含有测量误差与缺失值的纵向数据亚组分析方法的模拟研究*

复旦大学公共卫生学院生物统计学教研室(200032)　薛雅心　秦国友△

【提　要】 **目的**　研究可以同时处理协变量含有测量误差和响应变量含有缺失值的纵向数据下的亚组分析方法。**方法**　基于阈值回归模型进行亚组分析;利用重复测量之间的独立性来处理测量误差,并引入逆概率加权来处理缺失值,从而构造一个新的广义渐近无偏估计方程。**结果**　计算机随机模拟显示该估计方法在处理测量误差和缺失数据方面具有良好的效果,相比于未修正测量误差或缺失数据的广义估计方程方法具有更小的偏倚和均方误差。**结论**　亚组分析中,当协变量存在测量误差、响应变量存在缺失值时,通常需要考虑对测量误差和缺失值进行处理,以便得到可靠的参数估计。

【关键词】 亚组分析　纵向数据　广义估计方程　测量误差　缺失值
【中图分类号】 R195.1　　【文献标识码】 A　　**DOI** 10.11783/j.issn.1002-3674.2024.01.003

Simulation of Subgroup Analysis Methods with Longitudinal Data Containing Measurement Errors and Missingness

Xue Yaxin, Qin Guoyou (*Department of Biostatistics, School of Public Health, Fudan University* (200032), *Shanghai*)

【**Abstract**】 **Objective**　To develop a subgroup analysis method that can simultaneously deal with longitudinal data containing measurement errors and dropouts. **Methods**　Subgroup analysis was carried out based on a threshold regression model. A new generalized unbiased estimation equation is constructed by using the independence between repeated measurements to deal with measurement errors and introducing an inverse probability weighting matrix to deal with missing response. **Results**　The computer stochastic simulation shows that the proposed estimation method is effective in dealing with measurement errors and dropouts, and has smaller bias and mean square error than the generalized estimation equation method without correcting measurement errors or dropouts. **Conclusion**　In subgroup analysis, when there are measurement errors in covariables and missing values in response variables, it is usually necessary to deal with the measurement errors and missing values in order to obtain reliable parameter estimation.

【**Key words**】 Subgroup analysis; Longitudinal data; Generalized estimation equation; Measurement error; Missing value

在临床研究中评估新疗法对患者的效果时,由于人群疾病特征的异质性,不同患者对相同治疗的反应可能不同,从而产生了不同的亚组[1-2]。亚组分析,包括识别治疗敏感的亚组,检测亚组间治疗效果的差异,对于实现精准医疗具有重要意义[3-4]。现有文献中已提出许多亚组识别方法,如基于树的方法[5-6]、基于混合模型的方法[7-8]等。在实际应用中,可根据单个连续尺度的生物标志物是否超过某一阈值来定义两个亚组。对此 Ge 等[9]针对纵向数据提出了一类阈值线性边际模型,可用于确定形成亚组的一个连续基线协变量的分割点,并通过分析治疗与亚组间的相互作用来评估两个亚组间潜在疗效差异,极大似然方法被用于估计阈值参数及其他模型参数。

在实际问题中,含有测量误差和缺失值的数据较为常见,比如:在数据采集的过程中,存在由于测量仪器不准确、环境和研究人员不稳定而导致的测量误差问题;由于各种原因造成的失访而导致的数据缺失问题。然而,直接代入存在测量误差的协变量,或采用基于完全数据的模型来处理缺失数据,可能会导致有偏的结果[10]。关于如何处理纵向数据分析中的失访问题,人们提出了各种各样的方法。如逆概率加权方法[11]、双稳健估计[12-13]和多重稳健估计[14-15]。为了同时处理含有测量误差和缺失值的纵向数据,Qin 等[16]和 Lin 等[17]利用重复测量误差的独立性来处理测量误差,采用逆概率加权来处理响应变量的缺失值,并提出了一种新的广义估计方程来求得回归参数的估计。该方法的优点之一是,它可以在不假设测量误差分布的情况下构造出一个渐近无偏估计方程,可有效处理测量误差导致的偏性,且基于广义估计方程处理纵向数据,相比于混合效应模型,对模型的错误指定具有更强的稳健性[18]。

目前尚未有研究提出可以实现针对含有测量误差和缺失值的纵向数据的亚组分析方法。因此,本文将建立一个阈值线性边际模型,可同时确定一个可用于划分亚组的连续基线协变量的阈值,并根据亚组与干预间的交互作用来评估亚组间干预效果的差异。然后,利用重复测量误差之间的独立性来处理测量误差,并引入逆概率加权矩阵来处理响应变量的缺失

* 基金项目:国家自然科学基金(11871164)
△通信作者:秦国友,E-mail:gyqin@fudan.edu.cn

值，从而构造一个可同时处理测量误差和缺失值的渐近无偏广义估计方程。本文通过模拟研究，说明了本文提出的方法在处理同时含有测量误差和缺失值的纵向数据分析方面的优势。

模型及估计

1.阈值线性边际模型

考虑包含 n 个样本，各进行 m 次观测的纵向研究。令 Y_{ij} 和 $X_{ij}=(X_{ij,1},\cdots,X_{ij,p})'\in R^p$ 分别表示于 t_{ij} 时观测的响应变量和 p 维协变量，其中 $i=1,\cdots,n$；$j=1,\cdots,m$。不失一般性地，进一步令 $Y_i=(Y_{i1},\cdots,Y_{im})'$，$X_i=(X_{i1}',\cdots,X_{im}')'$。令 b_i 为第 i 个样本是否接受治疗的指示变量，如果第 i 个样本接受治疗，则 $b_i=1$，否则，$b_i=0$，并且 b_i 为协变量 X_i 的一部分。令 ω_i 为第 i 个样本的一个连续基线协变量，并假设可由 ω_i 是否超过未知阈值 c 来将样本分为两个亚组。假设各样本的响应变量间相互独立。

我们考虑如下阈值线性边际模型：

$$Y_i=X_i\beta+\eta_1 I(\omega_i>c)\vec{1}+\eta_2 b_i I(\omega_i>c)\vec{1}+\varepsilon_i \quad (1)$$

其中 $\beta\in R^p$，η_1，η_2 为未知回归参数，$I(\cdot)$ 为指示函数，即 $I(\omega_i>c)=\begin{cases}1 & if\ \omega_i>c\\ 0 & if\ \omega_i\leqslant c\end{cases}$，$\vec{1}$ 为由 1 组成的 m 维向量，$\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{im})'$ 为均值为零，协方差矩阵为 $V_i(\rho)$ 的独立随机误差。

为了简化模型，考虑

$$Y_i=X_i\beta+W_i\eta+\varepsilon_i \quad (2)$$

其中 $W_i=\begin{pmatrix} I(\omega_i>c) & b_i\times I(\omega_i>c)\\ \vdots & \vdots\\ I(\omega_i>c) & b_i\times I(\omega_i>c)\end{pmatrix}_{m\times 2}$，$\eta=(\eta_1,\eta_2)'$

2.测量误差模型

在实际应用中，协变量可能无法被精确观测。假设 X_{ij} 的观测存在测量误差，U_{ij} 为其观测值，可用如下经典可加误差模型[19]表示：

$$U_{ij}=X_{ij}+\delta_{ij} \quad (3)$$

其中测量误差 δ_{ij} 独立于 X_{ij} 和 ε_{ij}，假定 δ_{ij} 的均值为零，协方差矩阵为 V_m。

在实践中，通常为了得到更可靠的测量结果而进行重复测量，这里我们假设存在对协变量 X_{ij} 的两次重复观测，表示如下：

$$U_{ij(1)}=X_{ij}+\delta_{ij(1)},\ U_{ij(2)}=X_{ij}+\delta_{ij(2)} \quad (4)$$

其中 $\delta_{ij(1)}$ 与 $\delta_{ij(2)}$ 独立。

为了利用重复观测个体间的独立性构造一个渐近无偏估计方程，我们采用上述的两次独立可加测量误差模型。如果实际进行了超过两次的重复观测，可将这些观测分为两组，并分别定义两组的均值为 $U_{ij(1)}$ 和 $U_{ij(2)}$。

3.缺失模型

在实际研究中，除了协变量可能含有测量误差外，也常因失访而导致响应变量的缺失。考虑协变量 $D_i=\{U_i,\omega_i,b_i\}$ 可被完全观测，而响应变量 Y_{ij} 存在缺失。令 R_{ij} 为缺失的指示变量，即 $R_{ij}=1$ 表示 Y_{ij} 被观测到，$R_{ij}=0$ 表示 Y_{ij} 缺失。考虑由样本失访而导致的单调缺失数据的情况，即 $R_{ij}=0$ 意味着对任意的 $k>j$，有 $R_{ik}=0$。不失一般性地，假设 $R_{i1}=1$，$i=1,\cdots,n$。

由于协变量存在测量误差，我们通过观测数据对缺失数据过程进行建模。令 Y_i^{obs} 为 Y_i 的观测部分，$\tilde{R}_{ij}=\{R_{i1},\cdots,R_{i,j-1}\}$ 为第 i 个样本的第 j 个观测处表示历史缺失的指示变量，U_i 为包含 $U_{ij(1)}$ 和 $U_{ij(2)}$ 的协变量。我们假设缺失的条件概率仅取决于观测到的响应变量 Y_i^{obs} 和协变量 D_i，也可表示为

$$\Pr(R_{ij}=1\mid\tilde{R}_{ij},Y_i,D_i)=Pr(R_{ij}=1\mid\tilde{R}_{ij},Y_i^{obs},D_i)$$

在上述假设下，给定第 i 个样本的已观测到的协变量和响应变量，可观测到 Y_{ij} 的概率定义为 $\pi_{ij}=\Pr(R_{ij}=1\mid Y_i,D_i)$，在单调缺失的假设下，$\pi_{ij}=\prod_{k=1}^{j}\lambda_{ik}$，其中 $\lambda_{i1}=1$，$\lambda_{ij}=\Pr(R_{ij}=1\mid R_{i,j-1}=1,Y_i,D_i)$，$j=2,\cdots,m$。一个常用的估计 λ_{ij} 的方法是拟合 logistic 回归模型如下：

$$\ln\frac{\lambda_{ij}}{1-\lambda_{ij}}=Z'_{ij}\gamma \quad (5)$$

其中 $Z_{ij}=(Z_{ij1},\cdots,Z_{ijq})$ 包含协变量 $D_i=\{U_i,\omega_i,b_i\}$ 和已观测到的响应变量 Y_i^{obs} 的信息，γ 为 q 维回归参数向量。

4.参数估计

在上述假设下，由于指示函数 $I(\omega_i>c)$ 的存在，基于模型(2)给出的估计方程对 c 不连续，无法对 c 求导，传统的 Newton 迭代法不可行。基于 Brown 等[20]的光滑方法，我们采用 $\Phi[(\omega_i-c)/h]$ 作为指示函数 $I(\omega_i>c)$ 的光滑近似，其中 $\Phi(\cdot)$ 为标准正态分布函数，该函数对 c 连续且可微，其中 h 为带宽，随着样本量 n 逐渐增大，h 逼近于 0。根据 Lin 等[21]，我们选择带宽函数为 $h=\hat{d}(nm)^{-1/3}$，其中 $\hat{d}$ 为协变量 ω_i 的样本标准差。

将平滑近似代入模型(2)中，我们可得到如下平滑线性边际模型：

$$Y_i\approx X_i\beta+\tilde{W}_i\eta+\varepsilon_i \quad (6)$$

其中 $\tilde{W}_i=\begin{pmatrix}\Phi[(\omega_i-c)/h] & b_i\times\Phi[(\omega_i-c)/h]\\ \vdots & \vdots\\ \Phi[(\omega_i-c)/h] & b_i\times\Phi[(\omega_i-c)/h]\end{pmatrix}_{m\times 2}$

我们利用两次重复观测的测量误差之间的独立性来处理由协变量的测量误差所导致的偏倚，采用逆概率加权方法来修正由响应变量缺失所导致的选择偏倚。令 $\theta_1=(\beta',\eta')'$，$\theta=(\theta_1',c)$，未知参数 θ 可通

过求解下列估计方程来得到估计量：

$$E_{\theta,n}=\sum_{i=1}^{n}E_{\theta,i}=\sum_{i=1}^{n}(U_{i(1)},\widetilde{W}_i,\dot{W}_i\eta)'V_i^{-1}(\rho)S_i(\gamma)(Y_i-U_{i(2)}\beta-\widetilde{W}_i\eta)+(U_{i(2)},\widetilde{W}_i,\dot{W}_i\eta)'V_i^{-1}(\rho)S_i(\gamma)(Y_i-U_{i(1)}\beta-\widetilde{W}_i\eta)=0 \quad (7)$$

其中$\dot{W}_i=\frac{\partial W_i}{\partial c}$，$V_i(\rho)$为工作相关矩阵，其中相关系数$\rho$可表示各样本重复观测间的相关性，$S_i(\gamma)=diag\{\frac{R_{i1}}{\pi_{i1(\gamma)}},\cdots,\frac{R_{im}}{\pi_{im(\gamma)}}\}$为逆概率加权矩阵。

上述估计方程(7)包含了讨厌参数γ和ρ。我们采用最大似然估计方法来获得γ的稳健估计，似然函数如下：

$$\sum_{i=1}^{n}L_i(\gamma)=\prod_{i=1}^{n}\prod_{j=1}^{m}\{\lambda_{ij}(\gamma)^{R_{ij}}[1-\lambda_{ij}(\gamma)]^{1-R_{ij}}\}^{R_{i,j-1}}$$

或通过求解下述方程来得到$\hat{\gamma}$：

$$G_{\gamma,n}(\gamma)=\sum_{i=1}^{n}G_{\gamma,i}(\gamma)=\sum_{i=1}^{n}\frac{\partial\log L_i(\gamma)}{\partial_\gamma}=0 \quad (8)$$

对于相关参数ρ，基于给定的工作相关结构和各样本的两次观测值的残差，$\hat{e}_{i(1)}=(Y_{i1}-U_{i(1)}\beta-\hat{W}_i\eta)$，$\hat{e}_{i(2)}=(Y_{i2}-U_{i(2)}\beta-\hat{W}_i\eta)$，Wang 等[22]给出了完全数据下广义估计方程方法中求解讨厌参数的估计式。此处，相关参数ρ的估计应基于缺失数据的模式进行加权。例如，对 AR−1 相关结构，ρ可由下式进行估计：

$$\hat{\rho}=\frac{1}{n(m-1)-p-3}\sum_{i=1}^{n}\sum_{j\leqslant m-1}\frac{R_{ij}R_{i,j+1}}{\pi_{i,j+1}(\hat{\gamma})}\hat{e}_{ij(1)}\hat{e}_{i,j+1(2)}$$

基于上述讨论，估计程序说明如下：

步骤一：通过求解(8)获得估计量$\hat{\gamma}$并计算$\pi_{ij}(\hat{\gamma})$。

步骤二：假设重复观测相互独立，求解工作相关结构为独立相关结构下的广义估计方程(7)，计算出$\theta^{(0)}$作为回归参数初始值。

步骤三：基于给定的工作相关结构，按照上文提到的方法计算$\hat{\rho}$。

步骤四：进行下述迭代：

$$\theta^{(i+1)}=\theta^{(i)}-[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}E_{i,\theta|\theta=\theta^{(i)}}]^{-1}\sum_{i=1}^{n}E_{i,\theta|\theta=\theta^{(i)}} \quad (9)$$

步骤五：重复步骤三、四直到算法收敛，收敛条件如下：

$$\left\|[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}E_{i,\theta|\theta=\theta^{(i)}}]^{-1}\sum_{i=1}^{n}E_{i,\theta|\theta=\theta^{(i)}}\right\|<\epsilon$$

$\|.\|$表示欧几里得范数，$\epsilon=10^{-9}$。

模拟研究

1.模拟数据集的生成

考虑如下阈值线性边际模型

$$Y_{ij}=\beta_1X_{1,ij}+\beta_2b_i+\eta_1I(\omega_i>c)+\eta_2b_iI(\omega_i>c)+\varepsilon_{ij}$$

$i=1,\cdots,n;\ j=1,\cdots,m$

其中参数$\theta_0=(\beta_1,\beta_2,\eta_1,\eta_2,c)'=(1,1,1,1,0.8)'$，协变量$X_{1,ij}\sim$ Normal(0，1)，$b_i\sim$ Bernoulli(0.5)，$\omega_i\sim$ Uniform(0，1)。随机误差向量$\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{im})'$服从均值为零，协方差矩阵为$V_i(\rho)$的多元正态分布，其中$V_i(\rho)$为 AR−1 相关结构，$\rho=0.6$。本研究设置样本个数为$n=500$，1000，且各进行$m=5$次重复观测。

给定如下协变量测量误差模型，假设协变量$X_{1,ij}$存在两次重复观测，其替代变量$U_{ij(1)}$，$U_{ij(2)}$表示如下：

$$U_{ij(1)}=X_{1,ij}+\delta_{ij(1)},\ U_{ij(2)}=X_{1,ij}+\delta_{ij(2)}$$

其中$\delta_{ij(1)}$，$\delta_{ij(2)}$相互独立且服从均值为零，标准差为σ_m的正态分布。在模拟中，我们分别设定$\sigma_m=0.2$和$\sigma_m=0.4$，对应含误差的协变量X_1的方差σ_x^2可信度比$\frac{\sigma_x^2}{\sigma_x^2+\sigma_m^2}$分别为 0.961 和 0.862。

对于响应变量存在缺失的情况，缺失指示变量R_{ij}由下列 logistic 回归模型产生：

$$\ln\frac{\lambda_{ij}}{1-\lambda_{ij}}=\gamma_0+\gamma_1Y_{i,j-1}$$

其中$(\gamma_0,\gamma_1)'$取值分别设为$(1,0.5)'$和$(2,0.5)'$，分别将导致 35%和 18%的缺失。

2.评价标准

我们比较了本文提出的广义估计方程方法(GEEP)、考虑测量误差的完全数据下广义估计方程方法(C−GEEC)、未考虑测量误差的逆概率加权广义估计方程方法(NC−GEEW)和未考虑测量误差的完全数据下广义估计方程方法(NC−GEEC)。注意到在未考虑测量误差的形式中，含有测量误差的协变量的两次重复观测的均值被直接纳入到估计方程中。通过比较上述四种方法在协变量X_1所含测量误差的方差σ_m分别为 0.2 和 0.4，响应变量缺失率分别为 35%和 18%、样本量分别为 500 和 1000 的情况下的参数估计结果，可评估我们提出的方法在处理含有测量误差和缺失值的纵向数据亚组分析中的表现。

本研究将上述四种方法在不同样本量、不同响应缺失率、不同测量误差方差的情况下各进行 500 次模拟，计算每次模拟的参数估计值与真实值的差值，并根据收集来的偏倚值来绘制箱线图。箱线图的特质是可直观地显示出数据分布情况及离散程度，并可进行多组数据分布特征的比较，因此可以根据箱线图来对四种估计方法的表现进行分析比较。

箱子中间的一条线和菱形点分别表示 500 次模拟结果的偏倚的中位数和均值，可代表偏倚的平均水

平，根据中位数和菱形点与零值线的距离分析估计方法的效果，距离零值线越近，说明参数的估计值偏离真实值的程度越小，估计方法的效果越好。

箱子的上下边线分别表示 500 次模拟结果的偏倚的上四分位数和下四分位数，这意味着箱子部分代表了 50%的结果偏倚，因此箱子的宽度在一定程度上可反映参数估计偏倚的波动程度。箱子的宽度结合箱子在零值线上下分布的均衡程度，可直观展示出参数估计值均方误差的大小。宽度越窄，箱子在零值线上下分布越均衡，说明参数估计值越集中于真实值附近，估计方法的准确性越好。

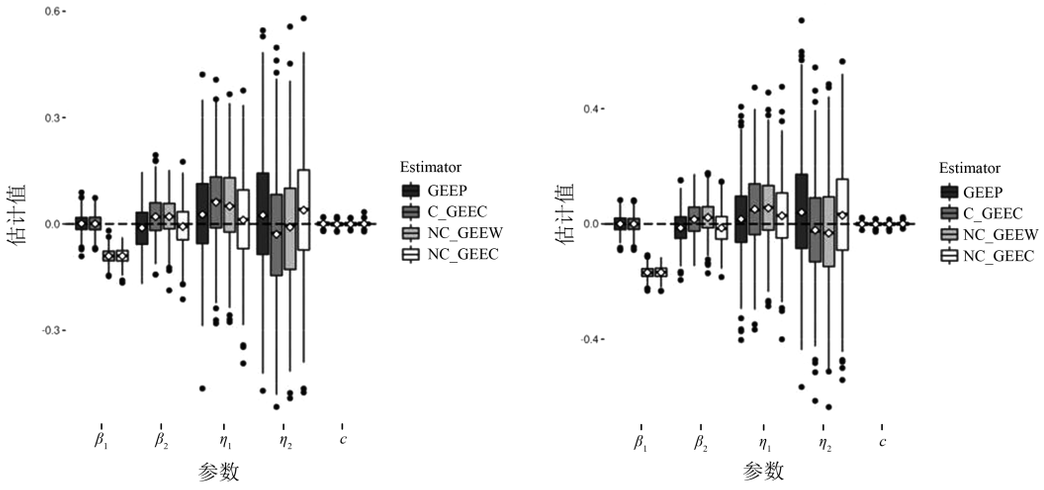
3.统计软件

整个模型的建立采用 R 软件进行实现，箱线图的绘制用 R 中的“ggplot2”程序包实现。

4.结果比较

图 1 和图 2 展示了样本量 $n=500$ 和 1000 情况下参数估计结果偏倚的箱线图。经过观察可明显看出，我们提出的方法（GEEP）所对应的箱子的中线和菱形点均靠近零值线，且箱子在零值线上下分布较为均衡且宽度较窄，说明该方法所得到的参数估计结果具有更小的偏倚，更加集中于真实值附近。

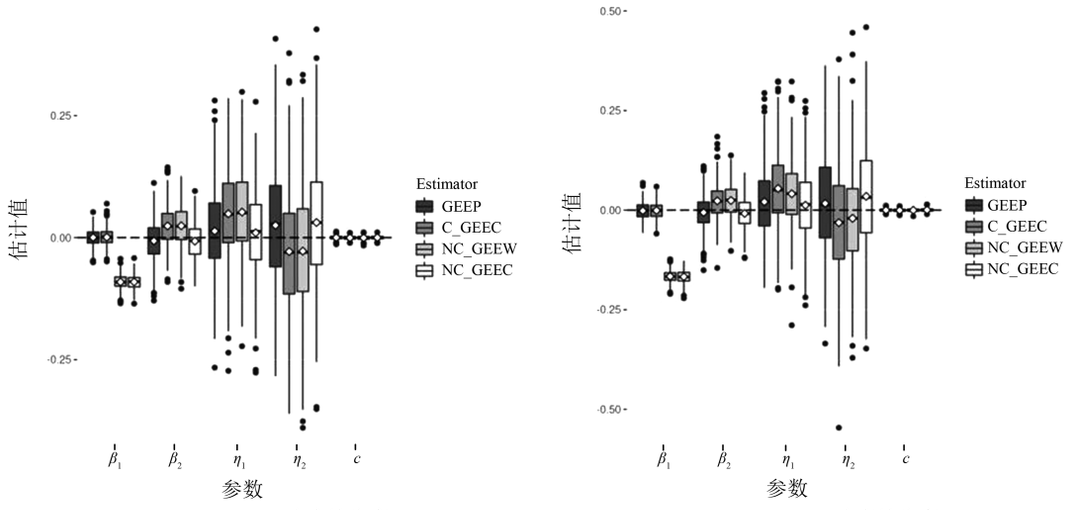
分析箱线图可看出，未考虑测量误差的方法（NC-GEEW，NC-GEEC）对应的箱子中线及菱形点距离零值线较远，尤其在本模拟中设定为存在测量误差的协变量X_1对应的回归参数β_1对应的箱子已完全偏离零值线，显示出极为明显的偏倚，且当测量误差的标准差σ_m增大时，NC-GEEW 和 NC-GEEC 方法的估计结果偏倚更为明显，但 GEEP 估计结果的偏倚仍较小。模拟数值结果显示，当样本量 $n=500$，响应缺失率为 18%，$\sigma_m=0.2$ 时，对参数β_1的估计方面，GEEP 估计出的$\hat{\beta}_1$的相对误差不足 1%，而 NC-GEEW 和 NC-GEEC 估计出$\hat{\beta}_1$的相对误差均达到了 9%；而其他条件不变，将σ_m增加到 0.4 时，两种未修正测量误差的方法（NC-GEEW，NC-GEEC）估计出的$\hat{\beta}_1$的相对误差达到约 17%，但 GEEP 估计出的$\hat{\beta}_1$的相对误差仍不足 1%。



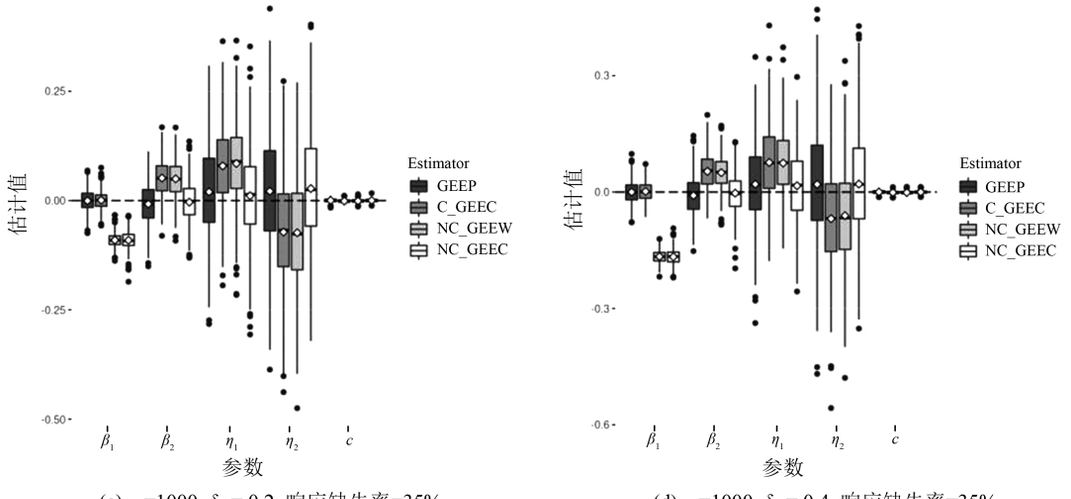
(a) n=500,δ_m=0.2,响应缺失率=18%　(b) n=500,δ_m=0.4,响应缺失率=18%

(c) n=500,δ_m=0.2,响应缺失率=35%　(d) n=500,δ_m=0.4,响应缺失率=35%

图 1　参数估计结果偏倚的箱线图（样本量 $n=500$）



(a) n=1000, δ_m = 0.2,响应缺失率=18%　　(b) n=1000, δ_m = 0.4,响应缺失率=18%



(c) n=1000, δ_m = 0.2, 响应缺失率=35%　　(d) n=1000, δ_m = 0.4, 响应缺失率=35%

图 2　参数估计结果偏倚的箱线图，样本量 n = 1000

此外，由箱线图还可看出，相比于 GEEP 方法，完全数据下的广义估计方程方法（C-GEEC，NC-GEEC）对应的箱子中线及菱形点偏离零值线较远，且箱子在零值线上下分布不均衡，说明 C-GEEC 和 NC-GEEC 的估计结果具有较大的偏倚和均方误差，特别是当缺失率变大的时候。模拟数值结果显示，当样本量 n = 500，σ_m = 0.2，响应缺失率为 35%时，GEEP 对各参数估计的相对误差均在 3%以下，而 C-GEEC 和 NC-GEEC 估计结果的相对误差较大，例如这两种方法估计出的$\hat{\beta}_2$，$\hat{\eta}_1$，$\hat{\eta}_2$的相对误差分别超过 5%、8%和 7%。

因此，我们提出的方法（GEEP）在四种方法中表现最佳，并体现了它在同时处理测量误差和缺失值的纵向数据下的优势。

讨　论

本文提出了一个阈值线性模型，既包含含有测量误差的协变量，也包含含有缺失值的响应变量，根据单个连续基线协变量是否超过某一阈值来定义亚组，并通过亚组与治疗间的交互作用来确定亚组间疗效差异；本文建立了一个新的渐近无偏广义估计方程，可估计阈值参数及回归参数，并将该方法与完全数据下的广义估计方程方法及未考虑测量误差的广义估计方程方法进行了对比。

在模拟研究部分，发现我们提出的方法（GEEP）的参数估计结果具有更小的偏倚和均方误差。协变量存在测量误差时，未修正测量误差的方法（NC-GEEW，NC-GEEC）会出现明显的偏倚，并随着测量误差的方差变大而愈加显著，但我们提出的方法偏倚始终较小；响应变量存在缺失值时，完全数据下的估计方程方法（C-GEEC，NC-GEEC）的偏倚较大，特别是当缺失率变大的时候，但我们提出的方法仍具有较小的偏倚和均方误差。因此，我们提出的方法在四种方法中表现最佳，并体现了它在同时处理含有测量误差和缺失值的纵向数据下的优势。

本文提出的方法适用于响应变量随机缺失（missing at random，MAR）的假设，但在实际应用中，观测值可能不是随机缺失的。而纵向数据分析中处理可能

非随机缺失的数据还需要进一步加以解决。另外，本文中只考虑了基于单一协变量来确定亚组的情况，而在实际中可能有多个协变量可用，将它们结合起来可能改善亚组的识别。

参 考 文 献

[1] Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA, 2007, 298(10): 1209-1212.

[2] Kent DM, Rothwell PM, Ioannidis J, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials, 2010, 11(1): 1-11.

[3] Lipkovich I, Dmitrienko A, B D'Agostino Sr R. Tutorial in biostatistics: data - driven subgroup identification and analysis in clinical trials. Statistics in Medicine, 2017, 36(1): 136-196.

[4] Dong J, Zhang JL, Zeng S, et al. Subgroup balancing propensity score. Statistical Methods in Medical Research, 2020, 29(3): 659-676.

[5] Lipkovich I, Dmitrienko A, Denne J, et al. Subgroup identification based on differential effect search--a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine, 2011, 30(21): 2601-2621.

[6] Foster JC, Taylor J, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in Medicine, 2011, 30(24): 2867-2880.

[7] Mcnicholas PD. Model-based classification using latent Gaussian mixture models. Journal of Statistical Planning & Inference, 2010, 140(5): 1175-1181.

[8] Shen J, He X. Inference for subgroup analysis with a structured logistic-normal mixture model. Journal of the American Statistical Association, 2015, 110(509): 303-312.

[9] Ge X, Peng Y, Tu D. A threshold linear mixed model for identification of treatmentsensitive subsets in a clinical trial based on longitudinal outcomes and a continuous covariate. Statistical Methods in Medical Research, 2020, 29(10): 2919-2931.

[10] Xu HX, Fan GL, Wang JF. Jackknife empirical likelihood for the error variance in linear errors-in-variables models with missing data. Communications in Statistics-Theory and Methods, 2022, 51(14): 4827-4840.

[11] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 1995, 90(429): 106-121.

[12] Lin H, Fu B, Qin G, et al. Doubly robust estimation of generalized partial linear models for longitudinal data with dropouts. Biometrics, 2017, 73(4): 1132-1139.

[13] Wei K, Qin G, Zhang J, et al. Doubly robust estimation in causal inference with missing outcomes: With an application to the Aerobics Center Longitudinal Study. Computational Statistics & Data Analysis, 2022, 168: 107399.

[14] Lu W, Qin G, Zhu Z, et al. Multiply robust subgroup identification for longitudinal data with dropouts via median regression. Journal of Multivariate Analysis, 2021, 181: 104691.

[15] Wei K, Zhu H, Qin G, et al. Multiply robust subgroup analysis based on a single - index threshold linear marginal model for longitudinal data with dropouts. Statistics in Medicine, 2022, 41(15): 2822-2839.

[16] Qin G, Zhang J, Zhu Z, et al. Robust estimation of partially linear models for longitudinal data with dropouts and measurement error. Statistics in Medicine, 2016, 35(29): 5401-5416.

[17] Lin H, Qin G, Zhang J, et al. Analysis of longitudinal data with covariate measurement error and missing responses: An improved unbiased estimating equation. Computational Statistics & Data Analysis, 2018, 121: 104-112.

[18] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika, 1986, 73(1): 13-22.

[19] Fuller WA. Measurement error models. John Wiley & Sons, 2009.

[20] Brown BM, Wang YG. Induced smoothing for rank regression with censored survival times. Statistics in Medicine, 2007, 26(4): 828-836.

[21] Lin H, Zhou L, Peng H, et al. Selection and combination of biomarkers using ROC method for disease classification and prediction. Canadian Journal of Statistics, 2011, 39(2): 324-343.

[22] Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. Advances in Statistics, 2014(1): 1-11.

（责任编辑：邓 妍）

（上接第 11 页）

[4] Zhou Y, Lee J, Yuan Y. A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. Statistics in Medicine, 2019, 38(28): 5299-5316.

[5] Lin R, Zhou Y, Yan F, et al. BOIN12: Bayesian optimal interval phase I/II trial design for utility-based dose finding in immunotherapy and targeted therapies. JCO Precision Oncology, 2020, 4: 1393-1402.

[6] Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics, 2000, 56: 1177-1182.

[7] Liu SY, Johnson VE. A robust Bayesian dose-finding design for phase I/II clinical trials. Biostatistics, 2015, 17(2): 249-263.

[8] Yuan Y, Yin G. Robust EM continual reassessment method in oncology dose finding. J Am Stat Assoc, 2011, 106: 818-831.

[9] Yuan Y, Lin R, Li D, et al. Time-to-event Bayesian optimal interval design to accelerate phase I trials. Clinical Cancer Research, 2018, 24(20): 4921-4930.

[10] Zhang YF, Zang Y. CWL: A conditional weighted likelihood method to account for the delayed joint toxicity-efficacy outcomes for phase I/II clinical trials. Statistical Methods in Medical Research, 2020, 30(3): 892-903.

[11] Hoering A, Thall PF, Nguyen H, et al. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. Biometrics, 66(2): 532-540.

（责任编辑：张 悦）