデータを所与とした賃金 W_i の期待値。}} = \mathbb{E}[W_i^* \mid s_i = 1, X_i] = \mathbb{E}[W_i^* \mid \gamma_0 + \gamma_1 Z_i + u_i > 0, X_i]$$

$$= \mathbb{E}[\beta_0 + \beta_1 X_i + \varepsilon_i \mid u_i > -\gamma_0 - \gamma_1 Z_i, X_i]$$

$$= \beta_0 + \beta_1 X_i + \mathbb{E}[\varepsilon_i \mid u_i > -\gamma_0 - \gamma_1 Z_i, X_i]$$

$$= \beta_0 + \beta_1 X_i + \rho_{\varepsilon, u}\, \mathbb{E}[u_i \mid u_i > -\gamma_0 - \gamma_1 Z_i]$$

（ただし設定より、$u_i \sim N(0,1)$） (*)

ここで一般に、$Z \sim N(0,1)$ のとき、$\mathbb{E}[Z \mid Z > c]$ の値を考える。

$$\mathbb{E}[Z \mid Z>c] = \int_{-\infty}^{\infty} z \, f_{Z|\mathbb{I}}(z \mid \mathbb{I}_{(Z>c)} = 1)\, dz \quad (\because \text{Conditional Expectation のテーギ})$$

$$= \int_{-\infty}^{\infty} z \cdot \frac{f_{Z,\mathbb{I}}(z, \mathbb{I}_{(Z>c)} = 1)}{\mathbb{P}(\{\mathbb{I}_{(Z>c)} = 1\})} dz \quad (\because \text{Conditional Prob のテーギ})$$

$$= \int_{-\infty}^{\infty} z \frac{f_{Z,\mathbb{I}}(z, Z>c)}{\mathbb{P}(\{Z>c\})} dz$$

$$= \frac{1}{\mathbb{P}(\{Z>c\})} \int_{c}^{\infty} z f_Z(z)\, dz$$

$$= \frac{1}{\mathbb{P}(\{Z<-c\})} \int_{c}^{\infty} z f_Z(z)\, dz \quad (\because Z \sim N(0,1) \text{ より対称性が成り立つ}) \quad \text{—①}$$

部分積分公式：$(f \cdot g)' = f' g + f \cdot g'$ について、$f(x) := -1$, $g(x) := \exp\left(-\frac{x^2}{2}\right)$ と考えると、

$$\int_{c}^{\infty} \underbrace{(-1)}_{f} \underbrace{(-z)\exp\left(-\frac{z^2}{2}\right)}_{g'} dz = \left[\underbrace{-}_{f}\underbrace{\exp\left(-\frac{z^2}{2}\right)}_{g}\right]_{c}^{\infty} - \int \underbrace{0}_{f'} \cdot \underbrace{\exp\left(-\frac{z^2}{2}\right)}_{g} dz$$

$$= \left[\exp\left(-\frac{z^2}{2}\right)\right]_{\infty}^{c} = \exp\left(-\frac{c^2}{2}\right) \text{ となる。} \quad \text{—②}$$

①, ② より一般に $Z \sim N(0,1)$ のとき、

$$\mathbb{E}[Z \mid Z>c] = \frac{1}{\Phi(-c)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{c^2}{2}\right) = \frac{\phi(-c)}{\Phi(-c)} \text{ を得る。}$$

これを※1に適用すると、

$$\mathbb{E}[W_i \mid S_i = 1, X_i] = \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \mathbb{E}[u_i \mid u_i > -\gamma_0 - \gamma_1 Z_i]$$

$$= \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \underbrace{\frac{\phi(\gamma_0 + \gamma_1 Z_i)}{\Phi(\gamma_0 + \gamma_1 Z_i)}}_{\text{L：逆ミズル比}} \text{ を得る。}$$

よって

$$\mathbb{E}[W_i \mid S_i = 1, X_i] = \underbrace{\beta_0 + \beta_1 X_i}_{\text{真値}} + \underbrace{\rho_{\varepsilon,u} \frac{\phi(\gamma_0 + \gamma_1 Z_i)}{\Phi(\gamma_0 + \gamma_1 Z_i)}}_{\text{Bias}}.$$

※ $\rho_{\varepsilon,u} = 0$ ならば Sample Selection Bias はない。

〈解釈〉Inv Mill Ratio $\dfrac{\phi(\gamma_0+\gamma_1 z_i)}{\Phi(\gamma_0+\gamma_1 z_i)}$ の図による解釈はRコード参照。

IMRが大きい(小さい) $\Leftrightarrow$ $\gamma_0+\gamma_1 z_i$ が小さい(大きい) (cf Rコード)

$\Leftrightarrow$ cut-off で切りすてられることが起りやすくなり(にくくなり)、

Cond Expe $\mathbb{E}[u_i \mid u_i > -\gamma_0-\gamma_1 z_i]$ も大きくなる(小さくなる)

$\Leftrightarrow$ Bias も大きく(小さく)なる。

(4)解決案：どうやって Sample Selection Bias をとり除いて、
Unbiased Estimator $\hat{\beta}(\text{Data})$ をつくるか？

Heckman's Two-step sample selection correction:

(I) 全ての観測値を用いて、まず、プロビットモデル

$$\mathbb{P}(\{s_i=1 \mid z_i\}) = \mathbb{P}(\{\gamma_0+\gamma_1 z_i + u_i > 0 \mid z_i\}) \quad (\because u_i \sim N(0,1))$$

の推定量 $\hat{\gamma}_{MLE}(\text{DATA})$ を得る。

(例)

i	s	z
1	1	2
2	0	1
3	1	0
4	0	1

から ML 法で求める

(例終)

これをもとに、Inverse Mills Ratio $\hat{\lambda}_i$ の実現値を計算する。

(例) $\hat{\gamma}_{MLE}$ の実現値：$\begin{bmatrix} c_{1MLE} \\ c_{2MLE} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ とすると、例えば上の表データでは

"推定値"

$$\equiv \hat{\lambda}_1 = \frac{\phi(0.5+2\cdot 0.5)}{\Phi(0.5+2\cdot 0.5)} \simeq 0.1387848$$

$$\hat{\lambda}_3 = \frac{\phi(0.5+0\cdot 0.5)}{\Phi(0.5+0\cdot 0.5)} \simeq 0.5091604$$

(例終)

・(Ⅱ). Selected Sample (つまり、$s_i=1$となっている個票データ)を対象とした、

OLS wage on x_i and $\hat{\lambda}_i$.

(例).

i	s	w^*	x	z	$\hat{\lambda}$	
1	1	9	16	2	0.1388	
~~2~~	~~0~~	~~・~~	~~12~~	~~1~~	~~0.2876~~	ムシ
3	1	10	20	0	0.5042	
~~4~~	~~0~~	~~・~~	~~14~~	~~1~~	~~0.2876~~	ムシ
⋮	⋮	⋮	⋮	⋮	⋮	

変数 $x, \hat{\lambda}$ を w^* に OLS。Unbiased Estimator $\hat{\beta}$ を得る。

(解釈).

これより、$\underbrace{W_{miss}}_{\text{働いていない人が、仮に働き出したときの賃金}} = b_0 + b_1 X + b_2(-\hat{\lambda})$ で求めると bias がなくなる。

⟹ 一般化

(Ⅰ). 全てのデータを用いて、

$\mathbb{P}(\{s_i=1 \mid z_i\}) = \mathbb{P}(\{s_i=1 \mid \gamma_0+\gamma_1 z_i+\varepsilon_i\})$ ($\varepsilon_i \sim N(0,1)$) について

$\hat{\gamma}_{MLE}$ (All Data) を求める。これはMLEの性質から一致性を満たす。

よって $\lambda_i = \dfrac{\phi(z'\gamma)}{\Phi(z'\gamma)}$ for $s_i=1$

も一致性をもつ推定量として定めることができる。

実際のデータをinputとして、推定値 λ_i を計算する。

(II) ここで (3) の Selection Bias 形成過程をふまえると、(1) の設定のもとでは、

$$\underbrace{\mathbb{E}[W_{S_i=1} \mid S_i = 1, X_i]}_{\text{今実際には働いている人が、}} = \beta_0 + \beta_1 X_i + \underbrace{\beta_2}_{\equiv \rho_{\varepsilon,u}} \frac{\phi(\gamma_0 + \gamma_1 Z_i)}{\Phi(\gamma_0 + \gamma_1 Z_i)}$$

もし、条件Xのもとで就業したら得る賃金の期待値

$$\underbrace{\mathbb{E}[W_{S_i=1} \mid S_i = 0, X_i]}_{\text{実際には今失業している人が、}} = \mathbb{E}[\beta_0 + \beta_1 X_i + \varepsilon_i \mid S_i = 0, X_i]$$

もし条件Xのもとで就業したら得る賃金の期待値

$$= \beta_0 + \beta_1 X_i + \mathbb{E}[\varepsilon_i \mid \gamma_0 + \gamma_1 Z_i + u_i < 0] \quad (u_i \sim N(0,1))$$

$$= \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \mathbb{E}[u_i \mid u_i < -\gamma_0 - \gamma_1 Z_i]$$

$$= \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \frac{\int_{-\infty}^{-\gamma_0-\gamma_1 Z_i} u_i f_u(u_i)\, du_i}{\mathbb{P}(\{u_i < -\gamma_0 - \gamma_1 Z_i\})}$$

$$= \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \frac{\int_{-\infty}^{-\gamma_0-\gamma_1 Z_i} u_i f(u_i)\, du_i}{1 - \mathbb{P}(\{u_i < \gamma_0 + \gamma_1 Z_i\})}$$

ここで、

$$\int_{-\infty}^{c} \underbrace{(-1)}_{f} \underbrace{(-u)\exp\left(-\frac{u^2}{2}\right)}_{g'} du = \Big[\underbrace{-}_{f}\ \underbrace{\exp\left(-\frac{u^2}{2}\right)}_{g}\Big]_{-\infty}^{c} - \int \underbrace{0}_{f'} \cdot \underbrace{\exp\left(-\frac{u^2}{2}\right)}_{g} du = \left[\exp\left(-\frac{u^2}{2}\right)\right]_{-c}^{\infty}$$

$$= -\exp\left(-\frac{c^2}{2}\right)$$ より、上式は、

$$\mathbb{E}[W_{S_i=1} \mid S_i = 0, X_i] = \beta_0 + \beta_1 X_i + \rho_{\varepsilon,u} \frac{-\phi(\gamma_0 + \gamma_1 Z_i)}{1 - \Phi(\gamma_0 + \gamma_1 Z_i)}$$

$$\equiv \beta_0 + \beta_1 X_i + \beta_2 \frac{-\phi(\gamma_0 + \gamma_1 Z_i)}{1 - \Phi(\gamma_0 + \gamma_1 Z_i)}$$

よって $S_i = 1$ となるデータに限定して OLS wage on X_i, λ_i をすると、
真のモデル(※)のパラメーター β_1 を推定できる。

これを用いて $\mathbb{E}[W_{S_i=1} \mid S_i = 0, X_i]$ もバイアスなく推定できる。