

# 微观计量经济学

## 第一次上机课：二元选择模型

赵震宇

[xishanyu2.github.io](https://xishanyu2.github.io)

# 概览

① 本章概要

② 数据案例：健康保险

③ 边际效应

④ 模型诊断

⑤ 扩展：内生回归元

⑥ 扩展：面板数据

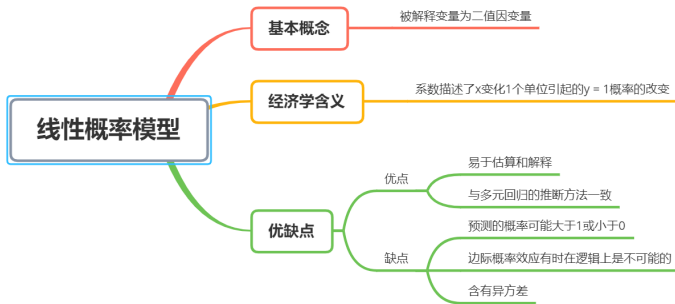
⑦ 数据案例：是否看医生

# 目录

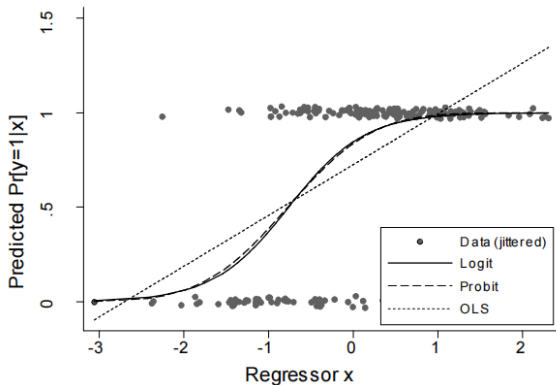
- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元
- 6 扩展：面板数据
- 7 数据案例：是否看医生

# 离散选择模型

- 离散选择或定性响应模型： $y$  只取有限个离散值（分类数据）。
- 二值选择模型：只有两个可能结果，记为 1 和 0；用 logit、probit 建模  $\Pr[y = 1 \mid \mathbf{x}]$ 。
- 二元 logit 与 probit 是非线性模型。



# 直观比较



OLS 的预测可能落在  $(0, 1)$  区间之外；logit 与 probit 的预测非常接近。

## 公式汇总

Table 14.1. Four commonly used binary outcome models

| Model                 | Probability $p = \Pr(y = 1 \mathbf{x})$                                                                                | Marginal effect $\partial p / \partial x_j$                                                   |
|-----------------------|------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| Logit                 | $\Lambda(\mathbf{x}'\boldsymbol{\beta}) = e^{\mathbf{x}'\boldsymbol{\beta}} / (1 + e^{\mathbf{x}'\boldsymbol{\beta}})$ | $\Lambda(\mathbf{x}'\boldsymbol{\beta})\{1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})\}\beta_j$ |
| Probit                | $\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz$                       | $\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$                                                  |
| Linear probability    | $F(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$                                                     | $\beta_j$                                                                                     |
| Complementary log-log | $C(\mathbf{x}'\boldsymbol{\beta}) = 1 - \exp\{-\exp(\mathbf{x}'\boldsymbol{\beta})\}$                                  | $\exp\{-\exp(\mathbf{x}'\boldsymbol{\beta})\} \exp(\mathbf{x}'\boldsymbol{\beta})\beta_j$     |

# Stata 命令

---

|                   |                                     |
|-------------------|-------------------------------------|
| 二项响应模型            | <code>probit; logit; cloglog</code> |
| 异方差 Probit 模型     | <code>hetprobit</code>              |
| 连续内生变量的 Probit 模型 | <code>ivprobit; eprobit</code>      |
| 面板随机效应 Probit 模型  | <code>xtprobit, re</code>           |
| 面板随机效应 Logit 模型   | <code>xtlogit, re</code>            |
| 面板固定效应 Logit 模型   | <code>xtlogit, fe</code>            |

---

# 目录

- 1 本章概要
- 2 数据案例：健康保险**
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元
- 6 扩展：面板数据
- 7 数据案例：是否看医生

# 私人健康保险数据

数据：2000 年 HRS，50–86 岁样本。ins=1 表示拥有私人健康保险。

```
use mus14data.dta, clear
describe ins retire age hstatusg hhincome educyear married hisp
summarize ins retire age hstatusg hhincome educyear married hisp
```

```
bysort ins: summarize retire age hstatusg hhincome educyear married hisp, sep(0)
```

ins=1 者更可能退休、年龄更大、健康状况更好、更富裕、受教育更多、已婚且非西班牙裔。

# Logit 模型

$$\Pr[y_i = 1 \mid \mathbf{x}_i] = \Lambda(\mathbf{x}_i' \beta) = \frac{\exp(\mathbf{x}_i' \beta)}{1 + \exp(\mathbf{x}_i' \beta)}$$

显然  $0 < \Pr[y_i = 1 \mid \mathbf{x}_i] < 1$ 。MLE  $\hat{\beta}$  满足一阶条件：

$$\sum_{i=1}^n (y_i - \Lambda(\mathbf{x}_i' \beta)) \mathbf{x}_i = \mathbf{0}$$

该方程关于  $\beta$  非线性，需迭代求解。

# Logit 估计结果

```
logit ins retire age hstatusg hhincome educyear married hisp
```

```
Iteration 0: Log likelihood = -2139.7712
Iteration 1: Log likelihood = -1996.7434
Iteration 2: Log likelihood = -1994.8864
Iteration 3: Log likelihood = -1994.8784
Iteration 4: Log likelihood = -1994.8784
```

Logistic regression

Number of obs = 3,206

LR chi2(7) = 289.79

Prob > chi2 = 0.0000

Pseudo R2 = 0.0677

Log likelihood = -1994.8784

|  | ins      | Coefficient | Std. err. | z     | P> z  | [95% conf. interval] |
|--|----------|-------------|-----------|-------|-------|----------------------|
|  | retire   | .1969297    | .0842067  | 2.34  | 0.019 | .0318875 .3619718    |
|  | age      | -.0145955   | .0112871  | -1.29 | 0.196 | -.0367178 .0075267   |
|  | hstatusg | .3122654    | .0916739  | 3.41  | 0.001 | .1325878 .491943     |
|  | hhincome | .0023036    | .000762   | 3.02  | 0.003 | .00081 .0037972      |
|  | educyear | .1142626    | .0142012  | 8.05  | 0.000 | .0864288 .1420963    |
|  | married  | .578636     | .0933198  | 6.20  | 0.000 | .3957327 .7615394    |
|  | hisp     | -.8103059   | .1957522  | -4.14 | 0.000 | -1.193973 -.4266387  |
|  | _cons    | -1.715578   | .7486219  | -2.29 | 0.022 | -3.18285 -.2483064   |

# Probit 模型

$$p_i = \Pr[y_i = 1 \mid \mathbf{x}_i] = \Phi(\mathbf{x}_i' \beta)$$

$\Phi(z) = \int_{-\infty}^z \phi(s) ds$  为标准正态 c.d.f., 导数  $\Phi'(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$  为标准正态 p.d.f. 与 logit 不同, probit 的 MLE 一阶条件不能简化:

$$\sum_{i=1}^N (y_i - \Phi(\mathbf{x}_i' \beta)) \frac{\phi(\mathbf{x}_i' \beta)}{\Phi(\mathbf{x}_i' \beta)(1 - \Phi(\mathbf{x}_i' \beta))} \mathbf{x}_i = \mathbf{0}$$

估计结果与 logit 尺度不同, 但  $t$  统计量相似。

# Probit 估计结果

```
probit ins retire age hstatusg hhincome educyear married hisp
```

```
Iteration 0: Log likelihood = -2139.7712
Iteration 1: Log likelihood = -1994.4552
Iteration 2: Log likelihood = -1993.624
Iteration 3: Log likelihood = -1993.6237
Iteration 4: Log likelihood = -1993.6237
```

Probit regression

Number of obs = 3,206

LR chi2(7) = 292.30

Prob > chi2 = 0.0000

Pseudo R2 = 0.0683

Log likelihood = -1993.6237

|  | ins      | Coefficient | Std. err. | z     | P> z  | [95% conf. interval] |           |
|--|----------|-------------|-----------|-------|-------|----------------------|-----------|
|  | retire   | .1183567    | .0512678  | 2.31  | 0.021 | .0178736             | .2188397  |
|  | age      | -.0088696   | .006899   | -1.29 | 0.199 | -.0223914            | .0046521  |
|  | hstatusg | .1977357    | .0554868  | 3.56  | 0.000 | .0889835             | .3064878  |
|  | hhincome | .001233     | .0003866  | 3.19  | 0.001 | .0004754             | .0019907  |
|  | educyear | .0707477    | .0084782  | 8.34  | 0.000 | .0541308             | .0873647  |
|  | married  | .362329     | .0560031  | 6.47  | 0.000 | .252565              | .4720931  |
|  | hisp     | -.4731099   | .1104393  | -4.28 | 0.000 | -.689567             | -.2566529 |
|  | _cons    | -1.069319   | .4580794  | -2.33 | 0.020 | -1.967139            | -.1715002 |

# 模型比较：系数不可直接比较

```

quietly logit ins retire age hstatusg hhincome educyear married hisp
estimates store blogit
quietly probit ins retire age hstatusg hhincome educyear married hisp
estimates store bprobit
quietly regress ins retire age hstatusg hhincome educyear married hisp
estimates store bols
quietly logit ins retire age hstatusg hhincome educyear married hisp, robust
estimates store blogitr
quietly probit ins retire age hstatusg hhincome educyear married hisp, robust
estimates store bprobitr
quietly regress ins retire age hstatusg hhincome educyear married hisp, robust
estimates store bolsr
estimates table blogit bprobit bols blogitr bprobitr bolsr, ///
      stats(N ll) b(%7.3f) t(%7.2f) stfmt(%8.2f)

```

# 预测概率比较

```
quietly logit ins retire age hstatusg hhincome educyear married hisp
predict plogit, p
quietly probit ins retire age hstatusg hhincome educyear married hisp
predict pprobit, p
quietly regress ins retire age hstatusg hhincome educyear married hisp
quietly predict pOLS
summarize ins plogit pprobit pOLS
```

| Variable | Obs          | Mean            | Std. dev.       | Min              | Max             |
|----------|--------------|-----------------|-----------------|------------------|-----------------|
| ins      | <b>3,206</b> | <b>.3870867</b> | <b>.4871597</b> | <b>0</b>         | <b>1</b>        |
| plogit   | <b>3,206</b> | <b>.3870867</b> | <b>.1418287</b> | <b>.0340215</b>  | <b>.9649615</b> |
| pprobit  | <b>3,206</b> | <b>.3861139</b> | <b>.1421416</b> | <b>.0206445</b>  | <b>.9647618</b> |
| pOLS     | <b>3,206</b> | <b>.3870867</b> | <b>.1400249</b> | <b>-.1557328</b> | <b>1.197223</b> |

# 目录

- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元
- 6 扩展：面板数据
- 7 数据案例：是否看医生

# 边际效应

不同模型的系数不能直接比较，比较边际效应：

$$ME_j = \frac{\partial \Pr[y = 1 \mid \mathbf{x}]}{\partial x_j} = F'(\mathbf{x}'\beta)\beta_j = \begin{cases} \Lambda(\mathbf{x}'\beta)(1 - \Lambda(\mathbf{x}'\beta))\beta_j, & \text{logit} \\ \phi(\mathbf{x}'\beta)\beta_j, & \text{probit} \\ \beta_j, & \text{OLS} \end{cases}$$

边际效应取决于：

- $F$  的函数形式；
- 评估点  $\mathbf{x}$ ；
- 参数  $\beta$ 。

# 三类边际效应

- ① 平均边际效应 (AME):

$$AME_j = \frac{1}{N} \sum_{i=1}^N ME_j = \frac{1}{N} \sum_{i=1}^N F'(\mathbf{x}'_i \hat{\beta}) \hat{\beta}_j$$

- ② 均值处边际效应 (MEM):

$$MEM_j = ME_j(\mathbf{x} = \bar{\mathbf{x}}) = F'(\bar{\mathbf{x}}' \hat{\beta}) \hat{\beta}_j$$

- ③ 评估点边际效应 (MER):

$$MER_j = ME_j(\mathbf{x} = \mathbf{x}^*) = F'(\mathbf{x}^{*'} \hat{\beta}) \hat{\beta}_j$$

# Stata 示例

只有当  $F(\mathbf{x}'\beta) = \mathbf{x}'\beta$  (线性模型) 时三者相同, 三者不同因为在不同点评估  $F'(\cdot)$ 。

Stata 命令: `margins, dydx(*)` (AME)、`atmean` (MEM)、`at()` (MER)

```
quietly logit ins retire age hstatusg hhincome educyear married hisp
margins, dydx(*)
margins, dydx(*) atmean
```

# Stata 示例

```
. margins, dydx(*)
```

Average marginal effects  
Model VCE: OIM

Number of obs = 3,206

Expression: Pr(ins), predict()

dy/dx wrt: retire age hstatusg hhincome educyear married hisp

|          | Delta-method |           |       |       |                      |           |
|----------|--------------|-----------|-------|-------|----------------------|-----------|
|          | dy/dx        | std. err. | z     | P> z  | [95% conf. interval] |           |
| retire   | .0427616     | .018228   | 2.35  | 0.019 | .0070354             | .0784878  |
| age      | -.0031693    | .0024486  | -1.29 | 0.196 | -.0079686            | .00163    |
| hstatusg | .0678058     | .0197778  | 3.43  | 0.001 | .0290419             | .1065696  |
| hhincome | .0005002     | .0001646  | 3.04  | 0.002 | .0001777             | .0008228  |
| educyear | .0248111     | .0029705  | 8.35  | 0.000 | .0189891             | .0306332  |
| married  | .1256459     | .0198205  | 6.34  | 0.000 | .0867985             | .1644933  |
| hisp     | -.175951     | .0421962  | -4.17 | 0.000 | -.258654             | -.0932481 |

# Stata 示例

```
. margins, dydx(*) atmean
```

Conditional marginal effects  
Model VCE: OIM

Number of obs = 3,206

Expression: Pr(ins), predict()  
dy/dx wrt: retire age hstatusg hhincome educyear married hisp  
At: retire = .6247661 (mean)  
age = 66.91391 (mean)  
hstatusg = .7046163 (mean)  
hhincome = 45.26391 (mean)  
educyear = 11.89863 (mean)  
married = .7330006 (mean)  
hisp = .0726762 (mean)

|          | Delta-method |           |       |       |                      |           |
|----------|--------------|-----------|-------|-------|----------------------|-----------|
|          | dy/dx        | std. err. | z     | P> z  | [95% conf. interval] |           |
| retire   | .0460479     | .0196856  | 2.34  | 0.019 | .0074648             | .084631   |
| age      | -.0034129    | .0026389  | -1.29 | 0.196 | -.008585             | .0017592  |
| hstatusg | .0730168     | .021412   | 3.41  | 0.001 | .0310499             | .1149836  |
| hhincome | .0005386     | .0001785  | 3.02  | 0.003 | .0001888             | .0008885  |
| educyear | .0267179     | .0033025  | 8.09  | 0.000 | .0202452             | .0331907  |
| married  | .135302      | .0217469  | 6.22  | 0.000 | .0926789             | .1779251  |
| hisp     | -.1894732    | .045563   | -4.16 | 0.000 | -.2787749            | -.1001714 |

# 边际效应的近似

一般地,  $ME_j = F'(\mathbf{x}'\beta)\beta_j$ 。

- OLS:  $ME_j = \hat{\beta}_j$ ;
- Logit:  $ME_j \leq 0.25\hat{\beta}_j$ , 因为

$$F'(\mathbf{x}'\beta) = \Lambda(\mathbf{x}'\beta)(1 - \Lambda(\mathbf{x}'\beta)) \leq 0.25$$

- Probit:  $ME_j \leq 0.40\hat{\beta}_j$ , 因为

$$F'(\mathbf{x}'\beta) = \phi(\mathbf{x}'\beta) \leq \frac{1}{\sqrt{2\pi}} \approx 0.40$$

由此得到经验法则:  $\hat{\beta}_{\text{Logit}} \simeq 4\hat{\beta}_{\text{OLS}}$ ,  $\hat{\beta}_{\text{Probit}} \simeq 2.5\hat{\beta}_{\text{OLS}}$ ,  $\hat{\beta}_{\text{Logit}} \simeq 1.6\hat{\beta}_{\text{Probit}}$ .

对 logit, 可以近似  $ME_j \simeq \bar{y}(1 - \bar{y})\hat{\beta}_j$ .

# Logit 的几率比解释

**几率比** (odds ratio):  $p/(1-p)$  度量  $y = 1$  相对于  $y = 0$  的概率。例如  $y = 1$  表示生存、 $y = 0$  表示死亡，几率比为 2 表示生存几率是死亡的两倍。Logit 模型：

$$p = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)} \Rightarrow \frac{p}{1-p} = \exp(\mathbf{x}'\beta)$$

$$\frac{\partial(p/(1-p))}{\partial x_j} = \exp(\mathbf{x}'\beta)\beta_j = \frac{p}{1-p}\beta_j$$

$\hat{\beta}_j = 0.1$  表示  $x_j$  增加 1 单位使几率比乘以  $\exp(\hat{\beta}_j)$ 。

```
logistic ins retire age hstatusg hhincome educyear married hisp
logistic ins retire age hstatusg hhincome educyear married hisp, coef
```

# 目录

- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断**
- 5 扩展：内生回归元
- 6 扩展：面板数据
- 7 数据案例：是否看医生

# McFadden's $R^2$

非线性模型的诊断不如线性模型清晰。有多种度量的方式，很多是二元选择模型特有的，没有单一最佳度量方式。

- ①  $R^2$  度量；
- ② 比较  $\hat{y}$  与  $y$ ；
- ③ 比较预测的  $\hat{\Pr}[y = 1]$  与实际的  $\Pr[y = 1]$ 。

$$Pseudo - R^2 = 1 - \frac{\ln L_{fit}}{\ln L_0}$$

其中  $\ln L_{fit}$  为拟合模型的对数似然， $\ln L_0$  为仅含截距模型的对数似然。

# Hosmer-Lemeshow gof test

```
quietly logit ins retire age hstatusg hhincome educyear married hisp  
estat gof, group(4)
```

note: obs collapsed on 4 quantiles of estimated probabilities.

Goodness-of-fit test after logistic model

Variable: **ins**

```
Number of observations = 3,206  
Number of groups = 4  
Hosmer-Lemeshow chi2(2) = 14.04  
Prob > chi2 = 0.0009
```

# 目录

- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元**
- 6 扩展：面板数据
- 7 数据案例：是否看医生

# 数据案例

详见课件：

- 生育率与劳动力供给：Angrist et al.(1998)
- 健康保险与医疗需求：@Han2019

# 目录

- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元
- 6 扩展：面板数据**
- 7 数据案例：是否看医生

# 混合 logit

直接把所有面板数据当成截面数据（Pooled）来跑普通的 Logit 回归。然后仅对标准误进行聚类调整，以解决同一个体在不同时间点上误差项相关的问题。

**类比：**等价于线性回归中的 OLS + 聚类稳健标准误。

```
logit y x, vce(cluster id)
```

等相关/可交换（Equicorrelation / exchangeable）模型还假设了同一个体不同时间点上的相关性结构，即假设同一个体任意两个不同时间的误差项相关系数都是常数：

$$\text{Cor}[y_{it}, y_{is} \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = E[(y_{it} - \Lambda(\mathbf{x}'_{it}\beta))(y_{is} - \Lambda(\mathbf{x}'_{is}\beta)) \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}] = \rho$$

**类比：**等价于线性回归中的可行广义最小二乘法（Feasible GLS）+ 聚类稳健标准误。

```
xtlogit y x, pa corr(exch) vce(cluster id)
```

# 随机效应 logit

在 logit 模型中引入个体效应  $\alpha_i$ :  $\Pr[y_{it} = 1 \mid \alpha_i, \mathbf{x}_{it}] = E[y_{it} \mid \alpha_i, \mathbf{x}_{it}] = \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)$

$$f(y_{i1}, \dots, y_{iT} \mid \alpha_i, \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = \prod_{t=1}^T \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)^{y_{it}} [1 - \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)]^{1-y_{it}}$$

$$E[y_{it} \mid \mathbf{x}_{it}] = E[E[y_{it} \mid \alpha_i, \mathbf{x}_{it}]] = \int \Lambda(\alpha_i + \mathbf{x}_{it}'\beta) g(\alpha_i) d\alpha_i$$

假设  $\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$ , 将其积分掉:

$$f(y_{i1}, \dots, y_{iT} \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = \int \left[ \prod_{t=1}^T \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)^{y_{it}} [1 - \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)]^{1-y_{it}} \right] g(\alpha_i) d\alpha_i$$

- Stata 命令: **xtlogit y x, re**

# 固定效应 logit

$\Pr[y_{it} = 1 \mid \alpha_i, \mathbf{x}_{it}] = E[y_{it} \mid \alpha_i, \mathbf{x}_{it}] = \Lambda(\alpha_i + \mathbf{x}_{it}'\beta)$ , 其中  $\alpha_i$  与  $\mathbf{x}_{it}$  相关。

- 不能像线性模型那样对  $y_{it}$  关于  $\mathbf{x}_{it}$  和  $N$  个个体虚拟变量做 logit;
- 难以消去  $\alpha_i$  (一阶差分无效);
- 但有办法: 条件 MLE 使用  $f(y_{i1}, \dots, y_{iT} \mid \sum_{t=1}^T y_{it})$ ;
- Stata 命令: `xtlogit y x, fe`

# 相关效应 logit

固定效应的替代方案，假设  $\alpha_i = \bar{\mathbf{x}}_i' \pi + \eta_i$ ,  $\eta_i \sim \mathcal{N}[0, \sigma_\eta^2]$ , 则

$$\Pr[y_{it} = 1 | \alpha_i, \mathbf{x}_{it}] = E[y_{it} | \alpha_i, \mathbf{x}_{it}] = \Lambda(\bar{\mathbf{x}}_i' \pi + \eta_i + \mathbf{x}_{it}' \beta)$$

对  $y_{it}$  关于  $\mathbf{x}_{it}$  和  $\bar{\mathbf{x}}_i$  做随机效应 logit:

```
xtlogit y x xbar, re
```

# 面板 probit

$$\Pr[y_{it} = 1 \mid \alpha_i, \mathbf{x}_{it}] = \Phi(\alpha_i + \mathbf{x}_{it}'\beta)$$

对随机效应 probit,  $\alpha_i \sim \mathcal{N}[0, \sigma_\alpha^2]$ , 可证明:

$$E[y_{it} \mid \mathbf{x}_{it}] = \int \Phi(\alpha_i + \mathbf{x}_{it}'\beta) g(\alpha_i) d\alpha_i = \Phi(\mathbf{x}_{it}'\beta^*), \quad \beta^* = \frac{\beta}{\sqrt{1 + \sigma_\alpha^2}}$$

因此 RE probit 的系数可解释, 且是混合 probit  $E[y_{it} \mid \mathbf{x}_{it}] = \Phi(\mathbf{x}_{it}'\beta)$  系数的倍数。  
固定效应 probit 无法消去  $\alpha_i$ , 因此没有 xtprobit, fe。

# 目录

- 1 本章概要
- 2 数据案例：健康保险
- 3 边际效应
- 4 模型诊断
- 5 扩展：内生回归元
- 6 扩展：面板数据
- 7 数据案例：是否看医生**

# 兰德健康保险实验

数据：样本为 Rand HIS 3–5 年的非平衡面板，`dmdu=1` 表示过去一年看过医生。

```
use mus18data.dta, clear
summarize dmdu lcoins ndisease female age lfam child id year
```

```
xtset id year
xtdescribe
```

```
xtsum dmdu
xttrans dmdu
corr dmdu L.dmdu L2.dmdu
xtsum dmdu lcoins ndisease female age lfam child id year
```

因变量存在一定组内变异，但回归元的组内变异很小：`lcoins`、`ndisease`、`female` 等的组内变异为零或很小，这会影响 FE 估计。

# 模型比较

```

global xlist lcoins ndisease female age lfam child
quietly logit dmdu $xlist
estimates store WRONGSE
quietly logit dmdu $xlist, vce(cluster id)
estimates store POOLED
quietly xtlogit dmdu $xlist, pa corr(exch) vce(robust)
estimates store PA
quietly xtlogit dmdu $xlist, re
estimates store RE
quietly xtlogit dmdu $xlist, fe
estimates store FE
estimates table WRONGSE POOLED PA RE FE, ///
    equations(1) se b(%8.4f) stats(N ll) stfmt(%8.0f)
  
```

## 结果比较

| Variable          | WRONGSE           | POOLED            | PA                | RE                | FE                |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| <b>#1</b>         |                   |                   |                   |                   |                   |
| lcoins            | -0.1572<br>0.0078 | -0.1572<br>0.0109 | -0.1603<br>0.0108 | -0.2404<br>0.0163 | (omitted)         |
| ndisease          | 0.0503<br>0.0027  | 0.0503<br>0.0040  | 0.0515<br>0.0039  | 0.0782<br>0.0055  | (omitted)         |
| female            | 0.3092<br>0.0318  | 0.3092<br>0.0446  | 0.2977<br>0.0438  | 0.4631<br>0.0663  | (omitted)         |
| age               | 0.0043<br>0.0017  | 0.0043<br>0.0022  | 0.0046<br>0.0021  | 0.0073<br>0.0032  | -0.0342<br>0.0184 |
| lfam              | -0.2048<br>0.0324 | -0.2048<br>0.0470 | -0.2044<br>0.0455 | -0.3022<br>0.0645 | 0.4788<br>0.2597  |
| child             | 0.0922<br>0.0553  | 0.0922<br>0.0728  | 0.1185<br>0.0674  | 0.1935<br>0.1002  | 0.2705<br>0.1685  |
| _cons             | 0.6039<br>0.0799  | 0.6039<br>0.1108  | 0.5777<br>0.1066  | 0.8630<br>0.1569  |                   |
| /lnsig2u          |                   |                   |                   | 1.2257<br>0.0491  |                   |
| <b>Statistics</b> |                   |                   |                   |                   |                   |
| N                 | 20186             | 20186             | 20186             | 20186             | 9025              |
| ll                | -11973            | -11973            |                   | -10879            | -3396             |

Legend: b/se

# 结果解读

- 混合 logit 应使用聚类稳健标准误；
- PA 模型比混合 logit 略微更有效；
- RE 模型的  $\beta$  与混合模型的  $\beta$  尺度不同；
- FE 模型估计精度差得多。

# 自测题

Q1. 线性概率模型（LPM）的主要缺陷是什么？

- A. 无法处理连续自变量
- B. 预测概率可能超出  $[0, 1]$ ，且误差天然具有异方差性
- C. 系数无法解释为概率变化
- D. 只能用于大样本

# 自测题

Q2. 在 Logit 模型中，系数  $\hat{\beta}_k = 0.8$  的正确解读是什么？

- A.  $X_k$  增加 1 单位， $Y = 1$  的概率增加 0.8（即 80 个百分点）
- B.  $X_k$  增加 1 单位，对数几率（log-odds）增加 0.8；需要乘以密度函数  $f(X\beta)$  后才能得到概率变化
- C.  $X_k$  增加 1 单位， $Y$  的取值增加 0.8
- D.  $X_k$  的边际效应恒为 0.8，与  $X$  的取值无关

# 自测题

Q3. 在实践中，Logit 和 Probit 模型应当如何选择？

- A. 通过比较两者的对数似然值选择拟合更好的模型
- B. 严格根据误差项的真实分布（Logistic 或正态）选择
- C. 在 AME 尺度上两者几乎总是给出相同结论，选择主要遵循学科惯例（经济学偏 Probit，公共卫生偏 Logit）
- D. Probit 总是优于 Logit，因为正态分布更常见

# 思考题

- ① 使用 Python 或 Stata 生成一个二元因变量的模拟数据集。分别用 LPM、Logit 和 Probit 估计，比较：
  - 系数的符号和显著性是否一致？
  - AME 是否接近？
  - LPM 预测值是否出现在  $[0, 1]$  之外？
- ② 对上一题的数据，在 Logit 模型中计算每个自变量的 AME，并根据 AME 的大小排序。哪个变量在概率尺度上的影响最大？与直接比较 Logit 系数大小的排序有何不同？
- ③ 使用 Python 或 Stata 的真实或模拟数据，做一个完整的二元选择分析：
  - 运行 Logit 和 Probit
  - 计算并比较两者的 AME
  - 画一个关键变量的预测概率图（预测概率随该变量的变化曲线）
  - 报告模型的正确分类率和 ROC 曲线

# 思考题

- 1 假设你研究“是否上大学对工资的影响”，你的因变量是“工资是否高于 5000 元”（二元变量）。有位审稿人建议你直接使用 LPM，“因为系数好解读”。你会如何回应？在什么情况下 LPM 确实是一个合理的（甚至被推荐的）选择？
- 2 在医学和公共卫生领域，Logit 回归的结果通常以几率比（OR）的形式报告；在经济学中则更偏好边际效应（AME）。讨论这两种做法的利弊；为什么在经济学中 OR 有时会误导政策讨论？