Public Policy Evaluation
Lecture 5: Maximum likelihood estimation, Binary Outcome models, and Propensity score matching

Yao Thibaut Kpegli

Licence 3 in Economics, ENS Paris-Saclay
2023–2024

April 2024

Outline

Intuition

Context

1. You have a sample $\boldsymbol{y} = \{y_1, ..., y_n\}$
2. You know that the sample comes from a random variable Y with a vector of parameters $\boldsymbol{\theta} \in R^K$ whose true value is $\boldsymbol{\theta}_0$
3. You don't know the true value $\boldsymbol{\theta}_0$

Objective

Provide an estimate of $\boldsymbol{\theta}_0$

Intuition

$\hat{\boldsymbol{\theta}}_{MLE}$ = the value of $\boldsymbol{\theta}$ that is such that the probability of having observed $\boldsymbol{y}$ is the highest possible.

Intuition

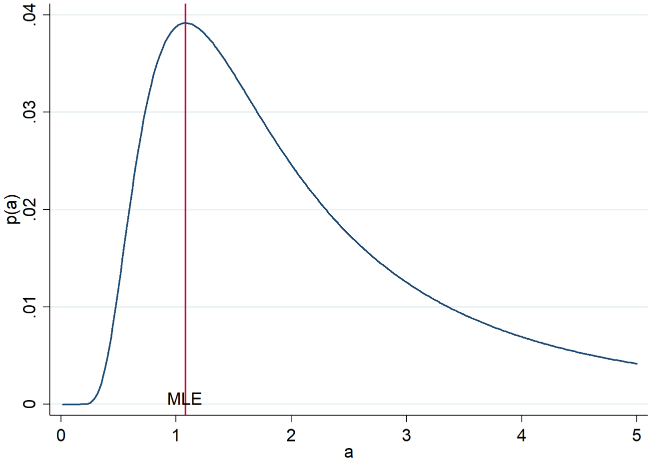
Example (a continuous case)

1. Assume that lifetime of an electronic equipment is a $r.v$ following an exponential distribution with parameter $a > 0$.
2. The density of exponential distribution with parameter $a > 0$ is $f(y; a) = \frac{1}{a}\exp\left(-\frac{y}{a}\right)$
3. We have observed randomly the lifetime 3 times, thereby constituting a sample $y_1 = 1, y_2 = 0.5$ and $y_3 = 1.75$.
4. You want to estimate a, that is the vector of parameters is simply $\boldsymbol{\theta} = a$
5. The joint density of having observed $\{y_1, y_2, y_3\}$ is

$$p(a) = \frac{1}{a}\exp\left(-\frac{1}{a}\right) \times \frac{1}{a}\exp\left(-\frac{0.5}{a}\right) \times \frac{1}{a}\exp\left(-\frac{1.75}{a}\right) = \frac{1}{a^3}\exp\left(-\frac{3.25}{a}\right)$$

Intuition

Figure: probability density of observing the sample as function of a



Intuition

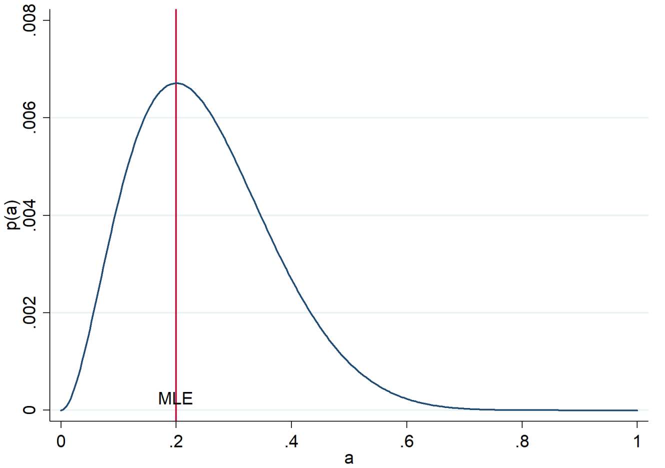
Example (a discrete case)

1. Assume that the fact that a resident of a town has a specific disease is a $r.v$ following a Bernoulli distribution with parameter $a \in (0,1)$, i.e. the probability of having the disease is a
2. We have randomly 10 people, thereby constituting a sample $\{y_1, y_2, ..., y_{10}\} = \{0, 1, 0, 0, 1, 0, 0, 0, 0, 0\}$.
3. You want to estimate a, that is the vector of parameters is simply $\boldsymbol{\theta} = a$
4. The joint probability of observing $\{y_1, ..., y_{10}\}$ is

$$p(a) = (1-a)a(1-a)(1-a)a(1-a)(1-a)(1-a)(1-a)(1-a) = a^2(1-a)^8$$

Intuition

Figure: probability of observing the sample as function of a



Notations and Definitions

Definition (Likelihood function)

The likelihood function $\mathcal{L}$ is:

$$\begin{array}{ccc} R^K & \longrightarrow & [0,1] \\ \theta & \mapsto & \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) = g(\boldsymbol{y};\boldsymbol{\theta}) \end{array}$$

Definition (Log-likelihood function)

The log-likelihood function is:

$$log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n} log f(y_i;\boldsymbol{\theta})$$

Notations and Definitions

Example

1. if $Y_i \rightsquigarrow \mathcal{B}(p)$, then $\boldsymbol{\theta} = p$ and
$$log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) = (\sum_{i=1}^{n} y_i)log(p) + (n - \sum_{i=1}^{n} y_i)log(1-p)$$
2. if $Y_i \rightsquigarrow \mathcal{N}(m, \sigma^2)$, then $\boldsymbol{\theta} = [m, \sigma^2]'$ and
$$log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) = -\frac{1}{2}\sum_{i=1}^{n}\left(log\sigma^2 + log(2\pi) + \frac{(y_i - m)^2}{\sigma^2}\right)$$

Notations and Definitions

Definition (Score function)

The score function is:

$$S(y;\boldsymbol{\theta}) = \frac{\partial log\ f(y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Example

1. For $Y \rightsquigarrow \mathcal{B}(p)$
$$S(y,\boldsymbol{\theta}) = \frac{y}{p} - \frac{1-y}{1-p}$$
2. For $Y \rightsquigarrow \mathcal{N}(m,\sigma^2)$
$$S(y,\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{y-m}{\sigma^2} \\ \\ \dfrac{1}{2\sigma^2}\left(\left(\dfrac{y-m}{\sigma}\right)^2 - 1\right) \end{bmatrix}$$

Notations and Definitions

Proposition 1 (Score expectation)

The expectation of the score is zero

Proof:

$$E[S(Y,\boldsymbol{\theta})] = \int \frac{\partial \, log \, f(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta})dy = \int \frac{\partial \, f(y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{f(y;\boldsymbol{\theta})} f(y;\boldsymbol{\theta})dy =$$

$$\frac{\partial \int f(y;\boldsymbol{\theta})dy}{\partial \boldsymbol{\theta}} = \frac{\partial(1)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Definition (Information matrix)

$$\mathcal{I}_Y(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \, log \, f(Y,\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)$$

Notations and Definitions

Remark (additivity of the information matrix)

The information matrix of two independent experiments is:

$$\mathcal{I}_{X,Y}(\boldsymbol{\theta}) = \mathcal{I}_X(\boldsymbol{\theta}) + \mathcal{I}_Y(\boldsymbol{\theta})$$

Proposition 2 (Variance of the score)

The variance of the score is equal to the information matrix

$$V\Big(S(Y;\boldsymbol{\theta})\Big) \equiv E\left(\left(\frac{\partial log\ f(Y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial log\ f(Y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'\right) = \mathcal{I}_Y(\boldsymbol{\theta})$$

Proof:

Note that $\dfrac{\partial^2\ log\ f(y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \dfrac{\partial^2 f(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\dfrac{1}{f(y;\boldsymbol{\theta})} - \dfrac{\partial log\ f(y,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\dfrac{\partial log\ f(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$

Also $E\left[\dfrac{\partial^2 f(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\dfrac{1}{f(y;\boldsymbol{\theta})}\right] = \dfrac{\partial^2(1)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{0}$. This leads to the result.

Notations and Definitions

Example

1. For $Y \rightsquigarrow \mathcal{B}(p)$

$$\mathcal{I}_Y(\boldsymbol{\theta}) = E\left[\frac{Y}{p^2} + \frac{1-Y}{(1-p)^2}\right] = \frac{1}{p(1-p)}$$

2. For $Y \rightsquigarrow \mathcal{N}(m, \sigma^2)$

$$\mathcal{I}_Y(\boldsymbol{\theta}) = E\begin{bmatrix} \frac{1}{\sigma^2} & \frac{y-m}{\sigma^4} \\ \frac{y-m}{\sigma^4} & \frac{(y-m)^2}{\sigma^6} - \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

Notations and Definitions

Theorem 1 (Cauchy-Schwarz inequality)

Let X and Y be two random variables. Then:

$$|Cov(X,Y)| \leq \sqrt{V(X)V(Y)}$$

Proof: Let $Z = Y - \frac{Cov(X,Y)}{V(X)}X$. Then,

$Cov(X,Z) = Cov(X,Y) - Cov(X,X)\frac{Cov(X,Y)}{V(X)} = 0$. Then,

$V(Y) = V\left(Z + \frac{Cov(X,Y)}{V(X)}X\right) = V(Z) + \left(\frac{Cov(X,Y)}{V(X)}\right)^2 V(X) \geq \frac{(Cov(X,Y))^2}{V(X)}$.

Hence, $V(Y)V(X) \geq (Cov(X,Y))^2$ which leads to the result.

Notations and Definitions

Theorem 2 (Fréchet-Darmois-Cramér-Rao bound)

Consider an unbiased estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ of $\boldsymbol{\theta}$. The variance of the estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ has a lower bound:

$$V\left(\hat{\boldsymbol{\theta}}(\boldsymbol{Y})\right) \geq \mathcal{I}_Y(\boldsymbol{\theta})^{-1} \equiv B_F(\boldsymbol{\theta})$$

Proof: First, $Cov\left(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}), S(Y;\boldsymbol{\theta})\right) = E\left[(\hat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})S(Y;\boldsymbol{\theta})\right] = E\left[\hat{\boldsymbol{\theta}}(\boldsymbol{Y})S(Y;\boldsymbol{\theta})\right] = \int \hat{\boldsymbol{\theta}}(\boldsymbol{y})S(Y;\boldsymbol{\theta})f(y;\boldsymbol{\theta})dy = \int \hat{\boldsymbol{\theta}}(\boldsymbol{y})\frac{\partial\ f(y;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{1}{f(y;\boldsymbol{\theta})}f(y;\boldsymbol{\theta})dy = \frac{\partial \int \hat{\boldsymbol{\theta}}(\boldsymbol{y})f(y;\boldsymbol{\theta})dy}{\partial\boldsymbol{\theta}} = \frac{\partial E\left(\hat{\boldsymbol{\theta}}(\boldsymbol{Y})\right)}{\partial\boldsymbol{\theta}} = \frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\theta}} = \mathbf{1}$. Hence, by Cauchy-Schwarz inequality we have $V\left(\hat{\boldsymbol{\theta}}(\boldsymbol{Y})\right)V\left(S(Y;\boldsymbol{\theta})\right) \geq \mathbf{1}$. The result follows as $V\left(S(Y;\boldsymbol{\theta})\right) = \mathcal{I}_Y(\boldsymbol{\theta})$.

Maximum Likelihood Estimation (MLE)

Definition (Identification)

The vector of parameters $\boldsymbol{\theta}$ is identifiable if, for any other vector $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta}^* \neq \boldsymbol{\theta} \Longrightarrow log\mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{y}) \neq log\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$$

Definition (Likelihood equation)

Necessary condition for maximizing the likelihood function:

$$\frac{\partial log\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} S(y_i; \boldsymbol{\theta}) = \mathbf{0}$$

Definition (Maximum Likelihood Estimator (MLE))

The maximum likelihood estimator $\boldsymbol{\theta}_{MLE}$ is the vector θ that maximizes the likelihood function. Formally:

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} log\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$$

Maximum Likelihood Estimator (MLE)

Example

1. For $Y \rightsquigarrow \mathcal{B}(p)$

$$\boldsymbol{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

2. For $Y \rightsquigarrow \mathcal{N}(m, \sigma^2)$

$$\boldsymbol{\theta}_{MLE} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i \\ \frac{1}{n}\sum_{i=1}^{n} (y_i - m_{MLE})^2 \end{bmatrix}$$

Regularity conditions

Regularity conditions

1. The support of Y does not depends on $\boldsymbol{\theta}$
2. $\boldsymbol{\theta}_0$ is identified
3. The log-likelihood function is continuous in $\boldsymbol{\theta}$
4. $E(log f(Y;\boldsymbol{\theta}_0))$ exists
5. The log-likelihood function is twice continuously differentiable
6. The information matrix at $\boldsymbol{\theta}_0$ $\mathcal{I}_Y(\boldsymbol{\theta}_0) = -E\left(\frac{\partial^2 \; log \; f(Y,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)$ exists and is nonsingular

Proposition 3 (Properties of MLE)

Under regularity conditions, the MLE is (i) consitent, (ii) asymptotically normally distributed and (iii) asymptotically efficient.

Consistency

Kullback-Liebler divergence

If $f_{\boldsymbol{\theta}_0}(y)$ and $f_{\boldsymbol{\theta}_1}(y)$ are two densities, the Kullback-Leibler divergence of $f_{\boldsymbol{\theta}_1}$ w.r.t $f_{\boldsymbol{\theta}_0}$ is

$$KL(f_{\boldsymbol{\theta}_1} \| f_{\boldsymbol{\theta}_0}) = E_{\boldsymbol{\theta}_0}\left[log \frac{f(Y,\boldsymbol{\theta}_0)}{f(Y,\boldsymbol{\theta}_1)}\right] = \int f(y,\boldsymbol{\theta}_0) log \frac{f(y,\boldsymbol{\theta}_1)}{f(y,\boldsymbol{\theta}_0)} dy$$

Proposition 3-0

1. $KL(f_{\boldsymbol{\theta}_1} \| f_{\boldsymbol{\theta}_0}) \geq 0$
2. $KL(f_{\boldsymbol{\theta}_1} \| f_{\boldsymbol{\theta}_0}) = 0 \qquad \text{iff} \qquad f_{\boldsymbol{\theta}_0} = f_{\boldsymbol{\theta}_1}$

Proof: First, $-log(x)$ is a convex function. By Jensen's inequality,
$KL(f_{\boldsymbol{\theta}_1} \| f_{\boldsymbol{\theta}_0}) = E_{\boldsymbol{\theta}_0}\left[-log \frac{f(Y,\boldsymbol{\theta}_1)}{f(Y,\boldsymbol{\theta}_0)}\right] \geq -log E_{\boldsymbol{\theta}_0}\left[\frac{f(Y,\boldsymbol{\theta}_1)}{f(Y,\boldsymbol{\theta}_0)}\right] =$
$-log \int f(y,\boldsymbol{\theta}_1) dy = 0.$
Second, because $-log(x)$ is strictly convex, equality holds if and only if $f(y,\boldsymbol{\theta}_1)/f(y,\boldsymbol{\theta}_0)$ is constant.

Consistency

Proposition 3-1 (Consistency of MLE)

Under regularity conditions, $\boldsymbol{\theta}_{MLE}$ converge in probability to the true value $\boldsymbol{\theta}_0$:

$$\text{plim}\ \boldsymbol{\theta}_{MLE} = \boldsymbol{\theta}_0$$

Informal argument:

$$\begin{aligned}
\boldsymbol{\theta}_{MLE} &= \arg\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} log f(Y_i;\boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}} -\frac{1}{n}\sum_{i=1}^{n} log f(Y_i;\boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} log f(Y_i;\boldsymbol{\theta_0}) - \frac{1}{n}\sum_{i=1}^{n} log f(Y_i;\boldsymbol{\theta}) \\
&\simeq_{n\longrightarrow+\infty} \arg\min_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}_0} log f(Y;\boldsymbol{\theta}_0) - E_{\boldsymbol{\theta}_0} log f(Y;\boldsymbol{\theta}) \\
&\simeq_{n\longrightarrow+\infty} \arg\min_{\boldsymbol{\theta}} KL(f_{\boldsymbol{\theta}_0} \| f_{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0 \\
&\simeq_{n\longrightarrow+\infty} \boldsymbol{\theta}_0
\end{aligned}$$

with $\simeq_{n\longrightarrow+\infty}$ stands for Law of Large Numbers

Consistency

Example

1. For $Y \rightsquigarrow \mathcal{B}(p)$

$$\boldsymbol{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \xrightarrow{p} \quad E[Y] = p$$

2. For $Y \rightsquigarrow \mathcal{N}(m, \sigma^2)$

$$\boldsymbol{\theta}_{MLE} = \begin{bmatrix} m_{MLE} \\ \sigma^2_{MLE} \end{bmatrix}$$

$$m_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \xrightarrow{p} \quad E[Y] = m$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \quad \xrightarrow{p} \quad E[Y^2] - E[Y]^2 = \sigma^2$$

Consistency

Remark

1. Regularity conditions are sufficient but not necessary conditions to have consistency
   - It is therefore possible for the MLE to be consistent even in situations that do not meet regularity conditions
   - e.g: if $Y \rightsquigarrow \mathcal{U}[0, \theta]$, we have $\boldsymbol{\theta}_{MLE} = \max\{Y_1, ..., Y_n\}$

$$\lim_{n \longrightarrow +\infty} P(|\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}| > \epsilon) = \lim_{n \longrightarrow +\infty} \left(1 - \frac{\epsilon}{\theta}\right)^n = 0$$

Asymptotic normality

Proposition 2 (Asymptotic normality)

The MLE estimator $\boldsymbol{\theta}_{MLE}$ is normally distributed asymptotically:

$$\sqrt{n}\left(\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_0\right) \xrightarrow{dist.} \mathcal{N}\left(0, \mathcal{I}_Y(\boldsymbol{\theta}_0)^{-1}\right)$$

Proof:

- First order Taylor expansion of the first derivative of $log\ \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$
$$\frac{\partial log\ \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}} \simeq \frac{\partial log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}} + \frac{\partial^2 log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$
- Evaluate at $\boldsymbol{\theta} = \boldsymbol{\theta}_{MLE}$
$$0 \simeq \frac{\partial log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}} + \frac{\partial^2 log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_0)$$

Asymptotic normality

- Rearrange

$$\sqrt{n}(\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_0) \simeq \sqrt{n}\frac{\frac{\partial log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}}}{-\frac{\partial^2 log\ \mathcal{L}(\boldsymbol{\theta}_0;\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}} = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial log\ f(y_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}}{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 log\ f(y_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}}$$

- Central Limit Theorem

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial log\ f(y_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} \xrightarrow{dist.} \mathcal{N}\left(E[S(Y,\boldsymbol{\theta}_0)], \frac{V\left(S(Y,\boldsymbol{\theta}_0)\right)}{n}\right)$$

- Law of Large Number

$$B_n = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 log\ f(y_i;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \xrightarrow{p} -E\left[\frac{\partial^2 log\ f(Y;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \mathcal{I}_Y(\boldsymbol{\theta}_0)$$

- Remembering that $E[S(Y,\boldsymbol{\theta}_0)] = 0$ and $V\left(S(Y,\boldsymbol{\theta}_0)\right) = \mathcal{I}_Y(\boldsymbol{\theta}_0)$, Slutsky Theorem leads to

$$\sqrt{n}(\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_0) = \sqrt{n}\frac{A_n}{B_n} = \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \mathcal{I}_Y(\boldsymbol{\theta}_0)^{-1}\right)$$

Asymptotically efficient

Proposition (Asymptotically efficient)

$\boldsymbol{\theta}_{MLE}$ is asymptotically efficient, *i.e.* achieves the FDCR lower bound for consistent estimators.

Proof: We have

$$\sqrt{n}(\boldsymbol{\theta}_{MLE} - \boldsymbol{\theta}_0) \xrightarrow{dist.} \mathcal{N}\left(0, \mathcal{I}_Y(\boldsymbol{\theta}_0)^{-1}\right)$$

This means that

$$V\left(\boldsymbol{\theta}_{MLE}\right) = I(\boldsymbol{\theta}_0)^{-1}$$

with $I(\boldsymbol{\theta}_0) = n\mathcal{I}_Y(\boldsymbol{\theta}_0)$ the information matrix associated to $\{Y_1, ..., Y_n\}$.

Remark:

- The asymptotic variance-covariance matrix $I(\boldsymbol{\theta}_0)$ of the MLE depends on the unknown value of $\boldsymbol{\theta}_0$
- In practice, the matrix is evaluated at $\boldsymbol{\theta}_{MLE}$

Sum up

Sum up

1. You have a sample $\boldsymbol{y} = (y_1, ..., y_n)$
2. You know that the sample comes from a random variable Y with a vector of parameters $\boldsymbol{\theta} \in R^K$ whose true value $\boldsymbol{\theta}_0$ is unknown
3. The log-likelihood of one observation y_i is computed analytically (as a function of $\boldsymbol{\theta}$): $l_i(\boldsymbol{\theta}; y_i) = log f(y_i; \boldsymbol{\theta})$
4. The log-likelihood of the sample is $log\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_i l_i(\boldsymbol{\theta}; y_i)$
5. The The MLE estimator results from the optimization problem
$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} log\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$$
6. We have: $\boldsymbol{\theta}_{MLE} \rightsquigarrow \mathcal{N}\left(0, I(\boldsymbol{\theta}_0)^{-1}\right)$, where $I(\boldsymbol{\theta}_0)$ is estimated as $I(\boldsymbol{\theta}_{MLE})$.

Outline

Context and Objectives

Context

1. You have observed a sample $\boldsymbol{y} = \{y_1, ..., y_n\}$ and y_i has only two possible values (say 0 and 1)

$$y_i = \begin{cases} 1 & \text{with probability} \quad p_i \\ 0 & \text{with probability} \quad 1 - p_i \end{cases} \tag{1}$$

2. You have also observed a vector of characteristics $\boldsymbol{x}_i$ $(K, 1)$ associated to the individual i
3. You suspect that $\boldsymbol{x}_i$ determines p_i :

- and you assume this is through a function g depending on vectors of parameters $\boldsymbol{\theta}$ $(K,1)$

$$p_i = g(\boldsymbol{\theta}'\boldsymbol{x}_i)$$

- $E[y_i|\boldsymbol{x}_i] = p_i = g(\boldsymbol{\theta}'\boldsymbol{x}_i)$
- $V(y_i|\boldsymbol{x}_i) = p_i(1 - p_i) = g(\boldsymbol{\theta}'\boldsymbol{x}_i)\,(1 - g(\boldsymbol{\theta}'\boldsymbol{x}_i))$

Context and Objectives

Objectives

Provide an estimate of $\boldsymbol{\theta}$

Example

Examples of binary context:

- treated or untreated

Context and Objectives

Objectives

Provide an estimate of $\boldsymbol{\theta}$

Example

Examples of binary context:

- treated or untreated
- buy or not a transportation ticket
- declares to tax administration the right level of income or not
- living in the city or in the countryside
- wear a mask or not / has covid or not
- trust or not in trust interaction
- bet or not
- success/failure (exams)
- ...

Linear Probability Model (LPM)

Linear Probability Model (LPM)

1. OLS regression of y on $\boldsymbol{x}$

$$y_i = \boldsymbol{\theta}'\boldsymbol{x}_i + \epsilon_i$$

- In LPM, we then have $g(\boldsymbol{\theta}'\boldsymbol{x}_i) = \boldsymbol{\theta}'\boldsymbol{x}_i$

2. Under the the assumptions of conditional-mean-zero and non-correlated errors, such a regression could be consistent
3. But, at least three problems

- heteroskedasticity: $V(\epsilon_i|\boldsymbol{x}_i) = V(y_i|\boldsymbol{x}_i) = \boldsymbol{\theta}'\boldsymbol{x}_i\,(1 - \boldsymbol{\theta}'\boldsymbol{x}_i)$
- discrete error: $\epsilon_i|\boldsymbol{x}_i = (1 - \boldsymbol{\theta}'\boldsymbol{x}_i, -\boldsymbol{\theta}'\boldsymbol{x}_i; \boldsymbol{\theta}'\boldsymbol{x}_i, 1 - \boldsymbol{\theta}'\boldsymbol{x}_i)$, so error cannot be normal
- unrestricted probability: estimated probability $\widehat{p}_i = \widehat{\boldsymbol{\theta}}'\boldsymbol{x}_i$ may be outside the range $[0, 1]$

Alternative models

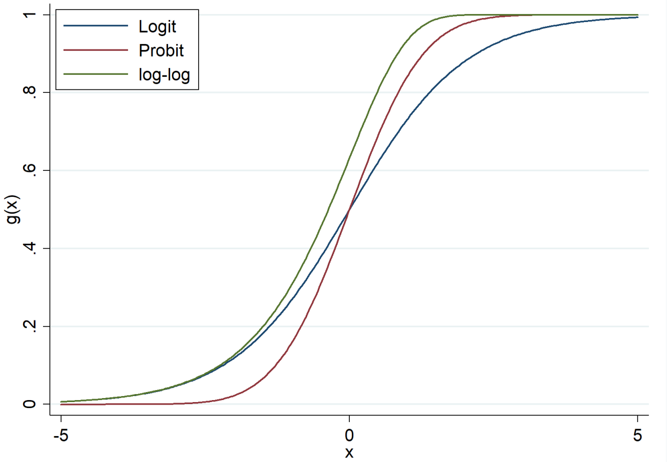
Table: Four common specifications of $g(.)$

Model	Function $g(z)$	Derivative
LPM	z	1
Logit	$\dfrac{\exp(z)}{1+\exp(z)}$	$\dfrac{\exp(z)}{(1+\exp(z))^2}$
Probit	$\int_{-\infty}^{z} \phi(t)dt$	$\phi(z)$
log-log	1 - $\exp(\exp(z))$	$\exp(\exp(z))\ \exp(z)$

$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$ is the density function of $\mathcal{N}(0,1)$

Alternative models

Figure: Probit, Logit and log-log functions



Interpretation in terms of latent variable

Interpretation in terms of latent variable

1. A continuous but unobservable variable y_i^*:

$$y_i^* = \boldsymbol{\theta}'\boldsymbol{x}_i + \epsilon_i$$

with ϵ_i following normal or logistic distribution

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases} \tag{2}$$

2. $p_i = P(y_i^* > 0) = P(\epsilon_i > -\boldsymbol{\theta}'\boldsymbol{x}_i) = 1 - g(-\boldsymbol{\theta}'\boldsymbol{x}_i) = g(\boldsymbol{\theta}'\boldsymbol{x}_i)$

Relationship with Random Utility Models

Relationship with Random Utility Models

1. What precedes relates to Random Utility Models (RUM)
2. Agent (i) chooses $y_i = 1$ if the utility associated with this choice $(U_{i,1})$ is greater than the one of $y_i = 0$ $(U_{i,0})$
3. The random utility:

$$U_{i,j} = V_{i,j} + \epsilon_{i,j}$$

   - where $V_{i,j}$ is the deterministic component of the utility associated with choice $j \in \{0, 1\}$ and $\epsilon_{i,j}$ is a random (agent-specific) component.

4. considering that $g(.)$ is the c.d.f of $\epsilon_{i,0} - \epsilon_{i,1}$, then:

$$p_i = P\left(V_{i,1} + \epsilon_{i,1} > V_{i,0} + \epsilon_{i,0}\right) = g\left(V_{i,1} - V_{i,0}\right)$$

5. In the simple case, $V_{i,j} = \boldsymbol{\theta}_j' \boldsymbol{x}_i$, we have

$$p_i = g\left(\boldsymbol{\theta}' \boldsymbol{x}_i\right) \qquad \text{with} \qquad \boldsymbol{\theta}' = \boldsymbol{\theta}_1' - \boldsymbol{\theta}_0'$$

Estimation

Estimation

1. These models can be estimated by Maximum Likelihood approaches (see previous seance).
2. $(y_i, \boldsymbol{x}_i)$ are assumed to be independent across entities i
3. We have y_i that follows a Bernoulli distribution:

$$P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with} \quad p_i = g(\boldsymbol{\theta}'\boldsymbol{x}_i) \tag{3}$$

4. Log-likelihood of entity i:

$$l_i(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = y_i \, log\left(g(\boldsymbol{\theta}'\boldsymbol{x}_i)\right) + (1-y_i) log\left(1 - g(\boldsymbol{\theta}'\boldsymbol{x}_i)\right) \tag{4}$$

5. Log-likelihood of the sample:

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) \tag{5}$$

Estimation

Estimation

1. F.O.C of the optimization program

$$\frac{\partial log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}} = \mathbf{0}, \text{ that is:}$$

$$\frac{\partial log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{g'(\boldsymbol{\theta}'\boldsymbol{x}_i)\,(y_i - g(\boldsymbol{\theta}'\boldsymbol{x}_i))}{g(\boldsymbol{\theta}'\boldsymbol{x}_i)\,(1 - g(\boldsymbol{\theta}'\boldsymbol{x}_i))}\boldsymbol{x}_i = \mathbf{0}$$

- Nonlinear equation that generally has to be numerically solved

2. We have

$$\boldsymbol{\theta}_{MLE} \xrightarrow{dist.} \mathcal{N}\left(0, I(\boldsymbol{\theta}_0)^{-1}\right)$$

where $I(\boldsymbol{\theta}_0) \approx -\dfrac{\partial^2 log\ \mathcal{L}(\boldsymbol{\theta}_{MLE};\boldsymbol{y},\boldsymbol{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$

Marginal Effects

Marginal Effects

1. The marginal effect is the effect on the probability that $y_i = 1$ of a marginal increase of $x_{i,k}$:

$$m_{i,k} \equiv \frac{\partial p_i}{\partial x_{i,k}} = \underbrace{g'(\boldsymbol{\theta}'\boldsymbol{x}_i)}_{>0}\theta_k$$

2. The sign of the marginal effect $m_{i,k}$ is the one of θ_k
3. It can be estimated by $\widehat{m}_{i,k} = g'(\boldsymbol{\theta}'_{MLE}\boldsymbol{x}_i)\theta_{MLE,k}$
   - the marginal effect $m_{i,k}$ depends on depends on $\boldsymbol{x}_i$, then it is specific to each entity i
4. Two solutions to have "aggregate" marginal effects
   - Marginal Effect at the Mean (MEM) : $g'(\boldsymbol{\theta}'_{MLE}\overline{\boldsymbol{x}})\theta_{MLE,k}$
   - Average Marginal Effect (AME) : $\frac{1}{n}\sum_{i=1}^{n}\widehat{m}_{i,k}$

Goodness of fit

McFadden's **pseudo** $- R^2$

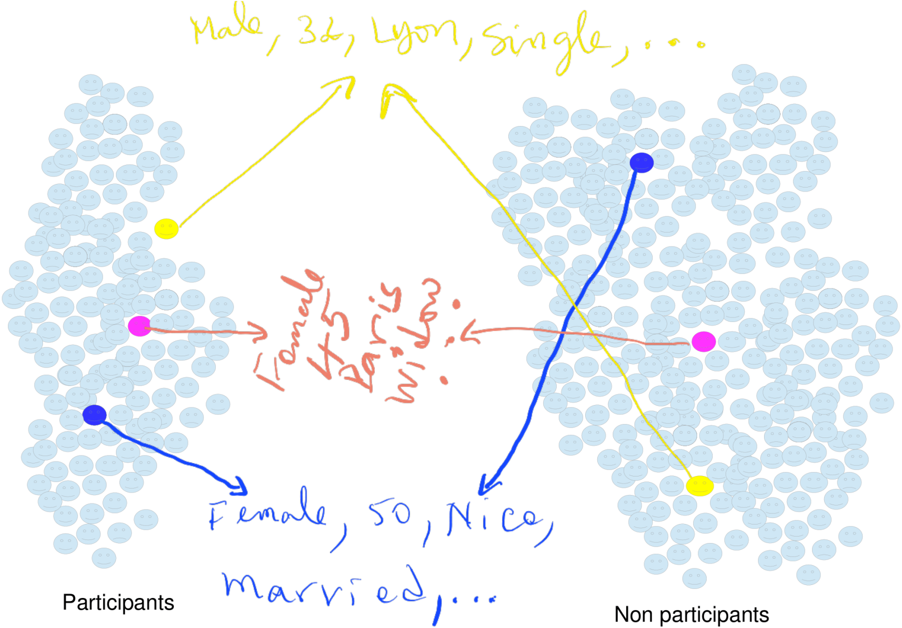
- The **pseudo** $- R^2$:

$$\textbf{pseudo} - R^2 = 1 - \frac{\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x})}{\mathcal{L}_0(\boldsymbol{y})}$$

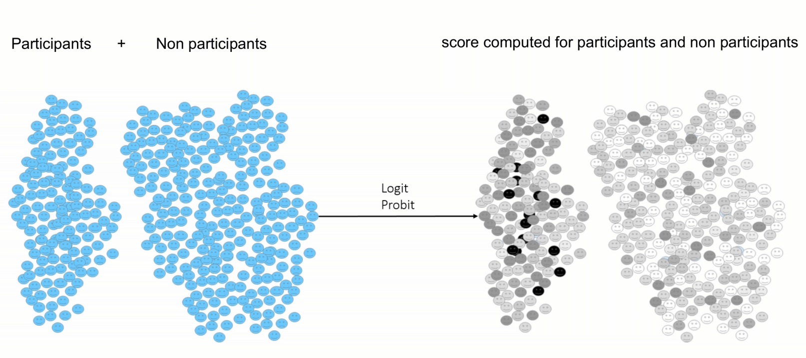
  - with $\mathcal{L}_0(\boldsymbol{y})$ the (maximum) log-likelihood that would be obtained for a model containing only a constant term (i.e. with $x_i = 1$ for all i).
- Intuitively, **pseudo** $- R^2$ will be 0 if the explanatory variables do not allow to predict the outcome variable (y).

Outline

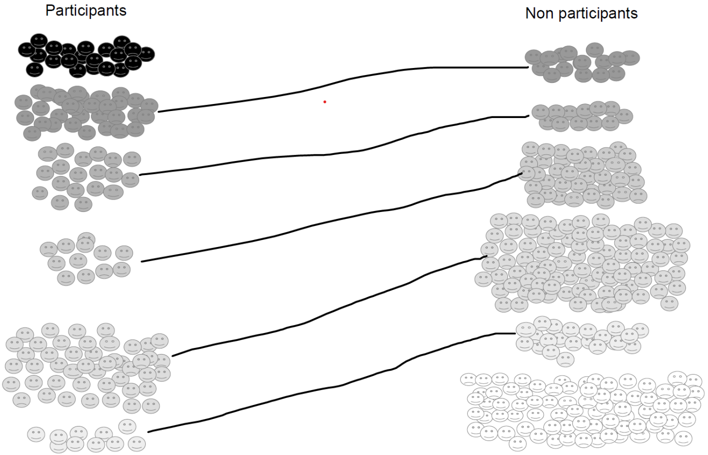
- Applicable when:
  - Treated and untreated have not been randomly assigned
  - Rich set of available information on both treated and untreated
- Aims to approximate the results of random assignment by searching within the sample of untreated individuals for those who are similar to the treated individuals


Male, 36, Lyon, single, ...
Female 45 Paris widow...
Female, 50, Nice, married, ...
Participants
Non participants

- **X**: rich set of information (age, education, sex, place of residence of observation, etc...)
- **Score**: $p_i \equiv P(\boldsymbol{T}_i = 1) = g(\boldsymbol{\theta}' \boldsymbol{X}_i)$

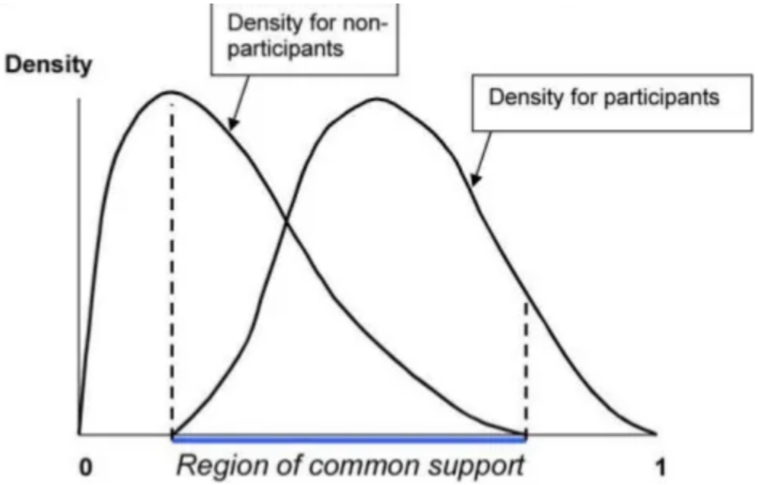


The probability of participation increases with the darkness of the dots.



Treatment effect is computed as the difference in means between matched treated and untreated

Condition of common support



Condition of sample balancing

The average values of observable characteristics $\boldsymbol{X}$ calculated for treated and untreated should be close for values close to the propensity score

Condition of CIA

Conditional Independence Assumption (CIA): treatment assignment is independent of potential outcomes after conditioning on the set of observed characteristics:

$$E[Y_1|T=1,X] = \underbrace{E[Y_1|T=0,X]}_{\text{Unobserved}}$$

$$\underbrace{E[Y_0|T=1,X]}_{\text{Unobserved}} = E[Y_0|T=0,X]$$

- Difference in means between matched treated and untreated individuals:

$$\underbrace{E[Y_1|T=1,X]-E[Y_0|T=0,X]}_{\text{Observed}} = \underbrace{E[Y_1|T=1,X]-{\color{red}E[Y_0|T=1,X]}}_{\text{ATT}} {\color{red}+} \underbrace{{\color{red}E[Y_0|T=1,X]}-E[Y_0|T=0,X]}_{\text{Bias}}$$

- Under CIA, the bias is null and the simple difference corresponds to ATT

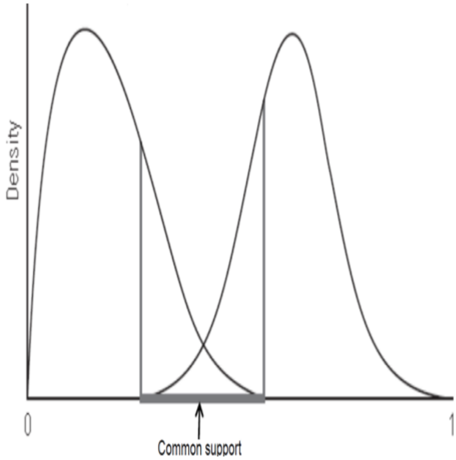
Implementation

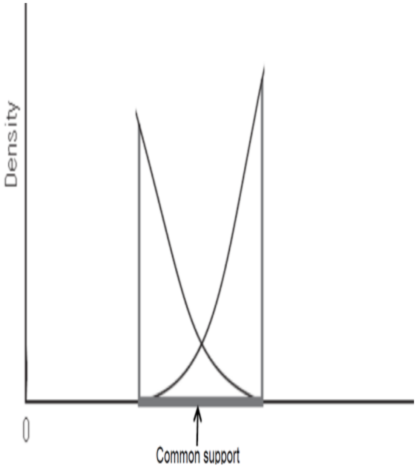
- Select a set of conditioning variables and compute the propensity score (logit, probit, LPM, etc.)
- Check the common support
- Match individuals and check the sample balancing
- Estimate the average treatment effects of interest

Limits: unobservable characteristics

- Selection issues are only due to observable characteristics
  - After controlling for observable characteristics, no selection issues.
  - What about unobservable characteristics (namely self-selection) ?

Limits: common support and local effect





Outline

Application

Blattman , C., Annan, J. (2010). The consequences of child soldiering. The review of economics and statistics, 92(4), 882-898.

Thank you for your attention !