SIMEX estimation for single-index model with covariate measurement error

姓　名：　张宇靖

中国人民大学统计学院

2021 年 12 月 21 日

SIMEX estimation for single-index model with covariate measurement error

考虑单指标测量误差模型

$$\begin{cases} Y = g\left(\beta^{\mathrm{T}} X\right) + \varepsilon \\ W = X + U \end{cases}$$

其中$g(\cdot)$是一元的未知联系函数

$\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$是$p \times 1$的未知指标参数向量，满足$\|\beta\| = 1$

ε是随机误差，满足$E(\varepsilon \mid X) = 0$

测量误差$U \sim N(0, \Sigma_{uu})$，且独立于(X, Y)，且假设Σ_{uu}已知

SIMEX estimation for single-index model with covariate measurement error

(1) 模拟步: 对于每一个$i=1,\cdots,n$, 产生随机变量序列

$$W_{ib}(\lambda)=W_i+(\lambda\Sigma_{uu})^{1/2}U_{ib},\quad b=1,\cdots,B,$$

其中$U_{ib}\sim N(0,I_p)$, I_p是$p\times p$的单位阵.

(2) 估计步: 假设联系函数$g(\cdot)$有连续的二阶导数, 对于t_0邻域内的点t, $g(t)$能够用一个线性函数进行逼近, 即

$$g(t)\approx g(t_0)+g'(t_0)(t-t_0)\equiv a+b(t-t_0)$$

下面介绍基于数据集$\{(Y_i,W_{ib}(\lambda)),i=1,\cdots,n,b=1,\cdots,B\}$, 关于未知指标参数$\beta$和联系函数$g(\cdot)$的估计问题.

步骤1. 对于给定的t_0和β. 定义下面的加权最小二乘目标函数

$$\sum_{i=1}^{n}\left\{Y_i - a - b\left[\beta^T W_{ib}(\lambda) - t_0\right]\right\}^2 K_h\left(\beta^T W_{ib}(\lambda) - t_0\right),$$

其中$K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$是核函数, h是窗宽. 简单计算, 可得

$$\hat{g}\left(\beta,\lambda;t_0\right) = \sum_{i=1}^{n} M_{ni}\left(\beta,\lambda;t_0\right) Y_i \hat{g}'\left(\beta,\lambda;t_0\right) = \sum_{i=1}^{n} \tilde{M}_{ni}\left(\beta,\lambda;t_0\right) Y_i$$

步骤2. 采用“去一分量”方法转换$\mathbb{R}^p$空间中单位超球面上的点到$\mathbb{R}^p$空间单位球的内点. 具体思路是: $\beta^{(r)}=(\beta_1,\cdots,\beta_{r-1},\beta_{r+1},\cdots,\beta_p)^{\mathrm{T}}$表示去掉$\beta$的第$r$个分量$\beta_r$以后的$p-1$维参数向量. 为了简单, 不妨假定$\beta$的第$r$个分量$\beta_r$是一个正的分量. 这时真参数$\beta^{(r)}$满足$\left\|\beta^{(r)}\right\|<1$. 则$\beta$在真参数$\beta^{(r)}$的某个邻域内是有限可微的, 其Jacobian矩阵是

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

其中$\gamma_s(1\le s\le p, s\ne r)$是一个$(p-1)$维的单位向量, 当$s<r$时, 其第$s$个元素为1, 当$s>r$时, 其第$s-1$个元素为1, 而$\gamma_r=-\left(1-\left\|\beta^{(r)}\right\|^2\right)^{-1/2}\beta^{(r)}$.

定义关于$\beta^{(r)}$的估计方程如下

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{ib}(\beta,\lambda)=0,$$

上式等价于极小化$Q(\beta)=E[Y-\hat{g}(\beta,\lambda;\beta^{T}W)]^{2}$

其中$\hat{\eta}_{ib}(\beta,\lambda)=\left[Y_i-\hat{g}\left(\beta,\lambda;\beta^{\mathrm{T}}W_{ib}(\lambda)\right)\right]\hat{g}'_{h_1}\left(\beta,\lambda;\beta^{\mathrm{T}}W_{ib}(\lambda)\right)J^{\mathrm{T}}_{\beta^{(r)}}W_{ib}(\lambda).$

使用Newton-Raphson迭代算法求解方程，得到$\beta^{(r)}$的估计，记为$\hat{\beta}_b^{(r)}(\lambda)$. 进而可得指标参数$\beta$的估计，记为$\hat{\beta}_b(\lambda)$.

步骤3. 对$\hat{\beta}_b(\lambda)$关于$b=1,\cdots,B$进行平均，则$\hat{\beta}(\lambda)$定义为

$\hat{\beta}(\lambda)=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b(\lambda)$

(3) 外推步: 定义外推函数$G(\lambda, \Gamma)$. 基于外推函数$G(\lambda, \Gamma)$, 对$\{\hat{\beta}(\lambda), \lambda \in \Lambda\}$在$\{\lambda \in \Lambda\}$上关于$\Gamma$作拟合回归模型, 可以获得$\Gamma$的估计, 记为$\hat{\Gamma}$. 最后可得到$\beta$的SIMEX估计定义为

$$\hat{\beta}_{\mathrm{SIMEX}} = G(-1, \hat{\Gamma}).$$

注意到: 如果λ退化到0, $\hat{\beta}_{\text{Naive}} = G(0, \hat{\Gamma})$就表示忽略测量误差后得到的估计, 即直接用$W$代替$X$得到的估计量.

在估计步中的步骤1中，用SIMEX估计$\hat{\beta}_{\text{SIMEX}}$代替$\beta$，并且用窗宽$h_2$获得估计$\hat{g}_b(\lambda;t_0)$，然后关于$b=1,\cdots,B$作平均，则有

$$\hat{g}(\lambda;t_0)=\frac{1}{B}\sum_{b=1}^{B}\hat{g}_b(\lambda;t_0)$$

关于$\mathbb{A}$，极小化

$$\sum_{\lambda\in\Lambda}\{\hat{g}(\lambda;t_0)-G(\lambda;\mathbb{A})\}^2$$

则可获得估计$\hat{\mathbb{A}}$. 进一步，可获得联系函数$g(\cdot)$的SIMEX估计为

$$\hat{g}_{\text{SIMEX}}(t_0)=G(-1,\hat{\mathbb{A}})$$

渐近性质

定理 1

假设正则条件$(C1)$一条件$(C7)$成立。当$n \to \infty$时，则有$\sqrt{n}\left(\hat{\beta}_{\mathrm{SIMEX}} - \beta\right)$ 是渐近正态的，均值为0，渐近协方差阵为

$$G_{\Gamma}(-1, \mathbf{\Gamma})\Sigma(\mathbf{\Gamma})\left\{G_{\mathbf{\Gamma}}(-1, \mathbf{\Gamma})\right\}^{\mathrm{T}}$$

其中$G_{\Gamma}(\lambda, \mathbf{\Gamma}) = \left\{\partial/\partial(\mathbf{\Gamma})^{\mathrm{T}}\right\} G(\lambda, \mathbf{\Gamma})$.

渐近性质

定理 2

假设正则条件(C1)一条件(C7) 成立. 如果$nh_2^5=O(1)$, 当$n\to\infty$和$B\to\infty$ 时, 则联系函数的$SIMEX$估计$\hat{g}_{\mathrm{SIMEX}}(t_0)$ 的渐近偏差和渐近方差分别为

$$C(\Lambda,\mathbb{A})\sum_{\lambda\in\Lambda}\frac{1}{2}h_2^2\mu_2 g''(\lambda;t_0)\gamma(\lambda,\mathbb{A})$$

和

$$\left[nh_2 f_0(t_0)\right]^{-1}\nu_2\operatorname{Var}\left(Y\mid\beta^{\mathrm{T}}W=t_0\right)C(\Lambda,\mathbb{A})DC^{\mathrm{T}}(\Lambda,\mathbb{A}),$$

其中$g(\lambda;t)=E\left(Y\mid\beta^{\mathrm{T}}W_b(\lambda)=t\right)$.

1. 单指标测量误差模型
2. SIMEX估计程序
3. 渐近性质
4. 模拟研究

考虑下面的单指标测量误差模型

$$\begin{cases} Y_i = -2\left(\beta^{\mathrm{T}} X_i - 1\right)^2 + 1 + \varepsilon_i, \\ W_i = X_i + U_i, \quad i = 1, \cdots, n, \end{cases}$$

其中

- $\beta = (\beta_1, \beta_2)^{\mathrm{T}} = (\sqrt{3}/3, \sqrt{6}/3)^{\mathrm{T}}$
- X_i是一个二维的随机向量，每个分量独立的来自于$N(0,1)$
- $\varepsilon_i \sim N\left(0, 0.2^2\right)$
- $U_i \sim N\left(0, \mathrm{diag}\left(\sigma_u^2, 0\right)\right)$，取$\sigma_u = 0.2, 0.4, 0.6$表示不同的测量误差水平

- $n=50,100$和150
- 重复模拟次数为$N=500$
- 对于SIMEX算法，取$\lambda=0,0.2,\cdots,2$和$B=50$
- Epanechnikov核函数$K(u)=0.75\left(1-u^2\right)_+$
- h,h_1和h_2可分别取为$cn^{-1/4}(\log n)^{-1/2}, cn^{-1/5}$ 和$cn^{-1/5}$，其中c为$\beta_{\text{int}}^{\text{T}}X$的标准差

表1: β_1 和β_2的SIMEX估计和自然估计的偏差(Bias)和标准差(SD)

		SIMEX		Naive	
		β_1	β_2	β_1	β_2
n	σ_u	Bias(SD)	Bias(SD)	Bias(SD)	Bias(SD)
	0.2	−0.0084(0.0520)	−0.0078(0.0377)	−0.0177(0.0291)	0.0146(0.0203)
50	0.4	−0.0405(0.0875)	0.0171(0.0638)	−0.0764(0.0537)	0.0546(0.0388)
	0.6	−0.0508(0.1253)	0.0342(0.0821)	−0.1207(0.0680)	0.0700(0.0330)
	0.2	−0.0083(0.0384)	−0.0074(0.0321)	−0.0126(0.0203)	0.0084(0.0142)
100	0.4	−0.0381(0.0581)	0.0158(0.0334)	−0.0761(0.0397)	0.0434(0.0224)
	0.6	0.0394(0.0719)	−0.0206(0.0567)	−0.1154(0.0383)	0.0632(0.0210)
	0.2	−0.0059(0.0169)	0.0039(0.0118)	−0.0187(0.0136)	0.0127(0.0093)
150	0.4	0.0160(0.0341)	−0.0126(0.0258)	−0.0497(0.0279)	0.0324(0.0177)
	0.6	−0.0279(0.0599)	0.0163(0.0394)	−0.1088(0.0315)	0.0563(0.0171)

联系函数$g(t)$估计$\hat{g}(t)$的完成情况，用均方根误差(root mean squared error, RMSE)进行评价

$$\mathrm{RMSE}=\left[n_{\text{grid}}^{-1}\sum_{k=1}^{n_{\text{grid}}}\left\{\hat{g}\left(t_{k}\right)-g\left(t_{k}\right)\right\}^{2}\right]^{1/2}$$

其中n_{grid}是格子点数，且$\{t_k, k=1,2,\ldots,n_{\text{grid}}\}$是等间距的格子点，模拟计算中取$n_{\text{grid}}=15$.

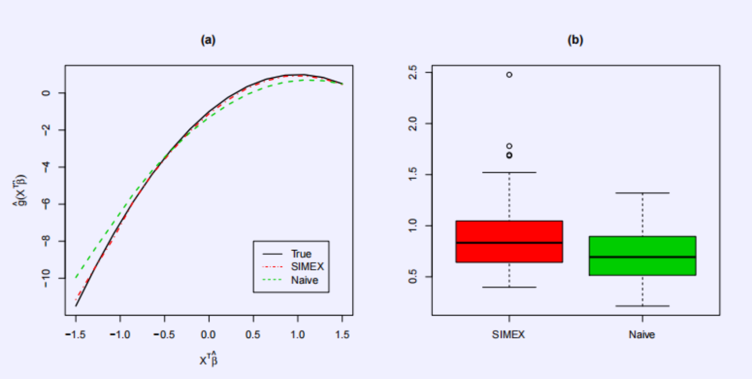

图2: (a) 实线表示联系函数$g(t)$的真实曲线, 断线表示自然估计的拟合曲线, 点断线表示SIMEX估计的拟合曲线, 其中$n = 100$和$\sigma_u = 0.4$; (b) 基于500次重复试验的联系函数$g(t)$两种估计的RMSE的箱线图

谢 谢!

2021.12.21