CENSORED ZERO-INFLATED MODEL

Xingxing Du and Yahui Wei

Southwest University

March 20, 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:

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

- Censored Poisson Regression:

$$Y_i|X_i \sim i.d.Poisson(\mu_i)$$

$$z_i = min\left\{Y_i, L\right\} = \begin{cases} y_i & , y_i < L \\ L & , y_i \geq L \end{cases} ;$$

$$Z_i|X_i \sim i.d.CensoredPoisson(\mu_i, L), log(\mu_i) = X_i^T\beta$$

- Censored Poisson Model:

$$P(Z_i = z_i) = p\left\{min(Y_i, L) = z_i\right\} = \begin{cases} Pr(Y_i = z_i) & , z_i < L \\ Pr(Y_i \geq L) & , z_i = L \\ 0 & , z_i > L \end{cases}$$

$$= \begin{cases} \frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!} & , z_i < L \\ \sum_{y_i=L}^{\infty} Pr(Y_i = y_i) & , z_i = L \\ 0 & , z_i > L \end{cases}$$

- define an indicator variable of censoring occurs:

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

- for $Y_i \geq y_i$,let:

$$c_i = Pr(Y_i \geq y_i) = \sum_{y_i=L}^{\infty} P(Y_i = y_i)$$

- Likelihood function of censored Poisson:

$$L_1 = \prod_{i=1}^{n}\left\{\left[\frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!}\right]^{(1-d_i)}\left[\sum_{y_i=L}^{\infty}P(Y_i=y_i)\right]^{d_i}\right\}$$

- Log-likelihood function of censored Poisson:

$$l_1 = log(L_1) = \sum_{i=1}^{n}\left\{(1-d_i)log\left[\frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!}\right]+d_i log(c_i)\right\}$$

- Censored ZIP Regression:

$$Y_i|X_i \sim i.d.ZIP(\omega_i, \mu_i)$$

$$Z_i = min\{Y_i, L\} = \begin{cases} y_i & , y_i < L \\ L & , y_i \geq L \end{cases}$$

$$Z_i|X_i \sim i.d.CensoredZIP(\omega_i, \mu_i, L), logit(\omega_i) = \mu_i^T \beta_w, log(\mu_i) = x$$

- Censored ZIP Model:

$$P(Z_i = z_i) = p\{min(Y_i, L) = z_i\} = \begin{cases} Pr(Y_i = z_i) & , z_i < L \\ Pr(Y_i \geq L) & , z_i = L \\ 0 & , z_i > L \end{cases}$$

$$= \begin{cases} [\omega + (1-\omega)e^{-\mu_i}]^{r_i} \left[(1-\omega)\frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!}\right]^{1-r_i} & , z_i < L \\ \sum_{y_i=L}^{\infty} P(Y_i = y_i) & z_i = L \\ 0 & , z_i > L \end{cases}$$

- define an indicator variable of zero occurs:

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

- Likelihood function of censored ZIP model:

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

- Log-likelihood function of censored ZIP model:

$$l_2 = log(L_2)$$

$$l_2 = \sum_{i=1}^{n} \left\{ (1-d_i) \left[r_i \bullet log(\omega + (1-\omega)e^{-\mu_i}) \right] + (1-r_i) log \left[(1-\omega) \left(\frac{e^{-\mu_i} \mu_i^{z_i}}{z_i!} \right) \right] + d_i \bullet log(c_i) \right\}$$

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

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

- statistics:

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

- 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.

- statistics:

$$S_{LR} = 2\left[l_1(\hat{\omega}, \hat{\beta}) - l_2(0, \hat{\beta})\right] \sim \chi^2(1)$$

- statistics:

$$S_{LR} = 2\left[l_1(\hat{\omega}, \hat{\beta}) - l_2(0, \hat{\beta}) \right] \sim \chi^2(1)$$

- statistics:

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

- method:
  - 在CZIP中的MLE中，令$\theta = \frac{\omega}{1-\omega}$，$\omega = 0$，
  - 得到$\hat{\theta}, \hat{\beta}, \frac{\hat{\theta}_0}{\sigma_{\theta_0}} \sim N(0,1)$

- score：

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

- 令$P_i=e^{-\mu_i}$：

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

- 令$\omega=0$：

$$\frac{\partial l_2}{\partial \omega}|_{\omega=0}=\sum_{i=1}^{n}\left\{(1-d_i)\left[\frac{r_i(1-P_i)}{P_i}-(1-r_i)\right]\right\}$$

$$\frac{\partial l_2}{\partial \omega}|_{\omega=0}=\sum_{i=1}^{n}\left\{(1-d_i)\left[\frac{r_i-P_i}{P_i}\right]\right\}$$

- 令$\theta=\frac{\omega}{1-\omega}$， 则 $\omega=\frac{\theta}{1+\theta},1-\omega=\frac{1}{1+\theta}$：

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

- $log(\mu_i)=x_i^T\beta,\mu_i=e^{-\mu_i}$

$$\frac{\partial log(\mu_i)}{\partial \beta}=x_i^T$$

$$\frac{\partial \mu_i}{\partial \beta}=x_i^T\mu_i$$

- $$\frac{\partial c_i}{\partial \beta} = \frac{\partial \sum_{y_i=L}^{\infty} P(Y_i = y_i)}{\partial \beta} = \frac{\partial \left[1 - \sum_{y_i=0}^{L-1} P(Y_i = y_i)\right]}{\partial \beta}$$

$$\frac{\partial c_i}{\partial \beta} = \frac{\partial \sum_{y_i=0}^{L-1} P(Y_i = y_i)}{\partial \beta} = -\sum_{y_i=0}^{L-1} \frac{\partial P(Y_i = y_i)}{\partial \beta}$$

$$\frac{\partial c_i}{\partial \beta} = -\sum_{y_i=0}^{L-1} \frac{\partial \frac{e^{-\mu_i} \mu_i^{y_i}}{(y_i)!}}{\partial \beta} = -\sum_{y_i=0}^{L-1} (y_i - \mu_i) P(Y_i = y_i) X_i^T$$

- $$\frac{\partial c_i}{\partial \beta} = \mu_i \sum_{k=0}^{L-1} P(Y_i = k) - \mu_i \sum_{k=1}^{L-1} P(Y_i = k)$$
  $$\frac{\partial c_i}{\partial \beta} = \mu_i P(Y_i = 0) = \mu_i e^{-\mu_i}$$
- $$\frac{\partial l_2}{\partial \beta} = \sum_{i=1}^{n} \left\{ (1 - d_i) \left[\frac{-r_i \mu_i p_i}{\theta_i + p_i} + (1 - r_i)(z_i - \mu_i) \right] + \frac{d_i}{c_i} \mu_i p_i \right\} x_i^T$$

- 令$\theta = 0$，得分数方程组如下：

$$\psi_1 = \frac{\partial l_2}{\partial \theta}|_{\theta=0} = \sum_{i=1}^{n} \left\{ (1-d_i) \left[\frac{r_i - p_i}{p_i} \right] \right\}$$

$$\psi_2 = \frac{\partial l_2}{\partial \beta}|_{\theta=0} = \sum_{i=1}^{n} \left\{ (1-d_i) \left[-r_i\mu_i + (1-r_i)(z_i - \mu_i) \right] + \frac{d_i}{c_i}\mu_i p_i \right\} x_i^T$$

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

- $$E(r_i) = P(z_i = 0) = P(Y_i = 0) = p_i$$
  $$S = \frac{1}{n}\sum_{i=1}^{n}(r_i - E(r_i)) = \frac{1}{n}\sum_{i=1}^{n}(r_i - p_i)$$
  $$s = E(s) = 0$$

- Estimation Equation(EE):

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

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

- 方差计算:

$$A = E(\frac{\partial\psi}{\partial\theta}) = E\begin{bmatrix} \frac{\partial\psi_1}{\partial s} & \frac{\partial\psi_1}{\partial\beta} \\ \frac{\partial\psi_2}{\partial s} & \frac{\partial\psi_2}{\partial\beta} \end{bmatrix} = E\begin{bmatrix} -1 & \frac{\partial\psi_1}{\partial\beta} \\ 0 & \frac{\partial\psi_2}{\partial\beta} \end{bmatrix}$$

$$A = E(\frac{\partial\psi}{\partial\theta}) = \begin{bmatrix} -1 & L \\ 0 & J \end{bmatrix}$$

$$B = E(\psi\psi^T) = E\begin{bmatrix} \psi_1^T\psi_1 & \psi_1^T\psi_2 \\ \psi_2^T\psi_2 & \psi_2^T\psi_2 \end{bmatrix}$$

- $$\frac{\partial \psi_1}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} p_i \mu_i x_{is}$$
$$\frac{\partial \psi_2}{\partial \beta} = \frac{\partial {l_1}^2}{\partial \beta^2}$$
$$\frac{\partial l_1}{\partial \beta_s \beta_r} = \sum_{i=1}^{n} \{-(1-d_i)\mu_i + d_i \mu_i p_i [c_i(1-\mu_i) - mu_i p_i]\} x_{is} x_{ir}$$

- 令：

$$J_{r_1 s} = \frac{\partial l_1}{\partial \beta_s \beta_r}$$

$$L = E(\frac{\partial \psi_1}{\partial \beta}) = -\left(\frac{1}{n}\sum_{i=1}^{n} p_i \mu_i x_{i0}, \frac{1}{n}\sum_{i=1}^{n} p_i \mu_i x_{i1}, \ldots, \frac{1}{n}\sum_{i=1}^{n} p_i \mu_i x_{ip}\right)^T$$

$$L = \frac{1}{n} X^T diag(p_i \mu_i) I$$

- $$A^{-1} = \begin{bmatrix} -1 & LJ^{-1} \\ 0 & J^{-1} \end{bmatrix}$$

$$E(d_i) = P(z_i \geq L) = \sum_{y_i=L}^{\infty} Y_i = y_i = c_i$$

$$\frac{\partial l_1{}^2}{\partial \beta_s \beta_r} = \frac{1}{n} \sum_{i=1}^{n} \left\{ -(1-c_i)\mu_i + \frac{d_i \mu_i p_i [c_i(1-\mu_i) - \mu_i p_i]}{c_i^2} \right\} x_{is} x_{ir}$$

$$E(\frac{\partial l_1{}^2}{\partial \beta_s \beta_r}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ c_i - 1 - \frac{\mu_i p_i^2}{c_i} + \mu_i p_i (1-\mu_i) \right\} \mu_i x_{is} x_{ir}$$

$$E(\frac{\partial l_1{}^2}{\partial \beta_s \beta_r}) = -J_{s,r}$$

$$J = -E(\frac{\partial l_1{}^2}{\partial \beta_s \beta_r})$$

- $$E(\psi^2) = -E(\frac{\partial\psi}{\partial\theta})$$

$$B = E(\psi\psi^T) = E\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i) & -L \\ -L^T & -J \end{bmatrix}$$

- 令$\lambda = \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i)$:

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

- 则：

$$A^{-1}BA^{-T} = \begin{pmatrix} -1 & -LJ^{-1} \\ 0 & J^{-1} \end{pmatrix} \begin{pmatrix} \lambda & -L \\ -L^T & -J \end{pmatrix} \begin{pmatrix} -1 & 0 \\ LJ^{-1} & J^{-1} \end{pmatrix}$$

$$A^{-1}BA^{-T} = \begin{pmatrix} -\lambda - LJ^{-1}L^T & 0 \\ -J^-1L^T & -I \end{pmatrix} \begin{pmatrix} -1 & 0 \\ LJ^{-1} & J^{-1} \end{pmatrix}$$

$$A^{-1}BA^{-T} = \begin{pmatrix} \lambda + LJ^{-1}L^T & 0 \\ 0 & -J^{-1} \end{pmatrix}$$

- statistics:

$$S_{HE} = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{p}_i)}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i) + LJ^{-1}L^T\right)^{1/2}} \sim N(0,1)$$