A NEW STATISTICAL TEST FOR ZMP MODEL

杜兴兴 魏雅慧

数学与统计学院

January 17, 2021

目录

模型设定

泊松模型：

- $p(y_i|\mu) = \frac{\mu^{y_i} * e^{-\mu}}{y_i!}$, $y_i \geq 0$

泊松回归：

- $y_i|x_i \sim i.d.\ Piosson(\mu_i)$, $log(\mu_i) = x_i^T\beta$
- 当在零存在过多时，为零膨胀模型(ZIP)，由于膨胀和紧缩仅在正负号存在区别，我们在此仅讨论膨胀现象。

模型设定

零膨胀泊松(ZIP)模型：

- 假设过多零的概率为ω，则原分布的概率为1-ω。
- 混合概率分布为

$$\begin{cases} P(Y_i = 0) = \omega + (1-\omega)e^{-\lambda_i} \\ P(Y_i = y_i) = (1-\omega)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad (y_i > 0)! \end{cases} \tag{1.1}$$

零膨胀泊松回归：

- ω 为动态时，
  - $y_i|x_i \sim i.d.\ ZIP(\omega_i, \mu_i)$, $logit(\omega_i) = x_i^T\beta_1$, $log(\mu_i) = v_i^T\beta_2$
- ω 为常数时，
  - $y_i|x_i \sim i.d.\ ZIP(\omega_i, \mu_i)$, $\omega_i = c_i$, $log(\mu_i) = v_i^T\beta_2$

目录

模拟设置

- 检验是否为零膨胀泊松（ZIP）：

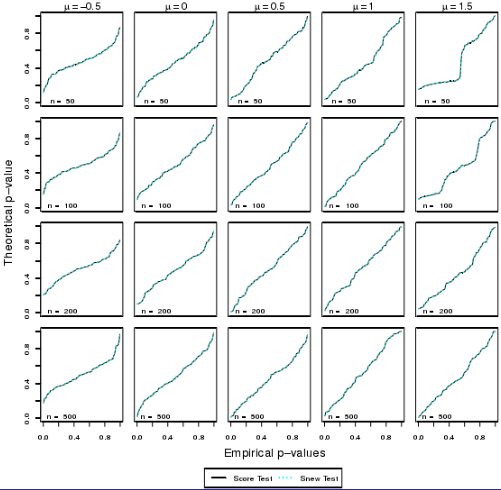
$$H_0 : \omega = 0 \quad vs \quad \omega > 0$$

- 均值变动：$y \sim Poisson(\mu)$
- 考虑三种情形：
  - 无协变量，$\mu = c$
  - 协变量服从均匀分布，$\mu = exp(\alpha - 1.4 * x)$
  - 协变量服从正态分布，$\mu = exp(\alpha - x)$
- 样本量：50，100，200，500
- Monte Carlo(蒙特卡罗)样本量：100

No Covariate

- R Code
- Type Error I:p-p plot
- Results

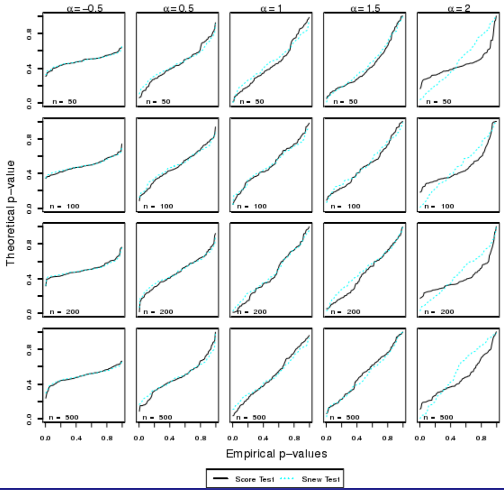
No Covariate (Type Error I:p-p plot)



Covariate x $\sim U(0,1)$

- R Code
- Type Error I:p-p plot
- Results

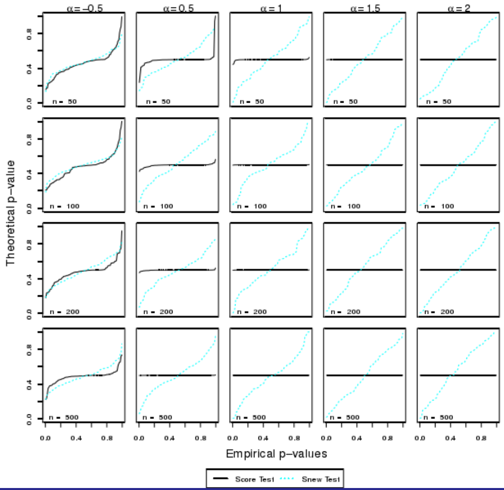
Covariate x $\sim U(0,1)$(Type Error I:p-p plot)



Covariate x $\sim N(0,1)$

- R Code
- Type Error I:p-p plot
- Results

Covariate x $\sim N(0,1)$(Type Error I:p-p plot)



模型设定

- ZIP统计模型：

$$\begin{cases} P(Y_i = 0) = \omega + (1-\omega)e^{-\lambda_i} \\ P(Y_i = y_i) = (1-\omega)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad (y_i > 0)! \end{cases} \tag{3.1}$$

- 令：
  - $\theta = \frac{\omega}{1-\omega} =$ 常数 $(-f(0) \leq \theta < \inf)$
  - $ln(\lambda_i) = X_i\beta$

- 示性函数:

$$I_{(condition)} = \begin{cases} 1 & \text{if condition is True} \\ 0 & \text{if condition is False} \end{cases} \tag{3.2}$$

- 对数似然:

$$l(\lambda, \theta; y) = \sum_i \left\{ -log(1+\theta) + I_{y_i=0} log(\theta + e^{-\lambda_i}) + +I_{y_i>0} \left[-\lambda_i + y_i log(\lambda_i) - log(y_i!) \right] \right\} \tag{3.3}$$

- 偏导方程:

$$\frac{dl}{d\beta_r} = \sum_i \left\{ I_{(y_i=0)} \left(\frac{-e^{-\lambda_i}}{\theta + e^{-\lambda_i}} \right) \lambda_i x_{ir} + I_{y_i>0} (y_i - \lambda_i) x_{ir} \right\} \quad r = 1, 2, .., p$$

$$\frac{dl}{d\theta} = \sum_i \left\{ \frac{-1}{(1+\theta)} + I_{y_i=0} \left(\frac{1}{\theta + e^{-\lambda_i}} \right) \right\}$$

- ZIP的对数似然函数基础上构建得分向量：

$$U(\beta,\theta)=\begin{bmatrix}\frac{\partial l}{\partial\beta}\\ \frac{\partial l}{\partial\theta}\end{bmatrix}$$

- Fisher信息矩阵：

$$J(\beta,\theta)=\begin{bmatrix}\frac{\partial^2 l}{\partial\beta\partial\beta'} & \frac{\partial^2 l}{\partial\beta\partial\theta}\\ \frac{\partial^2 l}{\partial\beta'\partial\theta} & \frac{\partial^2 l}{\partial\theta\partial\theta}\end{bmatrix}$$

- 在原假设成立下,根据Fisher信息矩阵构建Score检验统计量：

$$S(\hat{\beta})=S(\hat{\beta},0)=U^T(\hat{\beta},0)[J(\hat{\beta},0)]^{-1}U(\hat{\beta,0})\sim\chi^2_{(1)}$$

- 所以在原假设条件下，$\theta = 0$:

$$\frac{dl}{d\theta}\Big|_{\theta=0} = \sum_i \left\{ \frac{I_{(y_i=0)}}{e^{-\hat{\lambda}_i}} - 1 \right\}$$

- 计算得分向量得：

$$U^T(\beta, 0) = \left(0, ..., 0, \sum_i \left\{ \frac{I_{(y_i=0)}}{e^{-\hat{\lambda}_i}} - 1 \right\} \right)$$

- 计算二阶偏导:

$$\frac{d^2 l(.)}{d\beta_r d\beta_s} = \sum_i \left\{ I_{y_i=0} \left(\frac{-e^{-\lambda_i}[(1-\lambda_i)\theta + e^{-\lambda_i}]}{[\theta + e^{-\lambda_i}]^2} \right) \lambda_i x_{ir} x_{is} - I_{(y_i>0)} \lambda_i x_i r x_i s \right\} \quad r = 1 \ldots p, s = 1 \ldots p$$
$$\frac{d^2 l(.)}{d\beta_r d\theta} = \sum_i \left\{ I_{y_i=0} \left(\frac{-e^{-\lambda_i}}{[\theta + e^{-\lambda_i}]^2} \right) \lambda_i x_{ir} \right\} \quad r = 1 \ldots p$$
$$\frac{d^2 l(.)}{d\theta^2} = \sum_i \left\{ \frac{1}{(1+\theta)^2} - I_{y_i=0} \left(\frac{1}{[\theta + e^{-\lambda_i}]^2} \right) \right\} \tag{3.4}$$

- 可以得到$J(\hat{\beta}, 0)$:
  - $\hat{J}_{r,s} = \sum_i \hat{\lambda}_i x_{ir} x_{is} \quad r = 1 \ldots p, s = 1 \ldots p$
  - $\hat{J}_{r,p+1} = -\sum_i \hat{\lambda}_i x_i r \quad r = 1 \ldots p$
  - $\hat{J}_{p+1,p+1} = \sum_i \frac{1 - e^{-\hat{\lambda}_i}}{e^{-\hat{\lambda}_i}}$

- $J(\hat{\beta},0)$矩阵分块得:

$$J(\hat{\beta},0)=\begin{bmatrix}\hat{J_{11}} & \hat{J_{12}}\\ \hat{J_{21}} & \hat{J_{22}}\end{bmatrix}$$

- $\hat{J_{11}}=X^T diag(\hat{\lambda}X)$
- $\hat{J_{12}}=-X^T\hat{\lambda}$
- $\hat{J_{21}}=-\hat{\lambda}^T X^T$
- $\hat{J_{22}}=\sum_i\left(\frac{1}{e^{-\hat{\lambda}_i}}-1\right)$

- $J(\hat{\beta},0)^{-1}$定义为C:

$$J(\hat{\beta},0)^{-1} = C(\hat{\beta},0) = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

- 由于得分向量$U(\hat{\beta},0)$的结构只需要C_{22}即可：

$$C_{22}^{-1} = \hat{J}_{22} - \hat{J}_{21}\hat{J}_{11}^{-1}\hat{J}_{12}$$

$$C_{22}^{-1} = \sum_i \left(\frac{1}{e^{-\hat{\lambda}_i}} - 1\right) - \hat{\lambda}^T X[X^T diag(\hat{\lambda})X]^{-1}X^T\hat{\lambda}$$

- Score统计量最终形式为:

$$S(\hat{\beta}) = S(\hat{\beta}, 0) = U^T(\hat{\beta}, 0)[J(\hat{\beta}, 0)]^{-1}U(\hat{\beta}, 0) \sim \chi^2_{(1)}$$

$$S(\hat{\beta}, 0) = \frac{\left\{\sum_{i=1}^{n}\left(\frac{I_{y_i=0}}{e^{-\hat{\lambda}_i}} - 1\right)\right\}^2}{\sum_i\left(\frac{1}{e^{-\hat{\lambda}_i}} - 1\right) - \hat{\lambda}^T X[X^T diag(\hat{\lambda})X]^{-1}X^T\hat{\lambda}}$$

- 可得:

$$S_{score} = \frac{\left\{\sum_{i=1}^{n}\left(\frac{I_{y_i=0}}{e^{-\hat{\lambda}_i}} - 1\right)\right\}}{\left\{\sum_i\left(\frac{1}{e^{-\hat{\lambda}_i}} - 1\right) - \hat{\lambda}^T X[X^T diag(\hat{\lambda})X]^{-1}X^T\hat{\lambda}\right\}^{1/2}}$$

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