文章编号:1674-7070(2013)05-0466-06

响应变量随机缺失下部分线性单指标模型的非参数判别检验

来鹏[1]

摘要

研究了响应变量随机缺失情况下部分线性单指标模型的非参数部分检验问题,检验非参数部分预测变量同响应变量之间是否存在非线性关系. 用参数和非参数函数的借补估计对缺失响应变量进行插值,并基于借补估计构造了广义似然比检验统计量,证明了其渐近分布性质.

关键词

部分线性单指标模型;随机缺失;借补估计;广义似然比统计量

中图分类号 O211.61

文献标志码 A

0 引言

回归分析作为统计学中一个非常实用的技术方法被用在很多不同的领域,但是在回归分析中维数过高和变量之间的共线性问题常常会给分析工作者带来困扰. 为解决分析研究中碰到的高维问题,避免所谓的"维数祸根",许多降维方法以及一些参数或者半参数模型被提出,例如可加模型、部分线性模型以及广义部分线性单指标模型等. 部分线性单指标模型由于其本身既包含线性部分又包含非线性部分,使得其不仅可以拟合预测变量与响应变量之间的线性关系,又可以拟合它们之间的非线性关系,因而可以被运用到许多领域,能够处理比线性回归模型更广泛的问题,并且能在一定程度上避免协变量过多情况下非参数估计常碰到的"维数祸根"问题. 其模型形式为

$$Y = g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon, \tag{1}$$

其中 Y 是刻度响应变量,$\boldsymbol{X}$ 和 $\boldsymbol{Z}$ 别是 p 维和 q 维协变量,$g(\cdot)$ 是一个未知可测函数,随机统计误差 ε 满足 $E(\varepsilon \mid \boldsymbol{X},\boldsymbol{Z}) = 0$ 和 $\mathrm{Var}(\varepsilon \mid \boldsymbol{X},\boldsymbol{Z}) = \sigma^2$. 考虑到模型的可识别性,要求参数 $\boldsymbol{\beta}$ 满足 $\|\boldsymbol{\beta}\| = 1$. 从此模型的形式可以发现,当 $\beta = 1, p = 1$ 时,模型就变成一般的部分线性模型;而当 $\theta = 0$ 时,模型为单指标模型. 对模型(1)相应的研究可参见文献[1-4].

模型(1)包含线性部分和非线性部分,如何去判定模型中哪些变量同响应变量有线性关系,哪些变量同响应变量有非线性关系,这决定了模型建构是否正确,是否能有效地解决实际问题. Zhang[5] 在完全数据的情况下通过构造广义似然比统计量来检验部分线性单指标模型的非参数部分是否同响应变量存在线性关系. 在实际中,数据可能由于设计或者偶然性等各种原因而出现缺失. 例如可能由于响应变量 Y 的测量费用非常昂贵,而导致仅有部分数据被观测到;又可能响应变量 Y 代表一些问题的问答,而有些抽样的个体拒绝回答问题等. 数据经常为不完全的随机样本:

$$(Y_i, \delta_i, \boldsymbol{X}_i, \boldsymbol{Z}_i), \quad i = 1, 2, \cdots, n,$$

其中如果 Y_i 缺失,则 $\delta_i = 0$,否则 $\delta_i = 1$. 假定 Y 是随机缺失(MAR)的,也就是说 $p(\delta = 1 \mid Y, \boldsymbol{X}, \boldsymbol{Z}) = p(\delta = 1 \mid \boldsymbol{X}, \boldsymbol{Z})$. MAR 假定是缺失数据统计分析中的一种比较常见的假定[6].

收稿日期 2012-10-15

资助项目 国家自然科学基金(11226222,11301279);江苏省高校自然科学研究基金(12KJB110016);江苏省自然科学基金(BK2012459)

作者简介

来鹏,男,博士,讲师,主要研究半参数统计及复杂数据分析. laipengnuist@163.com

1 南京信息工程大学 数学与统计学院,南京,210044

对于响应变量随机缺失的部分线性单指标模型的估计问题,Lai 等[7]提出了借补估计的方法,能够有效地解决模型的参数和非参数估计. 在缺失数据的情况下,Zhang[5]的方法不再适用于区分预测变量与响应变量之间的线性和非线性关系,本文将推广 Zhang[5]的方法,使其能够应用于响应变量随机缺失的情况.

1 广义似然比检验统计量及其渐近性质

判断协变量同响应变量关系的方法基本上集中在通过刻画变量之间的关系图形,利用图形来直观地判断,或者建立检验统计量,利用假设检验的理论来判断. 利用图形的方法虽然简单直观,但是存在很大的不确定性和主观性,而检验统计量能够在基于图形判断的基础上给出客观的判断.

Zhang[5]在完全样本下对于部分线性单指标模型(1)引出了假设检验问题: $H_0: g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$ 是关于 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0$ 的线性函数,$\leftrightarrow H_1: g(\cdot)$ 是非线性函数;判断模型中单指标部分是同响应变量之间存在线性关系,还是某种非线性关系,等价的可表示为

$$H_0: g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0) = \alpha_0 + \alpha_1 \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \leftrightarrow H_1: g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \neq \alpha_0 + \alpha_1 \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0, \tag{2}$$

其中 α_0, α_1 是未知参数,在检验过程中可用它们的估计值代替.

注意到响应变量随机缺失,由此可以利用 Lai 等[7]提到的借补估计来对缺失响应变量进行插值借补,从而利用 Zhang[5]以及 Fan 等[8]提出的广义似然比方法来构造检验统计量:

$$\lambda_n^I = \frac{n}{2}\frac{R_{SS_0^I} - R_{SS_1^I}}{R_{SS_1^I}}, \tag{3}$$

其中 $R_{SS_0^I}$ 和 $R_{SS_1^I}$ 分别表示在原假设和备择假设下模型估计的残差平方和,即:

$$R_{SS_0^I} = \sum_{i=1}^{n} (\breve{Y}_i - \bar{\boldsymbol{\theta}}_{nI}^{\mathrm{T}}\boldsymbol{Z}_i - \bar{\alpha}_{0nI} - \bar{\alpha}_{1nI}\bar{\boldsymbol{\beta}}_{nI}^{\mathrm{T}}\boldsymbol{X}_i)^2, \tag{4}$$

$$R_{SS_1^I} = \sum_{i=1}^{n} (\breve{Y}_i - \hat{\boldsymbol{\theta}}_{nI}^{\mathrm{T}}\boldsymbol{Z}_i - \hat{g}_{nI}(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{nI}))^2, \tag{5}$$

其中 $\bar{\boldsymbol{\theta}}_{nI}, \bar{\alpha}_{0nI}, \bar{\alpha}_{1nI}$ 和 $\bar{\boldsymbol{\beta}}_{nI}$ 是基于原假设即线性假设下利用借补插值所得到的参数估计,可以通过最小二乘方法得到,而 $\hat{\boldsymbol{\theta}}_{nI}, \hat{g}_{nI}(\cdot), \hat{\boldsymbol{\beta}}_{nI}$ 是基于备择假设下利用借补插值所得到的参数和函数估计,可由 Lai 等[7]的方法得到. 在式(4)中,

$$\breve{Y}_i = \delta_i Y_i + (1-\delta_i)(\bar{\boldsymbol{\theta}}_{nc}^{\mathrm{T}}\boldsymbol{Z}_i + \bar{\alpha}_{0nc} + \bar{\alpha}_{1nc}\bar{\boldsymbol{\beta}}_{nc}^{\mathrm{T}}\boldsymbol{X}_i),$$

其中所包含的估计 $\bar{\boldsymbol{\theta}}_{nc}, \bar{\alpha}_{0nc}, \bar{\alpha}_{1nc}, \bar{\boldsymbol{\beta}}_{nc}$ 为在原假设条件下得到的 CC(Complete Case)估计,而在式(5)中,

$$\breve{Y}_i = \delta_i Y_i + (1-\delta_i)(\tilde{\boldsymbol{\theta}}_{nc}^{\mathrm{T}}\boldsymbol{Z}_i + \tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc})),$$

其中所包含的估计 $\tilde{\boldsymbol{\theta}}_{nc}, \tilde{\boldsymbol{\beta}}_{nc}, \tilde{g}_{nc}(\cdot)$ 为 Lai 等[7]所提出的初始估计. 利用最小二乘估计的思想,令

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_n = {} & \left\{\frac{1}{n}\sum_{i=1}^{n}\delta_i[\hat{p}_1(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{Z}_i - \hat{\boldsymbol{m}}_2(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})]\cdot\right. \\ & \left.[\hat{p}_1(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{Z}_i - \hat{\boldsymbol{m}}_2(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})]^{\mathrm{T}}\right\}^{-1} \times \\ & \left\{\frac{1}{n}\sum_{i=1}^{n}[\hat{p}_1(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})Y_i - \hat{m}_1(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})]\cdot\right. \\ & \left.[\hat{p}_1(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{Z}_i - \hat{\boldsymbol{m}}_2(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})]\delta_i\right\},\end{aligned} \tag{6}$$

其中

$$\begin{aligned}\hat{m}_1(t) &= \frac{\sum_{i=1}^{n}\delta_i Y_i K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}{\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}, \\ \hat{\boldsymbol{m}}_2(t) &= \frac{\sum_{i=1}^{n}\delta_i \boldsymbol{Z}_i K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}{\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}, \\ \hat{p}_1(t) &= \frac{\sum_{i=1}^{n}\delta_i K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}{\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}-t}{h_1}\right)}\end{aligned} \tag{7}$$

分别是 $m_1(t) = E[\delta Y \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t]$, $\boldsymbol{m}_2(t) = E[\delta \boldsymbol{Z} \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t]$, $p_1(t) = E[\delta \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t]$ 的非参数核估计,则利用局部线性近似的思想,在给定 $\boldsymbol{\beta}$ 的情况下定义 $g(t)$ 和 $g'(t)$ 的初始估计,分别记为 $\tilde{g}_n(t)$ 和 $\tilde{g}_n'(t)$. 令 $\hat{a}$ 和 $\hat{b}$ 为式(8)的解:

$$\min_{a,b}\sum_{i=1}^{n}\delta_i\{Y_i - \boldsymbol{Z}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_n - a - b(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - t)\}^2 K_{h_2}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - t), \tag{8}$$

使 $\tilde{g}_n(t) = \hat{a}, \tilde{g}'_n(t) = \hat{b}$. 利用前文所给出的估计 $\tilde{\boldsymbol{\theta}}_n, \tilde{g}_n(\cdot)$ 和 $\tilde{g}'_n(\cdot)$,求解在观测样本下使得残差平方和最小的 $\boldsymbol{\beta}$ 的估计,记为 $\tilde{\boldsymbol{\beta}}_{nc}$,其为在条件 $\|\boldsymbol{\beta}\| = 1$ 下最小化式(9)的解:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\delta_i\{Y_i - \boldsymbol{Z}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_n - \tilde{g}_n(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2, \tag{9}$$

那么将 $\tilde{\boldsymbol{\theta}}_n, \tilde{g}_n(\cdot)$ 和 $\tilde{g}'_n(\cdot)$ 中的 $\boldsymbol{\beta}$ 替换为 $\tilde{\boldsymbol{\beta}}_{nc}$,可以得到初始估计 $\tilde{\boldsymbol{\theta}}_{nc}, \tilde{g}_{nc}(\cdot), \tilde{g}'_{nc}(\cdot)$ 和 $\tilde{\boldsymbol{\beta}}_{nc}$. 采用类似的方法,将缺失的响应变量进行借补插值,则借补估计 $\hat{\boldsymbol{\theta}}_{nI}, \hat{g}_{nI}(\cdot), \hat{g}'_{nI}(\cdot)$ 和 $\hat{\boldsymbol{\beta}}_{nI}$ 可以通过将式(6)—(9)中的 δ_i 和 Y_i 分别用 1 和 $\breve{Y}_i$ 替换求解得到,所采用的窗

宽分别记为 h_3 和 h_4,那么所构造的广义似然比检验统计量(式(3))具有下面的性质:

定理 1 假设下一节所给的条件满足,那么在原假设 H_0 下,有

$$v_n^{-1}(\lambda_n^I - \mu_n) \xrightarrow{D} N(0,1),$$

其中 μ_n 和 v_n^2 的表达式可分别参见式(14)和(15).

2 定理的条件及证明

设 $\boldsymbol{X}$ 和 $\boldsymbol{Z}$ 具有有界支撑分别为 A_X 和 A_Z,令 $B = \{\boldsymbol{\beta} \in \mathbf{R}^p : \|\boldsymbol{\beta}\| = 1\}$,那么 $\boldsymbol{\beta}_0$ 是这个紧集中的一个内点,此外令 $B_n = \{\boldsymbol{\beta} \in B : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leqslant Cn^{-\frac{1}{2}}\}$,其中 C 为任意的一个正常数. 为了得到前面所提的定理,本文给出下面的条件:

1) $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ 的密度函数 $f(t)$,$t = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ 满足具有紧支撑,Lipschitz 连续,且在定义域 T 上满足 $\inf_{t\in T} f(t) > 0$,其中 $T = \{t = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{X} \in A, \boldsymbol{\beta} \in B_n\}$. 给定 t 的条件密度函数 $f_{X|t}(\boldsymbol{X}\mid t)$,$f_{Z|t}(\boldsymbol{Z}\mid t)$ 具有关于 $t \in T$ 上直到二阶的有界连续导数.

2) 函数 $g(t)$,$m_1(t)$,$m_{2j}(t)$,$m_{10}(t)$ 和 $m_{20j}(t)$ 在 T 上具有直到二阶的有界连续导数,其中 $m_{2j}(t)$ 和 $m_{20j}(t)$ 分别是 $\boldsymbol{m}_2(t)$ 和 $\boldsymbol{m}_{20}(t)$ 的第 j 个成分. 其中 $m_{10}(t) = E(\tilde{Y}\mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 = t)$,$\boldsymbol{m}_{20}(t) = E(\boldsymbol{Z}\mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 = t)$,$\tilde{Y}_i = \delta_i Y_i + (1-\delta_i)(g(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\theta}_0)$.

3) $E[\delta = 1 \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t, \boldsymbol{Z} = \boldsymbol{z}] = p(t,\boldsymbol{z})$ 具有直到二阶的有界偏导且满足 $\inf_{t\in T, \boldsymbol{z}\in A_Z} p(t,\boldsymbol{z}) > 0$.

4) 核函数 $K_i(\cdot) = K(\cdot)$,$i = 1,2,3,4$ 是满足 Lipschitz 连续的有界对称密度函数,且其支撑集为 $(-1,1)$.

5) 参数 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,其中 $\boldsymbol{\Theta}$ 是一个有界闭集.

6) 窗宽满足 $0 < h_\gamma \to 0$,且对某些 $\iota < 1 - s^{-1}$ 有 $n^{2\iota-1}h_\gamma \to \infty$,其中 $s > 2$,$\gamma = 1,3$. $\frac{\log n}{nh_1} \to 0$,$\frac{\log n}{nh_4^3} \to 0$,$nh_1^2 \to \infty$,$nh_2^3 \to \infty$,$nh_2h_3 \to \infty$,$nh_2h_4 \to \infty$,$\frac{h_3^4}{h_2} \to 0$,$\frac{h_4}{h_2} \to 0$,$nh_2^4 \to 0$.

7) $E\{\delta[p_1(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{Z} - \boldsymbol{m}_2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)]^{\otimes 2}\}$,$E\{[\boldsymbol{Z} - \boldsymbol{m}_{20}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)]^{\otimes 2}\}$,$E\{[g'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)]^2\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}$ 都为正定矩阵,其中 $\boldsymbol{B}^{\otimes 2} = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$.

8) 函数 $E(\boldsymbol{Z}\mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t)$,$E(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t)$ 和 $E[(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}) * (\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}) \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t)]$ 满足 Lipschitz 连续性,其中 $\boldsymbol{A} * \boldsymbol{B}$ 为关于矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的 Hadamard 乘积.

9) $E[|\varepsilon|^4] < \infty$.

证明 在原假设 H_0 下,模型为线性模型,则可以利用最小二乘的方法得到模型的初始估计以及最终的借补估计,利用最小二乘估计的性质,可以知道在响应变量随机缺失的条件下,所得到的初始估计和借补估计为 $\sqrt{n}$ 相合的. 因此在原假设 H_0 下,将估计代入可以得到

$$\begin{aligned}
&\frac{1}{2}(R_{SS_0^I} - R_{SS_1^I}) = \\
&\frac{1}{2}\sum_{i=1}^n \{\delta_i(\alpha_0 - \bar{\alpha}_{0nc}) + \delta_i(\alpha_1\boldsymbol{\beta} - \bar{\alpha}_{1nc}\bar{\boldsymbol{\beta}}_{nc})^{\mathrm{T}}\boldsymbol{X}_i + \\
&\delta_i(\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}_{nc})^{\mathrm{T}}\boldsymbol{Z}_i + \delta_i\varepsilon_i + (\bar{\alpha}_{0nc} - \bar{\alpha}_{0nI}) + \\
&(\bar{\alpha}_{1nc}\bar{\boldsymbol{\beta}}_{nc} - \bar{\alpha}_{1nI}\bar{\boldsymbol{\beta}}_{nI})^{\mathrm{T}}\boldsymbol{X}_i + (\bar{\boldsymbol{\theta}}_{nc} - \bar{\boldsymbol{\theta}}_{nI})^{\mathrm{T}}\boldsymbol{Z}_i\}^2 - \\
&\frac{1}{2}\sum_{i=1}^n \{\delta_i[g(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) - \tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc})] + \delta_i(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_{nc})^{\mathrm{T}}\boldsymbol{Z}_i + \\
&\delta_i\varepsilon_i + (\tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc}) - \hat{g}_{nI}(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{nI}) + (\tilde{\boldsymbol{\theta}}_{nc} - \hat{\boldsymbol{\theta}}_{nI})^{\mathrm{T}}\boldsymbol{Z}_i\}^2 := \\
&A_1 + A_2 \qquad (10)
\end{aligned}$$

注意到 A_1 中所包含的估计为原假设线性模型的条件下利用最小二乘方法所得到的估计,利用估计的 $\sqrt{n}$ 相合性,式(10)可以化简为

$$\begin{aligned}
&\frac{1}{2}(R_{SS_0^I} - R_{SS_1^I}) = \\
&-\frac{1}{2}\sum_{i=1}^n \{(1-\delta_i)[\tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc}) - g(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)] + \\
&(1-\delta_i)(\tilde{\boldsymbol{\theta}}_{nc} - \boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{Z}_i + \delta_i\varepsilon_i + [g(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) - \hat{g}_{nI}(\boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{nI})] + \\
&(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{nI})^{\mathrm{T}}\boldsymbol{Z}_i\}^2 + \frac{1}{2}\sum_{i=1}^n \delta_i\varepsilon_i^2 + O_p(1). \qquad (11)
\end{aligned}$$

由于 $nh_2^4 \to 0$,$nh_4^4 \to 0$,并且利用 Lai 等[7] 所给的初始估计的性质可得:

$$\begin{aligned}
&\tilde{g}_{nc}(t) - g(t) = \\
&\frac{1}{n}\sum_{i=1}^n \frac{1}{p_1(t)f_1(t)}\delta_i K_{h_2}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - t)\varepsilon_i + o_p\left(\frac{1}{\sqrt{nh_2}}\right) := \\
&e_{h2}(t) + o_p\left(\frac{1}{\sqrt{nh_2}}\right),
\end{aligned}$$

其中 $t = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0$,$p_1(t) = P(\delta = 1\mid t)$,$f_1(t)$ 是关于 t 的密度函数. 类似的,可以证明

$$\begin{aligned}
&\hat{g}_{nI}(t) - g(t) = \\
&\frac{1}{n}\sum_{i=1}^n \frac{1}{f_1(t)}\delta_i K_{h_4}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - t)\varepsilon_i + o_p\left(\frac{1}{\sqrt{nh_4}}\right) := \\
&e_{h4}(t) + o_p\left(\frac{1}{\sqrt{nh_4}}\right),
\end{aligned}$$

此外,经过简单的证明还可得:

$$\tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc}) - \tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) =$$

$g'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}}_{nc}-\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{X}_i+[\tilde{g}'_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-g'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)](\tilde{\boldsymbol{\beta}}_{nc}-\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{X}_i+o_p(n^{-\frac{1}{2}})$.

利用 $\tilde{g}_{nc}(t)-g(t)$,$\hat{g}_{nl}(t)-g(t)$ 和 $\tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{nc})-\tilde{g}_{nc}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)$ 的展开表达式,就可以得到式(11)中的平方项展开后各项的阶,则式(11)可以化简为

$$(11)=\sum_{i=1}^{n}\delta_i\varepsilon_i e_{h_4}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-\frac{1}{2}\sum_{i=1}^{n}(1-\delta_i)e_{h_2}^2(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)-\frac{1}{2}\sum_{i=1}^{n}e_{h_4}^2(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)+\sum_{i=1}^{n}(1-\delta_i)e_{h_4}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)e_{h_2}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)+o_p(h_4^{-\frac{1}{2}}):=T_1-T_2-T_3+T_4+o_p(h_4^{-\frac{1}{2}}),\quad(12)$$

其中 $e_{h_2}(t)=\frac{1}{np_1(t)f_1(t)}\sum_{i=1}^{n}K_{h_2}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0-t)\delta_i\varepsilon_i$,

$e_{h_4}(t)=\frac{1}{nf_1(t)}\sum_{i=1}^{n}K_{h_4}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0-t)\delta_i\varepsilon_i$.

则由式(12)可知:

$$T_1=\sum_{i=1}^{n}\delta_i\varepsilon_i\frac{1}{nf_1(t_i)}\sum_{k=1}^{n}K_{h_4}(t_k-t_i)\delta_k\varepsilon_k=\frac{\sigma^2K(0)}{h_4}\int p_1(t)\mathrm{d}t+\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\frac{\delta_i\delta_k\varepsilon_i\varepsilon_k}{f_1(t_i)}K_{h_4}(t_k-t_i)+O_p\left(\frac{1}{\sqrt{nh_4^2}}\right),$$

$$T_2=\frac{1}{2n^2}\sum_{i=1}^{n}\frac{1-\delta_i}{p_1^2(t_i)f_1^2(t_i)}\Big[\sum_{k=1}^{n}K_{h_2}(t_k-t_i)\delta_k\varepsilon_k\Big]^2=\frac{1}{2n^2}\sum_{i=1}^{n}\frac{1-\delta_i}{p_1^2(t_i)f_1^2(t_i)}K_{h_2}^2(t_k-t_i)\delta_k\varepsilon_k^2+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\frac{1-\delta_i}{p_1^2(t_i)f_1^2(t_i)}K_{h_2}(t_k-t_i)K_{h_2}(0)\delta_i\delta_k\varepsilon_i\varepsilon_k+\frac{1}{2n^2}\sum_{i,j,k=1;i\neq k,k\neq j,j\neq i}^{n}\frac{1-\delta_i}{p_1^2(t_i)f_1^2(t_i)}K_{h_2}(t_k-t_i)K_{h_2}(t_j-t_i)\delta_j\delta_k\varepsilon_j\varepsilon_k:=T_{21}+T_{22}+T_{23},$$

$$T_{21}=\frac{n(n-1)}{4n^2}\frac{2}{n(n-1)}\sum_{i<k}\Big[\frac{(1-\delta_i)\delta_k\varepsilon_k^2}{p_1^2(t_i)f_1^2(t_i)}\times K_{h_2}^2(t_k-t_i)+\frac{(1-\delta_k)\delta_i\varepsilon_i^2}{p_1^2(t_k)f_1^2(t_k)}K_{h_2}^2(t_i-t_k)\Big]=\frac{n(n-1)}{4n^2}U_n,$$

其中 U_n 是一个 U 统计量,那么可以利用 U 统计量的性质得到

$$T_{21}=\frac{\sigma^2}{2h_2}\int\frac{1-p_1(t)}{p_1(t)}\mathrm{d}t\int K^2(u)\mathrm{d}u+O_p\left(\frac{1}{\sqrt{nh_2^2}}\right).$$

从 T_{22} 的表达式可以看出 $T_{22}=O_p\left(\frac{1}{\sqrt{n^2h_2^3}}\right)$. 令

$$T_{23}=\frac{1}{2n^2}\sum_{i,j,k=1;i\neq k,k\neq j,j\neq i}^{n}\Delta_{ijk}\delta_j\delta_k\varepsilon_j\varepsilon_k,$$

$$\Delta_{ijk}\delta_j\delta_k\varepsilon_j\varepsilon_k=\frac{1-\delta_i}{p_1^2(t_i)f_1^2(t_i)}K_{h_2}(t_k-t_i)K_{h_2}(t_j-t_i),$$

则根据

$$E\Big(\frac{1}{n}\sum_{i\neq j,k}^{n}\Delta_{ijk}\,\Big|\,t_j,t_k\Big)=\frac{1-p_1(t_k)}{h_2p_1^2(t_k)f_1(t_k)}\int K(u)K\Big(u-\frac{t_j-t_k}{h_2}\Big)\mathrm{d}u(1+O_p(h_2)),$$

$$E\Big\{\frac{1}{n}\sum_{i\neq j,k}^{n}[\Delta_{ijk}-E(\Delta_{ijk}\,|\,t_j,t_k)]\Big\}^2\leqslant\frac{1}{n^2}\sum_{i\neq j,k}^{n}E(\Delta_{ijk}^2)=O\Big(\frac{1}{nh_2^2}\Big),$$

经过直接计算可以得到:

$$T_{23}=\frac{1}{2n}\sum_{j,k=1,j\neq k}^{n}\frac{1-p_1(t_k)}{h_2p_1^2(t_k)f_1(t_k)}\delta_j\delta_k\varepsilon_j\varepsilon_k\int K(u)K\Big(u-\frac{t_j-t_k}{h_2}\Big)\mathrm{d}u+O_p\Big(1+\frac{1}{\sqrt{nh_2^2}}\Big).$$

本文证得:

$$T_2=\frac{1}{2n}\sum_{j,k=1,j\neq k}^{n}\frac{1-p_1(t_k)}{h_2p_1^2(t_k)f_1(t_k)}\delta_j\delta_k\varepsilon_j\varepsilon_k\int K(u)K\Big(u-\frac{t_j-t_k}{h_2}\Big)\mathrm{d}u+\frac{\sigma^2}{2h_2}\int\frac{1-p_1(t)}{p_1(t)}\mathrm{d}t\int K^2(u)\mathrm{d}u+O_p\Big(1+\frac{1}{\sqrt{nh_2^2}}\Big).$$

类似于化简 T_2 的方法,可以得到:

$$T_3=\frac{\sigma^2}{2h_4}\int p_1(t)\mathrm{d}t\int K^2(u)\mathrm{d}u+O_p\Big(1+\frac{1}{\sqrt{nh_4^2}}\Big)+\frac{1}{2n}\sum_{j,k=1,j\neq k}^{n}\frac{1}{h_4f_1(t_k)}\delta_j\delta_k\varepsilon_j\varepsilon_k\int K(u)K\Big(u-\frac{t_j-t_k}{h_4}\Big)\mathrm{d}u$$

以及

$$T_4=\sum_{i=1}^{n}(1-\delta_i)\frac{1}{n}\sum_{j=1}^{n}\frac{\delta_j\varepsilon_j}{f_1(t_i)}K_{h_4}(t_j-t_k)\frac{1}{n}\sum_{k=1}^{n}\frac{\delta_k\varepsilon_k}{p_1(t_i)f_1(t_i)}K_{h_2}(t_k-t_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\frac{(1-\delta_i)\delta_k\varepsilon_k^2}{p_1(t_i)f_1^2(t_i)}K_{h_4}(t_k-t_i)K_{h_2}(t_k-t_i)+\frac{1}{n^2}\sum_{i,j,k=1;i\neq k,k\neq j,j\neq i}^{n}\frac{(1-\delta_i)\delta_j\delta_j\varepsilon_k\varepsilon_k}{p_1(t_i)f_1^2(t_i)}K_{h_4}(t_j-t_i)K_{h_2}(t_k-t_i)=\frac{1}{n}\sum_{j,k=1,j\neq k}^{n}\frac{1-p_1(t_j)}{h_2p_1(t_j)f_1(t)}\delta_j\delta_k\varepsilon_j\varepsilon_k\int K(u)K\Big(\frac{h_4}{h_2}u-\frac{t_k-t_j}{h_2}\Big)\mathrm{d}u+\frac{\sigma^2}{h_2}\int[1-p_1(t)]\mathrm{d}t\int K(u)K\Big(\frac{h_4}{h_2}u\Big)\mathrm{d}u+O_p\Big(1+\frac{1}{\sqrt{nh_2h_4}}\Big).$$

由此,根据 T_1,T_2,T_3,T_4 的表达式,(12)可以变换为

$$\frac{1}{2}(R_{SS_0^l}-R_{SS_1^l})=\sigma^2\mu_n+W_n+o_p(h_4^{-\frac{1}{2}}),\quad(13)$$

其中

$$\mu_n = \frac{1}{h_4}K(0)\int p_1(t)\mathrm{d}t - \frac{1}{2h_4}\int p_1(t)\mathrm{d}t\int K^2(u)\mathrm{d}u - \frac{1}{2h_2}\int\frac{1-p_1(t)}{p_1(t)}\mathrm{d}t\int K^2(u)\mathrm{d}u + \frac{1}{h_2}\int(1-p_1(t))\mathrm{d}t\int K(u)K\left(\frac{h_4}{h_2}u\right)\mathrm{d}u, \quad (14)$$

$$W_n = \sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\left\{\frac{\delta_i\delta_k\varepsilon_i\varepsilon_k}{nf_1(t_k)}\left[K_{h_4}(t_k-t_i) - \frac{1}{2}K_{h_4}*K_{h_4}(t_i-t_k)\right] - \frac{[1-p_1(t_k)]\delta_i\delta_k\varepsilon_i\varepsilon_k}{2nf_1(t_k)p_1^2(t_k)}K_{h_2}*K_{h_2}(t_i-t_k) + \frac{[1-p_1(t_k)]\delta_i\delta_k\varepsilon_i\varepsilon_k}{np_1(t_k)f_1(t_k)}\int K(u)K_{h_2}[h_4u-(t_i-t_k)]\mathrm{d}u\right\},$$

且 $K_{h_s}*K_{h_s}(t_i-t_k) = \frac{1}{h_s}\int K(u)K\left(u-\frac{t_i-t_k}{h_s}\right)\mathrm{d}u$, $s=2,4$. 令 $W_n = \sum_{i,k=1,i\neq k}^{n} w(i,k)$, 其中

$$w(i,k) = \frac{\delta_i\delta_k\varepsilon_i\varepsilon_k}{nf_1(t_k)}\left[K_{h_4}(t_k-t_i) - \frac{1}{2}K_{h_4}*K_{h_4}(t_i-t_k)\right] - \frac{[1-p_1(t_k)]\delta_i\delta_k\varepsilon_i\varepsilon_k}{2nf_1(t_k)p_1^2(t_k)}K_{h_2}*K_{h_2}(t_i-t_k) + \frac{[1-p_1(t_k)]\delta_i\delta_k\varepsilon_i\varepsilon_k}{np_1(t_k)f_1(t_k)}\int K(u)K_{h_2}[h_4u-(t_i-t_k)]\mathrm{d}u$$

注意到 $\{\varepsilon_i\}_{i=1}^n$ 是独立同分布的随机序列,且满足 $E(\varepsilon_i)=0$,所以有 $E(W_n)=0$. 经过直接计算可得 $\mathrm{Var}(W_n)=\sigma^4v_n^2[1+o(1)]$,其中

$$v_n^2 = \frac{1}{h_4}\int p_1^2(t)\mathrm{d}t\int\left[\left[K(v)-\frac{1}{2}K*K(v)\right]^2 + \left[K(v)-\frac{1}{2}K*K(v)\right]\left[K(v)-\frac{1}{2}K*K(-v)\right]\right]\mathrm{d}v + \frac{1}{4h_2}\int\frac{[1-p_1(t)]^2}{p_1^2(t)}\mathrm{d}t\int[[K*K(v)]^2 + [K*K(v)][K*K(-v)]]\mathrm{d}v + \frac{1}{h_2}\int(1-p_1(t))^2\mathrm{d}t\int\left[\left[\int K(u)K\left(\frac{h_4}{h_2}u-v\right)\mathrm{d}u\right]^2 + \left[\int K(u)K\left(\frac{h_4}{h_2}u-v\right)\mathrm{d}u\right]\left[\int K(u)K\left(\frac{h_4}{h_2}u+v\right)\mathrm{d}u\right]\right]\mathrm{d}v - \frac{1}{h_2}\int[1-p_1(t)]\mathrm{d}t\int\left[\left[K(v)-\frac{1}{2}K*K(v)\right]\times\left[K*K\left(\frac{h_4}{h_2}v\right)+K*K\left(-\frac{h_4}{h_2}v\right)\right]\right]\mathrm{d}v + \frac{2}{h_2}\int[1-p_1(t)]p_1(t)\mathrm{d}t\int\left[K(v)-\frac{1}{2}K*K(v)\right]\times\left[\int K(u)K\left(\frac{h_4}{h_2}u-\frac{h_4}{h_2}v\right)\mathrm{d}u\right]\mathrm{d}v - \frac{1}{h_2}\int\frac{[1-p_1(t)]^2}{p_1(t)}\mathrm{d}t\int[[K*K(v)]\times\left[\int K(u)K\left(\frac{h_4}{h_2}u-v\right)\mathrm{d}u+\int K(u)K\left(\frac{h_4}{h_2}u+v\right)\mathrm{d}u\right]]\mathrm{d}v + \frac{1}{h_2}\int[1-p_1(t)]\mathrm{d}t\int\left[K(v)-\frac{1}{2}K*K(v)\right]\times\left[\int K(u)K\left(\frac{h_4}{h_2}u+\frac{h_4}{h_2}v\right)\mathrm{d}u\right]\mathrm{d}v, \quad (15)$$

那么由式(14),可知

$$W_n = O_p\left(\frac{1}{\sqrt{h_4}}+\frac{1}{\sqrt{h_2}}\right) = O_p\left(\frac{1}{\sqrt{h_4}}\right),$$

回想式(13),有

$$\frac{R_{SS_1^I}}{n} = \frac{R_{SS_0^I}}{n} - \frac{2\sigma^2\mu_n}{n} - \frac{2W_n}{n} + o_p\left(\frac{1}{n\sqrt{h_4}}\right) = \sigma^2(1+o_p(1)),$$

也就是说

$$\lambda_n^I - \mu_n = \frac{W_n}{\sigma^2} + o_p(h_4^{-\frac{1}{2}}).$$

那么类似于 Fan 等[8]和 Zhang[5]关于 W_n 渐近分布的讨论,通过验证文献[9]中的命题 3.2 的条件满足,由此命题可以得到 W_n 是渐近正态的,即:

$$\sigma^{-2}v_n^{-1}W_n \xrightarrow{D} N(0,1),$$

由此得到定理结果:

$$v_n^{-1}(\lambda_n^I-\mu_n) \xrightarrow{D} N(0,1).$$

参考文献
References

[1] Carroll R J, Fan J Q, Gijbels I, et al. Generalized partially linear single-index models[J]. Journal of the American Statistical Association, 1997, 92(438): 477-489
[2] Xia Y C, Härdle W. Semi-parametric estimation of partially linear single-index models[J]. Journal of Multivariate Analysis, 2006, 97(5): 1162-1184
[3] Zhu L X, Xue L G. Empirical likelihood confidence regions in a partially linear single-index model[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2006, 68(3): 549-570
[4] Liang H, Liu X, Li R. Z, et al. Estimation and testing for partially linear single-index models[J]. Annals of Statistics, 2010, 38(6): 3811-3836
[5] Zhang R Q. Tests for nonparametric parts on partially linear single index models[J]. Science in China Series A: Mathematics, 2007, 50(3): 439-449
[6] Little R J A, Rubin D B. Statistical analysis with missing data[M]. New York: Wiley, 2002
[7] Lai P, Wang Q H. Partially linear single-index model with

missing responses at random[J]. Journal of Statistical Planning and Inference,2011,141(2):1047-1058
[8] Fan J Q,Zhang C M,Zhang J. Generalized likelihood ratio statistics and Wilks phenomenon[J]. Annals of Statistics,2001,29(1):153-193
[9] de Jong P. A central limit theorem for generalized quadratic forms[J]. Probability Theory Related Fields,1987,75(2):261-277

Tests for nonparametric parts on partially linear single index model with responses missing at random

LAI Peng[1]

1 School of Mathematics & Statistics,Nanjing University of Information Science & Technology,Nanjing 210044

Abstract This paper studies the hypothesis problem of partially linear single index model with responses missing at random. The nonlinear relationship between the predictors and response variable is tested. To test the nonparametric part of this model,the missing responses are imputed based on the imputation estimators of the parameters and nonparametric function. The generalized likelihood ratio statistic is constructed based on the imputation estimators, and the asymptotic property of the statistic is also proved.

Key words partially linear single index model;missing at random;imputation estimator;generalized likelihood ratio statistic