治理技术专题

定量政治分析方法

Quantitative Analysis II

苏 毓 淞

清华大学社会科学院政治学系

第二讲　Logistic 和 Probit 回归分析

二元因变量

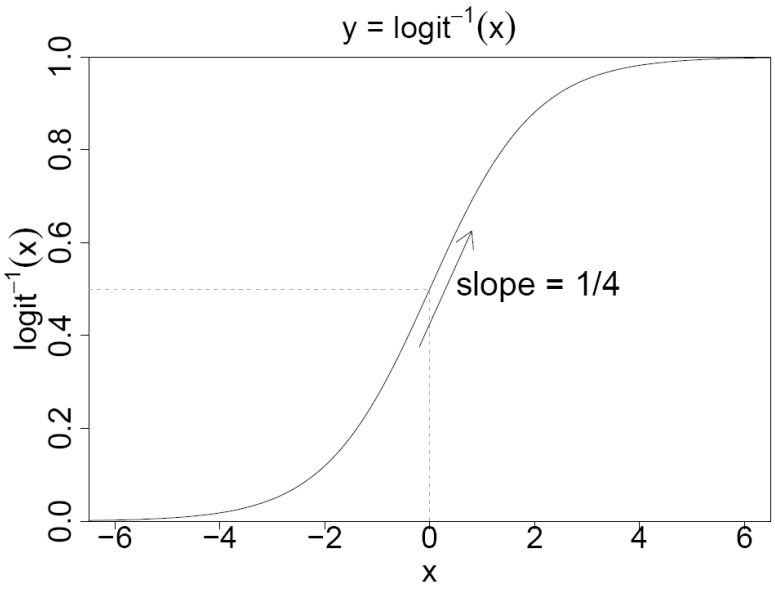
- 当因变量为二元变量时（哑变量、虚拟变量、dummy variable），也就是 y 的取值是 0 和 1，我们使用 Logistic 回归或 Probit 回归。
- 基本原理：$\Pr(y_i = 1) = f(\mathbf{X}_i\boldsymbol{\beta})$，透过 f 函数的转换，$\mathbf{X}_i\boldsymbol{\beta}$ 会成为 0 和 1 之间的数值。
- Logistic 回归和 Probit 的回归就是使用对应的 f 函数，将 $\mathbf{X}_i\boldsymbol{\beta}$ 转换的线性回归。
- 令 $z_i = \mathbf{X}_i\boldsymbol{\beta}$，我们称 z 为 y 之于 x 的潜变量 (latent variable)，而 z 为介于 0 和 1 之间的连续变量，透过 f 转换后，成为取值为 0 和 1 的 y。

logistics 回归

- logistic 回归，又称 logit 回归模型。
- 数学表达式：

$$\begin{aligned}\Pr(y=1) &= \text{logit}^{-1}\left(\mathbf{X}_i\boldsymbol{\beta}\right) \\ &= \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{\exp(\mathbf{X}_i\boldsymbol{\beta})+1} \\ &= \frac{1}{1+\exp(-\mathbf{X}_i\boldsymbol{\beta})}\end{aligned}$$

logistics 回归



-
- logit 曲线在中心时，斜率最陡，斜率是 logit 回归系数除以 4。

logistics 回归案例

```
. logit vote  if year==1992 & presvote < 3

Iteration 0:   log likelihood =  -795.6188
Iteration 1:   log likelihood =  -795.6188

Logistic regression                                     Number of obs   =       1179
                                                        LR chi2(0)      =      -0.00
                                                        Prob > chi2     =          .
Log likelihood =  -795.6188                             Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.3864169   .0593375    -6.51   0.000    -.5027163   -.2701176
------------------------------------------------------------------------------
```

logistics 回归案例

```
. logit vote income if year==1992 & presvote < 3

Iteration 0:   log likelihood =  -795.6188
Iteration 1:   log likelihood = -778.49911
Iteration 2:   log likelihood = -778.45807
Iteration 3:   log likelihood = -778.45807

Logistic regression                                     Number of obs   =       1179
                                                        LR chi2(1)      =      34.32
                                                        Prob > chi2     =     0.0000
Log likelihood = -778.45807                             Pseudo R2       =     0.0216

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .3259947   .0568807     5.73   0.000     .2145106    .4374788
       _cons |   -1.40213   .1894595    -7.40   0.000    -1.773464   -1.030796
------------------------------------------------------------------------------
```

logistics 回归案例

- 回归系数显著性：$\frac{\beta}{SE_\beta} > 2(1.96)$
- 偏差和似然比检验 (Deviance and likelihood ratio tests)：

$$D_{\text{model}} - D_{\text{null}} = -2\log\left(\frac{\text{Likelihood of fitted model}}{\text{Likelihood of null model}}\right)$$

  - null：没有自变量的情形下
  - model: 有自变量的情形下
  - 自由度：model 的自变量数-1
  - χ^2 显著表示有自变量的模型较没有自变数的模型可以解释 y 更多的偏差，拟合优度 (goodness of fit) 显著性改善。
- Psuedo$R^2 = R_L^2 = \frac{D_{\text{null}} - D_{\text{model}}}{D_{\text{null}}}$
  - 表示有自变量的模型较没有自变数的模型可以解释的偏差比。
  - $PRE = \frac{E1 - E2}{E1}$

解释 logit 回归系数

- 由于 logit 回归拟合线是曲线，所以 x 的期望差对应的 y 的期望差是不固定的，斜率最大值在曲线中点。
- $\text{logit}(0.5) = 0, \text{logit}(0.6) = 0.4$，所以在 logit 的刻度上加上 0.4，会让概率从 50% 增加到 60%。
- $\text{logit}(0.9) = 2.2, \text{logit}(0.93) = 2.6$，所以在 logit 的刻度上加上 0.4，会让概率从 90% 增加到 93%

解释 logit 回归系数

- $\Pr(\text{Bush Support}) = \text{logit}^{-1}\left(-1.40 + 0.33\texttt{income}\right)$
- 收入每增加 1 单位，Bush 支持度增加 0.33 logit 概率。
- $\text{logit}^{-1}\left(-1.40 + 0.33 \times 3\right) - \text{logit}^{-1}\left(-1.40 + 0.33 \times 2\right) = 0.08$，收入每增加 1 单位，Bush 支持度增加 8% 概率。
- 增加的概率大约是系数除以 4。

潜变量模型

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$

$$z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

潜变量模型方差无法识别

$$z_i = -1.40 + 0.33x_i + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1.6^2)$$
$$z_i = -14.0 + 3.3x_i + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 16^2)$$
$$z_i = -140 + 33x_i + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 160^2)$$

- 方差 σ^2 在模型中是无法确定的，所以通常我们都把它设定为 1.6^2。
- 1.6 就是 logistic 分布的元单位。
- logistic 分布的方差是 $\frac{\pi}{3} = 1.81^2$，但是我们使用 1.6^2，详见 R 代码。
- $z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon \sim \text{logistic}(0, \sigma^2)$
- 把方差设定为 1，就成为了 probit 回归了！

案例分析

- 因变量：switch 是否改变取水井。
- 自变量：
  - dist：距离最近安全井的距离（公尺）
  - arsenic：自家水井的砷含量
  - assoc：家庭成员中是社区委员会成员
  - educ：教育程度

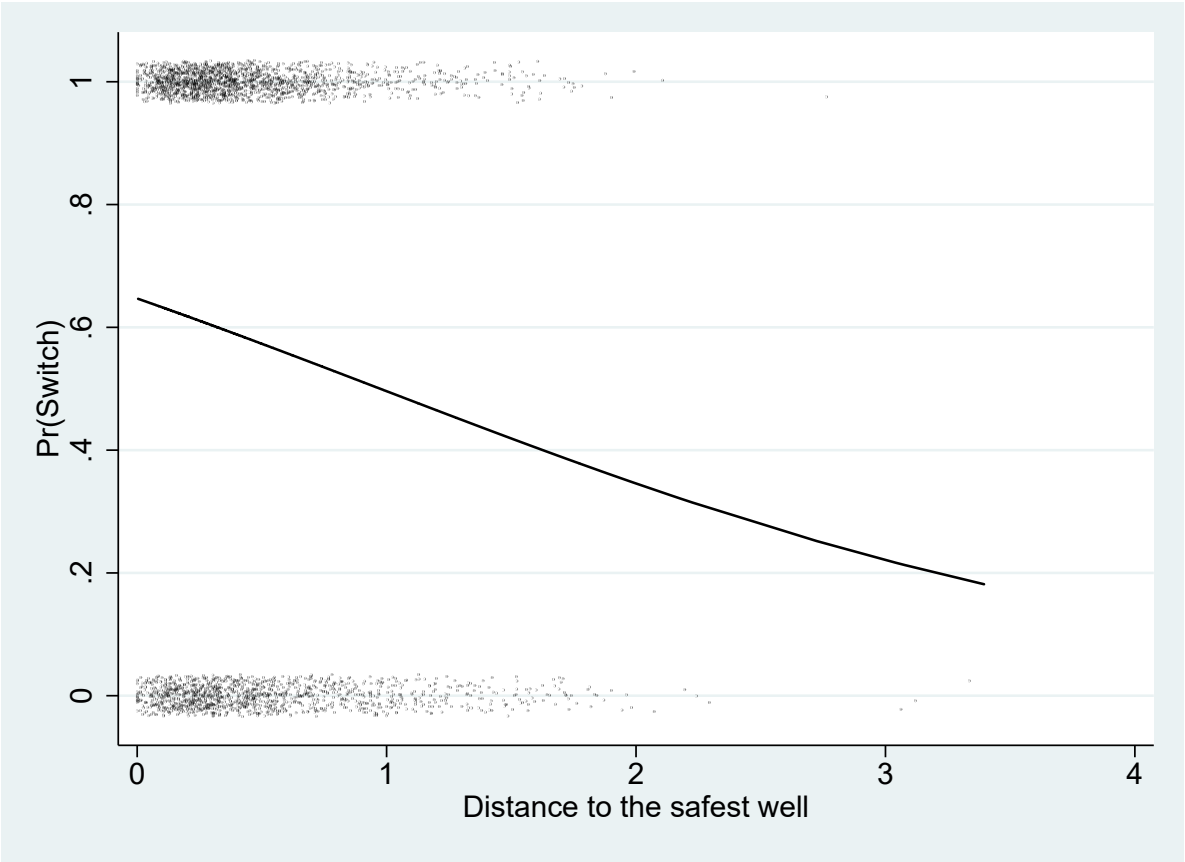
案例分析

```
. logit switch dist

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood = -2038.1212
Iteration 2:   log likelihood = -2038.1189
Iteration 3:   log likelihood = -2038.1189

Logistic regression                                     Number of obs   =       3020
                                                        LR chi2(1)      =      41.86
                                                        Prob > chi2     =     0.0000
Log likelihood = -2038.1189                             Pseudo R2       =     0.0102

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0062188   .0009743    -6.38   0.000    -.0081283   -.0043093
       _cons |   .6059594   .0603102    10.05   0.000     .4877535    .7241652
------------------------------------------------------------------------------
```

案例分析

```
. gen dist100 = dist/100
dist100 already defined
r(110);

.
. logit switch dist100

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood = -2038.1212
Iteration 2:   log likelihood = -2038.1189
Iteration 3:   log likelihood = -2038.1189

Logistic regression                                     Number of obs   =       3020
                                                        LR chi2(1)      =      41.86
                                                        Prob > chi2     =     0.0000
Log likelihood = -2038.1189                             Pseudo R2       =     0.0102

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dist100 |  -.6218819   .0974259    -6.38   0.000    -.8128331   -.4309307
       _cons |   .6059594   .0603102    10.05   0.000     .4877535    .7241652
------------------------------------------------------------------------------
```

案例分析：图示



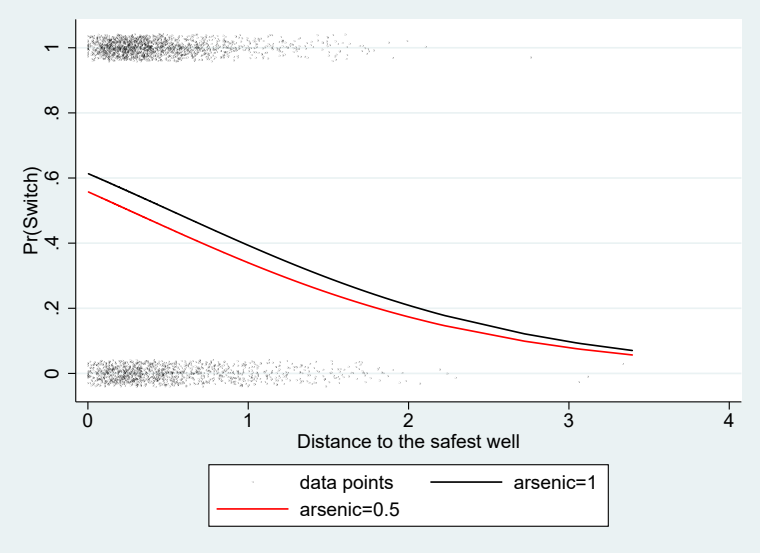
案例分析：系数解读

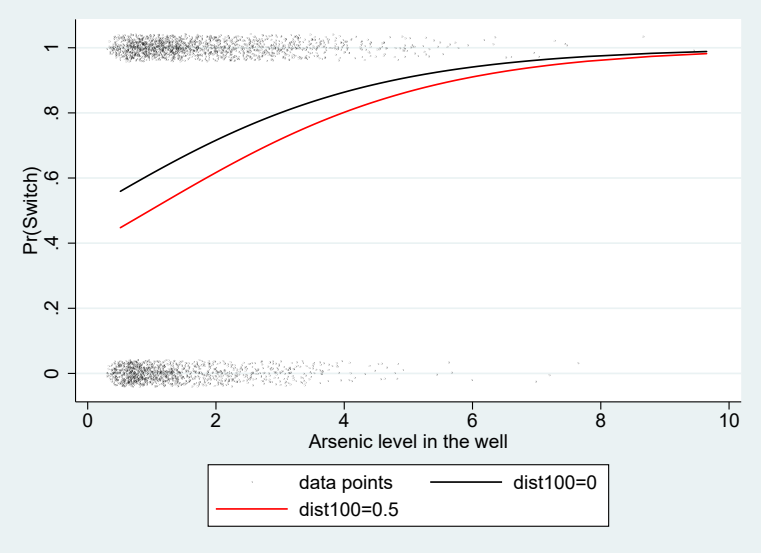
- $\Pr(\texttt{switch}) = \text{logit}^{-1}\left(0.61 - 0.62\texttt{dist100}\right)$
- 常数项：当 $\texttt{dist100} = 0$ 时，换井的概率为 $\text{logit}^{-1}\left(0.61\right) = 0.65$，如果你就住在安全的井旁边，你会换井的概率为 65%
- 回归系数：如果你家每远安全井 100 公尺，你会换井的概率就下降 $0.61/4 \approx 15\%$。

案例分析：两个变量

```
. logit switch dist100 arsenic;

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood =  -1965.863
Iteration 2:   log likelihood = -1965.3343
Iteration 3:   log likelihood = -1965.3341

Logistic regression                                     Number of obs   =       3020
                                                        LR chi2(2)      =     187.43
                                                        Prob > chi2     =     0.0000
Log likelihood = -1965.3341                             Pseudo R2       =     0.0455

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dist100 |  -.8966439   .1043469    -8.59   0.000     -1.10116   -.6921277
     arsenic |   .4607747   .0413848    11.13   0.000     .3796619    .5418875
       _cons |   .0027489   .0794477     0.03   0.972    -.1529657    .1584635
------------------------------------------------------------------------------
```

案例分析：图示


Pr(Switch)
0
.2
.4
.6
.8
1
0
1
2
3
4
Distance to the safest well
data points
arsenic=1
arsenic=0.5


Pr(Switch)
0
.2
.4
.6
.8
1
0
2
4
6
8
10
Arsenic level in the well
data points
dist100=0
dist100=0.5

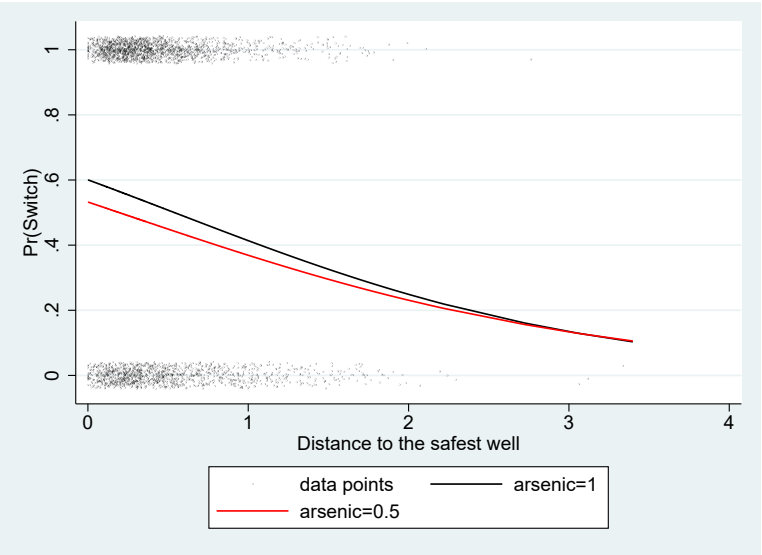
案例分析：系数解读

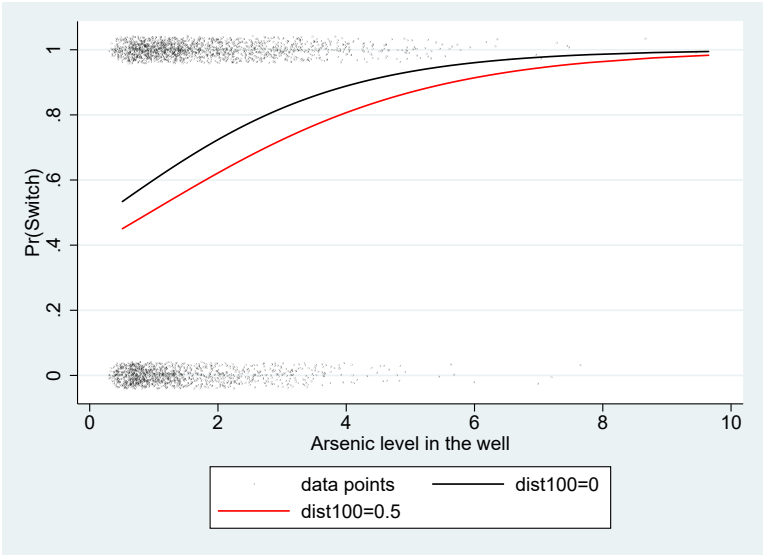
- $\Pr(\texttt{switch}) = \text{logit}^{-1}\left(0.002 - 0.90\texttt{dist100} + 0.46\texttt{arsenic}\right)$
- 回归系数 dist100：如果你家每远安全井 100 公尺，你会换井的概率就下降 $0.90/4 \approx 22\%$；距离每增加 1 个标准差，你会换井的概率就下降 $0.90 \times 0.38/4 \approx 8\%$
- 回归系数 arsenic：如果你家水井砷含量每多 1 单位，你会换井的概率就增加 $0.46/4 \approx 11\%$；砷含量每增加 1 个标准差，你会换井的概率就增加 $0.46 \times 1.1/4 \approx 13\%$

案例分析：两个变量

```
. logit switch dist100 arsenic dist100ars;

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood = -1964.7519
Iteration 2:   log likelihood =  -1963.815
Iteration 3:   log likelihood = -1963.8142
Iteration 4:   log likelihood = -1963.8142

Logistic regression                               Number of obs   =       3020
                                                  LR chi2(3)      =     190.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -1963.8142                       Pseudo R2       =     0.0463

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dist100 |  -.5772179   .2091793    -2.76   0.006    -.9872017    -.167234
     arsenic |   .5559767   .0693194     8.02   0.000     .4201133    .6918402
  dist100ars |   -.178906   .1023282    -1.75   0.080    -.3794656    .0216536
       _cons |  -.1478681   .1175381    -1.26   0.208    -.3782385    .0825023
------------------------------------------------------------------------------
```

案例分析：图示


Pr(Switch)
0
.2
.4
.6
.8
1
0
1
2
3
4
Distance to the safest well
data points
arsenic=1
arsenic=0.5


Pr(Switch)
0
.2
.4
.6
.8
1
0
2
4
6
8
10
Arsenic level in the well
data points
dist100=0
dist100=0.5

案例分析：系数解读

- $\Pr(\texttt{switch}) = \text{logit}^{-1}\left(-0.15 - 0.58\texttt{dist100} + 0.56\texttt{arsenic} - 0.18\texttt{dist100} : \texttt{arsenic}\right)$
- 常数项：当 $\texttt{dist100} = 0$ 和 $\texttt{arsenic} = 0$ 时，换井的概率为 $\text{logit}^{-1}\left(-0.15\right) = 0.47$。
- 回归系数 dist100：将 arsenic 定在均值 1.66，如果你家每远安全井 100 公尺，而你家水井砷含量为 1.66，你会换井的概率就下降 $-0.58 + 0.18 * 1.66 = 0.88/4 \approx 22\%$
- 回归系数 arsenic：将 dist100 定在均值 0.48，如果你家水井砷含量每多 1 单位，你家距离安全井为 48 公尺你会换井的概率就增加 $0.56 - 0.18 \times 0.48 = 0.47/4 \approx 12\%$；

案例分析：系数解读

- $\Pr(\texttt{switch}) = \text{logit}^{-1}\left(-0.15 - 0.58\texttt{dist100} + 0.56\texttt{arsenic} - 0.18\texttt{dist100} : \texttt{arsenic}\right)$
- 交叉项：
  - 从 dist100 的角度来看，每增加 1 单位的 dist100，也就是说距离安全井的距离每多 100 公尺，会减少 arsenic 的回归系数 0.18；自家井含砷量越高，距离预测是否换井概率的重要性随之增加（因为黑线（arsenic$= 1$）在红线 (arsenic$= 0.5$) 之上)；但是这样的关系随着自家井含砷量的增加而越不显著（黑线和红线的差距随着自家井含砷量的增加而越来越不明显)。
  - 从 arsenic 的角度来看，每增加 1 单位的 arsenic，也就是说自家井含砷量每增加 1 单位，会减少 dist100 的回归系数 0.18；距离安全井越远，自家井含砷量预测是否换井概率的重要性随之减少（因为黑线（dist100$= 0$）在红线 (dist100$= 0.5$) 之上)；但是这样的关系随着自家井含砷量的增加而越不显著 (黑线和红线的差距随着安全井距离的增加而越来越不明显)。

案例分析：多个变量

```
. logit switch cdist100 carsenic cdisars assoc educ;

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood = -1953.7595
Iteration 2:   log likelihood = -1952.6766
Iteration 3:   log likelihood = -1952.6755
Iteration 4:   log likelihood = -1952.6755

Logistic regression                               Number of obs   =       3020
                                                  LR chi2(5)      =     212.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -1952.6755                       Pseudo R2       =     0.0517

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cdist100 |  -.8752828   .1050702    -8.33   0.000    -1.081217    -.669349
    carsenic |   .4753105   .0422936    11.24   0.000     .3924165    .5582044
     cdisars |  -.1612339   .1022485    -1.58   0.115    -.3616372    .0391695
       assoc |   -.123188   .0769771    -1.60   0.110    -.2740604    .0276843
        educ |   .0419477   .0095941     4.37   0.000     .0231436    .0607518
       _cons |   .2025163   .0693009     2.92   0.003      .066689    .3383436
------------------------------------------------------------------------------
```

潜变量模型：Probit

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$

$$z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i, \qquad \epsilon \sim \mathcal{N}(0, 1)$$

所以 probit 的回归系数是 logit 的回归系数除以 1.6

案例分析：多个变量

```
. probit switch cdist100 carsenic cdisars assoc educ

Iteration 0:   log likelihood = -2059.0496
Iteration 1:   log likelihood = -1954.2447
Iteration 2:   log likelihood = -1954.0525
Iteration 3:   log likelihood = -1954.0525

Probit regression                                 Number of obs   =       3020
                                                  LR chi2(5)      =     209.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -1954.0525                       Pseudo R2       =     0.0510

------------------------------------------------------------------------------
      switch |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cdist100 |  -.5329807   .0642019    -8.30   0.000     -.658814   -.4071473
    carsenic |   .2787411   .0239218    11.65   0.000     .2318553    .3256269
     cdisars |  -.0774047   .0604616    -1.28   0.200    -.1959071    .0410978
       assoc |  -.0792548   .0474158    -1.67   0.095    -.1721882    .0136785
        educ |   .0263465   .0058917     4.47   0.000     .0147989    .0378941
       _cons |   .1179523   .0425845     2.77   0.006     .0344882    .2014164
------------------------------------------------------------------------------
```

事件发生比

- 事件发生的概率：p
- 事件发生比 (Odds)：$\pi = \frac{p}{1-p}$
- logit $(p) = \log(\pi)$
- 事件发生比率 (odds ratio)：$\frac{\pi^*}{\pi}$

非线性到线性的转换

$$p = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

$$\pi = \frac{p}{1-p} = \frac{\frac{1}{1+e^{-(\alpha+\beta x)}}}{1 - \frac{1}{1+e^{-(\alpha+\beta x)}}} = e^{\alpha+\beta x}$$

$$\text{logit}\,(p) = \log \pi = \alpha + \beta x$$

发生比率解读

$$\pi = e^{(\alpha+\beta x)}$$

$$\pi^* = e^{(\alpha+\beta(x+1))}$$

$$\frac{\pi^*}{\pi} = \frac{e^{(\alpha+\beta(x+1))}}{e^{(\alpha+\beta x)}} = e^{\beta}$$

发生比率解读

- 当 $\beta = 0.693$，则 $e^{\beta} = 2$ 表示自变量变化一个单位，导致新的发生比率是原来的 2 倍。
- $e^{\beta} = 0.8$ 表示自变量变化一个单位，导致新的发生比率是原来的 80%。
- 如果自变量是虚拟变量，$e^{\beta} = 1.6$，则表示取值为 1 的哪一类的发生比是参照类的 1.6 倍。