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

第七讲　计数因变量分析

计数变量

- 人一生中犯罪的次数。
- 家庭中小孩的个数。
- 去年新建商场个数。
- 过去一周，叫外卖的次数。
- 上个月广州发生上访事件次数。
- 每年国家间发生战争的次数。
- 各个国家 2020 年 COVID-2019 的确诊案例数。

计数变量

- 由时间区间内事件的发生次数构成的变量
- 非负数的整数变量；离散的；分布非正态；偏态严重
- 很多时候计数变量的观测值为零

建模考量

- 可以使用 OLS 回归分析计数变量，但：
  - OLS 回归后预测值会出现负数，但是计数变量是非负数的。
  - 计数变量是偏态严重的变量，违反了 OLS 关于正态分布的假设。
- 因此，建议使用 Poisson Regression Model（泊松回归）或是 Negative Binomial Regression Model(负二项回归)。
- 进阶模型：零膨胀泊松回归 (zero inflated poisson) 或者零膨胀负二项回归 (zero inflated negative binomial Regression)。

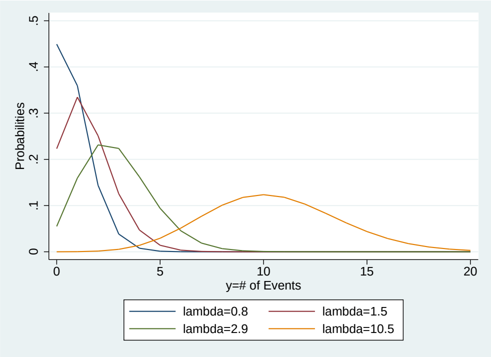
泊松回归

- $$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \text{ for } y = 0, 1, 2, \ldots$$
  $$\Rightarrow y_i \sim \text{Poisson}(\lambda_i) \qquad i = 1, 2, \ldots, n$$
- λ：次数的期望值 (均值)，同时也是方差（这是泊松分布重要的前提假设）。
- y：是次数的观测值。
- λ：y 可一看做与自变量的关系：

$$\log(\lambda_i) = \sum \mathbf{X}_i\boldsymbol{\beta}$$
$$\Rightarrow \lambda_i = \exp\left(\sum \mathbf{X}_i\boldsymbol{\beta}\right)$$

- 很像 logit 回归，log 的使用可以避免负数出现。

泊松分布



- 当 λ 增加，分布的峰态就往右偏去。
- 当 λ 增加，0 的次数就减少。
- 当 λ 增加，会趋近于正态分布，例如当 $\lambda = 10.5$。

偏移 (offset)

- 如果比较不同单元间的事件估计率（均值 λ），有时候必须除以 N 标准化，才能比较不同的 y，因为 N 越大，发生的次数就可能越多。

$$
\begin{aligned}
\log\left(\frac{\lambda}{N}\right) &= \mathbf{X}\beta \\
\Rightarrow \log(\lambda) &= \log(N) + \mathbf{X}\beta \\
\Rightarrow \lambda &= N \times \exp(\mathbf{X}\beta) \\
\Rightarrow \lambda &= \exp(\mathbf{X}\beta + \log(N))
\end{aligned}
$$

- N 在这称作偏移 (offset)，$\log(N)$ 称作偏移量，当所有协变量都无法解释 y 时，$\lambda = N$。

曝险 (exposure)

- 如果比较不同单元间的事件估计率（均值 λ），有时候还必须考虑曝险 (exposure)，才能比较不同的 y，例如时间越长，发生次数的可能就越多。
-

$$\lambda \times t = \exp(\mathbf{X}\beta) \times t$$
$$\lambda \times t = \exp(\mathbf{X}\beta + \log(t))$$

- t 是每个单元的曝险时长，年龄、时间从开始到结束的观测时长都可作为曝险变量。
- exposure 与 offset 的比较：exposure 是以 log 形式出现在回归方程式右侧，如果把时间变量当成 offset，则必须以 log 方式提供。

泊松分布

```
. poisson art fem mar phd kid5 ment

Iteration 0:   log likelihood = -1651.4574
...
Iteration 3:   log likelihood = -1651.0563

Poisson regression                                Number of obs   =        915
                                                  LR chi2(5)      =     183.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -1651.0563                       Pseudo R2       =     0.0525

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2245942   .0546138    -4.11   0.000    -.3316352   -.1175532
         mar |   .1552434   .0613747     2.53   0.011     .0349512    .2755356
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        kid5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3739677   .1665859     2.24   0.025     .0474654    .7004699
------------------------------------------------------------------------------
```

- 系数检验：Coef/Std.Err.; Likelihood ratio test; Pseudo R^2

系数解读

- 泊松回归中，y 通常被理解为事件发生率、频次，因此 β 为正，表示估计率 $\hat{y}$ 越大，负则越小。
- 泊松回归与自变量也是非线性关系，所以系数无法直接解读。
- 必须透过事件发生率比 (incidence rate ratios, irr) 来解读，类似胜算比 (odds ratios)

泊松回归，事件发生率比

```
. poisson art fem mar phd kid5 ment, irr

Poisson regression                                Number of obs   =        915
                                                  LR chi2(5)      =     183.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -1651.0563                       Pseudo R2       =     0.0525

------------------------------------------------------------------------------
         art |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |   .7988403   .0436277    -4.11   0.000     .7177491    .8890932
         mar |   1.167942   .0716821     2.53   0.011     1.035569    1.317236
         phd |   1.012905   .0267379     0.49   0.627     .9618325     1.06669
        kid5 |   .8312018   .0333538    -4.61   0.000     .7683342    .8992134
        ment |   1.025872    .002058    12.73   0.000     1.021846    1.029913
       _cons |    1.45349   .2421309     2.24   0.025      1.04861    2.014699
------------------------------------------------------------------------------
```

- β of mar: 1.17，结婚的比没结婚的发 paper 的发生率的 1.17 倍，结婚的比没结婚的发 paper 多 17%。
- β of fem: 0.80，女生比男生发 paper 的的发生率 0.8 倍，女生比男生发 paper 少 20%。

泊松回归解读

- 最好的解读方式还是透过估计值或边际次数估计值来处理。
- STATA 命令：`prcount`，`prvalue`，`margins`，`mfx`
- R：作图和使用 fake data simulation

泊松回归的基本假设

- 最大的假设，均值等于方差（等离散假设, equi-dispersion）
- 现实数据中，几乎不可能。
- 通常方差会大于均值，也就是过度离散 (overdispersion)
- 如果忽略此点，poisson 回归求得的系数标准误是低估的，但对于预测次数值不会有太大的影响。

泊松回归的基本假设

- 过度离散可以从变量的偏态观察得之。
- 变量如果有太多的 0，则很有可能会造成过度离散。
- 例如每个月机动车违章记录，大部分人是 0，但有些疯狂的人会有 50，所以均值会很小，但标准差会很大。

泊松回归的基本假设

- 如果 violation 等离散假设是轻度的，可以考虑使用三明治法求得 Robust Standard Error：

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

  Omega 是重新构建一个余数协方差矩阵，其中考虑到离群值的影响。

- 或者使用 Bootstrap!

负二项回归

- 解决过度离散的另一个选择是使用负二项回归。
- 在泊松分布中加入余数项：

$$\lambda = \exp(\mathbf{X}\beta + \epsilon)$$

- 但必须增加其他假设：
  - $\exp(\epsilon)$ 的期望值为 1;
  - $\exp(\epsilon)$ 为 Gamma 分布
  - 以上两个假设比泊松分布的等离散假设合理，但也有可能不符合现实。

负二项回归数学表达式

- $$\mathrm{Pr}(y|\lambda,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}}\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y}$$
  $$= \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})}\frac{(\alpha^{-1})^{\alpha^{-1}}\times\lambda^{y}}{(\alpha^{-1}+\lambda)^{\alpha^{-1}+y}}$$
  $$\Rightarrow \mathrm{Pr}(y|\lambda,\theta) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\frac{\theta^{\theta}\times\lambda^{y}}{(\theta+\lambda)^{\theta+y}} \qquad \text{Let } \theta = \alpha^{-1}$$
  $$y_i \sim \text{NegBinomial}(\lambda_i,\theta), \text{for } i = 1,2,\ldots,N$$
- 其中，α 决定了数据的离散度，越大则离散度越大。
- 当 $\alpha = 0$，上面公式则变为泊松分布概率函数。

负二项回归 α 检验

- α 似然比检验：$H_0 : \alpha = 0$。
- 如果检验统计量 (G^2) 显著，则拒绝 H_0，数据离散度不等于 0，负二项回归合适。
- 如果检验统计量 (G^2) 不显著，则拒绝 H_0，数据离散度等于 0，泊松回归合适。
- $G^2 = 2(\log L_{\text{NB}} - \log L_{\text{Poisson}})$
- 自由度 1（多了个 α 这个参数）。

计数中有过多的 0

- 计数变量取值有过多的 0，而这些 0 跟其他的计数反映了完全不同的情况
- 泊松或者负二项回归均不能很好的解释过多的 0.
- 所以可以使用 Hurdle 栅栏回归、ZIP (零膨胀泊松回归) 或者 ZINB(零膨胀负二项回归)

栅栏 (Hurdle) 回归

- 计数数据中零过多的情况。
- Gary King （1989）使用 Hurdle model 估计国际关系中的冲突发生的次数。
- Hurdle 模型将事件的发生看作两个不同的数据发生过程（零事件 vs 非零事件）。
  1. 第一个过程决定零事件的发生过程，令其服从二项分布（Probit or logit）
  2. 第二个过程决定当跨越了栅栏 (Hurdle) 进入到第二个事件发生过程，令其服从计数分布（Poisson or Negative Binomial），但是这个事件发生的取值 > 0，由于该事件发生是基于第一个过程的基础上，因此是截断的，截点为 1，称之为 Zero truncated Poisson or Negative Binomial)

栅栏 (Hurdle) 回归

- 第一个事件发生过程的概率函数为：

$$\text{Pr}(Y=y)=\begin{cases}\pi, & y=0\\ 1-\pi, & y=1,2,3,\ldots\end{cases}$$

- 第二个事件发生过程的概率函数为：

$$\text{Pr}(Y=y|Y\neq 0)=\begin{cases}0, & \text{otherwise}\\ \frac{\lambda^y}{(e^\lambda-1)y!}, & y=1,2,3,\ldots\end{cases}$$

- 改写上述两个概率函数为无条件概率函数：

$$\text{Pr}(Y=y)=\begin{cases}\pi, & y=0\\ (1-\pi)\times\frac{\lambda^y}{(e^\lambda-1)y!}, & y=1,2,3,\ldots\end{cases}$$

零膨胀计数回归原理

- 使用两个回归模型：
  1. 第一个回归模型建模预测样本是否永远取值为 0(组 A)，或者并不永远取值为 0(组 B)；
  2. 第二个模型建模预测组 B 的计数（包含 0）。
- 第一个回归中，y 重新编码为二元变量，0 为 1，其他为 0，所以回归系数为正，代表更可能取值为 0：$\Pr(Y = 0) = f(\boldsymbol{X})$
- 第二个回归则是复合模型 (Mixture model)，结合 logit 与 poisson (or negative binomial)

零膨胀计数回归数学原理

- 第一个事件发生过程的概率函数为：

$$\Pr(Y=y)=\begin{cases}\pi, & y=0\\ 1-\pi, & y=1,2,3,\ldots\end{cases}$$

- 第二个事件发生过程的概率函数为：

$$\Pr(Y=y|Y\neq 0)=\begin{cases}1-\pi, & \text{otherwise}\\ \frac{\lambda^y}{(e^\lambda-1)y!}, & y=1,2,3,\ldots\end{cases}$$

- 改写上述两个概率函数为复合概率函数：

$$\Pr(Y=y)=\begin{cases}\pi+(1-\pi)\exp(-\lambda), & y=0\\ (1-\pi)\times\frac{\lambda^y}{(e^\lambda-1)y!}, & y=1,2,3,\ldots\end{cases}$$

如何从这六个模型选择

- 数据过于离散就采用负二项回归
- 数据等离散就采用泊松回归
- 零过多时，考虑栅栏回归、零膨胀计数回归

ZIP/ZINB vs ZAP/ZANB

- ZAP/ZANB 是零膨胀计数回归的一种，Zero Altered Poisson/ Negative Binomial。
- 与 ZIP/ZINB 的差别在于如何处理 0 与非 0 之间的关系。
- 举例：我决定是否买苹果，决定买后，买几个（正整数），在这个情况下，就是 ZAP/ZANB；如果我决定买后，我可以买 $0-\infty$，在这个情况下，就是 ZIP/ZINB，这里之所以会买 0 个，可能是我就是可以买 0 个，或者是市场缺货。
- VGAM 包里的 `vglm()` 可以实现。