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

第五讲　定序型因变量分析

类别变量

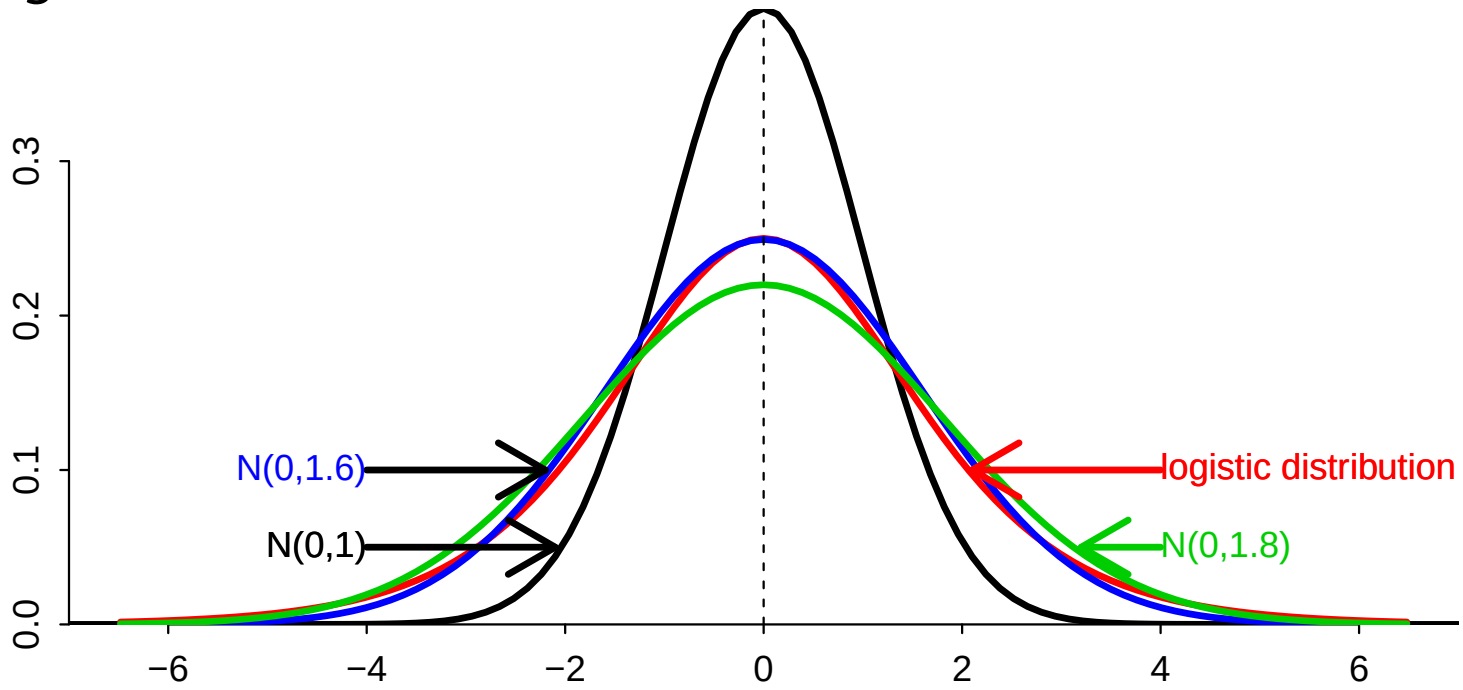
- 定序型：
  - 官员职等：科级 $<$ 处级 $<$ 厅级 $<$ 部级
  - 满意度：非常不满意 $<$ 不满意 $<$ 满意 $<$ 非常满意
- 非定序型
  - 地区：东北、华北、华中、华南、西南、西北、边疆

如何针对定序型类别因变量建模？

- 把它当成连续变量，使用 OLS 回归即可，尤其是它有 5 个类别以上时（最好使用条状图检验下是否近似正态分布，否则不适用 OLS）。
- 拆解成几个二元变量，使用 logistic 或 Probit 回归。
- 使用 ordered logistic/probit 回归。

logistic 和 probit 的差别

- 对于余数的假设不同：logisitic 回归的余数呈标准 logisitic 分布（均值为 0，方差为 3.29），probit 回归的余数呈现标准正态分布 $N(0,1)$。
- logistic 回归的系数是 Probit 系数的 1.6（1.814）倍！



logistic 和 probit 的数学表示式

- logit

$$\begin{aligned}\Pr(y=1) &= \text{logit}^{-1}(\alpha+\beta x)\\ &= \text{logit}^{-1}(\mathbf{X}'\boldsymbol{\beta})\end{aligned}$$

$$y^* = \mathbf{X}'\boldsymbol{\beta}+\epsilon, \epsilon \sim \mathcal{N}(\mu=0, \sigma^2=\frac{\pi^2}{3}=3.29)$$

$$\text{logit} = \log\left(\frac{x}{1-x}\right), \text{logit}^{-1} = \frac{1}{1+\exp(-x)}$$

- probit

$$\begin{aligned}\Pr(y=1) &= \Phi(\alpha+\beta x)\\ &= \Phi(\mathbf{X}'\boldsymbol{\beta})\\ y^* &= \mathbf{X}'\boldsymbol{\beta}+\epsilon, \epsilon \sim N(\mu=0, \sigma=1)\end{aligned}$$

- $y=1$ if $y^*>0, y=0$ if $y^*<0$

Logistic Regression 回顾

```
. 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
------------------------------------------------------------------------------
```

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。
- $\text{Psuedo}_R^2 = R_L^2 = \frac{D_{\text{null}} - D_{\text{model}}}{D_{\text{null}}}$
  - 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。
  - $PRE = \frac{E1 - E2}{E1}$

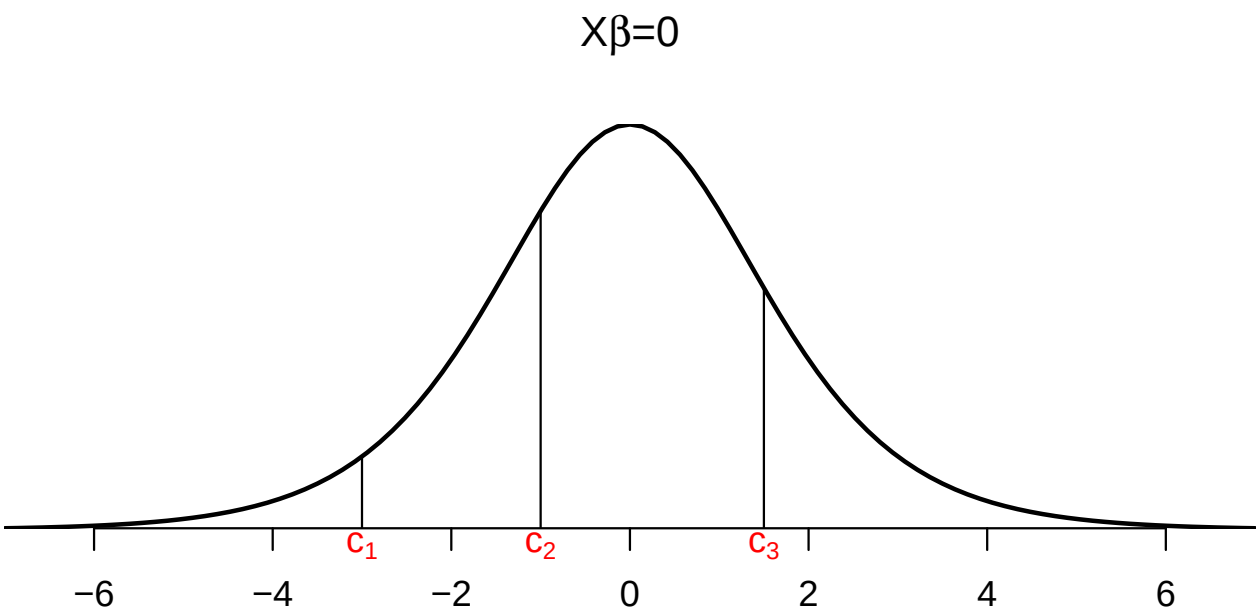
Ordered Logistic Regression

- 又称比例优势模型 (proportional odds model)、平行线模型（parallel lines）model、平行回归模型（parallel regression model）。
- 假设 y 是定序变量，取值 $1, 2, \ldots, K$，则可以表示为：

$$
\begin{aligned}
\Pr(y > 1) &= \text{logit}^{-1}\left(X\beta\right) \\
\Pr(y > 2) &= \text{logit}^{-1}\left(X\beta - c_2\right) \\
\Pr(y > 3) &= \text{logit}^{-1}\left(X\beta - c_3\right) \\
&\vdots \\
\Pr(y > K-1) &= \text{logit}^{-1}\left(X\beta - c_{k-1}\right) \\
\Pr(y = k) &= \Pr(y > k-1) - \Pr(y > k) \\
&= \text{logit}^{-1}\left(X\beta - c_{k-1}\right) - \text{logit}^{-1}\left(X\beta - c_k\right)
\end{aligned}
$$

- 其中 c 代表断点 (cutting points)。

断点的意涵



$$y_i = \begin{cases} 1 & \text{if } z_i < c_1 \\ 2 & \text{if } c_1 < z_i < c_2 \\ 3 & \text{if } z_i > c_2 \end{cases}$$

$$z_i = X\beta + \epsilon_i$$

断点的意涵（续）

$$y_i = \begin{cases} 1 & \text{if } z_i < c_1 \\ 2 & \text{if } c_1 < z_i < c_2 \\ 3 & \text{if } c_2 < z_i < c_3 \\ & \dots \\ K-1 & \text{if } c_{k-2} < z_i < c_{k-1} \\ K & \text{if } z_i > c_{k-1} \end{cases}$$

$$z_i = X\beta + \epsilon_i$$

Ordered Logistic Regression

```
> summary(M2)

Re-fitting to get Hessian

Call:
polr(formula = ordered(happy) ~ male + age + ccpmember + loghinc,
    data = dat2, method = "logistic")

Coefficients:
              Value Std. Error t value
male      -0.117903   0.038521  -3.061
age        0.003698   0.001273   2.904
ccpmember  0.445792   0.059165   7.535
loghinc    0.239434   0.014741  16.242

Intercepts:
    Value   Std. Error t value
1|2 -1.3228  0.1801    -7.3444
2|3  0.3265  0.1714     1.9055
3|4  1.6028  0.1715     9.3480
4|5  4.2958  0.1762    24.3771

Residual Deviance: 24617.98
AIC: 24633.98
(1394 observations deleted due to missingness)
```

预测概率 (predictive probability)

$$
\begin{aligned}
\Pr(y = 1) &= \Pr(y > 0) - \Pr(y > 1) \\
&= 1 - \text{logit}^{-1}\left(Z - c_1\right) \\
&= 1 - \text{logit}^{-1}\left[4 - (-1.3228)\right] \\
\Pr(y = 2) &= \Pr(y > 1) - \Pr(y > 2) \\
&= \text{logit}^{-1}\left(Z - c_1\right) - \text{logit}^{-1}\left(Z - c_2\right) \\
&= \text{logit}^{-1}\left[4 - (-1.3228)\right] - \text{logit}^{-1}\left(4 - 0.3265\right) \\
&\vdots \\
\Pr(y = 5) &= \Pr(y > 4) - \Pr(y > 5) \\
&= \text{logit}^{-1}\left(Z - c_4\right) - 0 \\
&= \text{logit}^{-1}\left(4 - 4.2958\right)
\end{aligned}
$$

胜算比 (odds ratios

```
> exp(coef(M2))
     male       age ccpmember   loghinc
0.8887819 1.0037047 1.5617260 1.2705297
```

- 比较除了性别以外相同的两个人，男性比女性少 ($1 - 0.89 \approx$=0.11）11% 的概率表达他们比较快乐。
- 比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他表达他们比较快乐的概率多 0.4%。
- 比较除了党员身份以外相同的两个人，党员比非党员表达他们比较快乐的概率多 56%。
- 比较除了家庭收入以外相同的两个人，收入每增加 1%，他表达他们比较快乐的概率多 27%。

平行性假设

- Ordinal logistic/robit) regression 假设结果变量各类之间的关系是一样的。
- 也就是说，回归系数 β 解释第 1 类和第 K 类的关系与解释第 2 类和第 K 类的关系是一样的，如果不一样，就得要不同的 β。
- Brant (1990) 平行性检验：比较 Ordinal regression 和各组单独的 binary regression 系数和方差的关系，如果差异显著，则平行性检验不通过，如果不显著，则平行性检验通过。

平行性检验

```
> brant(M2)
------------------------------------------
Test for        X2      df      probability
------------------------------------------
Omnibus         55.5    12      0
male            1.83    3       0.61
age             8.77    3       0.03
ccpmember       16.55   3       0
loghinc         20.51   3       0
------------------------------------------

H0: Parallel Regression Assumption holds
```

只有 male 通过检验，其他变量都没有，说明他们对于幸福每个类别的影响不是等比例的。在这个检验中，我们希望统计值不显著，接受 H0。

无法通过平行性检验的对策

- 不管它，因为对于推论影响不大。
- 使用 Multinomial logit model.
- 将因变量类别合并编码。
- 将因变量类别拆分成数个二元变量。

Pseudo R^2 拟合度评价指标

- 这里的 Pseudo R^2 即 McFadden R^2。
- Pseudo $R^2 = \frac{D_{\text{null}} - D_{\text{model}}}{D_{\text{null}}}$
- 表示有自变量的模型较没有自变量的模型可以解释的偏差比。