ZERO-INFLATED POISSON REGRESSION MODELS WITH RIGHT CENSORED COUNT DATA

Xingxing Du and Yahui Wei

Southwest University

February 6, 2021

目录

- Censoring:
  - whether the lower or upper bound had been passed and the value of the bound are available.
- Truncation:
  - a sample with all values outside the bounds entirely omitted, not even the count to those omitted were kept.
- Zero-Inflated Poisson(ZIP) models with censored:
  - count data are censored from above (right) or below (left) a specific point or a combination of them.
  - the number of zero counts are much greater than expected for the Poisson distribution.

- There is a large body of literature on zero-inflated Poisson models:LR,Wald,Score test
- Hua He et al.(2019) develop a new approach for testing inflated zero under the Poisson model.
- Yi Tang et al.(2019) compare the new test with the Wald,score,and likelihood ratio tests.
- Yuhan Zou et al.(2020)applied the new approach for latent class in censored data due to detection limit.

- Poisson Model:

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

- Poisson Regression:

$$y_i|x_i \sim i.d.Poisson(\mu_i), \;\; log(\mu_i) = x_i^T\beta$$

- ZIP Model:

$$P(Y_i = y_i) = \begin{cases} \omega + (1-\omega)e^{-\mu_i} & y_i = 0 \\ (1-\omega)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & y_i > 0 \end{cases}$$

- ZIP Regression:

$$y_i|x_i \sim i.d.ZIP(\omega_i, \mu_i), \;\; logit(\omega_i) = \mu_i^T\beta_\omega, \; log(\mu_i) = x_i^T\beta$$

- Right Censoring:

$$Y_i = min\{y_i, L\} = \begin{cases} L & , y_i \geq L. \\ y_i & , otherwise. \end{cases}$$

$$P(Y_i = y_i) = \begin{cases} P(Y_i \geq L) & , y_i \geq L. \\ P(Y_i = y_i) & , otherwise. \end{cases}$$

- Right-Censored Poisson Model:

$$P(Y_i = y_i) = \begin{cases} \sum_{j=y_i}^{\infty} P(y_i = j) & , y_i \geq L. \\ (e^{-\mu_i})^{r_i} * \left(\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right)^{1-r_i} & , otherwise. \end{cases}$$

- Censored Poisson Regression:

$$y_i|x_i \sim CensorePoisson(\mu_i, L), \;\; log(\mu_i) = x_i^T\beta$$

- Right Censoring:

$$Y_i = min\left\{y_i, L\right\} = \begin{cases} L & , y_i \geq L. \\ y_i & , otherwise. \end{cases}$$

$$P(Y_i = y_i) = \begin{cases} P(Y_i \geq L) & , y_i \geq L. \\ P(Y_i = y_i) & , otherwise. \end{cases}$$

- Censored ZIP Model:

$$P(Y_i = y_i) = \begin{cases} \sum_{j=y_i}^{\infty} P(y_i = j) & , Y_i \geq y_i. \\ \left(\omega + (1-\omega)e^{-\mu_i}\right)^{r_i} \left((1-\omega)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right)^{1-r_i} & , otherwise. \end{cases}$$

- Censored ZIP Regression:

$$y_i|x_i \sim CZIP(\omega_i, \mu_i, L),\;\; logit(\omega_i) = \mu_i^T \beta_\omega,\;\; log(\mu_i) = x_i^T \beta_\mu$$

- H_0 : $Y_i|x_i$ follows the censored Poisson model
- H_1 : $Y_i|x_i$ follows the censored ZIP model

$$H_0 : \omega = 0 \quad vs. \quad H_1 : \omega > 0$$

- define an indicator variable of zero occurs:

$$r_i = I_{(Y_i=0)} = \begin{cases} 1, & Y_i = 0 \\ 0, & otherwise \end{cases}$$

- define an indicator variable of censoring occurs:

$$d_i = I_{(Y_i \geq y_i)} = \begin{cases} 1, & Y_i \geq y_i \\ 0, & otherwise \end{cases}$$

- for $Y_i \geq y_i$:

$$Pr(Y_i \geq y_i) = \sum_{j=y_i}^{\infty} P(y_i = j) = 1 - \sum_{j=0}^{y_i-1} P(y_i = j)$$

- Likelihood function of censored Poisson:

$$L_1 = \prod_{i=1}^{n} \left\{ \left[\left(e^{-\mu_i}\right)^{r_i} \left(\frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right)^{1-r_i} \right]^{(1-d_i)} \left[Pr(Y_i \geq y_i)\right]^{d_i} \right\}$$

- Log-likelihood Function:

$$l_1 = log(L_1) = \sum_{i=1}^{n} \left\{ (1-d_i)log\left[-r_i\mu + (1-r_i)log\left(\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \right) \right] + d_i log\left[\sum_{j=y_i}^{\infty} P(y_i = j) \right] \right\}$$

- Likelihood function of censored zero-inflated Poisson model:

$$L_2 = \prod_{i=1}^{n} \left\{ [Pr(Y_i \geq y_i)]^{d_i} \left[\left(\omega + (1-\omega)e^{-\mu_i}\right)^{r_i} \left((1-\omega)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right)^{1-r_i} \right]^{1-d_i} \right\}$$

- Log-likelihood Function:

$$l_2 = log(L_2) = \sum_{i=1}^{n} (1-d_i) \left[r_i \left[log\omega + (1-\omega)e^{-\mu_i} \right] + (1-r_i) \left[log(1-\omega) + y_i log(\mu_i) - log(y_i!) - \mu_i \right] \right] + d_i log \sum_{j=y_i}^{\infty} Pr(Y_i = j)$$

- statistics:

$$\frac{\hat{\omega}}{\hat{\sigma}} \sim N(0,1) \quad or \quad \frac{\hat{\omega}^2}{\hat{\sigma}^2} \sim \chi(1)$$

- method:
    - ω can be obtained by the maximum likelihood function.
    - the variance of the MLE of ω can further be estimated by the fisher information matrix.

- let.

$$c_i = \sum_{j=y_i}^{\infty} P(y_i = j), \quad \theta_i = \frac{\omega_i}{1-\omega_i}$$

- then

$$\frac{\partial c_i}{\partial \beta} = -\sum_{j=0}^{y_i-1} \frac{\partial Pr(Y_i = j)}{\partial \beta} = -\sum_{j=0}^{y_i-1} Pr(Y_i = j)(y_i - \mu_i)x_i$$

- solve the score equations:

$$\frac{\partial l_2}{\partial \beta} = \sum_{i=1}^{n} \left\{ (1-d_i) \left[r_i \frac{-\theta_i^{-1} e^{-\mu_i}}{1+\theta_i e^{-\mu_i}} x_i \mu_i + (1-r_i)(y_i - \mu_i)x_i \right] + \frac{d_i}{c_i} \frac{\partial c_i}{\partial \beta} \right\}$$

$$\frac{\partial l_2}{\partial \omega} = \sum_{i=1}^{n} \left\{ (1-d_i) \left[r_i \frac{1-e^{-\mu_i}}{\theta_i e^{-\mu_i}} - (1-r_i) \right] \frac{\theta_i}{1+\theta_i} \right\}$$

- statistics:

$$S_{LR} = 2\left[l_1(\hat{\omega}, \hat{\beta}) - l_2(0, \hat{\beta})\right] \sim \chi^2$$

- statistics:

$$S_{Score} \sim N(0,1)$$

- let

$$\theta_i = \frac{\omega_i}{1-\omega_i}$$

- then

$$l_2 = \sum_{i=1}^{n} (1-d_i)\left[-\log(\theta_i+1) + r_i log\theta_i + e^{-\mu_i} - (1-r_i)\left[-\mu_i + y_i log(\mu_i) - log(y_i!)\right]\right] + d_i log \sum_{j=y_i}^{\infty} Pr(Y_i = j)$$

- and

$$\frac{\partial l_2}{\partial \beta} = \sum_{i=1}^{n} \left\{ (1-d_i)\left[r_i \frac{-\theta_i^{-1} e^{-\mu_i}}{1+\theta_i e^{-\mu_i}} x_i^T \mu_i + (1-r_i)(y_i-\mu_i) x_i^T \right] + \frac{d_i}{c_i}\frac{\partial c_i}{\partial \beta} \right\}$$

$$\frac{\partial l_2}{\partial \theta_i} = \sum_{i=1}^{n} \left\{ (1-d_i)\left[-\frac{1}{1+\theta_i} + \frac{r_i}{\theta_i + e^{-\mu_i}} \right] \right\}$$

- when

$$\theta_i = 0$$

- then score equation are:

$$\frac{\partial l_2}{\partial \beta} = \sum_{i=1}^{n} \left\{ (1-d_i) \left[-r_i(\mu_i + (1-r_i)(y_i - \mu_i)) - \frac{d_i c_i}{1-c_i} \right] x_i^T \right\}$$

$$\frac{\partial l_2}{\partial \theta_i} = \sum_{i=1}^{n} \left\{ (1-d_i) \frac{r_i - e^{-\mu_i}}{e^{-\mu_i}} \right\}$$

- the variance of $\frac{\partial l_2}{\partial \theta_i}$ can be obtained from Fisher Information matrix

- statistics:

$$S_{Score} = \frac{\sum_{i=1}^{n}(1-c_i)\frac{r_i-p_i}{p_i}}{\left\{\sum_{i=1}^{n}\left[(1-c_i)\frac{1-p_i}{p_i} - (c_i\mu_i)^T X\left(diag(dU)\right)^{-1} X^T(c_i\mu_i)\right]\right\}^{\frac{1}{2}}}$$

  - $c_i = \sum_{j=0}^{y_i-1} P(Y_i = y_i)$
  - $p_i = e^{-\mu_i}$
  - $dU = X^T\left[\sum_{i=1}^{n}\frac{c_i}{1-c_i}(y_i - c_i)^2\right]X$

$$S_{Score} \sim N(0,1)$$

- statistics: $\sqrt{n}\,(\hat{s}-0) \rightarrow N(0,\tau^2)$
- τ^2 是矩阵$A^{-1}BA^{-T}$ 的$(1,1)$ 项
- $A(\gamma) = E\left[\frac{\partial}{\partial\gamma}\Psi_i(Y_i,\gamma)\right]$
- $B(\gamma) = Var(\Psi_i(Y_i,\gamma))$
- Estimation Equation(EE):

$$\psi_1 = \frac{1}{n}\sum_{i=1}^{n}(r_i - p_i - s)$$

$$\psi_2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-d_i-c_i}{1-c_i}(y_i-\mu_i)x_i^T\right]$$

- statistics:

$$S_{new} = \frac{\sum_{i=1}^{n}(1-c_i)(r_i-\hat{p}_i)}{\left\{\sum_{i=1}^{n}\left[(1-c_i)(1-\hat{p}_i)-(c_i\hat{p}_i\mu_i)^T X\left(diag(dU)\right)^{-1}X^T(c_i\hat{p}_i\mu_i)\right]\right\}^{\frac{1}{2}}}$$

$$S_{new} \sim N(0,1)$$

- 检验是否为零膨胀：

$$H_0 : \omega = 0 \quad vs \quad \omega > 0$$

- L=5
- 均值变动，考虑三种情形：
  - 无协变量
  - 协变量服从均匀分布
  - 协变量服从正态分布
- 样本量：50，100，200，500
- Monte Carlo(蒙特卡罗)样本量：100

模拟介绍

- 详情见R程序
  - R Code
  - Type Error I

- 检验是否为零膨胀：

$$H_0 : \omega = 0 \quad vs \quad \omega > 0$$

- L=5
- 均值变动，考虑三种情形：
  - 无协变量
  - 协变量服从均匀分布
  - 协变量服从正态分布
- 样本量：50，100，200，500
- Monte Carlo(蒙特卡罗)样本量：100

模拟介绍

- 详情见R程序
  - R Code
  - Power
  - Results