Statistics and Application 统计学与应用, 2026, 15(4), 290-300
Published Online April 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.154091

Hans汉斯

响应变量缺失下部分线性单指标模型的经验似然推断

黄培耘*，韩知良，郑小平，郑广豪

重庆工商大学数学与统计学院，重庆

收稿日期：2026年3月22日；录用日期：2026年4月12日；发布日期：2026年4月29日

摘　要

本文研究响应变量随机缺失下部分线性单指标模型的经验似然推断问题。首先，基于响应变量随机缺失机制，分别采用逆概率加权方法和增广借补方法，构造了两类纠偏的参数向量的经验对数似然比统计量。然后，在适当的正则条件下，证明了上述两类统计量均渐近服从标准卡方分布，进而建立了参数向量的经验似然置信域。最后，给出了参数向量的极大经验似然估计量和链接函数估计量的渐近分布。Monte Carlo模拟研究表明，与逆概率加权经验似然方法相比，增广借补经验似然方法在有限样本条件下表现更好。

关键词

随机缺失，部分线性单指标模型，经验似然，增广借补方法，置信域

Empirical Likelihood Inference for Partially Linear Single-Index Models with Missing Responses

Peiyun Huang*, Zhiliang Han, Xiaoping Zheng, Guanghao Zheng

School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing

Received: March 22, 2026; accepted: April 12, 2026; published: April 29, 2026

Abstract

This paper investigates empirical likelihood inference for partially linear single-index models with

*通讯作者。

文章引用: 黄培耘, 韩知良, 郑小平, 郑广豪. 响应变量缺失下部分线性单指标模型的经验似然推断[J]. 统计学与应用, 2026, 15(4): 290-300. DOI: 10.12677/sa.2026.154091

randomly missing response variables. First, under the missing-at-random mechanism, two types of bias-corrected empirical log-likelihood ratio statistics for the parameter vector are constructed based on the inverse probability weighting method and the augmented imputation method, respectively. Then, under appropriate regularity conditions, it is shown that both types of statistics asymptotically follow the standard chi-square distribution, and consequently empirical likelihood confidence regions for the parameter vector are established. Finally, the asymptotic distributions of the maximum empirical likelihood estimator of the parameter vector and the estimator of the link function are derived. Monte Carlo simulation studies indicate that, compared with the inverse probability weighting empirical likelihood method, the augmented imputation empirical likelihood method performs better in finite samples.

Keywords

Missing at Random, Partially Linear Single-Index Model, Empirical Likelihood, Augmented Imputation Method, Confidence Region

1. 引言

部分线性单指标模型是一种重要的半参数模型，在经济金融，生物医药及人文社科等领域被广泛应用。考虑如下部分线性单指标模型：

$$Y = g\left(\beta^{\mathrm{T}} X\right) + \theta^{\mathrm{T}} Z + \varepsilon \tag{1.1}$$

其中 $X \in R^p$, $Z \in R^q$ 是协变量，Y 是响应变量，$\beta = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^{\mathrm{T}}$，$\theta = \left(\theta_1, \theta_2, \cdots, \theta_q\right)^{\mathrm{T}}$ 分别是 p 维和 q 维未知参数向量，$g(\cdot)$ 是一元未知链接函数，ε 是随机误差，$E\left(\varepsilon \middle| X, Z\right) = 0$。为了模型的可识别性，假定 $\|\beta\| = 1$ 且 β 的第一个非零元素大于 0。Ichimur [1]研究了单指标模型的可识别性；Härdle 等[2]和 Weisberg 等[3]提出了基于半参数最小二乘的指标系数估计方法，并在一定正则条件下给出了该估计的收敛速度；Zhu 和 Xue [4]研究了部分线性单指标模型中的经验似然推断问题，提出了一种用于校正经验似然偏差的纠偏方法；Xue [5]提出了一种两阶段方法用于估计部分线性单指标变系数模型的未知参数和系数函数，并证明了估计量的大样本性质。

在医学分析、社会经济调查和其它科学研究领域，由于调查实施过程中存在许多不可控因素，数据缺失现象普遍存在。Robins 等[6]在 Rubin 提出的随机缺失机制下，基于逆概率加权估计方程构建了新的半参数高效估计方法，实现了在缺失概率已知条件下参数向量的一致估计；Wang 和 Rao [7]在响应变量缺失情形下，提出了基于借补方法的非参数回归的经验似然推断，并给出了相应估计量的渐近性质；Wang 和 Sun [8]分别利用半参数回归替代方法、逆概率加权方法和借补方法，估计了部分线性模型的回归系数和非参数函数；Xue [9]在协变量随机缺失的情形下，提出了基于单指标模型的经验似然方法，并研究了估计量的渐近性质；陈盼盼等[10]基于逆概率加权方法，研究了协变量缺失下半参数变系数模型的估计问题。

Xue 和 Lian [11]基于响应变量随机缺失下的单指标模型，提出了修正偏差的加权经验似然推断方法，有效解决了缺失响应导致的估计偏差问题。Xue 和 Zhang [12]则进一步将研究拓展到部分线性单指标模型，在数据随机缺失的情形下构造了两类纠偏的经验似然比统计量，并证明了两类统计量均渐近服从卡

方分布。受他们的启发，本文在响应变量随机缺失的情形下，分别基于逆概率加权方法和增广借补方法，构造参数分量的两类纠偏的经验对数似然比统计量，并在适当的条件下证明所提出的两类统计量均渐近服从标准卡方分布。

2. 估计方法与主要结论

2.1. 逆概率加权经验似然

假设 $\{(Y_i,X_i,Z_i,\delta_i)\}_{i=1}^n$ 为一组来自模型(1.1)的不完全观测随机样本。当 $\delta_i=0$ 时，表示响应变量 Y_i 缺失，当 $\delta_i=1$ 时，表示响应变量 Y_i 被观测。考虑 lai 等[13]提出的 Y_i 的随机缺失机制：

$$P(\delta_i=1|Y_i,X_i,Z_i)=P(\delta_i=1|X_i,Z_i)=\pi(X_i,Z_i) \tag{2.1}$$

上式表明，在给定 X_i,Z_i 的条件下，Y_i 和 δ_i 是条件独立的。其中称 $\pi(X_i,Z_i)$ 为选择概率函数。

在构造经验似然比统计量时，需要使用 $g(\beta^{\mathrm{T}}X)$ 关于 β 每一分量的偏导数，而 $\|\beta\|=1$ 表明 β 位于单位超球面上，从而 $g(\beta^{\mathrm{T}}X)$ 在 β 处不存在偏导数。为此，本文采取广泛使用的“去一分量法”对 β 再参数化。设 $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$，$\beta^{(r)}=(\beta_1,\cdots,\beta_{r-1},\beta_{r+1},\cdots,\beta_p)^{\mathrm{T}}$ 是去掉 β 的第 r 个非零分量 β_r 后的 $p-1$ 维向量。不失一般性，设 $\beta_r>0$，则 $\|\beta^{(r)}\|<1$。

这意味着 β 在 $\beta^{(r)}$ 的某个邻域内无穷多次可微，记其 Jacobian 矩阵为

$$J_{\beta^{(r)}}=\frac{\partial\beta}{\partial\beta^{(r)}}=(\gamma_1,\gamma_2,\cdots,\gamma_p)^{\mathrm{T}},$$

其中 $\gamma_s\ (1\le s\le p,s\ne r)$ 是第 s 个元素为 1 的 $p-1$ 维单位向量，$\gamma_r=-\left(1-\|\beta^{(r)}\|^2\right)^{-1/2}\beta^{(r)}$.

为表示方便，令 $V^{\mathrm{T}}=(X^{\mathrm{T}},Z^{\mathrm{T}})^{\mathrm{T}}$，$d=p+q$，受 Xue 和 Zhang [12]的启发，定义纠偏的加权辅助随机向量为

$$\eta_{i,\mathrm{W}}(\beta^{(r)},\theta)=\frac{\delta_i}{\pi(V_i)}\xi_i(\beta^{(r)},\theta), \tag{2.2}$$

其中

$$\xi_i(\beta^{(r)},\theta)=\omega(\beta^{\mathrm{T}}X_i)\{Y_i-g(\beta^{\mathrm{T}}X_i)-\theta^{\mathrm{T}}Z_i\}\times\begin{pmatrix}g'(\beta^{\mathrm{T}}X_i)J_{\beta^{(r)}}^{\mathrm{T}}\{X_i-g_1(\beta^{\mathrm{T}}X_i)\}\\ Z_i-g_2(\beta^{\mathrm{T}}X_i)\end{pmatrix}. \tag{2.3}$$

$\omega(\cdot)$ 是一个具有紧支撑 u_w 的非负有界权重函数，$g'(\cdot)$ 是 $g(\cdot)$ 的导数，$g_1(u)=E(X|\beta^{\mathrm{T}}X=u)$，$g_2(u)=E(Z|\beta^{\top}X=u)$.

若 $(\beta^{(r)},\theta)$ 为参数的真实值，那么对于 $1\le i\le n$，有 $E\{\eta_{i,\mathrm{W}}(\beta^{(r)},\theta)\}=0$ 成立。因此，定义纠偏的加权经验对数似然比函数为

$$R_W(\beta^{(r)},\theta)=-2\max\left\{\sum_{i=1}^n\log(np_i):p_i\ge0,\sum_{i=1}^n p_i=1,\sum_{i=1}^n p_i\eta_{i,W}(\beta^{(r)},\theta)=0\right\}. \tag{2.4}$$

然而，(2.4)式中的 $\pi(\cdot),g(\cdot),g'(\cdot),g_1(\cdot)$ 和 $g_2(\cdot)$ 均为未知函数，不能直接使用 $R_W(\beta^{(r)},\theta)$ 来构造参数向量 $(\beta^{(r)},\theta)$ 的置信域，常用它们的估计量来替代上面的未知函数。首先，$\pi(v)$ 的估计量被定义为

$$\hat{\pi}(v)=\sum_{i=1}^n W_{ni}^*(v)\delta_i, \tag{2.5}$$

$$W_{ni}^{*}(x)=\frac{K^{*}\left(\dfrac{V_i-v}{h_1}\right)}{\sum_{j=1}^{n}K^{*}\left(\dfrac{V_j-v}{h_1}\right)}, \tag{2.6}$$

其中 $K^{*}(\cdot)$ 是定义在 R^{d} 上的核函数，$h_1=h_1(n)$ 是合适的带宽。

然后，基于 Fan 和 Gijbels [14]提出的局部线性估计方法，给出 $g(\cdot)$ 和 $g'(\cdot)$ 的估计量。对于任意固定的 (β,θ)，我们利用加权最小二乘方法计算 a,b, 即最小化如下目标函数

$$\min_{a,b}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i(V_i)}\left[Y_i-\theta^{\mathrm{T}}Z_i-a-b\left(\beta^{\mathrm{T}}X_i-u\right)\right]^{2}K_{h2}\left(\beta^{\mathrm{T}}X_i-u\right), \tag{2.7}$$

其中 $K_{h_2}(\cdot)=h_2^{-1}K(\cdot/h_2)$，$K(\cdot)$ 是定义在 R 上的核函数，$h_2=h_2(n)$ 是合适的带宽。记 $(\hat{a},\hat{b})$ 是使(2.7)式达到最小的解，经过一系列计算，不难得到

$$\hat{g}(u;\beta,\theta)=\hat{a}=\sum_{i=1}^{n}W_{ni}(u;\beta)\left(Y_i-\theta^{\mathrm{T}}Z_i\right), \tag{2.8}$$

$$\hat{g}'(u;\beta,\theta)=\hat{b}=\sum_{i=1}^{n}\widetilde{W_{ni}}(u;\beta)\left(Y_i-\theta^{\mathrm{T}}Z_i\right), \tag{2.9}$$

其中

$$W_{ni}(u;\beta)=\frac{n^{-1}\left\{\delta_i/\hat{\pi}(V_i)\right\}K_{h_2}\left(\beta^{\mathrm{T}}X_i-u\right)\left\{S_{n,2}(u;\beta)-\left(\beta^{\mathrm{T}}X_i-u\right)S_{n,1}(u;\beta)\right\}}{S_{n,0}(u;\beta)S_{n,2}(u;\beta)-S_{n,1}^{2}(u;\beta)}, \tag{2.10}$$

$$\widetilde{W_{ni}}=\frac{n^{-1}\left\{\delta_i/\hat{\pi}(V_i)\right\}K_{h_2}\left(\beta^{\mathrm{T}}X_i-u\right)\left\{\left(\beta^{\mathrm{T}}X_i-u\right)S_{n,0}(u;\beta)-S_{n,1}(u;\beta)\right\}}{S_{n,0}(u;\beta)S_{n,2}(u;\beta)-S_{n,1}^{2}(u;\beta)}, \tag{2.11}$$

$$S_{n,l}(u;\beta)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(V)_i}\left(\beta^{\mathrm{T}}X_i-u\right)^{l}K_{h_2}\left(\beta^{\mathrm{T}}X_i-u\right),\quad l=0,1,2. \tag{2.12}$$

进而，分别给出 $g_1(\cdot)$ 和 $g_2(\cdot)$ 的估计量

$$\hat{g}_1(u;\beta)=\sum_{i=1}^{n}W_{ni}(u;\beta)X_i, \tag{2.13}$$

和

$$\hat{g}_2(u;\beta)=\sum_{i=1}^{n}W_{ni}(u;\beta)Z_i. \tag{2.14}$$

最后，$\omega(\cdot)$ 的估计量为 $\hat{\omega}\left(\beta^{\mathrm{T}}x\right)=I\left\{\hat{p}\left(\beta^{\mathrm{T}}x\right)\geq c\right\}$，其中 $\hat{p}(\cdot)$ 是 $\beta^{\mathrm{T}}X$ 的概率密度函数 $p(\cdot)$ 的核估计量，定义为

$$\hat{p}(u)=\frac{1}{nh_3}\sum_{i=1}^{n}K'\left(\frac{\beta^{\mathrm{T}}X_i-u}{h_3}\right), \tag{2.15}$$

其中 $K'(\cdot)$ 是定义在 R 上的核函数，$h_3=h_3(n)$ 是合适的带宽。在实际应用中，由于估计结果对于权重函数的选择并不敏感，通常取 $\hat{\omega}(\cdot)=1$ [11]。

把上述估计量分别代入(2.2)式，(2.3)式和(2.4)式，重新定义纠偏的加权辅助随机向量和纠偏的加权经验对数似然比函数为

$$\hat{\eta}_{i,\mathrm{W}}\left(\beta^{(r)},\theta\right)=\frac{\delta_i}{\hat{\pi}\left(V_i\right)}\hat{\xi}_i\left(\beta^{(r)},\theta\right), \tag{2.16}$$

和

$$\hat{R}_W\left(\beta^{(r)},\theta\right)=-2\max\left\{\sum_{i=1}^{n}\log\left(np_i\right):p_i\geq 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\hat{\eta}_{i,W}\left(\beta^{(r)},\theta\right)=0\right\}. \tag{2.17}$$

其中

$$\hat{\xi}_i\left(\beta^{(r)},\theta\right)=\hat{\omega}\left(\beta^{\mathrm{T}}X_i\right)\left\{Y_i-\theta^{\mathrm{T}}Z_i-\hat{g}\left(\beta^{\mathrm{T}}X_i;\beta,\theta\right)\right\}\times\begin{pmatrix}\hat{g}'\left(\beta^{\mathrm{T}}X_i;\beta,\theta\right)J_{\beta^{(r)}}^{\mathrm{T}}\left\{X_i-\hat{g}_1\left(\beta^{\mathrm{T}}X_i;\beta\right)\right\}\\ Z_i-\hat{g}_2\left(\beta^{\mathrm{T}}X_i;\beta\right)\end{pmatrix}. \tag{2.18}$$

2.2. 增广借补经验似然

在第 2.1 节中，由于定义的纠偏加权经验对数似然比函数仅用到了可以完全观测的数据，而没有考虑缺失数据所包含的信息。因此，当大量数据缺失时，基于上述纠偏加权经验似然方法得到的置信域往往具有相对较低的置信域精度。然而，若采用常规借补方法加以改进，则会破坏经验似然比统计量的 Wilks 性质，使其渐近分布不再服从标准卡方分布。综上，参考 Robins 等[6]和 Qin 等[15]处理缺失数据与构造经验似然的思想，定义如下增广函数

$$m(v)=m\left(v;\beta^{(r)},\theta\right)=E\left\{\hat{\xi}_i\left(\beta^{(r)},\theta\right)|V_i=v\right\}, \tag{2.19}$$

其估计量为

$$\hat{m}\left(v\right)=\hat{m}\left(v;\beta^{(r)},\theta\right)=\sum_{i=1}^{n}W_{ni}\left(v\right)\hat{\xi}_i\left(\beta^{(r)},\theta\right). \tag{2.20}$$

其中

$$W_{ni}\left(v\right)=\frac{\delta_i\mathcal{K}\left(\dfrac{V_i-v}{h_4}\right)}{\sum_{j=1}^{n}\delta_j\mathcal{K}\left(\dfrac{V_j-v}{h_4}\right)}, \tag{2.21}$$

$\mathcal{K}(\cdot)$ 是定义在 R^d 上的核函数，$h_4=h_4(n)$ 是合适的带宽。为此，利用 $\hat{m}\left(V_i;\beta^{(r)},\theta\right)$ 对 $\hat{\xi}_i\left(\beta^{(r)},\theta\right)$ 进行增广借补，定义一个调整的辅助随机向量

$$\hat{\eta}_{i,I}\left(\beta^{(r)},\theta\right)=\frac{\delta_i}{\hat{\pi}\left(V_i\right)}\hat{\xi}_i\left(\beta^{(r)},\theta\right)+\left\{1-\frac{\delta_i}{\hat{\pi}\left(V_i\right)}\right\}\hat{m}\left(V_i;\beta^{(r)},\theta\right). \tag{2.22}$$

因此，定义纠偏的增广借补经验对数似然比函数为

$$\hat{R}_{\mathrm{I}}\left(\beta^{(r)},\theta\right)=-2\max\left\{\sum_{i=1}^{n}\log\left(np_i\right)\middle|p_i\geq 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\hat{\eta}_{i,\mathrm{I}}\left(\beta^{(r)},\theta\right)=0\right\}. \tag{2.23}$$

由 Lagrange 乘数法，$\hat{R}_{\mathrm{I}}\left(\beta^{(r)},\theta\right)$ 可以表示为

$$\hat{R}_{\mathrm{I}}\left(\beta^{(r)},\theta\right)=2\sum_{i=1}^{n}\log\left(1+\lambda^{\top}\hat{\eta}_{i,\mathrm{I}}\left(\beta^{(r)},\theta\right)\right), \tag{2.24}$$

其中 $\lambda\in R^{p+q-1}$ 是下面方程的解

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\eta}_{i,\mathrm{I}}\left(\beta^{(r)},\theta\right)}{1+\lambda^{\mathrm{T}}\hat{\eta}_{i,\mathrm{I}}\left(\beta^{(r)},\theta\right)}=0. \tag{2.25}$$

2.3. 渐近性质

为证明本文的主要定理，参照文献[12]的设定，首先给出一些正则条件：

C1：随机变量 $\beta^{\mathrm{T}}X$ 的密度函数 $p(u)$ 在其支撑集 $\mathcal{U}_w$ 上远离 0 与无穷，且 $p(u)$ 和 $q(u)$ 均满足一阶 Lipschitz 连续条件，其中 $q(u)=E\left(\varepsilon^2/\pi(V)\middle|\beta^{\mathrm{T}}X=u\right)$。

C2：函数 $g(u),g_{1s}(u)$ 和 $g_{2s}(u)$ 在 $\mathcal{U}_w$ 上具有有界的二阶连续导数，其中 $g_{1s}(u),g_{2s}(u)$ 分别表示 $g_1(u),g_2(u)$ 的第 s 个分量。

C3：核函数 $K(u)$ 是具有有界导数的对称概率密度函数，其支撑集为 $(-1,1)$。

C4：$K^*(v)$ 和 $\mathcal{K}(v)$ 是 $b>d$ 阶的核函数，且存在 $c_1,c_2,\rho>0$，对任意 $L=K^*$ 或 $L=K$，满足：

$$c_1 I\left(\|v\|\le\rho\right)\le L(v)\le c_2 I\left(\|v\|\le\rho\right).$$

C5：对于条件 C4 中的 b，带宽 h_l 满足 $nh_l^{2d}\to\infty$，$nh_l^{2b}\to 0$，$b>d$，$l=1.4$，且对于任意固定的 $c>0$，带宽 h_2 满足 $h_2=cn^{-1/5}$.

C6：在一定的条件分布下，协变量与误差向量的四阶矩均有限，即：

$$\sup_u E\left(\|X\|^4\middle|\beta^{\mathrm{T}}X=u\right)<\infty,\quad \sup_u E\left(\|Z\|^4\middle|\beta^{\mathrm{T}}X=u\right)<\infty,\quad \sup_u E\left(\varepsilon^4\middle|\beta^{\mathrm{T}}X=u\right)<\infty$$

$$\sup_v E\left(\|X\|^4\middle|V=v\right)<\infty,\quad \sup_v E\left(\|Z\|^4\middle|V=v\right)<\infty,\quad \sup_v E\left(\varepsilon^4\middle|V=v\right)<\infty.$$

C7：随机向量 V 的概率密度函数 $f(v)$ 至多具有 b 阶有界偏导数，且存在 $a,b>0$，对任意的 $r_0\in[0,r_1]$ 和 $v_0\in\mathcal{V}$，满足

$$\int_{v\in S(v_0,r_0)\cap\mathcal{V}} f(v)\mathrm{d}v\ge ar_0^b,$$

其中 $\mathcal{V}$ 为 V 的支撑集，$S(v_0,r_0)$ 为以 v_0 为中心，r_0 为半径的闭球体.

C8：选择概率函数 $\pi(v)$ 和增广函数 $m(v)$ 均至多具有 b 阶有界偏导数，且存在 $c>0$，使得

$$\inf_v \pi(v)\ge c>0.$$

C9：矩阵 $\Omega=E\left[\left\{1/\pi(V)\right\}\mathrm{cov}\left(\xi\middle|V\right)\right]$ 和 $\Sigma=E\left[\omega\left(\beta^{\mathrm{T}}X\right)\left\{\Lambda-E\left(\Lambda\middle|\beta^{\mathrm{T}}X\right)\right\}^{\otimes 2}\right]$ 均是正定矩阵，其中 $\xi=\omega\left(\beta^{\mathrm{T}}X\right)\varepsilon\left\{\Lambda-E\left(\Lambda\middle|\beta^{\mathrm{T}}X\right)\right\},\Lambda=\left(g'\left(\beta^{\mathrm{T}}X\right)X^{\mathrm{T}}J_{\beta^{(r)}},Z^{\mathrm{T}}\right)^{\mathrm{T}}$.

定理 1：假设条件(C1)~(C9)成立，且 $\left(\beta^{(r)},\theta\right)$ 为真实参数向量，则

$$\hat{R}_{\mathrm{J}}\left(\beta^{(r)},\theta\right)\xrightarrow{D}\chi^2_{p+q-1},$$

$J=W$ 或 I。

注 1：根据定理 1 的结果，对于任意给定的 $0<\alpha<1$，可以构建参数向量 $\left(\beta^{(r)},\theta\right)$ 的经验似然置信域为

$$\mathrm{J}_{1-\alpha}\left(\beta^{(r)},\theta\right)=\left\{\left(\beta^{(r)},\theta\right)\middle|\hat{R}_{\mathrm{J}}\left(\beta^{(r)},\theta\right)\le\chi^2_{p+q-1}(1-\alpha),\left\|\beta^{(r)}\right\|<1\right\}.$$

定理 2：假设条件(C1)~(C9)成立，则

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_J^{(r)}-\beta\\ \hat{\theta}_J-\theta\end{pmatrix}\xrightarrow{D}N\left(0_{p+q-1},\Sigma^{-1}\Omega\Sigma^{-1}\right),$$

$J=W$ 或 I。其中 0_{p+q-1} 是一个 $(p+q-1)$ 维的零矩阵，Σ 和 Ω 的定义见 C9。

定理 3：假设条件(C1)~(C8)成立，则

$$\sqrt{nh_2}\left[\hat{g}_J\left(u;\hat{\beta}_J^{(r)},\hat{\theta}_J\right)-g(u)-b(u)\right]\xrightarrow{D}N\left(0,\gamma^2(u)\right),$$

$J=W$ 或 I。其中 $b(u)=\frac{1}{2}h_2^2g''(u)\int t^2K(t)\mathrm{d}t$，$\gamma^2(u)=\left\{q(u)/p(u)\right\}\int K^2(t)\mathrm{d}t$，$p(u)$ 是 $\beta^{\mathrm{T}}X$ 的概率密度函数，$q(u)=E\left(\varepsilon^2/\pi(V)\middle|\beta^{\mathrm{T}}X=u\right)$。

3. 数值模拟

在本节中，主要通过数值模拟来研究本文所提出方法的有限样本性质。模拟过程的数据由如下部分线性单指标模型产生：

$$Y=g\left(\beta^{\mathrm{T}}X\right)+\theta^{\mathrm{T}}Z+\varepsilon,$$

其中 $\beta=(\beta_1,\beta_2)^{\mathrm{T}}=(0.8,0.6)^{\mathrm{T}}$，$\theta=1$，$g(u)=15\exp(-u)$。协变量 X 服从二元标准正态分布，$Z\sim N(0,1)$，随机误差项 $\varepsilon\sim N(0,0.04)$。在模拟过程中，$\theta$ 和 β 的初始值分别由线性模型和广义线性模型得到，$\omega(u)=1$，核函数 $K^*(\cdot),K(\cdot)$ 和 $\mathcal{K}(\cdot)$ 均取为 $0.75\left(1-u^2\right)_+$，并采用“去一个体”最小二乘交叉验证方法选择带宽 $h_u(u=1,2,4)$。

在模拟过程中，一方面，由于“去一分量法”所对应的 Jacobian 矩阵依赖于被删除分量 $\beta^{(r)}$，而当该分量的取值接近于零时，Jacobian 矩阵可能出现数值不稳定甚至发散的问题。因此，为提高经验似然推断的稳定性，本文选取 $\left|\beta^{(r)}\right|>c>0$ 的分量进行删除，即优先删除绝对值较大的分量。另一方面，在估计选择概率函数 $\pi(\cdot)$，链接函数及其导数 $g(\cdot)$ 和 $g'(\cdot)$，条件期望函数 $g_1(\cdot)$ 和 $g_2(\cdot)$ 以及增广函数 $m(\cdot)$ 时，涉及多个带宽参数的选择问题。虽然可以通过最小二乘交叉验证法确定它们各自的最优带宽，但若在每一步均重新选择带宽并进行联合优化搜索，实际操作非常复杂。因此，为降低计算复杂度并保持各平滑步骤渐近阶的一致性，本文采用“主带宽–比例带宽”的带宽选择策略，即首先通过最小二乘交叉验证法确定主带宽，再按比例对其余带宽进行配置。具体地，首先最小化如下交叉验证目标函数

$$CV_g(h)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i(V_i)}\left\{Y_i-Z_i^{\mathrm{T}}\hat{\theta}-\hat{g}_{[i]}\left(\hat{\beta}^{\mathrm{T}}X_i\right)\right\}^2,$$

其中，$\hat{g}_{[i]}$ 为去掉第 i 个样本后 $g(\cdot)$ 的去一估计量。然后，将其余带宽按预设比例进行配置。同时，为考察所提方法对带宽选择的敏感性，本文以交叉验证法选取的主带宽为基准，在不同扰动带宽下重新计算了参数分量的偏差、标准差及平均区间长度等指标，结果表明，上述指标在适当扰动范围内变化较小，说明所提方法对带宽选择不敏感。

另外，为了代表响应变量的不同缺失水平，利用 logistic 回归模型设定以下三种不同的缺失机制：

情形 1：

$$\pi_1=\frac{1}{1+\exp(-3.15+X_1+X_2+Z)};$$

情形 2：

$$\pi_2 = \frac{1}{1+\exp\left(-2.05+X_1+X_2+Z\right)};$$

情形 3：

$$\pi_3 = \frac{1}{1+\exp\left(-1.27+X_1+X_2+Z\right)}.$$

以上三种情形对应的平均缺失概率大约为 10%，20%和 30%。

在不同的缺失概率下，分别取样本容量 n = 100，150 和 200，重复模拟 500 次，结果如下所示：

Table 1. The biases and standard deviations of β_1,β_2 and θ (×100)

表 1. β_1,β_2 和 θ 的偏差和标准差(×100)

π_i	n	β_1		β_2		θ	
		WEL	IEL	WEL	IEL	WEL	IEL
π_1	100	−0.0196	0.0076	0.0220	−0.0115	1.9585	−0.7828
		(0.0029)	(0.0009)	(0.0052)	(0.0016)	(0.3769)	(0.3490)
	150	−0.0194	0.0075	0.0218	−0.0113	1.7720	−0.5945
		(0.0017)	(0.0005)	(0.0030)	(0.0009)	(0.2487)	(0.2302)
	200	−0.0121	0.0073	0.0162	−0.0110	1.4874	−0.5705
		(0.0012)	(0.0003)	(0.0021)	(0.0006)	(0.1441)	(0.1335)
π_2	100	0.0393	0.0248	−0.0510	−0.0353	4.0933	−1.5433
		(0.0032)	(0.0010)	(0.0057)	(0.0017)	(0.5573)	(0.5160)
	150	−0.0203	0.0127	0.0253	−0.0175	3.3813	−1.1990
		(0.0021)	(0.0006)	(0.0037)	(0.0011)	(0.2753)	(0.2549)
	200	−0.0127	0.0080	0.0165	−0.0113	2.9996	−1.1750
		(0.0015)	(0.0004)	(0.0026)	(0.008)	(0.1597)	(0.1478)
π_3	100	−0.0463	0.0292	0.0598	−0.0414	6.1351	−2.3902
		(0.0037)	(0.0011)	(0.0066)	(0.0019)	(0.7161)	(0.6631)
	150	0.0204	−0.0128	−0.0256	0.0178	5.9800	−2.3662
		(0.0026)	(0.0007)	(0.0047)	(0.0012)	(0.3218)	(0.2980)
	200	0.0129	0.0082	−0.0167	−0.0115	5.9560	−2.3422
		(0.0019)	(0.0006)	(0.0034)	(0.0010)	(0.2057)	(0.1904)

Table 2. The average interval lengths and corresponding coverage probabilities of the confidence regions for β_1,β_2 and θ at the 95% confidence level

表 2. 置信水平为 95%的 β_1,β_2 和 θ 的平均区间长度以及对应的覆盖概率

π_i	n	β_1		β_2		θ	
		WEL	IEL	WEL	IEL	WEL	IEL
π_1	100	0.0174	0.0157	0.0236	0.0209	0.2424	0.2406
		(0.944)	(0.950)	(0.943)	(0.950)	(0.942)	(0.949)

续表

	150	0.0144	0.0128	0.0192	0.0171	0.1979	0.1964
		(0.950)	(0.956)	(0.950)	(0.956)	(0.948)	(0.955)
	200	0.0125	0.0111	0.0166	0.0148	0.1710	0.1698
		(0.956)	(0.961)	(0.955)	(0.960)	(0.954)	(0.960)
π_2	100	0.0201	0.0176	0.0268	0.0235	0.2811	0.2756
		(0.940)	(0.946)	(0.939)	(0.946)	(0.938)	(0.945)
	150	0.0165	0.0144	0.0220	0.0192	0.2301	0.2256
		(0.946)	(0.953)	(0.946)	(0.952)	(0.949)	(0.951)
	200	0.0142	0.0125	0.0190	0.0166	0.1988	0.1949
		(0.952)	(0.957)	(0.951)	(0.957)	(0.950)	(0.956)
π_3	100	0.0226	0.0196	0.0302	0.0262	0.3205	0.3112
		(0.935)	(0.942)	(0.934)	(0.941)	(0.933)	(0.940)
	150	0.0185	0.0160	0.0246	0.0214	0.2620	0.2544
		(0.942)	(0.948)	(0.941)	(0.948)	(0.940)	(0.947)
	200	0.0160	0.0138	0.0213	0.0185	0.2263	0.2179
		(0.947)	(0.953)	(0.946)	(0.952)	(0.945)	(0.951)

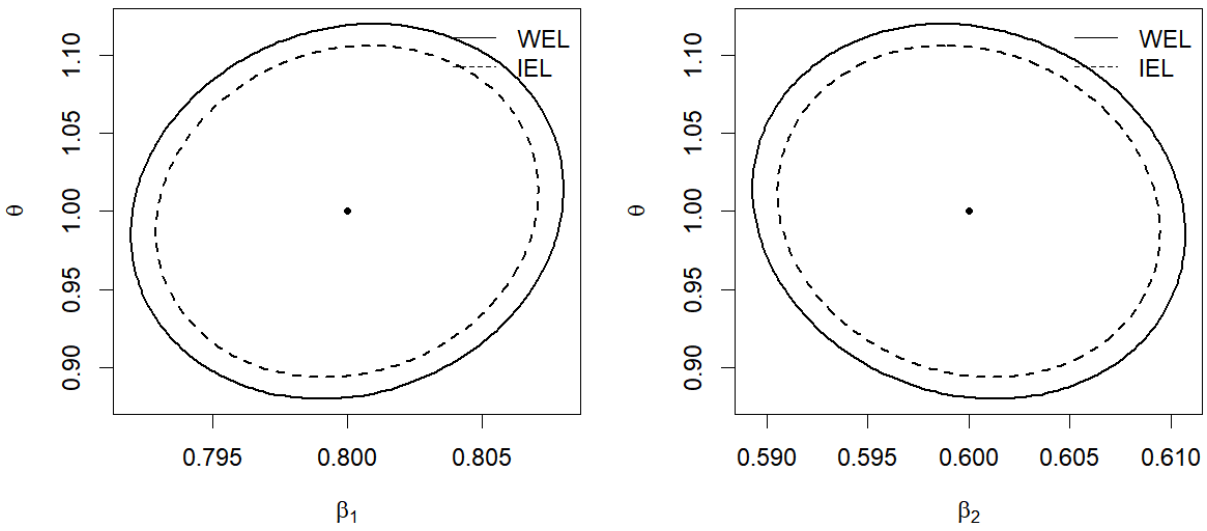

Figure 1. The 95% confidence region for (β_1,θ) and (β_2,θ)

图 1. 置信水平为 95%的(β_1,θ)和(β_2,θ)的置信域

从表 1 可以看出：

1) 在缺失率、样本量相同的情况下，WEL 方法的偏差和方差均大于 IEL 方法。

2) 在缺失率相同的情况下，样本量越大，WEL 方法和 IEL 方法的偏差和方差越小。

3) 在样本量相同的情况下，缺失率越大，WEL 方法和 IEL 方法的偏差和方差越大。

从表 2 可以看出：

1) 在缺失率、样本量相同的情况下，WEL 方法的平均区间长度大于 IEL 方法的平均区间长度，但 WEL 方法的覆盖概率低于 IEL 方法的覆盖概率。

2) 在缺失率相同的情况下，样本量越大，WEL 方法和 IEL 方法的平均区间长度越小，覆盖概率越大。

3) 在样本量相同的情况下，缺失率越大，WEL 方法和 IEL 方法的平均区间长度越大，覆盖概率越小。

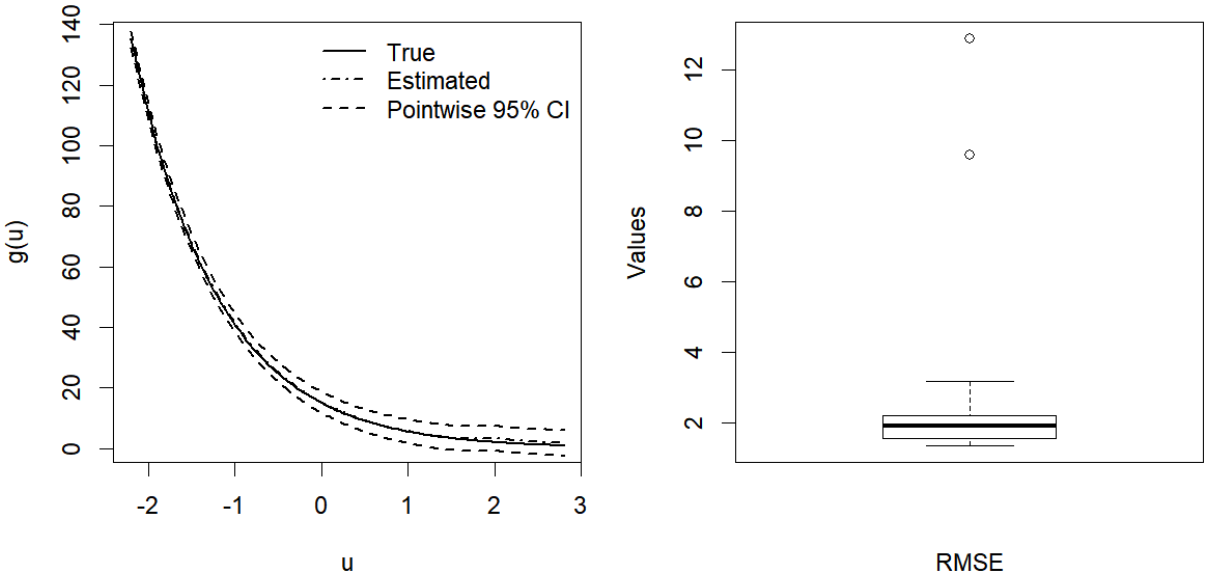

Figure 2. The estimation results and error performance of the link function
图 2. 链接函数的估计结果和误差表现

从图 1 可以看出，IEL 方法给出了较小的置信域，表明 IEL 方法具有更高的估计精度。从图 2 可以看出，估计函数与真实函数十分接近，且其在大部分区间内被 95%的平均逐点置信区间所覆盖，同时 RMSE 很小，表明所提经验似然方法是有效的。

4. 结论

本文主要通过采用逆概率加权方法和增广借补方法，构造了两类纠偏的参数向量的经验对数似然比统计量，并在适当正则条件下，证明了上述两类统计量均渐近服从标准卡方分布。根据构造的两类经验对数似然比统计量，建立了参数向量的经验似然置信域，并证明了参数向量极大经验似然估计量和链接函数估计量的渐近性质。最后通过数值模拟比较了两类方法的有限样本表现，结果表明，基于增广借补经验似然方法的模型表现较好。

致　谢

感谢袁德美老师的耐心指导与学术引领，同时也感谢审稿专家对论文提出的宝贵意见。

参考文献

[1] Ichimura, H. (1993) Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics*, **58**, 71-120. https://doi.org/10.1016/0304-4076(93)90114-k

[2] Hardle, W., Hall, P. and Ichimura, H. (1993) Optimal Smoothing in Single-Index Models. *The Annals of Statistics*, **21**,

157-178. https://doi.org/10.1214/aos/1176349020

[3] Weisberg, S. and Welsh, A.H. (1994) Adapting for the Missing Link. *The Annals of Statistics*, **22**, 1674-1700. https://doi.org/10.1214/aos/1176325749

[4] Zhu, L. and Xue, L. (2006) Empirical Likelihood Confidence Regions in a Partially Linear Single-Index Model. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **68**, 549-570. https://doi.org/10.1111/j.1467-9868.2006.00556.x

[5] Xue, L. (2023) Two-Stage Estimation and Bias-Corrected Empirical Likelihood in a Partially Linear Single-Index Varying-Coefficient Model. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **85**, 1299-1325. https://doi.org/10.1093/jrsssb/qkad060

[6] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1990) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**, 846-866. https://doi.org/10.1080/01621459.1994.10476818

[7] Wang, Q.H. and Rao, J.N.K. (2002) Empirical Likelihood-Based Inference under Imputation for Missing Response Data. *The Annals of Statistics*, **30**, 896-924. https://doi.org/10.1214/aos/1028674845

[8] Wang, Q.H. and Sun, Z.H. (2006) Estimation in Partially Linear Models with Missing Responses at Random. *Journal of Multivariate Analysis*, **98**, 1470-1493. https://doi.org/10.1016/j.jmva.2006.10.003

[9] Xue, L.G. (2013) Estimation and Empirical Likelihood for Single-Index Models with Missing Data in the Covariates. *Computational Statistics & Data Analysis*, **68**, 82-97. https://doi.org/10.1016/j.csda.2013.06.017

[10] 陈盼盼, 冯三营, 薛留根. 缺失数据下半参数变系数部分线性模型的统计推断[J]. 数学物理学报, 2015, 35(2): 345-358.

[11] Xue, L.G. and Lian, H. (2016) Empirical Likelihood for Single-Index Models with Responses Missing at Random. *Science China Mathematics*, **59**, 1187-1207. https://doi.org/10.1007/s11425-015-5097-y

[12] Xue, L.G. and Zhang, J.H. (2020) Empirical Likelihood for Partially Linear Single-Index Models with Missing Observations. *Computational Statistics & Data Analysis*, **144**, Article 106877. https://doi.org/10.1016/j.csda.2019.106877

[13] Lai, P. and Wang, Q.H. (2011) Partially Linear Single-Index Model with Missing Responses at Random. *Journal of Statistical Planning and Inference*, **141**, 1047-1058. https://doi.org/10.1016/j.jspi.2010.09.012

[14] Fan, J.Q. and Gijbels, I. (1996) Local Polynomial Modeling and Its Applications. Chapman & Hall.

[15] Qin, J. and Lawless, J. (1994) Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, **22**, 300-325. https://doi.org/10.1214/aos/1176325370