Topics in Econometrics— Discrete Choice Models
Binary Outcome Models

Yao Thibaut Kpegli

Master II Advanced Economics, ENS Lyon

2022–2023

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:

- 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

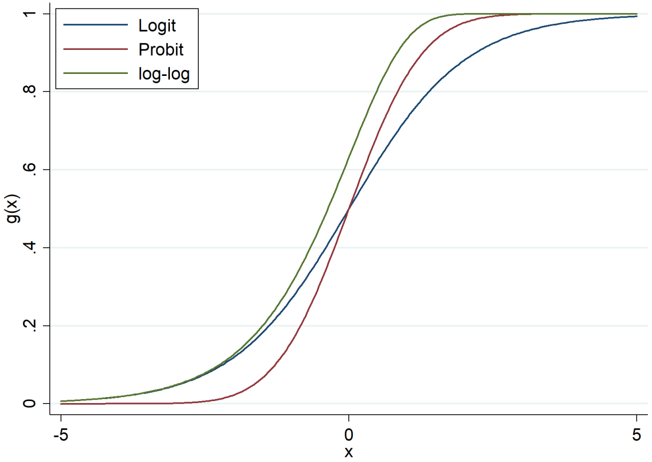
TABLE: Four common specifications of $g(.)$

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

$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$ 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. $p_i = \mathbb{P}(y_i^* > 0) = \mathbb{P}(\epsilon_i > -\boldsymbol{\theta}' \boldsymbol{x}_i) = 1 - g(-\boldsymbol{\theta}' \boldsymbol{x}_i) = g(\boldsymbol{\theta}' \boldsymbol{x}_i)$

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:

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

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

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

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

- Nonlinear equation that generally has to be numerically solved

2. We have

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

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

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).

Data

Data description (Cameron and Trivedi)

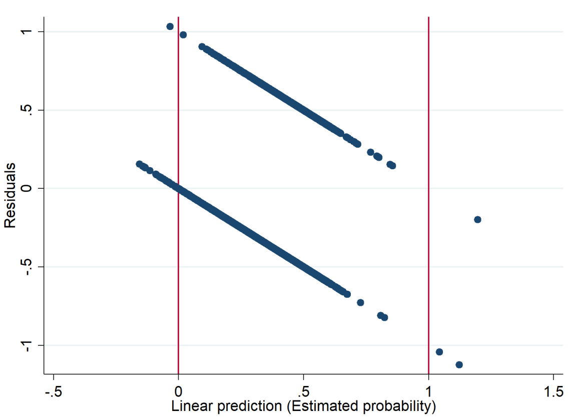
1. The data come from wave 5 (2002) of the Health and Retirement Study (HRS), a panel survey sponsored by the National Institute of Aging. The sample is restricted to Medicare beneficiaries.
2. The HRS contains information on a variety of medical service uses. The elderly can obtain supplementary insurance coverage either by purchasing it themselves or by joining employer-sponsored plans.
3. y: "ins" is the binary variable that indicates the purchase of private insurance from any source, including private markets or associations.
4. $\boldsymbol{x}$: the characteristics are "hstatusg" (self-assessed health-status), "retire", "age", "hhincome" (household income), "educyear", "married", "hisp" (ethnicity)

Data

	ins	retire	age	hstatusg	hhincome	educyear	married	hisp
mean	.3871	.6248	66.9139	.7046	45.2639	11.8986	.7330	.0727
sd	.4872	.4842	3.6758	.4563	64.3394	3.3046	.4425	.2596
min	0	0	52	0	0	0	0	0
max	1	1	86	1	1312.124	17	1	1

Linear Probability Model (LPM)

Figure: Residuals and estimated probabilities



Coeficients: LPM, Logit and Probit

	Logit		Probit		LPM	
retire	0.197*	(0.0842)	0.118*	(0.0513)	0.0409*	(0.0182)
age	-0.0146	(0.0113)	-0.00887	(0.007)	-0.00290	(0.00242)
hstatusg	0.312***	(0.0917)	0.198***	(0.0555)	0.0656***	(0.0195)
hhincome	0.00230**	(0.0008)	0.00123**	(0.0004)	0.0005***	(0.0001)
educyear	0.114***	(0.0142)	0.0707***	(0.00848)	0.0234***	(0.0029)
married	0.579***	(0.0933)	0.362***	(0.0560)	0.123***	(0.0194)
hisp	-0.810***	(0.196)	-0.473***	(0.110)	-0.121***	(0.0337)
_cons	-1.716*	(0.749)	-1.069*	(0.458)	0.127	(0.161)
N	3206		3206		3206	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

MEM: LPM, Logit and Probit

	Logit		Probit		LPM	
retire	0.0460*	(0.0197)	0.0449*	(0.0194)	0.0409*	(0.0182)
age	-0.00341	(0.0026)	-0.00336	(0.00261)	-0.00290	(0.00242)
hstatusg	0.0730***	(0.0214)	0.0749***	(0.0210)	0.0656***	(0.0195)
hhincome	0.0005**	(0.0002)	0.0005**	(0.0001)	0.0005***	(0.0001)
educyear	0.0267***	(0.0033)	0.0268***	(0.0032)	0.0234***	(0.0029)
married	0.135***	(0.0217)	0.137***	(0.0212)	0.123***	(0.0194)
hisp	-0.189***	(0.0456)	-0.179***	(0.0418)	-0.121***	(0.0337)
N	3206		3206		3206	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

AME: LPM, Logit and Probit

	Logit		Probit		LPM	
retire	0.0428*	(0.0182)	0.0420*	(0.0181)	0.0409*	(0.0182)
age	-0.00317	(0.00245)	-0.00315	(0.00245)	-0.0029	(0.00242)
hstatusg	0.0678***	(0.0198)	0.0701***	(0.0196)	0.0656***	(0.0195)
hhincome	0.0005**	(0.0002)	0.0004**	(0.000137)	0.0005***	(0.0001)
educyear	0.0248***	(0.0030)	0.0251***	(0.00291)	0.0234***	(0.0029)
married	0.126***	(0.0198)	0.129***	(0.0195)	0.123***	(0.0194)
hisp	-0.176***	(0.0422)	-0.168***	(0.0389)	-0.121***	(0.0337)
N	3206		3206		3206	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Next: Multiple Choice Models!