

## 第 13 章 排序与计数模型

### 13.1 排 序 模 型

有些离散数据有天然的排序。比如, 公司债券的评级(AAA, AA, A, B, C 级); 对“春节联欢晚会”的满意度(很满意、满意、不满意、很不满意);

Li and Zhou(2005)研究经济增长绩效对地方官员仕途的影响, 以 0 表示“卸任”, 1 表示“留任或平级调动”, 2 表示“提拔”。

这种数据称为“排序数据”(ordered data)。

如使用 multinomial logit, 将无视数据内在的排序, 而 OLS 又把排序视为基数来处理。

可用潜变量法推导 MLE 估计量。

假设  $y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$  ( $y^*$  不可观测), 而选择规则为

$$y = \begin{cases} 0, & \text{若 } y^* \leq r_0 \\ 1, & \text{若 } r_0 < y^* \leq r_1 \\ 2, & \text{若 } r_1 < y^* \leq r_2 \\ \dots\dots\dots \\ J, & \text{若 } r_{J-1} \leq y^* \end{cases}$$

$r_0 < r_1 < r_2 < \dots < r_{J-1}$  为待估参数, 称为“切点” (cutoff points)。

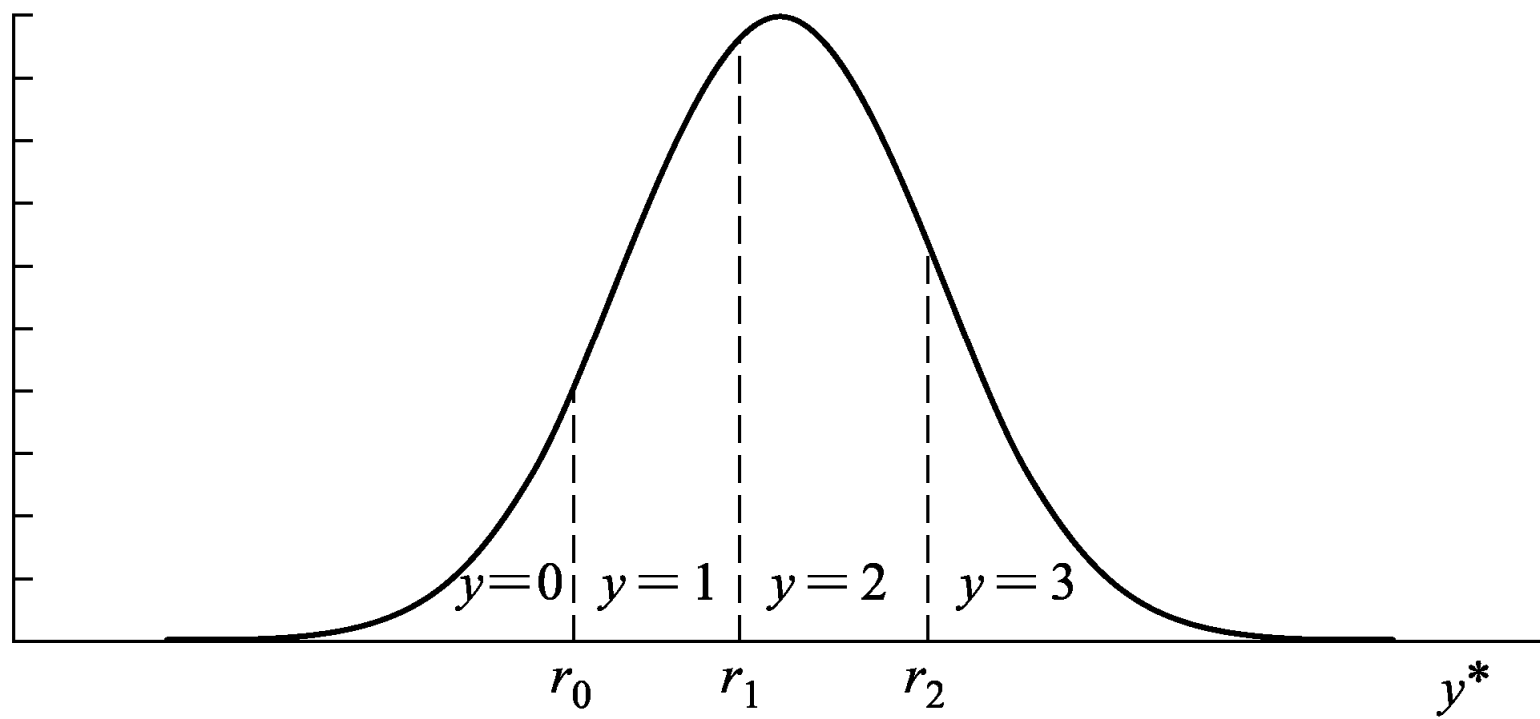


图 13.1 ordered logit 示意图

假设  $\varepsilon \sim N(0, 1)$  (将扰动项  $\varepsilon$  的方差标准化为 1), 则

$$\begin{aligned} P(y = 0 \mid \mathbf{x}) &= P(y^* \leq r_0 \mid \mathbf{x}) = P(\mathbf{x}'\boldsymbol{\beta} + \varepsilon \leq r_0 \mid \mathbf{x}) \\ &= P(\varepsilon \leq r_0 - \mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}) = \Phi(r_0 - \mathbf{x}'\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} P(y = 1 \mid \mathbf{x}) &= P(r_0 < y^* \leq r_1 \mid \mathbf{x}) \\ &= P(y^* \leq r_1 \mid \mathbf{x}) - P(y^* < r_0 \mid \mathbf{x}) \\ &= P(\mathbf{x}'\boldsymbol{\beta} + \varepsilon \leq r_1 \mid \mathbf{x}) - \Phi(r_0 - \mathbf{x}'\boldsymbol{\beta}) \\ &= P(\varepsilon \leq r_1 - \mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}) - \Phi(r_0 - \mathbf{x}'\boldsymbol{\beta}) \\ &= \Phi(r_1 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(r_0 - \mathbf{x}'\boldsymbol{\beta}) \end{aligned}$$

$$P(y = 2 | \mathbf{x}) = \Phi(r_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(r_1 - \mathbf{x}'\boldsymbol{\beta})$$

.....

$$P(y = J | \mathbf{x}) = 1 - \Phi(r_{J-1} - \mathbf{x}'\boldsymbol{\beta})$$

写出样本似然函数，可得 MLE 估计量，即 ordered probit 模型。

如果假设扰动项服从逻辑分布，则为 ordered logit 模型。

## 13.2 泊 松 回 归

有些被解释变量只能取非负整数，即  $0, 1, 2, \dots$ 。

比如，专利个数、奥运金牌个数、子女人数、看病次数。

对于这一类计数数据 (count data), 常使用“泊松回归” (Poisson regression)。

对于个体  $i$ , 记被解释变量为  $Y_i$ , 假设  $Y_i = y_i$  的概率由参数为  $\lambda_i$  的泊松分布决定:

$$P(Y_i = y_i | \mathbf{x}_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad (y_i = 0, 1, 2, \dots)$$

$\lambda_i > 0$  为“泊松到达率”, 表示事件发生的平均次数, 由解释变量  $\mathbf{x}_i$  所决定。

泊松分布的期望与方差都等于泊松到达率, 即  $E(Y_i | \mathbf{x}_i) = \text{Var}(Y_i | \mathbf{x}_i) = \lambda_i$ 。

为保证 $\lambda_i$ 非负，假设 $Y_i$ 的“条件期望函数”为

$$E(Y_i | \mathbf{x}_i) = \lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$$

假定样本 iid，则似然函数为

$$L(\boldsymbol{\beta}) = \frac{\exp(-\sum_{i=1}^n \lambda_i) \cdot \prod_{i=1}^n \lambda_i^{y_i}}{\prod_{i=1}^n y_i!}$$

对数似然函数为

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^n [-\lambda_i + y_i \ln \lambda_i - \ln(y_i!)] \\ &= \sum_{i=1}^n [-\exp(\mathbf{x}_i' \boldsymbol{\beta}) + y_i \mathbf{x}_i' \boldsymbol{\beta} - \ln(y_i!)] \end{aligned}$$

最大化的一阶条件为

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

根据 MLE 理论，如果似然函数正确，则  $\hat{\boldsymbol{\beta}}_{\text{MLE}}$  为一致估计量。

即使似然函数不正确，由于泊松分布属于线性指数分布族，故只要条件期望函数  $\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$  正确，则  $\hat{\boldsymbol{\beta}}_{\text{QMLE}}$  就是一致的。

如果似然函数不正确，则常规的标准误(比如，OIM 或 BHHH 法)不是真实标准误的一致估计量，必须使用在 QMLE 基础上计算的稳健标准误，它对于似然函数是否正确比较稳健。



$\hat{\beta}_{\text{MLE}}$  不表示边际效应。由于  $\ln \lambda_i = \mathbf{x}_i' \boldsymbol{\beta}$ , 故  $\frac{\partial \ln \lambda_i}{\partial x_k} = \beta_k$ 。

可将  $\beta_k$  解释为“半弹性” (semi-elasticity), 即当解释变量  $x_k$  增加微小量时, 事件的平均发生次数将增加多少百分点。

由于泊松到达率  $\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ , 故也可计算  $\exp(\beta_k)$ , 称为“发生率比” (Incidence Rate Ratio, 简记 IRR)

IRR 表示, 当  $x_k$  增加一单位时(从  $x_k$  增加到  $x_k + 1$ ), 事件的平均发生次数将是原来的多少倍, 因为

$$\exp[(x_k + 1)\beta_k] / \exp(x_k \beta_k) = \exp(\beta_k)$$

泊松分布描述，在给定时间内某观测单位的事件发生次数。

如果观测时间变长，或观测单位的空间规模变大，则事件发生的平均次数也应同比例增多。

记个体  $i$  在单位时间内事件发生的平均次数为  $\phi_i$ 。

如果不同个体的时间或空间规模不同，记为  $T_i$ ，称为“暴露期”(exposure)，则该事件发生的平均次数也须相应调整为  $\phi_i T_i$ 。

比如，考察某疾病在不同城市的发病人数，而各城市的人口基数不同：

$$P(Y_i = y_i \mid \mathbf{x}_i, T_i) = \frac{e^{-\phi_i T_i} (\phi_i T_i)^{y_i}}{y_i!} \quad (y_i = 0, 1, 2, \dots)$$

假设  $\phi_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ 。在上式中，令  $\lambda_i \equiv \phi_i T_i$ ，则  $\ln \lambda_i = \mathbf{x}_i' \boldsymbol{\beta} + \ln T_i$ 。

如果暴露期  $T_i$  随  $i$  而变，应把  $\ln T_i$  作为解释变量放入泊松回归，并且令其系数为 1。

### 13.3 负二项回归

泊松分布的期望与方差一定相等，称为“均等分散”(equidispersion)；此特征常与实际数据不符。

如果被解释变量的方差明显大于期望，即存在“过度分散”(overdispersion)。

在条件期望函数的对数表达式中加入一项：

$$\ln \lambda_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

$\varepsilon_i$ 表示不可观测部分或个体的异质性。可得：

$$\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}) \cdot \exp(\varepsilon_i) \equiv u_i v_i$$

$u_i \equiv \exp(\mathbf{x}_i' \boldsymbol{\beta})$ 为 $\mathbf{x}_i$ 的确定性函数，而 $v_i \equiv \exp(\varepsilon_i) > 0$ 为随机变量。

给定 $\mathbf{x}_i$ 与 $v_i$ ，则 $y_i$ 依然服从泊松分布：

$$P(Y_i = y_i \mid \mathbf{x}_i, v_i) = \frac{e^{-u_i v_i} (u_i v_i)^{y_i}}{y_i!} \quad (y_i = 0, 1, 2, \dots)$$

由于  $v_i$  不可观测，无法对此方程进行估计。

记  $v_i$  的概率密度函数为  $g(v_i)$ ，可将  $v_i$  积分掉，计算  $y_i$  的边缘密度：

$$P(Y_i = y_i | \mathbf{x}_i) = \int_0^{\infty} \frac{e^{-u_i v_i} (u_i v_i)^{y_i}}{y_i!} g(v_i) dv_i$$

由于  $v_i > 0$ ，通常选择  $v_i$  服从 Gamma 分布(指数分布与卡方分布为 Gamma 分布的特例)。

假设  $v_i \sim \text{Gamma}(1/\alpha, \alpha)$ ，其中  $\alpha > 0$ 。

对于  $\text{Gamma}(a, b)$ ，期望为  $ab$ ，方差为  $ab^2$ 。故  $E(v_i) = 1$ ，而  $\text{Var}(v_i) = \alpha$ 。

代入  $\text{Gamma}(1/\alpha, \alpha)$  的概率密度, 可得负二项分布的概率密度, 进行 MLE 估计, 称为“负二项回归”(negative binomial regression)。

有关负二项分布。

假设某事件在一次实验中成功的概率为  $\theta$  ( $0 < \theta < 1$ )。

记  $Y$  为在第  $J$  次成功前失败的总次数, 则  $Y$  的分布律为

$$P(Y = y | \theta, J) = C_{y+J-1}^{J-1} \theta^J (1-\theta)^y \quad (y = 0, 1, 2, \dots)$$

由于第  $(y+J)$  次一定为成功, 故只要在前面的  $(y+J-1)$  次中找出成功的  $(J-1)$  次的组合次数即可。

负二项回归模型的条件期望仍为 $E(Y_i | \mathbf{x}_i) = u_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ ，而条件方差为

$$\text{Var}(Y_i | \mathbf{x}_i) = u_i + \alpha u_i^2 > u_i = E(Y_i | \mathbf{x}_i)$$

条件方差的表达式包含条件期望 $u_i$ 的平方项，称为“NB2 模型”。

当 $\alpha \rightarrow 0$ 时，泊松回归变为负二项回归的特例。

进行负二项回归后，只要对原假设“ $H_0: \alpha = 0$ ”进行检验，即可确定应使用负二项回归还是泊松回归。

如果将 $\text{Gamma}(1/\alpha, \alpha)$ 中的 $\alpha$ 换为 $\delta/u_i$ ，即假设 $v_i \sim \text{Gamma}(u_i/\delta, \delta/u_i)$ ，其中 $\delta > 0$ ，则条件方差函数变为：

$$\text{Var}(Y_i | \mathbf{x}_i) = u_i + (\delta/u_i)u_i^2 = u_i + \delta u_i > u_i$$

条件方差为条件期望 $u_i$ 的一次函数，称为“NB1 模型”，是负二项回归的另一形式。

如果 $\delta \rightarrow 0$ ，则回到泊松回归的特例。Stata 会汇报对原假设“delta=0”的检验结果。

实践中，常使用 NB2 模型。

多数情况下，NB2 模型更符合数据特点。

使用 NB2 模型的另一好处是，即使似然函数不正确，只要条件期望函数正确，则 $\hat{\beta}_{\text{QMLE}}$ 依然是一致估计，NB1 模型则无此优点。



在 NB2 负二项回归中，条件方差函数主要由参数 $\alpha$ 来刻画。

作为推广，可让此参数依个体  $i$  而变，记为 $\alpha_i$ ，并让 $\ln \alpha_i$ 依赖于变量 $\mathbf{z}_i$ ( $\mathbf{z}_i$ 可与 $\mathbf{x}_i$ 重叠)。

然后使用 MLE 对条件均值方程与条件方差方程同时进行估计，称为“广义负二项回归”(generalized negative binomial regression)。

究竟何时使用泊松回归或负二项回归？

即使数据中存在过度分散，“泊松回归+稳健标准误”依然提供了对参数及标准误的一致估计。

另一方面，如果比较了解条件方差函数，则“负二项回归+稳健标准误”可提供更有效率的估计。

如果研究者只关心参数 $\beta$ 的估计值，或许泊松回归就足够了；但如果希望预测“ $Y = y$ ”的发生概率，则可考虑负二项回归。

另外，对“ $H_0: \alpha = 0$ ”的 LR 检验结果也提供了在泊松回归与负二项回归之间选择的参考依据。

## 13.4 零膨胀泊松回归与负二项回归

如果计数数据中含有大量的“0”值，可考虑使用“零膨胀泊松回归”(Zero-inflated Poisson Regression, 简记 ZIP)或“零膨胀负二项回归”(Zero-inflated Negative Binomial Regression, 简记 ZINB)。

决策可能分两阶段进行。

首先，决定“取零”(无)或“取正整数”(有)，相当于二值选择。

其次，如果决定“取正整数”，进一步确定具体选择哪个正整数。

假定被解释变量  $y_i$  服从以下“混合分布”(mixed distribution):

$$\begin{cases} P(y_i = 0 | \mathbf{x}_i) = \theta \\ P(y_i = j | \mathbf{x}_i) = \frac{(1 - \theta)e^{-\lambda_i} \lambda_i^j}{j!(1 - e^{-\lambda_i})} \quad (j = 1, 2, \dots) \end{cases}$$

$\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ , 而  $\theta > 0$  与  $\boldsymbol{\beta}$  为待估参数。

可以证明,  $\sum_{j=0}^{\infty} P(y_i = j | \mathbf{x}_i) = 1$ 。

也可让  $\theta$  依赖于解释变量  $\mathbf{z}_i$  ( $\mathbf{z}_i$  可与  $\mathbf{x}_i$  重叠)。

使用 MLE 估计以上模型, 即得到“零膨胀泊松回归”。

类似地, 可以定义“零膨胀负二项回归”。

究竟应该使用标准的泊松回归(standard Poisson)还是零膨胀泊松回归(ZIP)?

Stata 提供了一个“Vuong 统计量”(Vuong, 1989), 其渐近分布为标准正态。

如果 Vuong 统计量很大(为正数), 则应选择零膨胀泊松回归(或零膨胀负二项回归)。

反之, 如果 Vuong 统计量很小(为负数), 则应选择标准的泊松回归(或“标准的负二项回归”)。