Statistics and Application 统计学与应用, 2020, 9(1), 53-62
Published Online February 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.91007

Hans汉斯

Application of Partial Linear Variable Coefficient EV Model in Fresh Product Sales Forecast under Missing Data

Zhenjun Wei*, Yanhui Gong, Chunliu Li

Trial Retail Engineering Co. Ltd., Yantai Shandong
Email: *weizhenjun1991@163.com

Received: Dec. 30th, 2019; accepted: Jan. 14th, 2020; published: Jan. 21st, 2020

Abstract

In this paper, we mainly consider the statistical inference for partially linear varying coefficient errors in variables models in the nonparametric part and the responses are missing at random. Based on local linear smoothing techniques, profile least-squares and bias-corrected methods, we obtained estimators successfully about both parametric and nonparametric components. Besides, to avoid to estimate the asymptotic covariance in establishing confidence region of the parametric component with the normal-approximation method, we define an empirical likelihood based statistic. Then, the confidence regions of the parametric component with asymptotically correct coverage probabilities can be constructed by the result. The simulation results show that the empirical likelihood method has better finite sample properties compared with the normal approximation method. Finally, the method is applied to a real data analysis of the supermarket fresh sales volume data and gives better estimation.

Keywords

Missing Data, Partially Linear Varying Coefficient Model, Variable with Measurement Errors, Empirical-Likelihood

缺失数据下部分线性变系数EV模型在生鲜产品销售量预测中的应用

未振军*，宫妍慧，李春柳

烟台创迹软件有限公司，山东 烟台
Email: *weizhenjun1991@163.com

*通讯作者。

文章引用: 未振军, 宫妍慧, 李春柳. 缺失数据下部分线性变系数 EV 模型在生鲜产品销售量预测中的应用[J]. 统计学与应用, 2020, 9(1): 53-62. DOI: 10.12677/sa.2020.91007

收稿日期：2019年12月30日；录用日期：2020年1月14日；发布日期：2020年1月21日

摘　要

本文主要研究了在响应变量随机缺失同时非参数分量带测量误差的条件下，部分线性变系数模型的统计推断。利用局部线性光滑、profile最小二乘及偏差纠正方法，构造了模型中参数分量和非参数分量的估计；另外，为了避免使用近似正态方法构造参数置信域时估计渐近协方差，我们又利用经验似然方法研究了参数置信域的构造问题，进而给出了参数分量的置信区间，模拟研究表明经验似然方法相比正态近似方法具有更好的有限样本性质。最后使用超市生鲜产品销售量数据集进行了实例分析，得到了更好的结果。

关键词

缺失数据，部分线性变系数模型，协变量含误差，经验似然

1. 引言

半参数部分线性变系数模型是近几十年来发展起来的一种统计模型，既含有线性模型部分，又含有变系数模型部分，是线性模型与变系数模型的结合，因此其具有线性模型便于解释和非参数模型稳健的特征。部分线性变系数模型的形式如下：

$$Y = X^{\tau}\beta + Z^{\tau}\alpha(T) + \varepsilon, \tag{1.1}$$

其中 Y 是响应变量，X，Z 和 T 为协变量，为了避免“维数祸根”，我们一般假设 T 为单变量，$\beta = (\beta_1, \beta_2, \cdots, \beta_q)^{\tau}$ 是 $q\times 1$ 维的未知参数向量，$\alpha(\cdot) = (\alpha_1(\cdot), \alpha_2(\cdot), \cdots, \alpha_p(\cdot))^{\tau}$ 是 $p\times 1$ 维未知的系数函数向量，ε 为模型误差并且满足 $E(\varepsilon \mid X, Z, T) = 0$，$\mathrm{Var}(\varepsilon \mid X, Z, T) = \sigma^2$。这个模型是一般的且包含许多重要的统计模型，比如：当 $\alpha(\cdot) \equiv \alpha$，其中 α 是一个常数向量时，模型(1.1)就变成通常的线性回归模型，这个模型被用到实际生活的许多方面，现在已经研究的非常成熟；当 $q \equiv 1$ 且 $Z \equiv 1$ 时，模型(1.1)就变成了部分线性回归模型，这个模型也用的相当广泛，Engle 等[1]用此模型研究了气温与用电量关系之后，部分线性模型受到了统计学家的重点关注，在理论和应用中都得到了更深入的研究，并取得了一系列突破性成果；当 $X \equiv 0$ 时，模型(1.1)就变成著名的变系数模型，由于其良好的解释能力，变系数模型得到了广泛的应用，被成功应用到非线性时间序列建模、函数型数据和纵向数据分析、空间分析以及金融计量分析等相关问题的研究中。针对此模型，Fan 和 Huang [2]对参数分量提出了 Profile 最小二乘估计，并且基于广义似然比检验方法研究了该模型参数和非参数函数的假设检验问题；Huang 和 Zhang [3]则利用经验似然方法研究了非参数分量 $\alpha(\cdot)$ 的统计推断问题。

在实际问题分析中经常会遇到数据缺失的现象。因此，在数据缺失下的研究是非常有必要的。随机缺失(MAR)的基本思想是观测到响应变量的概率仅仅依赖于其他观测变量的值，而不依赖于那些缺失值。例如，假设得到来自 (X, Y, δ) 的一个不完全观测样本

$$\left\{\left(X_i,Y_i,\delta_i\right),1\le i\le n\right\},\tag{1.2}$$

其中 X_i 为可完全观测到的协变量，Y_i 为响应变量且带有随机缺失(MAR)，δ 为指示变量。即，如果 $\delta_i=0$，则表示 Y_i 缺失；如果 $\delta_i=1$，则表示 Y_i 不缺失。随机缺失机制意味着给定 X 的条件下 Y 和 δ 是相对独立的，即

$$Pr\left(\delta=1\mid Y,X\right)=Pr\left(\delta=1\mid X\right),\tag{1.3}$$

随机缺失是在分析缺失数据时的一种常用假设，这在许多实际情况下是较为合理的。针对缺失数据的研究和处理一直被统计学家所重视，对模型(1.1)，Wei [4]研究了在因变量随机缺失情形下参数的统计推断问题以及基于广义似然比检验方法研究了该模型参数的假设检验问题；Wei 和 Mei [5]是基于广义似然比检验对参数部分 X 带测量误差和因变量 Y 随机缺失下考虑了参数的估计问题。

在许多实际应用中，我们往往无法得到协变量的精确值，只能采集到含有误差的观测值，比如在工程计算、经济学、生物医学和流行病学等领域，由于实验仪器的原因采集到的数据经常含有测量误差。在模型(1.1)中，非参数协变量 Z 是可精确观测的，我们所考虑的问题是 Z_i 含有测量误差，即我们所观测到的是 W_i 而不是 Z_i，两者之间的关系满足

$$W_i=Z_i+U_i,\ \ i=1,2,\cdots,n.\tag{1.4}$$

其中 U_i 是测量误差，且独立于 $\left(X^\tau,Z^\tau,U,\delta\right)^\tau$，其协方差阵为 $\mathrm{Cov}\left(U\right)=\Sigma_U$。为了模型的可识别性，我们假设 Σ_U 是已知的，如果该协方差未知，则可以利用 W 的重复观测数据得到 Σ_U 的估计值。对于上述情况，Yang 等[6]研究了带测量误差的部分线性模型参数的两步估计以及参数的经验似然置信域；针对模型(1.1)，You 和 Chen [7]研究了当协变量 X 含误差时参数的估计问题；Zhang 和 li 等[8]研究了在约束条件下协变量 X 含误差时参数的统计推断问题；Fan 和 Liang 等[9]基于经验似然方法研究了当自变量 X 含误差时参数的估计问题，构造了参数的经验似然置信域；Feng 和 Xue 等[10]研究了在约束条件下协变量 Z 含误差时的参数的估计和假设检验问题；Fan 和 Xu 等[11]考虑了协变量 Z 含误差时基于带辅助信息的经验似然的参数和非参数函数的估计问题，并证明了估计的渐近性质。本文则针对模型(1.1)同时考虑响应变量随机缺失和非参数部分带测量误差。

2. 模型与参数估计

我们首先基于完整数据给出模型(1.1)中系数函数 $\alpha\left(\cdot\right)$ 的估计，假设 $\left\{Y_i,\delta_i,X_i,Z_i,T_i\right\}_{i=1}^n$ 为来自模型(1.1)的观测数据，则有

$$\delta_iY_i=\delta_iX_i^\tau\beta+\delta_iZ_i^\tau\alpha\left(T_i\right)+\delta_i\varepsilon_i,\ i=1,2,\cdots,n,\tag{2.1}$$

假定 β 已知，则模型变为如下形式的变系数模型

$$\delta_i\left(Y_i-X_i^\tau\beta\right)=\delta_i\alpha_1\left(T_i\right)Z_{i1}+\delta_i\alpha_2\left(T_i\right)Z_{i2}+\cdots+\delta_i\alpha_p\left(T_i\right)Z_{ip}+\delta_i\varepsilon_i,\tag{2.2}$$

利用局部线性光滑的局部最小二乘法来估计未知系数函数。给定 T 领域内的一点 t，对 $\left\{\alpha_j\left(T\right),j=1,2,\cdots,q\right\}$ 利用 Taylor 展开有

$$\alpha_j\left(T\right)\approx\alpha_j\left(t\right)+\alpha_j'\left(t\right)\left(T-t\right)\equiv a_j+b_j\left(T-t\right),\ j=1,2,\cdots,q,\tag{2.3}$$

其中 $\alpha_j'\left(t\right)=\partial_{\alpha_j(t)}\big/\partial_t$ 为 $\alpha_j\left(t\right)$ 的一阶导数，对 a_j，b_j 极小化

$$\sum_{i=1}^n\left\{\left(Y_i-X_i^\tau\beta\right)-\sum_{j=1}^p\left[a_j+b_j\left(T_i-t\right)\right]Z_{ij}\right\}^2K_h\left(T_i-t\right)\delta_i,\tag{2.4}$$

其中 $K_h(\cdot)=K(\cdot/h)/h$， $K(\cdot)$ 为核函数，h 为带宽。为了表示方便，引入下面记号 $a=(a_1,a_2,\cdots,a_q)^\tau$，$b=(b_1,b_2,\cdots,b_q)^\tau$，$Y=(Y_1,Y_2,\cdots,Y_n)^\tau$，$X=(X_1,X_2,\cdots,X_n)^\tau$，$W=(W_1,W_2,\cdots,W_n)^\tau$，$Z=(Z_1,Z_2,\cdots,Z_n)^\tau$，$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^\tau$，$\omega_t^\delta=\mathrm{diag}\left(K_h(T_1-t)\delta_1,K_h(T_2-t)\delta_2,\cdots,K_h(T_n-t)\delta_n\right)$，

$$D_t^Z=\begin{bmatrix} Z_1^\tau & \dfrac{T_1-t}{h}Z_1^\tau \\ \vdots & \vdots \\ Z_n^\tau & \dfrac{T_n-t}{h}Z_n^\tau \end{bmatrix},\quad M=\begin{bmatrix} Z_1^\tau\alpha(T_1) \\ \vdots \\ Z_n^\tau\alpha(T_n) \end{bmatrix},$$

则式(2.4)的解为

$$\left[\hat{a}^\tau,h\hat{b}^\tau\right]^\tau=\left\{\left(D_t^Z\right)^\tau\omega_t^\delta D_t^Z\right\}^{-1}\left(D_t^Z\right)^\tau\omega_t^\delta(Y-X\beta), \tag{2.5}$$

上面在估计时假定了数据 Z 可以精确观测。如果我们简单的用 W 来代替 Z，而没有考虑测量误差的情况，可以证明这样得到的估计是不相合的，为了解决这个问题，我们需要对估计进行“校正衰减”，纠偏之后的估计为

$$\left[\hat{a}^\tau,h\hat{b}^\tau\right]^\tau=\left\{\left(D_t^W\right)^\tau\omega_t^\delta D_t^W-\Omega^\delta\right\}^{-1}\left(D_t^W\right)^\tau\omega_t^\delta(Y-X\beta), \tag{2.6}$$

其中 D_t^W 和 D_t^Z 有相同的形式，仅仅用 W_i 替代 Z_i，并且

$$\Omega^\delta=\sum_{i=1}^n\Sigma_u\otimes\begin{bmatrix} 1 & \dfrac{T_i-t}{h} \\ \dfrac{T_i-t}{h} & \left(\dfrac{T_i-t}{h}\right)^2 \end{bmatrix}K_h(T_i-t)\delta_i,$$

其中 $\otimes$ 为 Kronecker 乘积。因此当 β 已知时，纠偏后的系数函数 $\alpha(t)$ 的估计为

$$\tilde{\alpha}(t)=\left(I_q\ \ 0_q\right)\left\{\left(D_t^W\right)^\tau\omega_t^\delta D_t^W-\Omega^\delta\right\}^{-1}\left(D_t^W\right)^\tau\omega_t^\delta(Y-X\beta), \tag{2.7}$$

其中 I_q 为 q 阶单位阵，0_q 是 $q\times q$ 的 0 矩阵。令 $Q_i=\left(I_q\ \ 0_q\right)\left\{\left(D_{T_i}^W\right)^\tau\omega_{T_i}^\delta D_{T_i}^W-\Omega^\delta\right\}^{-1}\left(D_{T_i}^W\right)^\tau\omega_{T_i}^\delta, i=1,2,\cdots,n$。$Q=\left(Q_1^\tau,Q_2^\tau,\cdots,Q_n^\tau\right)^\tau$，$S=\left(Q_1^\tau W_1,Q_2^\tau W_2,\cdots,Q_n^\tau W_n\right)^\tau$，$\tilde{Y}=(I-S)Y$，$\tilde{X}=(I-S)X$，通过极小化下式

$$\sum_{i=1}^n\delta_i\left\{Y_i-X_i^\tau\beta-W_i^\tau\hat{\alpha}(T_i)\right\}^2-\sum_{i=1}^n\delta_i\hat{\alpha}^\tau(T_i)\Sigma_u\hat{\alpha}(T_i), \tag{2.8}$$

可以得到参数 β 的估计

$$\hat{\beta}=\left\{\sum_{i=1}^n\delta_i\left(\tilde{X}_i\tilde{X}_i^\tau-X^\tau Q_i^\tau\Sigma_u Q_i X\right)\right\}^{-1}\left\{\sum_{i=1}^n\delta_i\left(\tilde{X}_i\tilde{Y}_i-X^\tau Q_i^\tau\Sigma_u Q_i Y\right)\right\}, \tag{2.9}$$

参数 β 估计的矩阵形式为

$$\hat{\beta}=\left\{\tilde{X}^\tau\Delta\tilde{X}-X^\tau Q^\tau\Delta\otimes\Sigma_u QX\right\}^{-1}\left\{\tilde{X}^\tau\Delta\tilde{Y}-X^\tau Q^\tau\Delta\otimes\Sigma_u QY\right\},$$

其中 $\Delta=\mathrm{diag}(\delta_1,\delta_2,\cdots,\delta_n)$。

再由(2.7)式，我们可以得到系数函数 $\alpha(t)$ 的估计为

$$\hat{\alpha}(t)=\left(I_q \;\; 0_q\right)\left\{\left(D_t^W\right)^{\tau}\omega_t^{\delta}D_t^W-\Omega^{\delta}\right\}^{-1}\left(D_t^W\right)^{\tau}\omega_t^{\delta}\left(Y-X\hat{\beta}\right). \tag{2.10}$$

3. 经验似然推断

下面我们给出模型(1.1)回归参数的经验似然比，由(2.9)式可知，$\hat{\beta}$ 是下面方程的解

$$\sum_{i=1}^{n}\delta_i\left[\tilde{X}_i\left(\tilde{Y}_i-\tilde{X}_i^{\tau}\beta\right)-X^{\tau}Q_i^{\tau}\Sigma_u Q_i\left(Y-X\beta\right)\right]=0, \tag{3.1}$$

由此我们引进辅助随机变量

$$\eta_i(\beta)=\delta_i\tilde{X}_i\left(\tilde{Y}_i-\tilde{X}_i^{\tau}\beta\right)-\delta_i X^{\tau}Q_i^{\tau}\Sigma_u Q_i\left(Y-X\beta\right), \tag{3.2}$$

通过 Qin 和 Lawless [12]，基于经验似然方法，关于 β 的经验对数似然比函数可以定义为

$$\mathcal{L}(\beta)=-2\max\left\{\sum_{i=1}^{n}\log(np_i)\mid p_i\geq 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\eta_i(\beta)=0\right\}, \tag{3.3}$$

使用 Lagrange 乘子法可以算出

$$p_i=\frac{1}{n}\frac{1}{1+\lambda^{\tau}\eta_i(\beta)}, \tag{3.4}$$

其中 λ 是下面方程的解

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\eta_i(\beta)}{1+\lambda^{\tau}\eta_i(\beta)}=0. \tag{3.5}$$

由上可知，经验对数似然比可以表示为

$$\mathcal{L}(\beta)=2\sum_{i=1}^{n}\log\left\{1+\lambda^{\tau}\eta_i(\beta)\right\}. \tag{3.6}$$

4. 模拟研究

本节通过数值模拟来研究所提方法的有限样本性质。考虑如下的部分线性变系数含误差模型

$$\begin{cases} y_i=x_i\beta+z_{1i}\alpha_1(T_i)+z_{2i}\alpha_2(T_i)+\varepsilon_i, \\ w_{1i}=z_{1i}+u_{1i}, \\ w_{2i}=z_{2i}+u_{2i}, \end{cases} \qquad i=1,2,\cdots,n,$$

其中 $x_i\sim N(0,1)$，$z_{1i}\sim N(0,1)$，$z_{2i}\sim N(0,1)$，$T_i\sim U(0,1)$，$\beta=2$，$\alpha_1(t)=2\sin(2\pi t)$，$\alpha_2(t)=-1.5\cos(1.5\pi t)$，$\left[u_1 \;\; u_2\right]^{\tau}\sim N_2(0,\Sigma_u)$，$\Sigma_u$ 分别取 $0.25I_2$ 和 $0.5I_2$ 两种情况。为了研究模型误差对结果的影响，我们设定误差 ε_i 有以下形式：$\varepsilon_i\sim N(0,0.5)$。模拟中我们选取 Epanechnikov 核函数，即 $K(x)=0.75\left(1-x^2\right)I_{|x|\leq 1}$。并采用“去一个体”交叉验证的方法来选择带宽 h_{CV}，使其满足下列式子达到最小

$$\mathrm{CV}(h)=\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{Y_i-X_i^{\tau}\hat{\beta}^{(-i)}-W_i^{\tau}\hat{\alpha}_{h,(-i)}(T_i)\right\}^2-\frac{1}{n}\sum_{i=1}^{n}\delta_i\hat{\alpha}_{h,(-i)}^{\tau}(T_i)\Sigma_u\hat{\alpha}_{h,(-i)}(T_i),$$

其中 $\hat{\alpha}_{h,(-i)}(T_i)$ 和 $\hat{\beta}^{(-i)}$ 分别为去掉第 i 个观测值后的 $\alpha(u)$ 和 β 的估计。分别考虑如下两种形式的缺失机制：

Case I $Pr(\delta=1|X=x,Z=z,T=t)=0.8$，对所有的 x,z,t；

Case II $Pr(\delta=1|X=x,Z=z,T=t)=0.8+0.2(|x|+|t-0.5|)$，当 $|x|+|t-0.5|\le 1$ 时，否则取 0.88。

下面我们分别给出不带缺失数据和非参数分量含测量误差不纠偏的 β 的估计公式，以便于与本文得出的结果作对比。

1) 不含缺失数据：

$$\bar{\beta}=\left\{\tilde{X}^{\tau}\tilde{X}-X^{\tau}Q^{\tau}I\otimes\Sigma_u QX\right\}^{-1}\left\{\tilde{X}^{\tau}\tilde{Y}-X^{\tau}Q^{\tau}I\otimes\Sigma_u QY\right\},$$

$$\bar{\alpha}(t)=\left(I_q\ \ 0_q\right)\left\{\left(D_t^W\right)^{\tau}\omega_t D_t^W-\Omega\right\}^{-1}\left(D_t^W\right)^{\tau}\omega_t\left(Y-X\bar{\beta}\right),$$

上述两个公式可以在 Feng 和 Xue [10]这篇文章中找到。

2) 非参数分量含测量误差不纠偏：

$$\tilde{\beta}=\left[\sum_{i=1}^{n}\delta_i\tilde{X}_i\tilde{X}_i^{\tau}\right]^{-1}\left[\sum_{i=1}^{n}\delta_i\tilde{X}_i\tilde{Y}_i\right],$$

$$\tilde{\alpha}(t)=\left(I_q\ \ 0_q\right)\left\{\left(D_t^W\right)^{\tau}\omega_t^{\delta}D_t^W\right\}^{-1}\left(D_t^W\right)^{\tau}\omega_t^{\delta}\left(Y-X\tilde{\beta}\right),$$

上述两个公式可以在[13]中找到相似的结果。

在上述两种缺失情形下，数据 Y 的平均缺失概率为别为 0.2 和 0.1。在模拟过程中，对于每种情况，样本容量分别取 $n=100,150,200$，并进行 1000 次模拟，得到了 $\hat{\beta},\bar{\beta}$ 和 $\tilde{\beta}$ 的均值、标准差和均方误差，结果如表 1、表 2、表 3 所示。在样本量 $n=200$ 及响应变量 Y 的平均缺失概率为 0.2 和 $\Sigma_u=0.5I_2$ 的情形下，我们在图 1 中给出了模型中非参数函数的估计曲线。此外，在给定两种不同的缺失水平下，分别用经验似然方法(EL)和正态逼近方法(NA)给出参数 β 的置信水平为 95%的置信区间，并计算出置信区间的平均长度及覆盖概率，结果如表 4 所示。

Table 1. Mean, SD and MSE of $\hat{\beta}$ under different conditions

表 1. 不同情况下 $\hat{\beta}$ 的均值(Mean)、标准差(SD)和均方误差(MSE)

Case	Σ_u	n	Mean	SD	MSE
I	$0.25I_2$	100	2.0224	0.1554	0.0210
		150	2.0187	0.1115	0.0117
		200	2.0013	0.0951	0.0099
	$0.5I_2$	100	1.9986	0.1962	0.0361
		150	2.0129	0.1624	0.0268
		200	2.0107	0.1346	0.0168
II	$0.25I_2$	100	2.0032	0.1503	0.0192
		150	2.0041	0.1086	0.0114
		200	2.0008	0.0895	0.0086
	$0.5I_2$	100	1.9956	0.1853	0.0323
		150	2.0082	0.1601	0.0259
		200	2.0023	0.1261	0.0154

Table 2. Mean, SD and MSE of $\bar{\beta}$ under different conditions

表 2. 不同情况下 $\bar{\beta}$ 的均值(Mean)、标准差(SD)和均方误差(MSE)

Case	Σ_u	n	Mean	SD	MSE
Null	$0.25I_2$	100	2.0178	0.3975	0.1581
		150	1.9982	0.1687	0.0284
		200	2.0027	0.0996	0.0099
	$0.5I_2$	100	1.9864	0.7580	0.5761
		150	1.9906	0.4992	0.2490
		200	2.0028	0.3664	0.1341

Table 3. Mean, SD and MSE of $\tilde{\beta}$ under different conditions

表 3. 不同情况下 $\tilde{\beta}$ 的均值(Mean)、标准差(SD)和均方误差(MSE)

Case	Σ_u	n	Mean	SD	MSE
I	$0.25I_2$	100	2.0191	0.1343	0.0183
		150	2.0165	0.1038	0.0108
		200	2.0008	0.0863	0.0085
	$0.5I_2$	100	1.9993	0.1864	0.0342
		150	2.0119	0.1576	0.0251
		200	2.0087	0.1219	0.0149
II	$0.25I_2$	100	2.0045	0.1332	0.0179
		150	1.9991	0.1043	0.0106
		200	2.0031	0.0852	0.0081
	$0.5I_2$	100	1.9985	0.1771	0.0315
		150	2.0061	0.1521	0.0224
		200	2.0013	0.1131	0.0134

Table 4. Average length and coverage probability of β confidence interval for 95% confidence level under different conditions

表 4. 不同情形下置信水平为 95%的 β 的置信区间的平均长度和覆盖概率

Case	Σ_u	n	平均长度		覆盖概率	
			NA	EL	NA	EL
I	$0.25I_2$	100	0.2916	0.2713	0.881	0.903
		150	0.2725	0.2593	0.904	0.919
		200	0.2396	0.2119	0.912	0.921
	$0.5I_2$	100	0.3325	0.3109	0.883	0.892
		150	0.3191	0.2867	0.890	0.910
		200	0.2758	0.2610	0.907	0.918
II	$0.25I_2$	100	0.2805	0.2693	0.891	0.911
		150	0.2716	0.2609	0.912	0.927
		200	0.2309	0.2231	0.919	0.932
	$0.5I_2$	100	0.3341	0.3096	0.890	0.907
		150	0.3082	0.2719	0.913	0.919
		200	0.2718	0.2576	0.917	0.921

从表 1 中我们可以得到三个结论：

1) 在缺失概率和测量误差协方差给定的情况下，随着样本的增加，参数估计量的标准差和均方误差都逐渐减小；

2) 在缺失概率和样本量给定的情况下，测量误差协方差越小，参数估计量的标准差和均方误差越小；

3) 在测量误差协方差和样本量给定的情况下，缺失概率越小，参数估计量的标准差和均方误差越小。

通过对表 1 与表 2 对比可以得知：

本文所用的处理缺失数据的方法与无缺失数据的结果相差不大，很好地说明了本文所论述方法的优良性。

通过对表 1 与表 3 对比可以得知：

1) 在对含测量误差不做任何处理时，得到的结果偏差较大，不能使人满意。

2) 本文所提的方法对含有测量误差数据的处理有很好的效果。

从表 4 中我们可以得到如下四个结论：

1) EL (经验似然法)比 NA (正态近似法)有更短的置信区间和更高的覆盖率；

2) 对给定的缺失概率，随着样本量的增加，经验似然和正态近似的置信区间均会缩短；

3) 对给定的样本量，随着缺失概率的增加，经验似然和正态近似的置信区间均会增长；

4) 对给定的样本量和缺失概率，误差方差越大，不论是经验似然还是正态近似的置信区间均会增长，且覆盖率会下降。

从图 1 可以看出我们所提出的非参数函数的估计量(虚线)与无缺失数据(点虚线)情形下是几乎重合的，非常接近于真实(实线)曲线，而不纠偏时估计(点线)的效果显然是较差的，这表明我们所提出的非参数部分的估计量是有效的。

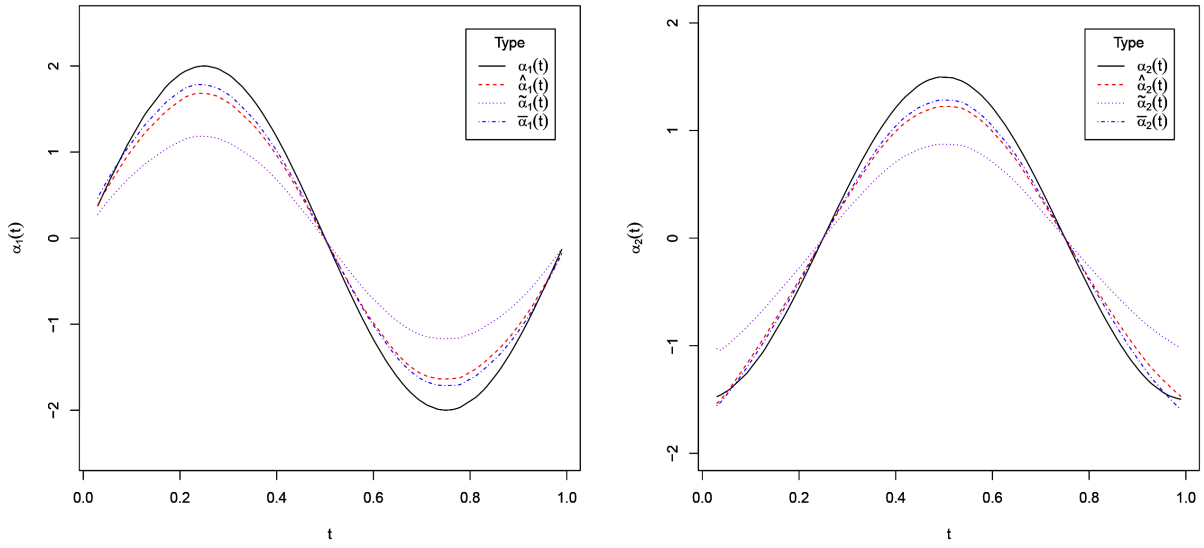

Figure 1. Real curve and all kinds of estimation curves of coefficient function (The left picture is $\alpha_1(t)$, right picture is $\alpha_2(t)$)

图 1. 系数函数(左图为 $\alpha_1(t)$，右图为 $\alpha_2(t)$)的真实曲线和各类估计曲线

5. 实例分析

下面通过分析超市生鲜产品销售量的数据进一步说明我们方法的有效性。数据源于日本某公司分析客流量、天气、产品价格等因素对超市生鲜产品销售量影响的工作。该数据集由生鲜产品销售量和客流

量、生鲜产品价格、天气情况、是否是节假日一些变量构成，其中生鲜产品销售量为 2017 年 11 月至 2019 年 7 月店铺中生鲜产品每天的销售量(包含打折数量)，九个感兴趣的协变量分别为：生鲜产品价格(PFP)、客流量(VOC)、恶劣天气程度(SWC)、平均气温(MT)、最高气温(MaxT)、平均湿度(MH)、平均降雨(MR)、最大风力(MWF)、是否是节假日(HD)。为了符号简单起见，协变量 PFP，MT，MaxT，MH，MR，MWF，VOC，HD 分别记为 $Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, X_1, X_2$。

取 $Z_1 = 1$ 为截距项，协变量 $T = \sqrt{\text{SWC}}$。研究讨论了 $Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, X_1, X_2$ 以及 SWC 对超市生鲜产品销售量的影响，并采用部分线性变系数模型

$$Y = \sum_{i=1}^{7} \alpha_i(T) Z_i + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

来拟合给定的数据。在进行分析之前，我们首先对协变量进行标准化变换，同时需要对变量 SWC 进行变换使其分布为 $U[0,1]$。为了证明我们所提方法的有效性，我们假定 Z_6 带有测量误差，即

$$W_6 = Z_6 + U_6$$

其中 $U_6 \sim N(0, 0.02)$，响应变量有 5%的缺失值，模拟中 δ 的产生是随机的，选取 Epanechnikov 核函数，即 $K(x) = 0.75(1 - x^2) I_{|x| \le 1}$，通过“去一个体”交叉验证的方法来选择带宽 $h = 0.0113$，通过本章提出的方法所得参数分量 β_1 的估计 $\hat{\beta}_1$ 的值为 2.9711，置信区间为[2.7546, 3.1876]；β_2 的估计 $\hat{\beta}_2$ 的值为 3.3058，置信区间为[−3.0865, −3.5251]。协变量客流量 X_1 和是否是节假日 X_2 的系数估计值 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 分别为 2.9711 和 3.3058，说明随着客流量的增加或当天为节假日时，生鲜产品的销售量也是递增的，销售量与协变量之间有正相关关系。这也与实际情况相符合。

参考文献

[1] Engle, R.F., Granger, C.W.J., Rice, J.J. and Weiss, A. (1986) Semiparametric Estimates of the Relation between Weather and Electricity Sales. *Journal of the American Statistical Association*, **81**, 310-320. https://doi.org/10.1080/01621459.1986.10478274

[2] Fan, J.Q. and Huang, T. (2005) Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models. *Bernoulli*, **11**, 1031-1057. https://doi.org/10.3150/bj/1137421639

[3] Huang, Z.S. and Zhang, R.Q. (2009) Empirical Likelihood for Nonparametric Parts in Semiparametric Varying-Coefficient Partially Linear Models. *Statistics and Probability Letters*, **79**, 1798-1808. https://doi.org/10.1016/j.spl.2009.05.008

[4] Wei, C.H. (2012) Statistical Inference in Partially Linear Varying-Coefficient Models with Missing Responses at Random. *Communications in Statistics-Theory and Methods*, **41**, 1284-1298. https://doi.org/10.1080/03610926.2010.542854

[5] Wei, C.H. and Mei, C.L. (2012) Empirical Likelihood for Partially Linear Varying-Coefficient Models with Missing Response Variables and Error-Prone Covariates. *Journal of the Korean Statistical Society*, **41**, 97-103. https://doi.org/10.1016/j.jkss.2011.06.004

[6] Yang, Y.P., Xue, L.G. and Cheng, W.H. (2011) Two-Step Estimators in Partial Linear Models with Missing Response Variables and Error-Prone Covariates. *Journal of Systems Science and Complexity*, **24**, 1165-1182. https://doi.org/10.1007/s11424-011-8393-9

[7] You, J.H. and Chen, G.M. (2006) Estimation of a Semiparametric Varying-Coefficient Partially Linear Errors-in-Variables Model. *Journal of Multivariate Analysis*, **97**, 324-341. https://doi.org/10.1016/j.jmva.2005.03.002

[8] Zhang, W.W., Li, G.R. and Xue, L.G. (2011) Profile Inference on Partially Linear Varying-Coefficient Errors-in-Variables Models under Restricted Condition. *Computational Statistics and Data Analysis*, **55**, 3027-3040. https://doi.org/10.1016/j.csda.2011.05.012

[9] Fan, G.L., Liang, H.Y. and Shen, Y. (2016) Penalized Empirical Likelihood for High-Dimensional Partially Linear Varying Coefficient Model with Measurement Errors. *Journal of Multivariate Analysis*, **147**, 183-201. https://doi.org/10.1016/j.jmva.2016.01.009

[10] Feng, S.Y. and Xue, L.G. (2014) Bias-Corrected Statistical Inference for Partially Linear Varying Coefficient Er-

rors-in-Variables Models with Restricted Condition. *Annals of the Institute of Statistical Mathematics*, **66**, 121-140. https://doi.org/10.1007/s10463-013-0407-z

[11] Fan, G.L., Xu, H.X. and Huang, Z.S. (2016) Empirical Likelihood for Semivarying Coefficient Model with Measurement Error in the Nonparametric Part. *AStA-Advances in Statistical Analysis*, **100**, 21-41. https://doi.org/10.1007/s10182-015-0247-7

[12] 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

[13] Wei, C.H. (2012) Statistical Inference for Restricted Partially Linear Varying Coefficient Errors-in-Variables Models. *Journal of Statistical Planning and Inference*, **142**, 2464-2472. https://doi.org/10.1016/j.jspi.2012.02.041