A New Statistical Test for Tobit Model and ZMP Model

杜兴兴 魏雅慧

数学与统计学院

April 13, 2021

目录

目录

- 截尾数据问题
- 潜在类数据问题

截尾数据问题

- 案例1：
  尿液、血清或其他生物基质中感兴趣的生物标志物通常具有检测的检测界限，在检测界限下通常存在过量的观测数据或删失数据。在公共卫生和医学研究中，由于检测限而导致的数据审查非常普遍。

潜在类数据问题

- 案例2:
  在实践中，通常数据会表现出模型预期的过度删失观察。一个常见的原因是研究人群的异质性，也就是说，存在一个缺乏这种生物标志物的亚群，其值总是低于或高于检测限，因此被限制。
- 问题：如何检测数据中是否存在这样的现象？

目录

- 传统方法
- 其他文献

前人的方法研究

- LR Test:
  - $S_{LR} = 2[l_2(\hat{\omega}, \hat{\beta}_\mu, \hat{\sigma}) - l_1(\hat{\beta}_\mu{}', \hat{\sigma}')] \sim \chi_1^2$
- Wald Test:
  - $S_{Wald} = \frac{\hat{\omega}}{\hat{\sigma_\omega}^2} \sim \chi_1^2$ 或 $Z_{Wald} = \frac{\hat{\omega}}{\hat{\omega}} \sim N(0,1)$
- Score Test:
  - $S_{score} = \frac{\sum_{i=1}^n \frac{1}{p_i}(r_i - p_i)}{[\sum_{i=1}^n \frac{(1-p_i)}{p_i} - (\hat{\mu})X(X^T diag(\hat{\mu})X)^{-1}X^T(\hat{\mu})]^{1/2}} \sim N(0,1)$

前人的方法研究(CONT.)

- Vuong Test:

$$V = \frac{\sqrt{n}\bar{g}}{s_g} \sim AN(0,1), \quad g_i = log\left[\frac{f_1(y_i|\theta_1^*)}{f_2(y_i|\theta_2^*)}\right]$$

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i; s_g^2 = \frac{1}{n-1}\sum_{n-1}^{n}(g_i - \bar{g})^2$$

- 用似然比构造统计量，其分布服从渐进正态分布
- 第一类错误大
- 需要确定分布形式

其它文献

- -Paul Wilson and Jochen Einbeck,(2019)
- 核心思想：
  基于泊松回归，通过“中P值”的方法检验数据中零的个数；
- 建立假设：
  H_0 : $Y_i|x_i$ 服从特定的分布
  $H_1^{(a)}$: $Y_i|x_i$ 是一个零修改模型，即$n_0 < n_{\alpha/2}$ or $n_0 > n_{1-\alpha/2}$
  $H_1^{(b)}$: $Y_i|x_i$ 是一个零膨胀模型，即$n_0 > n_{1-\alpha}$
  $H_1^{(c)}$: $Y_i|x_i$ 是一个零紧缩模型，即$n_0 < n_{\alpha}$

其它文献(CONT.)

- 检验步骤：
  - 通过数据计算出泊松参数$\hat{\mu}_i$，构造混合估计量:

$$\hat{\mu}_H = h\hat{\mu}_W + (1-h)\hat{\mu}_T \ , \ 0 \leq h \leq 1.$$

  - 对每个y_i 估计p_i:

$$p_i = p(0|\hat{\mu}_i, \hat{\phi}) = e^{-\hat{\mu}_i}$$

  - 根据$N_0 \sim Poisson - binomial(p_i)$，计算"中P 值":

$$P^*_{0.5} = 0.5(P[N_0 \geq t] + P[N_0 \geq t+1])$$

  - 根据"中P 值"判断假设是否拒绝原假设
- 评价：
  - $p_{0.5}$ 不准确，只是一种经验的判断
  - 需要确定分布形式

目录

- Tobit 模型与mTobit模型
- 泊松模型和零修改模型

- a new statistical test for latent class in censored data due to detection limit.(Zou et al.,2020)

模型设定1

- Tobit模型：
  - 截尾数据(检测限度L)，总体来自正态分布
  - $f(y_i) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_i-\mu_i)^2}{2\sigma^2}) & if \quad y_i \geq L \\ \Phi(\frac{L-\mu_i}{\sigma}) & if \quad y_i < L \end{cases}$
- Tobit回归：
  - $Y_i|x_i \sim i.d.\ Tobit(\mu_i, \sigma^2, L), \quad \mu_i = x_i^T\beta.$

模型设定2

- mTobit模型：
  - 已知Tobit 分布，检测界限为L
  - 实例中，通常表现出检测界限以下存在过量的观测数据，即潜在类，概率为ω
  - $$f(y_i) = \begin{cases} (1-\omega)\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_i-\mu_i)^2}{2\sigma^2}) & if\ exposed\ and\ y_i \geq L, \\ (1-\omega)\Phi(\frac{L-\mu_i}{\sigma}) & if\ exposed\ and\ y_i < L, \\ \omega & if\ nonexposed. \end{cases}$$
- mTobit回归：
  - $Y_i|x_i \sim i.d.\ mTobit(\omega_i, \mu_i, \sigma^2, L), logit(\omega_i) = \mu_i^T\beta_1, \mu_i = v_i^T\beta_2.$

模型设定

- 零的表示：

$$r_i = 1_{\{y_i < L\}}$$

- 似然函数：
  - Tobit 模型：

$$L_1 = \prod_{i=1}^{n} \left[\Phi\left(\frac{L - \mu_i}{\sigma} \right) \right]^{r_i} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2} \right) \right]^{(1-r_i)}.$$

  - mTobit 模型：

$$L_2 = \prod_{i=1}^{n} \left[\omega + (1-\omega)\, \Phi\left(\frac{L - \mu_i}{\sigma} \right) \right]^{r_i} \left[(1-\omega)\, \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2} \right) \right]^{(1-r_i)}.$$

假设检验

- 假设：

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

构建统计量

- 分布：

$$\sqrt{n}\left(\hat{s}-0\right) \rightarrow N(0,\tau^2)$$

- 构建思想：
  - MLE 估计得到Tobit 模型的参数；
  - 使用其构造新的无偏量；
  - 其服从渐进正态分布；

构建统计量

- 分布：

$$\sqrt{n}\left(\hat{s}-0\right) \rightarrow N(0,\tau^2)$$

- 构建思想：
  - MLE 估计得到Tobit 模型的参数；
  - 使用其构造新的无偏量；
  - 其服从渐进正态分布；

统计量的推导

- MLE估计得Tobit 模型的参数β 和σ:
  - 假设: $y_i|\mathbf{x}_i \sim$ i.d. Tobit(μ_i, σ^2, L), with $\mu_i = \mathbf{x}_i^T\beta$, whre $\beta = (\beta_0, \beta_1, .., \beta_p)$.
  - 令: $c_i = \Phi\left(\frac{L-\mu_i}{\sigma}\right)$, $d_i = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(L-\mu_i)^2}{2\sigma^2})$
  - 似然函数:

$$L_i = \prod_{i=1}^{n}\left[\Phi\left(\frac{L-\mu_i}{\sigma}\right)\right]^{r_i}\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)\right]^{(1-r_i)}.$$

  - 对数似然函数:

$$l(\beta,\sigma) = \sum_{i=1}^{n}\left\{r_i\log c_i + (1-r_i)\left[\log\left(-\sqrt{2\pi}\sigma\right) - \frac{(y_i-\mu_i)^2}{2\sigma^2}\right]\right\}. \tag{3.1}$$

  - 偏导方程:

$$\frac{dl(\cdot)}{d\beta_r} = \sum_{i=1}^{n}\left\{\frac{r_i}{c_i}\frac{dc_i}{d\beta_r} + (1-r_i)\frac{(y_i-\mu_i)}{\sigma^2}x_{ir}\right\} = 0, \quad r = 0, 1, 2, \ldots, p, \tag{3.2}$$

$$\frac{dl(\cdot)}{d\sigma} = \sum_{i=1}^{n}\left\{\frac{r_i}{c_i}\frac{dc_i}{d\sigma} + (1-r_i)\left[-\frac{1}{\sigma} + \frac{(y_i-\mu_i)^2}{\sigma^3}\right]\right\} = 0. \tag{3.3}$$

统计量的推导(CONT.)

- MLE估计得Tobit 模型的参数β 和σ (Cont.):
  - 由于:

$$\frac{dc_i}{d\beta_r} = \left(\int_{-\infty}^{\frac{L-\mu_i}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \right)' = -d_i \frac{x_{ir}}{\sigma}, \quad r = 0, 1, 2, \ldots, p,$$

$$\frac{dc_i}{d\sigma} = \left(\int_{-\infty}^{\frac{L-\mu_i}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \right)' = -d_i \frac{L-\mu_i}{\sigma^2},$$

  - 得到新的偏导方程:

$$\begin{cases} \sum_{i=1}^{n} \left\{ -r_i \frac{d_i}{c_i} + (1-r_i) \frac{(y_i-\mu_i)}{\sigma} \right\} \frac{x_{ir}}{\sigma} = 0, \quad r = 0, 1, 2, \ldots, p, \\ \sum_{i=1}^{n} \left\{ -r_i \frac{d_i}{c_i} \frac{L-\mu_i}{\sigma} + (1-r_i) \left[-1 + \frac{(y_i-\mu_i)^2}{\sigma 2} \right] \right\} = 0. \end{cases}$$

统计量的推导(CONT.)

- 构造新的无偏估计量:
  - 根据Tobit 模型下截尾数据的观测比例与期望比例的差异可以给出:

$$s = \frac{1}{n}\sum_{i=1}^{n}\left(r_i - E\left(r_i\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\left[r_i - \Phi\left(\frac{L-\mu_i}{\sigma}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[r_i - c_i\right]. \tag{3.4}$$

  - 根据上式构造出检验统计量s':

$$s' = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_i - c_i)}{\frac{1}{n}\left[\sum_{i=1}^{n} c_i(1-c_i)\right]^{1/2}} \sim AN(0,1) \tag{3.5}$$

  - s 是关于β, σ 的函数;s'也是关于β, σ 的函数
  - 用β, σ 的拟合值来表示s，s是有偏的
  - 加入约束条件，使得s为无偏估计量

$$\frac{1}{n}\sum_{i=1}^{n}(r_i - c_i - s) = 0 \tag{3.6}$$

  - 由于要考虑β, σ 的变异性，所以使用估计方程方法来求解统计量s

统计量的推导(CONT.)

- 构造新的无偏估计量:
  - 综上，可以得到一组方程:

$$\psi_{1r} = \frac{1}{n}\sum_{i=1}^{n}\left[(1-r_i)\frac{(y_i-\mu_i)}{\sigma} - r_i\frac{d_i}{c_i}\right]\frac{x_{ir}}{\sigma}, \quad r = 0, 1, 2, \ldots, p,$$

$$\psi_2 = \frac{1}{n}\sum_{i=1}^{n}\left\{(1-r_i)\left[-\frac{1}{\sigma} + \frac{(y_i-\mu_i)^2}{\sigma^3}\right] - r_i\frac{d_i}{c_i}\frac{L-\mu_i}{\sigma^2}\right\},$$

$$\psi_3 = \frac{1}{n}\sum_{i=1}^{n}(r_i - c_i - s). \tag{3.7}$$

  - 定义: $\Psi = (\psi_1, \psi_2, \psi_3)^T$, $\psi_1 = (\psi_{10}, \psi_{11}, \psi_{12}, \ldots, \psi_{1p})$ and $\gamma = (\beta, \sigma, s)$.

统计量的推导(CONT.)

- 统计量的渐进分布:
  - 由估计方程（EE）可以得到统计量的渐进分布:

$$\sqrt{n}\,(\hat{s}-0) \to N(0,\tau^2)$$

  - τ^2 是矩阵$A^{-1}BA^{-T}$ 的$(p+3)*(p+3)$ 项
  - $A(\gamma) = E\left[\frac{\partial}{\partial\gamma}\Psi_i(Y_i,\gamma)\right]$
  - $B(\gamma) = Var(\Psi_i(Y_i,\gamma)$

统计量的推导(CONT.)

- A，B矩阵的具体形式：

$$A(\gamma) = E\frac{\partial \Psi}{\partial \gamma} = E\begin{pmatrix} \frac{\partial \psi_1^T}{\partial \beta} & \frac{\partial \psi_1^T}{\partial \sigma} & \frac{\partial \psi_1^T}{\partial s} \\ \frac{\partial \psi_2}{\partial \beta} & \frac{\partial \psi_2}{\partial \sigma} & \frac{\partial \psi_2}{\partial s} \\ \frac{\partial \psi_3}{\partial \beta} & \frac{\partial \psi_3}{\partial \sigma} & \frac{\partial \psi_3}{\partial s} \end{pmatrix}, \tag{3.8}$$

$$B(\gamma) = E\Psi^T\Psi = E\begin{pmatrix} \psi_1^T\psi_1 & \psi_1^T\psi_2 & \psi_1^T\psi_3 \\ \psi_2\psi_1 & \psi_2^2 & \psi_2\psi_3 \\ \psi_3\psi_1 & \psi_3\psi_2 & \psi_3^2 \end{pmatrix}. \tag{3.9}$$

统计量的推导(CONT.)

- 计算A矩阵中的每个元素：

$$\frac{\partial \psi_1^T}{\partial \beta} = \frac{\partial^2 l}{\partial \beta^2} = \begin{pmatrix} \frac{d^2 l}{d\beta_0 d\beta_0} & \cdots & \frac{d^2 l}{d\beta_0 d\beta_p} \\ \vdots & \cdots & \vdots \\ \frac{d^2 l}{d\beta_p d\beta_0} & \cdots & \frac{d^2 l}{d\beta_p d\beta_p} \end{pmatrix},$$

$$\frac{\partial \psi_1^T}{\partial \sigma} = \frac{\partial^2 l}{\partial \beta \partial \sigma} = \left(\frac{\partial \psi_2}{\partial \beta}\right)^T = \left(\frac{d^2 l}{d\beta_0 d\partial \sigma}, \frac{d^2 l}{d\beta_1 d\partial \sigma}, \ldots, \frac{d^2 l}{d\beta_p d\partial \sigma}\right)^T,$$

$$\frac{\partial \psi_1^T}{\partial s} = \mathbf{0}_{(p+1)\times 1},$$

统计量的推导(CONT.)

- 计算A矩阵中的每个元素(cont.):

$$\frac{\partial \psi_2}{\partial \sigma} = \frac{\partial^2 l}{\partial \sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left\{-r_i \frac{d_i\,(L-\mu_i)}{c_i \sigma^3}\left[\frac{d_i}{c_i}\frac{L-\mu_i}{\sigma} + \frac{(L-\mu_i)^2}{\sigma^2} - 2\right] + (1-r_i)\left[\frac{1}{\sigma^2} - 3\frac{(y_i-\mu_i)^2}{\sigma^4}\right]\right\},$$

$$\frac{\partial \psi_2}{\partial s} = 0,$$

$$\frac{\partial \psi_3}{\partial \beta} == \left(\frac{1}{n}\sum_{i=1}^{n} d_i \frac{x_{i0}}{\sigma}, \frac{1}{n}\sum_{i=1}^{n} d_i \frac{x_{i1}}{\sigma}, \ldots, \frac{1}{n}\sum_{i=1}^{n} d_i \frac{x_{ip}}{\sigma}\right),$$

$$\frac{\partial \psi_3}{\partial \sigma} = \frac{1}{n}\sum_{i=1}^{n} d_i \frac{L-\mu_i}{\sigma^2},$$

$$\frac{\partial \psi_3}{\partial \mathbf{s}} = -1.$$

统计量的推导(CONT.)

- 对A矩阵进行分块：

$$A = \begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & \mathbf{0} \\ \mathbf{J}_{21} & \mathbf{J}_{22} & \mathbf{0} \\ \mathbf{J}_{31} & \mathbf{J}_{32} & -1 \end{pmatrix},$$

  - $\mathbf{J}_{11}$ 是：$(p+1) \times (p+1)$ 矩阵
  - $\mathbf{J}_{21}$、$\mathbf{J}_{31}$ 是：$1 \times (p+1)$ 矩阵
  - $\mathbf{J}_{12} = \mathbf{J}_{21}^T$
  - $\mathbf{J}_{22}$、$\mathbf{J}_{32}$ 是相同形状的矩阵

统计量的推导(CONT.)

- 对B矩阵进行计算：
  - 由(??)，ψ_1 和 ψ_2 是偏导方程，可以得到下面的推导：

$$
\begin{aligned}
E\psi^2 &= E\left(\frac{\partial \log f}{\partial \theta}\right)^2 = \int \left(\frac{\partial \log f}{\partial \theta}\right)^2 \cdot f dx = \int \frac{\partial \log f}{\partial \theta}\frac{\partial \log f}{\partial \theta} \cdot f dx \\
&= \int \frac{\partial \log f}{\partial \theta}\frac{\partial f}{\partial \theta}\frac{1}{f} \cdot f dx = \int \frac{\partial \log f}{\partial \theta}\frac{\partial f}{\partial \theta} dx \\
&= \int \left[\frac{\partial\left(\frac{\partial \log f}{\partial \theta} \cdot f\right)}{\partial \theta} - \frac{\partial^2 \log f}{\partial \theta^2} \cdot f\right] dx \\
&= \frac{\partial}{\partial \theta}\int \frac{\partial \log f}{\partial \theta} \cdot f dx - \int \frac{\partial^2 \log f}{\partial \theta^2} \cdot f dx = E\psi - E\frac{\partial \psi}{\partial \theta} \\
&= -E\frac{\partial \psi}{\partial \theta},
\end{aligned}
$$

统计量的推导(CONT.)

- B矩阵的表达形式:
  - 由(??)，ψ_1 和 ψ_2 是偏导方程，可以得到下面的推导:

$$B(\gamma) = E\Psi^T\Psi = \begin{pmatrix} -\mathbf{J}_{11} & -\mathbf{J}_{12} & -\mathbf{J}_{13} \\ -\mathbf{J}_{21} & -\mathbf{J}_{22} & -\mathbf{J}_{23} \\ -\mathbf{J}_{31} & -\mathbf{J}_{32} & \frac{1}{n}\sum_{i=1}^{n} c_i(1-c_i) \end{pmatrix} = \begin{pmatrix} -\mathbf{J} & -\mathbf{L}^T \\ -\mathbf{L} & \frac{1}{n}\sum_{i=1}^{n} c_i(1-c_i) \end{pmatrix}.$$

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

统计量的推导(CONT.)

- $A^{-1}BA^{-T}$ 的表达式：

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

- 所以可以得到τ^2 的表达式：

$$\begin{aligned}\mathbf{LJ}^{-1}\mathbf{L}^T+\lambda &= (\mathbf{J}_{31}, \mathbf{J}_{32}) \begin{pmatrix} W & -\mathbf{J}_{11}^{-1}\mathbf{J}_{12}V \\ -\mathbf{J}_{22}^{-1}\mathbf{J}_{21}W & V \end{pmatrix} \cdot \begin{pmatrix} \mathbf{J}_{31}^T \\ \mathbf{J}_{32} \end{pmatrix} + \lambda \\ &= \left(\mathbf{J}_{31}W - \mathbf{J}_{32}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}W, -\mathbf{J}_{31}\mathbf{J}_{11}^{-1}\mathbf{J}_{12}V + \mathbf{J}_{32}V\right) \cdot \begin{pmatrix} \mathbf{J}_{31} \\ \mathbf{J}_{32} \end{pmatrix} + \lambda \\ &= \lambda + \mathbf{J}_{31}W\mathbf{J}_{31}^T - \mathbf{J}_{32}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}W\mathbf{J}_{31}^T - \mathbf{J}_{31}\mathbf{J}_{11}^{-1}\mathbf{J}_{12}V\mathbf{J}_{32} + \mathbf{J}_{32}V\mathbf{J}_{32}.\end{aligned}$$

统计量的推导(CONT.)

- 最终统计量形式：

$$S_{\text{new}} = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{r_i - \widehat{c}_i\}}{\left(\frac{1}{n} \sum_{i=1}^{n} \widehat{c}_i(1-\widehat{c}_i) + \widehat{\mathbf{J}}_{31} \widehat{W} \widehat{\mathbf{J}}_{31}^T - \widehat{\mathbf{J}}_{32} \widehat{\mathbf{J}}_{22}^{-1} \widehat{\mathbf{J}}_{21} \widehat{W} \widehat{\mathbf{J}}_{31}^T - \widehat{\mathbf{J}}_{31} \widehat{\mathbf{J}}_{11}^{-1} \widehat{\mathbf{J}}_{12} \widehat{V} \widehat{\mathbf{J}}_{32} + \widehat{\mathbf{J}}_{32} \widehat{V} \widehat{\mathbf{J}}_{32} \right)^{\frac{1}{2}}}. \quad (3.10)$$

统计量的推导(CONT.)

- 回顾：

$$Y_i|x_i \sim i.d.\ Tobit(\mu_i, \sigma^2, L), \quad \mu_i = x_i^T\beta.$$

$$S_{score} = \frac{\sum_{i=1}^{n} \frac{1}{p_i}(r_i - p_i)}{[\sum_{i=1}^{n} \frac{(1-p_i)}{p_i} - (\hat{\mu})X(X^T diag(\hat{\mu})X)^{-1}X^T(\hat{\mu})]^{1/2}} \sim N(0,1)$$

- 补充：
  - 当μ 为一个常数时，上述推导的统计量S_{new} 等价于S_{score}

- A test of inflated zeros for Poisson regression models.(He et al.,2019)

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。

ZIP回归：

- $y_i|x_i \sim i.d.\ ZIP(\omega_i, \mu_i)$, $logit(\omega_i) = x_i^T\beta_1$, $log(\mu_i) = v_i^T\beta_2$

假设检验

假设：

- $H_0 : \omega = 0 \quad vs \quad H_1 : \omega \neq 0$

统计量

分布：

- $\sqrt{n}(\hat{s}-0) \rightarrow N(0,\tau^2)$

构建思想：

- LSE估计得泊松模型的参数；
- 使用其构建新的无偏量；
- 其服从渐进正态分布；

统计量的推导

统计量服从正态分布：

- 由估计方程，得到统计量：
  $\sqrt{n}(\hat{s}-0) \rightarrow N(0,\tau^2)$
- τ^2是矩阵$A^{-1}BA^{-T}$ 的$(p+2)*(p+2)$ 项；
- $A(\gamma)=E\frac{\partial\Psi}{\partial\gamma}$, $B=E\left(\Psi\Psi^T\right)$
- $\Psi_1=\frac{1}{n}\sum_{i=1}^{n}\left[r_i-exp(-exp(x_i^T\beta))-S\right]=0$
- $\Psi_2=\frac{1}{n}\sum_{i=1}^{n}X_i^T(y_i-exp(x_i^T\beta))=0$
- （去掉等于0）

目录

模拟设置

- 检验在删失数据中时候存在一个潜在类：

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

- 检测限制：L设为-1
- Tobit 回归模型的方差：$\sigma^2 = 4$
- 均值变动：$y \sim Tobit(\mu, 4, -1)$
- 考虑三种情形：
  - No covariate
  - Covariate x $\sim Uniform(0, 1)$
  - Covariate x $\sim N(0, 2)$
- 样本量：50，100，200，500，1000
- Monte Carlo(蒙特卡罗)样本量：1000

No Covariate

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

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

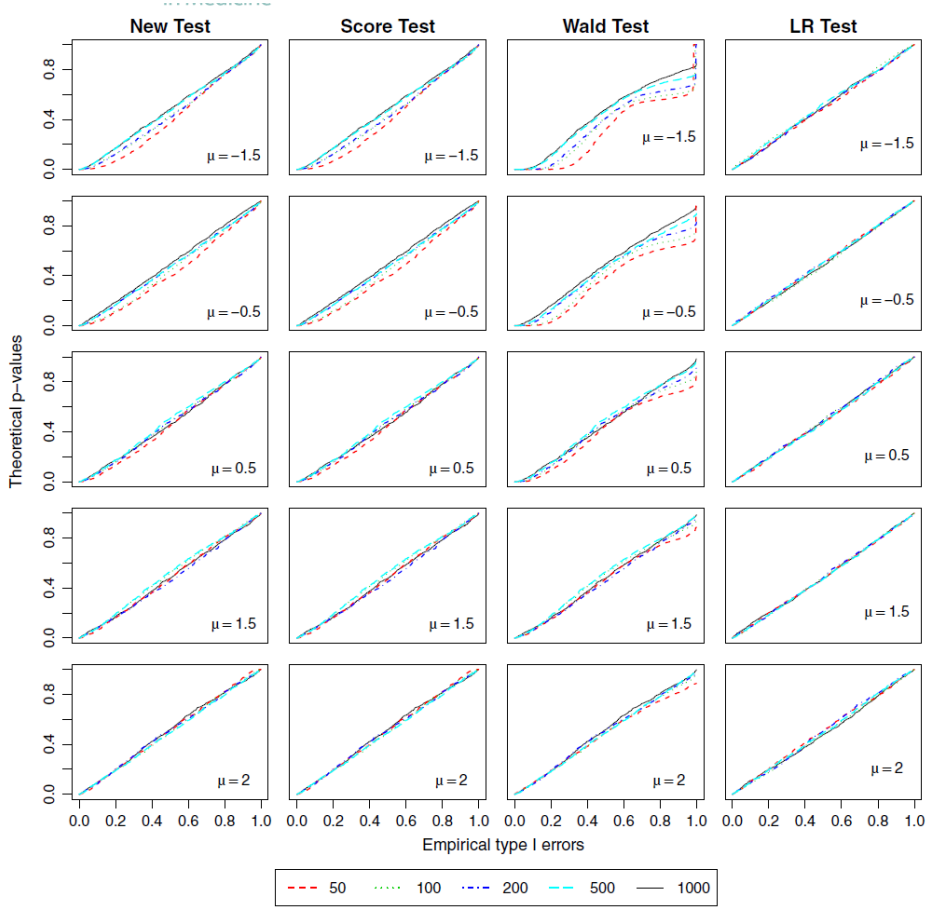

FIGURE 1 Plots of the theoretical *P*-values and the corresponding empirical type I errors for the Tobit model without covariates and sample sizes 50, 100, 200, 500, and 1000 [Color figure can be viewed at wileyonlinelibrary.com]

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)

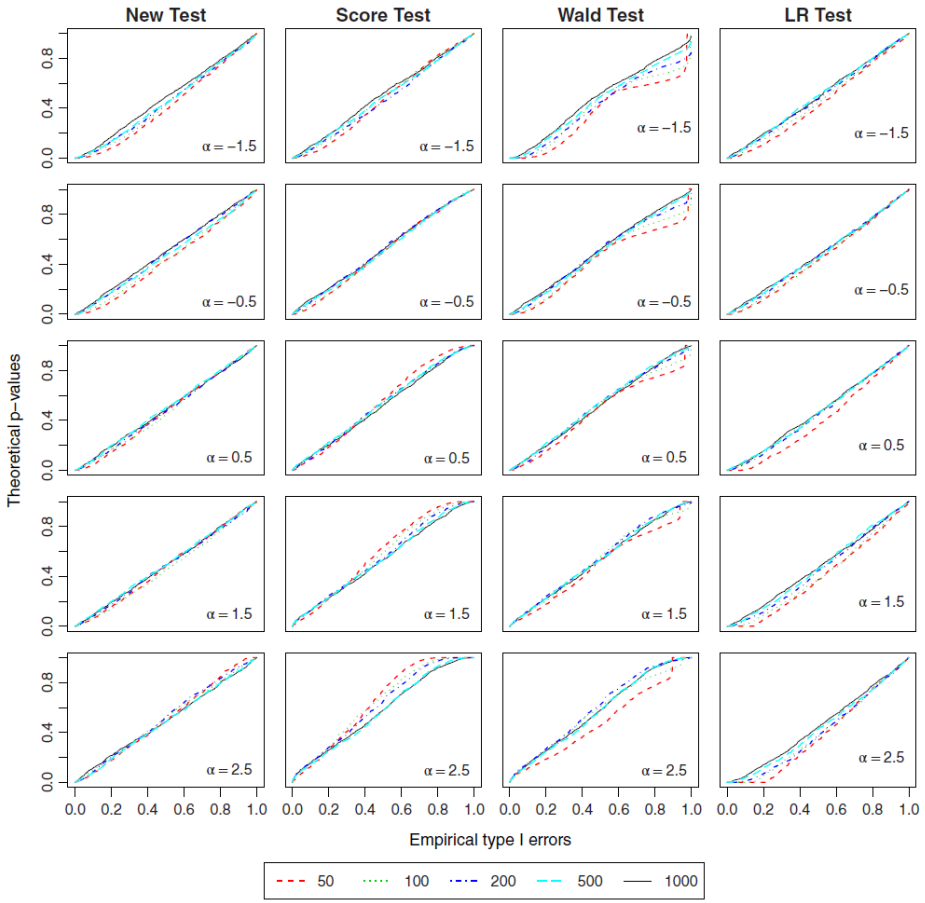

FIGURE 3 Plots of theoretical *P*-values and the corresponding empirical type I errors for the Tobit model with normally distributed predictors and sample sizes 50, 100, 200, 500, and 1000 [Color figure can be viewed at wileyonlinelibrary.com]

Covariate x $\sim N(0, 2)$

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

Covariate x $\sim N(0, 2)$(Type Error I:p-p plot)

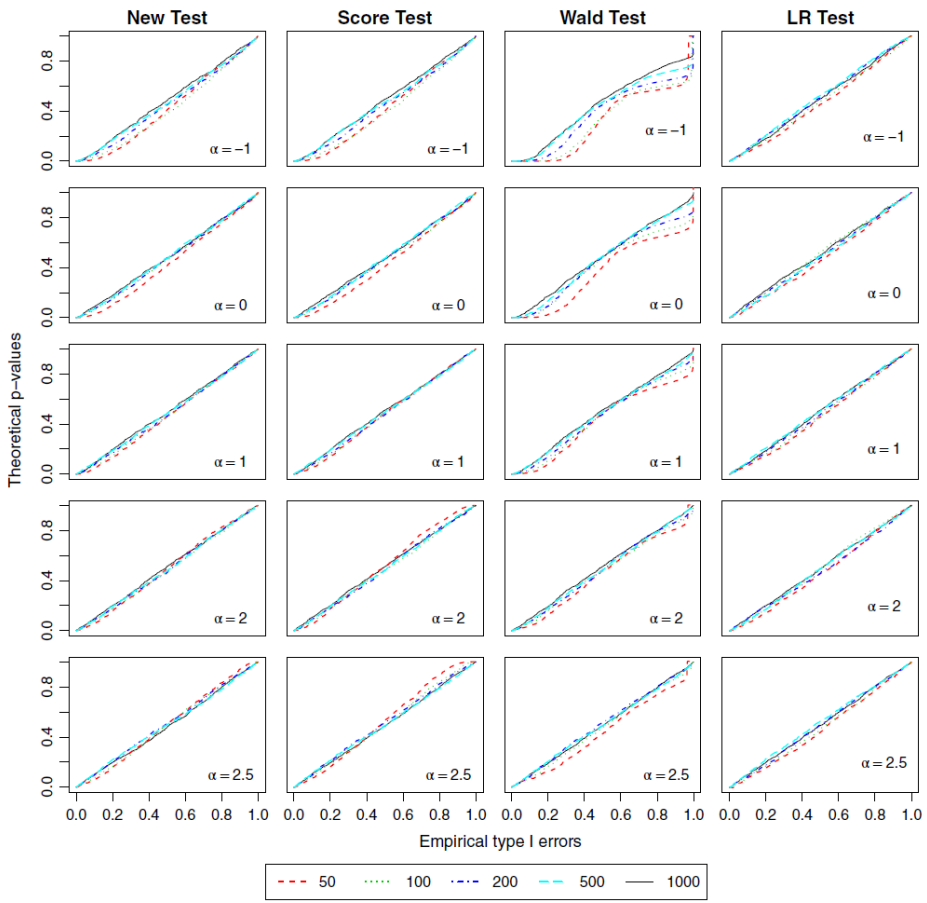

FIGURE 2 Plots of theoretical *P*-values and the corresponding empirical type I errors for the Tobit model with uniformly distributed predictors and sample sizes 50, 100, 200, 500, and 1000 [Color figure can be viewed at wileyonlinelibrary.com]

No Covariate

- $y \sim mTobit(\omega, \mu, 4, -1), \ \omega = 0.05k, \ k = 1, 2, ..., 6$
  - R Code
  - Power:Probabilities of latent class
  - Results

No Covariate(Power)

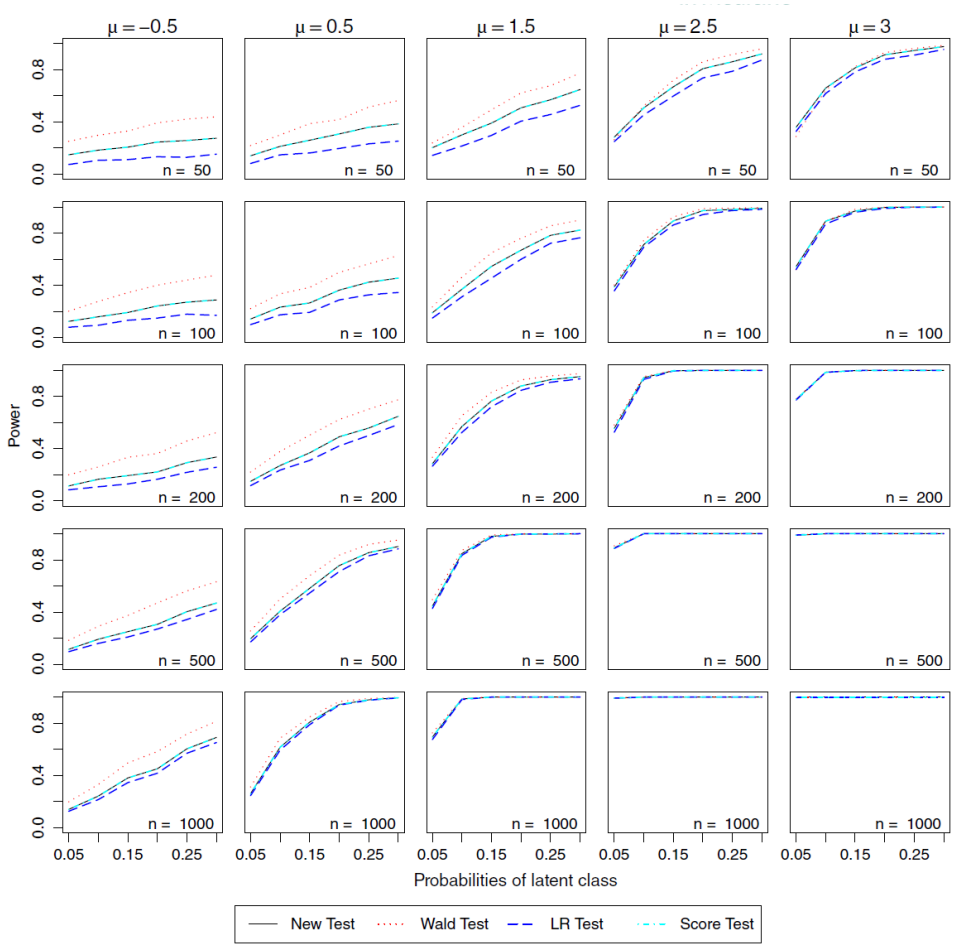

FIGURE 4 Power of detecting the latent class in mTobit model when there are no covariates for both Tobit model and the latent class [Color figure can be viewed at wileyonlinelibrary.com]

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

- $y \sim mTobit(\omega, \mu, 4, -1), \mu = \alpha - x, \ \omega = 0.05k$

$$k = 1, 2, ..., 6, \ \alpha = 0, 1, 2, 3, 4$$

  - R Code
  - Power:Probabilities of latent class
  - Results

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

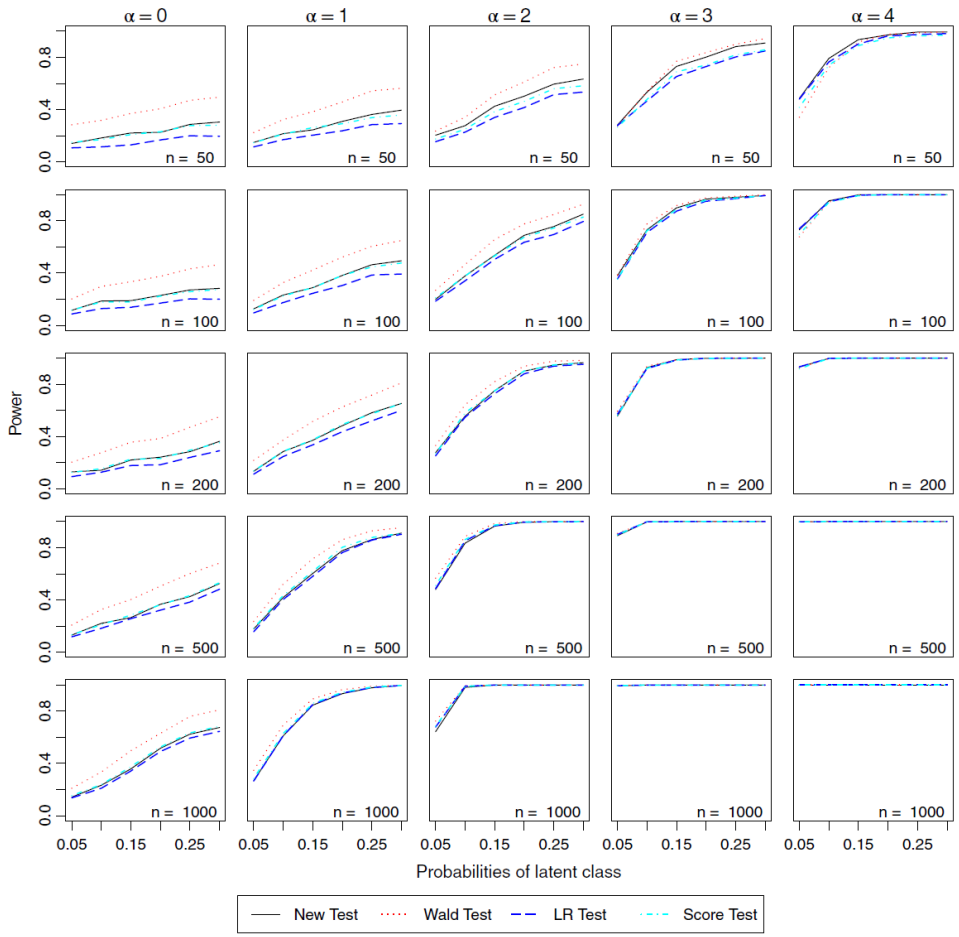

FIGURE 5 Power of detecting the latent class in mTobit model with uniformly distributed covariate for Tobit component only [Color figure can be viewed at wileyonlinelibrary.com]

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

- $y \sim mTobit(\omega, \mu, 4, -1), \ \ \mu = \alpha - x, \ \ logit(\omega = -b + x)$

$$\alpha = 0, 1, 2, 3, 4; \ \ b = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$$

  - R Code
  - Power:Probabilities of latent class
  - Results

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

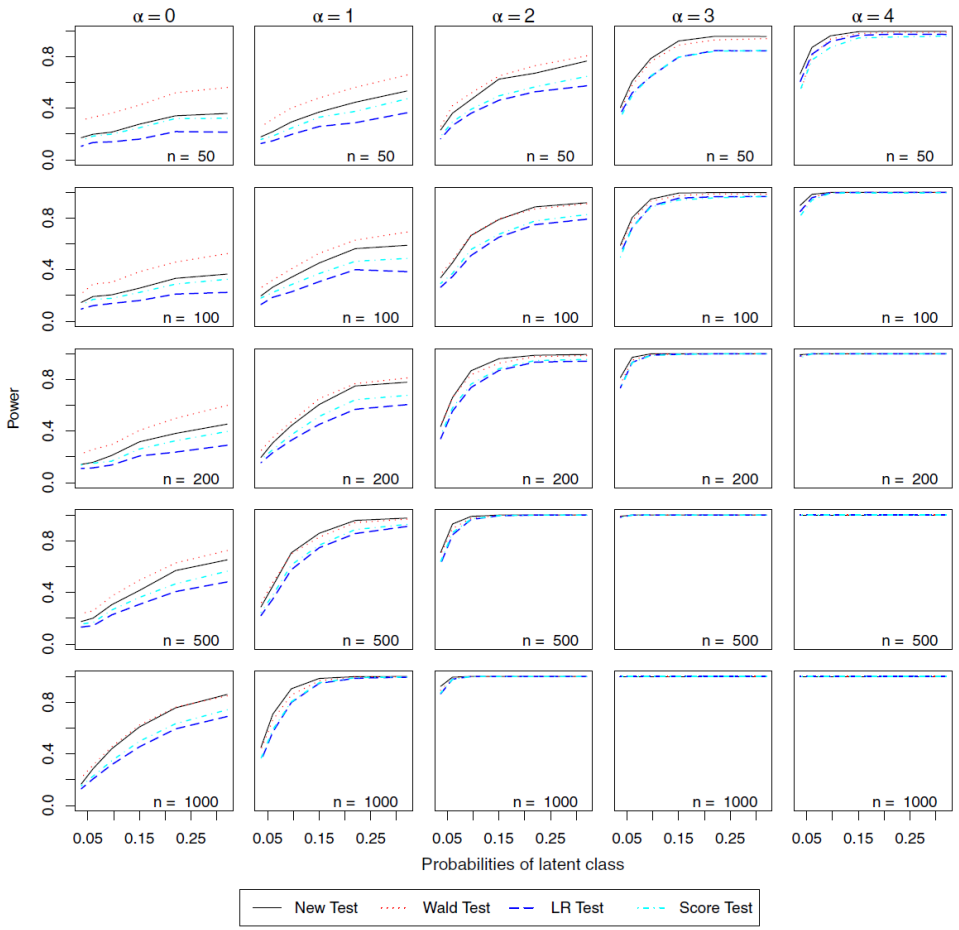

FIGURE 6 Power of detecting the latent class in mTobit model with uniformly distributed covariate for both the Tobit component and the latent class [Color figure can be viewed at wileyonlinelibrary.com]

Covariate x $\sim N(0, 2)$

- $y \sim mTobit(\omega, \mu, 4, -1), \mu = \alpha - x, \ \omega = 0.05k$

$$k = 1, 2, ..., 6, \ \alpha = 0, 1, 2, 3, 4$$

  - R Code
  - Power:Probabilities of latent class
  - Results

Covariate x $\sim N(0, 2)$(Power)

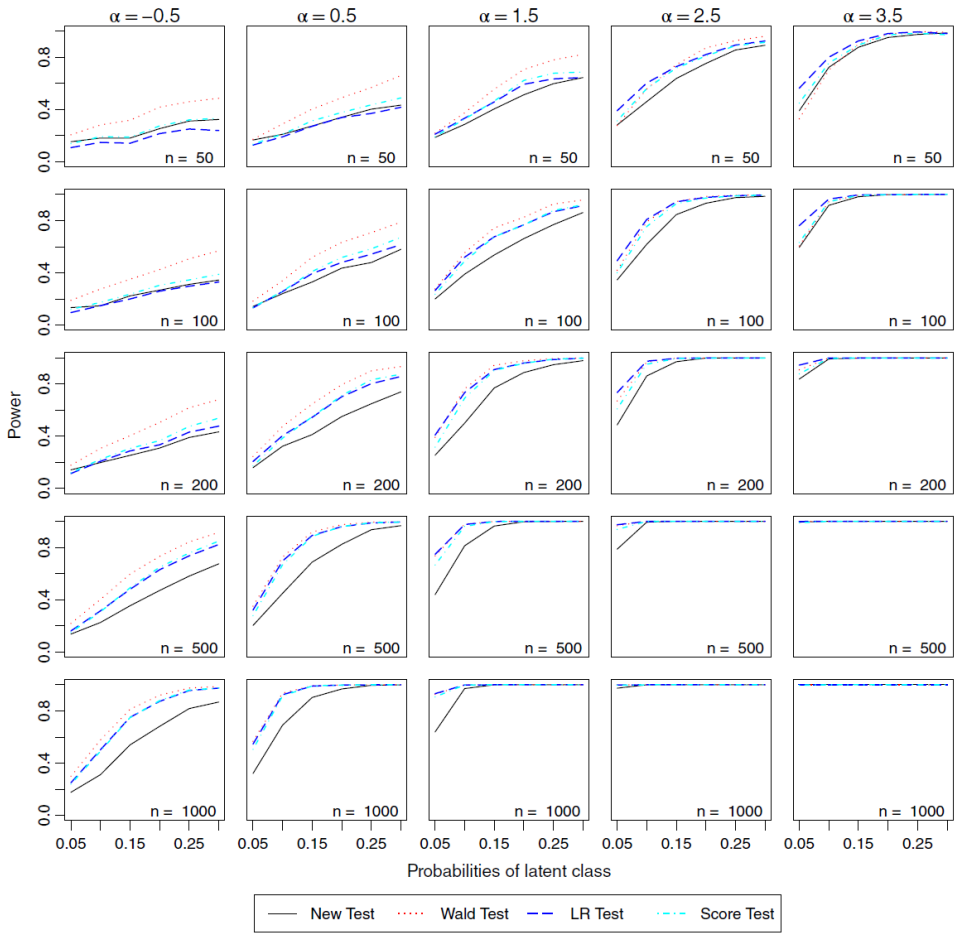

FIGURE 7 Power of detecting the latent class in mTobit model with normally distributed covariate for Tobit component only [Color figure can be viewed at wileyonlinelibrary.com]

Covariate x $\sim N(0, 2)$

- $y \sim mTobit(\omega, \mu, 4, -1), \ \ \mu = \alpha - x, \ \ logit(\omega = -b + x)$

$$\alpha = 0, 1, 2, 3, 4; \ \ b = 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$$

  - R Code
  - Power:Probabilities of latent class
  - Results

Covariate x $\sim N(0, 2)$(Power)

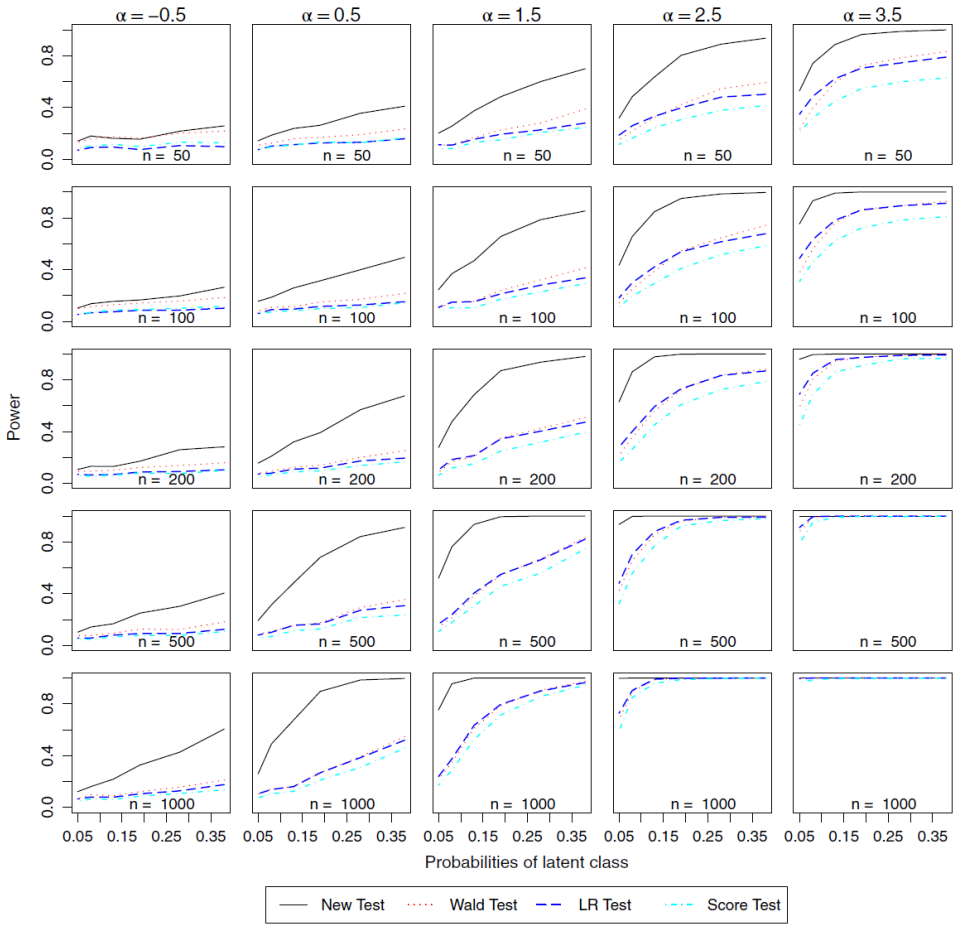

FIGURE 8 Power of detecting the latent class in mTobit model with normally distributed covariate for Tobit component and the latent class [Color figure can be viewed at wileyonlinelibrary.com]

目录

- 数据介绍
- 建立模型
- 假设检验
- 检验结果

数据介绍

- 使用从2003年至2010年的四次调查中收集的尿液三氯生浓度;
- 调查对象为3659 名儿童(6-19岁)和6566 名成人(20 岁);
- 以检验是否存在一个潜在的浓度类别，即是否存在一个尿液三氯生非接触组;

建立模型

- 对于转换后的数据，假设mTobit回归模型包含年龄、性别、种族、教育程度、BMI、尿可替宁和肌酐等协变量:

$$y_i \sim mTobit(\omega, \mu_i, \sigma^2, L = log(2.3))$$

$$\mu_i = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Race_i + \beta_4 Edu_i + \beta_5 BMI_i + \beta_6 Cotinine_i + \beta_7 Creatinine_i$$

假设检验

- 设立假设检验：

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

检验结果

- 对于$H_0 : \omega = 0 \ vs \ \omega > 0$:
  - Wald、LR、score 和new Test 分别代表的P值为：

$$0.5000, 0.5037, 0.768, 0.986$$

- 四个测试都得出了同样得结论，即没有显著证据表明能够拒绝H_0。
- 即尿中三氯生得浓度没有潜在得类别。

目录

- 方法评价
- 后期计划

方法评价

- 不需要确定的函数形式
- 犯第一类错误更少
- 统计功效更高

后期计划

- 对比研究
- 在纵向数据上的应用