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 定类型因变量

# 类别变量（nomial variable)

- 定类型：
  - 宗教信仰：佛教、基督教、伊斯兰教、道教、天主教。
  - 职业：农牧渔民、商业服务业、个体工商户、私营业主、工人、党政干部、管理人员、军警、专业技术人员、一般职员。
- 定序型变量未通过平行检验时，也可当做定类变量，如幸福变量（非常不幸福、不幸福、幸福、非常幸福)。

# 如何针对定类因变量建模？

- 拆解成几个二元变量，使用 logistic 或 Probit 回归。
- 使用 multinomial logistic/probit 回归。

# ologit 和 mlogit 建模的差别

- ologit 的 $y$ 和 $x$ 有单一的线性函数。
- mlogit 的 $y$ 和 $x$ 有 $c-1$ 的线性函数， $c$ 是类别个数。
- mlogit 就像多元回归一样，由许多回归组成。

# mlogit 数学表达式：以三类定类变量为例

- 假设有一个变量取值有三类，发生这三类的概率各是 $\pi_1, \pi_2, \pi_3$。
- 则他们之间的胜算比 (Odds ratio) 可用 logistic 回归表示：

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \mathbf{X}\boldsymbol{\beta}_1$$

$$\log\left(\frac{\pi_2}{\pi_3}\right) = \mathbf{X}\boldsymbol{\beta}_2$$

- 其中，只需要求解 $3-1$ 个 logistic 回归即可。

# mlogit 数学表达式：以三类定类变量为例

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \mathbf{X}\boldsymbol{\beta}_1 \rightarrow \frac{\pi_1}{\pi_3} = \exp(\mathbf{X}\boldsymbol{\beta}_1) \rightarrow \pi_1 = \pi_3 \exp(\mathbf{X}\boldsymbol{\beta}_1)$$

$$\log\left(\frac{\pi_2}{\pi_3}\right) = \mathbf{X}\boldsymbol{\beta}_2 \rightarrow \frac{\pi_2}{\pi_3} = \exp(\mathbf{X}\boldsymbol{\beta}_2) \rightarrow \pi_2 = \pi_3 \exp(\mathbf{X}\boldsymbol{\beta}_2)$$

# mlogit 数学表达式：以三类定类变量为例

$$
\begin{aligned}
\pi_1 + \pi_2 + \pi_3 &= 1 \\
\pi_3 \exp(\mathbf{X}\boldsymbol{\beta}_1) + \pi_3 \exp(\mathbf{X}\boldsymbol{\beta}_2) + \pi_3 &= 1 \\
\pi_3 &= \frac{1}{1 + \exp(\mathbf{X}\boldsymbol{\beta}_1) + \exp(\mathbf{X}\boldsymbol{\beta}_2)} \\
\pi_1 &= \frac{\exp(\mathbf{X}\boldsymbol{\beta}_1)}{1 + \exp(\mathbf{X}\boldsymbol{\beta}_1) + \exp(\mathbf{X}\boldsymbol{\beta}_2)} \\
\pi_2 &= \frac{\exp(\mathbf{X}\boldsymbol{\beta}_2)}{1 + \exp(\mathbf{X}\boldsymbol{\beta}_1) + \exp(\mathbf{X}\boldsymbol{\beta}_2)}
\end{aligned}
$$

# mlogit 数学表达式：一般式

$$
\begin{aligned}
\pi_1 &= \frac{\exp(\mathbf{X}\boldsymbol{\beta}_1)}{1+\sum_{j=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\beta}_j)} \\
\pi_2 &= \frac{\exp(\mathbf{X}\boldsymbol{\beta}_2)}{1+\sum_{j=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\beta}_j)} \\
&\vdots \\
\pi_{k-1} &= \frac{\exp(\mathbf{X}\boldsymbol{\beta}_{k-1})}{1+\sum_{j=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\beta}_j)} \\
\pi_k &= \frac{1}{1+\sum_{j=1}^{k-1}\exp(\mathbf{X}\boldsymbol{\beta}_j)}
\end{aligned}
$$

# IIA: Independent Irrelevance Assumption

- 上面的推导告诉我们，$\pi$ 之间是彼此独立的，也就是 Independent Irrelevance Assumption, IIA，独立不相关假定。
- 我搭公交和搭出租车上班的相对概率（胜算比），不会因为其他选择的加入而改变，例如加入新交通工具（自行车）。
- 这是很强的假定，现实中很难实现。所以可以考虑其他建模方式，例如 Multinomial Probit。

# mprobit 数学表达式：以三类定类变量为例

- 使用潜变量 (latent variable) 表示。
- $$\begin{aligned} Y_1^* &= \mathbf{X}\boldsymbol{\beta}_1 + \epsilon_1 \\ Y_2^* &= \mathbf{X}\boldsymbol{\beta}_2 + \epsilon_2 \\ Y_3^* &= \mathbf{X}\boldsymbol{\beta}_3 + \epsilon_3 \\ \boldsymbol{\epsilon} &\sim N(0, \boldsymbol{\Sigma}) \end{aligned}$$
- $$Y = \begin{cases} 1 & \text{if} \quad Y_1^* > Y_2^* > Y_3^* \\ 2 & \text{if} \quad Y_2^* > Y_1^* > Y_3^* \\ 3 & \text{otherwise} \end{cases}$$

# mprobit 数学表达式：以三类定类变量为例

- $$\epsilon \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}\right)$$
- 这表示余数项彼此间相互不独立，是相关的，所以不用考虑 IIA。
- 余数项彼此间相互独立，则称之为独立 probit。

# mprobit 数学表达式：一般式

$$
\begin{aligned}
Y_1^* &= \mathbf{X}\boldsymbol{\beta}_1 + \epsilon_1 \\
Y_2^* &= \mathbf{X}\boldsymbol{\beta}_2 + \epsilon_2 \\
&\vdots \\
Y_k^* &= \mathbf{X}\boldsymbol{\beta}_k + \epsilon_k \\
\boldsymbol{\epsilon} &\sim N(0, \boldsymbol{\Sigma})
\end{aligned}
$$

$$
Y = \begin{cases}
1 & \text{if} \quad \max(Y_j^*) = Y_1^* \quad j = 1, 2, \ldots, K \\
2 & \text{if} \quad \max(Y_j^*) = Y_2^* \quad j = 1, 2, \ldots, K \\
\vdots & \\
K & \text{otherwise}
\end{cases}
$$

Probit 的不可识别性！

# IIA test

- Hausman Test
- Small and Hsiao Test

## Hausman test for IIA

1. 使用 mlogit 估计基本模型，得到回归系数 $\hat{\beta}_F^*$(F=Full)。
2. 去掉一个类别（或数个类别）后，重新使用 mlogit 估计基本模型，得到回归系数 $\hat{\beta}_R$(R=Restricted)。
3. $H = \left(\hat{\beta}_R - \hat{\beta}_F^*\right)' \left[\text{var}(\hat{\beta}_R) - \text{var}(\hat{\beta}_F^*)\right]^{-1} \left(\hat{\beta}_R - \hat{\beta}_F^*\right)$
4. $H$ 是一个 $\chi^2$ 分布，自由度是 $\hat{\beta}_R$ 的个数（自变量数 $+1$）。
5. 原假设 $H_0$：IIA 成立。
6. 所以如果 $H$ 显著，则 IIA 不成立。
7. 【原理】：既然各类互相独立，那么他们之间的系数和方差差异应该不显著，应该是一样的，反之则各类不独立。

# Small and Hsiao test for IIA

1. 将数据随机分为两半，使用 mlogit 估计基本模型，得到二组回归系数，合并为（加权平均）$\hat{\beta}_u^{S_1 S_2}$(u=unrestricted)：

$$\hat{\beta}_u^{S_1 S_2} = \left(\frac{1}{\sqrt{2}}\right)\hat{\beta}_u^{S_1} + \left[1 - \left(\frac{1}{\sqrt{2}}\right)\right]\hat{\beta}_u^{S_2}$$

2. 使用其中一半数据，去掉一个类别后，重新使用 mlogit 估计基本模型，得到回归系数 $\hat{\beta}_R^{S_2}$，以及 $L(\hat{\beta}_R^{S_2})$
3. 把 $\hat{\beta}_u^{S_1 S_2}$ 带入第二步骤的数据 (当成初始值)，估计 $L(\hat{\beta}_u^{S_1 S_2})$
4. $SH = -2\left[L(\hat{\beta}_u^{S_1 S_2}) - L(\hat{\beta}_R^{S_2})\right]$
5. $SH$ 是一个 $\chi^2$ 分布，自由度是 $K+1$ 自变量个数 $+1$（截距)。
6. 原假设 $H_0$：IIA 成立。
7. 所以如果 $SH$ 显著，则 IIA 不成立。

# IIA test 小结

- 通常检验结果常会不一致。
- 所以建议以理论为指导前提。

# 那么该使用 mprobit 吗?

- mlogit 的 IIA 检验不通过，代表各类之间并非完全独立，所以原则上使用没有 IIA 假设的 mprobit 比较合适。
- 但是由于 mprobit 的余数方差矩阵无法识别，连带着各类之间的预测概率也无法容易的使用最大似然法进行预测，因此 mprobit 有其问题：
  - 不可识别的参数
  - MCMC 模拟发现，即便在 IIA 严重违反的情况下，mlogit 预测的系数仍然多数情况下比 mprobit 准确。
  - 当变量增加时，计算量以等比速度增加，mprobit 无法收敛的问题更为严重 (`STATA> dispaly e(converged)`)。
- go Bayesian!