

文章编号:1673-5005(2007)04-0162-06

回归误差方差的区间估计

扈慧敏¹, 芮杰², 徐兴忠¹

(1. 北京理工大学 理学院, 北京 100081; 2. 中国石油大学 数学与计算科学学院, 山东 东营 257061)

摘要: 首先将回归函数限制在一个有限维函数空间中, 得到一个近似的线性模型, 在此基础上, 由 Fiducial 推断给出了误差方差的区间估计。该区间估计形式简单, 易于计算。给出了这个区间估计的真实覆盖率和名义水平差异的一个上界, 该上界由回归函数与其近似的线性函数的距离界定, 并且对所给区间估计的真实覆盖率进行了数值模拟。

关键词: 非参数回归; 误差方差; 区间估计; 置信水平

中图分类号: O 212.7 **文献标识码:** A

An interval estimator of error variance in nonparametric regression model

HU Hui-min¹, RUI Jie², XU Xing-zhong¹

(1. School of Science, Beijing Institute of Technology, Beijing 100081, China;

2. College of Mathematics and Computational Science in China University of Petroleum, Dongying 257061, Shandong Province, China)

Abstract: Firstly an approximate linear model was obtained by confining the regression function in a functional space with finite dimension. Then the interval estimator was proposed by Fiducial inference. The interval estimator has a simple form and is easy to be calculated. An upper bound of the difference between the true and the nominal level of the interval estimator was given. It is dominated by the distance between the regression function and its approximating function. True percentage of coverage for the interval estimator was also simulated.

Key words: nonparametric regression; error variance; interval estimator; confidence level

1 问题的提出

考虑非参数回归模型

$$y = g(z) + \varepsilon, \quad (1)$$

其中 z 在 $[0, 1]^q$ 上取值, $g(z)$ 是未知的函数, $\varepsilon \sim N(0, \sigma^2)$, $\sigma > 0$ 未知。若进行 n 次独立观测, 有

$$Y = G + \varepsilon, \quad (2)$$

其中 $Y = (y_1, \dots, y_n)^T$, $G = (g(z_1), \dots, g(z_n))^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ 。本文中要给出误差方差 σ^2 的一个区间估计。

对于非参数问题的点估计的研究, 历史上已经有很多方法, 其中包括直接利用非参数方法例如核

估计^[1], 以及把非参数部分参数化, 然后用参数的方法估计, 如最小二乘估计^[2], M 估计^[3], Profile 估计^[4-6], Sieve 估计^[5-7], 局部线性化^[8] 和局部多项式^[9] 估计等。在非参数回归模型中, 对于回归误差方差的研究可参看文献[10]~[12]等。在这些文献中, 用 σ 的任意 $\sqrt{n}$ 相合估计作为 σ 的估计, 这是具有大样本性质的估计。笔者用 Fiducial 方法给出对于小样本的情况仍然适用的 σ^2 的区间估计。

2 误差方差的区间估计

有线性回归模型

$$Y = X\beta + \varepsilon. \quad (3)$$

收稿日期: 2007-03-21

基金项目: 国家自然科学基金(10271013)

作者简介: 扈慧敏(1981-), 女(汉族), 山东枣庄人, 博士研究生, 研究方向为统计推断。

式中, Y 为 n 维随机向量; X 为已知的 $n \times p$ 阶列满秩矩阵; β 为未知的回归系数矩阵; ε 为误差向量, 服从 $N(0, \sigma^2 I_n)$ 分布, σ^2 是未知的误差方差。在模型(3)下,

$$\begin{cases} \hat{\beta} = (X^T X)^{-1} X^T Y, \\ S^2 = Y^T (I - P_X) Y \end{cases} \quad (4)$$

是充分统计量, 其中 $P_X = X(X^T X)^{-1} X^T$, 所以以后的讨论基于它们进行。

注意 $\hat{\beta}$ 和 S^2 可表示成如下函数模型^[13-14]:

$$\begin{cases} \hat{\beta} = \beta + \sigma (X^T X)^{-1} X^T V, \\ S^2 = \sigma^2 W^2, \end{cases}$$

其中

$$\begin{cases} V \sim N(0, I_n), \\ W^2 = S^2 / \sigma^2 \sim \chi^2(n-p), \end{cases}$$

且 V 和 W^2 独立。由模型(4), 在给定 Y 和 S^2 的观测值 y 和 s^2 后, 可得 σ^2 的 Fiducial 模型 $\Sigma^2 = \frac{s^2}{W^2}$, 因此 σ^2 的区间估计为

$$\left[\frac{S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-p)}, \frac{S^2}{\chi^2_{\frac{\alpha}{2}}(n-p)} \right]. \quad (5)$$

其中 $\chi^2_t(n-p)$ 是自由度为 $n-p$ 的 χ^2 的 t 分位数。

考虑模型(1), 对非参数回归函数 $g(z)$ 用一类已知函数的线性组合近似。设 $g_j(z)$ ($j=1, \dots, m$) 是 $[0, 1]$ 上的已知函数, 对 $g(z)$ 作如下近似:

$$g(z) \approx \sum_{j=1}^m \beta_j g_j(z), \quad (6)$$

则模型(1)可近似地表示为 $y \approx \sum_{j=1}^m \beta_j g_j(z) + \varepsilon$, 模型(2)可近似地表示为

$$Y \approx X\beta + \varepsilon. \quad (7)$$

其中

$$\begin{aligned} \beta &= (\beta_1, \dots, \beta_m)^T, \\ X &= \begin{pmatrix} g_1(z_1) & \cdots & g_m(z_1) \\ \vdots & \cdots & \vdots \\ g_1(z_n) & \cdots & g_m(z_n) \end{pmatrix}. \end{aligned} \quad (8)$$

这样得到了一个在形式上与模型(3)一样的模型, 则 σ^2 的一个类似于式(5)的区间估计为

$$D_{1-\alpha}(Y) = \left[\frac{S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-m)}, \frac{S^2}{\chi^2_{\frac{\alpha}{2}}(n-m)} \right]. \quad (9)$$

其中 $1-\alpha$ 为给定的置信水平, S^2 和 X 由式(4)和式(8)给出。

当回归函数为多维时, 即 $z \in \mathbf{R}^r$ 时, 同一维情况一样, 若有式(6), 则 σ^2 的置信区间也为式(9)。为研究式(9)所给出的区间估计 $D_{1-\alpha}(Y)$ 的真实置信

水平, 先给出一个一般性结论。

引理1 设 ξ 和 η 为两个非负随机变量, 它们的分布函数分别为 $F(x)$ 和 $G(x)$ 。 ξ 的密度函数有上界 M , $|\frac{\eta}{\xi} - 1|$ 的数学期望存在, 则对给定的 x_0 , 有

$$|G(x_0) - F(x_0)| \leq 2 \sqrt{M x_0 E \left| \frac{\eta}{\xi} - 1 \right|} + E \left| \frac{\eta}{\xi} - 1 \right|.$$

由引理1可得定理1。

定理1 在模型(1)中, 由式(9)给出的 σ^2 的区间估计的真实覆盖率满足

$$\begin{aligned} & \left| P \left(\frac{S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-m)} \leq \sigma^2 \leq \frac{S^2}{\chi^2_{\frac{\alpha}{2}}(n-m)} \right) - (1-\alpha) \right| \\ & \leq 2 \frac{1}{\sqrt{\sigma}} \left(\sqrt{AB_{\sigma,d}} (\chi_{1-\frac{\alpha}{2}} + \chi_{\frac{\alpha}{2}}) \right) d^{\frac{1}{2}} + 2 \frac{1}{\sigma} B_{\sigma,d}. \end{aligned}$$

其中

$$A = \frac{\left(\frac{n-m-2}{2} \right)^{\frac{n-m-2}{2}-1}}{2 \Gamma \left(\frac{n-m-2}{2} \right)} e^{-\frac{n-m-2}{2}}, \quad (10)$$

$$B_{\sigma,d} = \frac{1}{\sigma} \frac{n}{n-m-2} d + \sqrt{2} \frac{\sqrt{n} \Gamma \left(\frac{n-m-1}{2} \right)}{\Gamma \left(\frac{n-m}{2} \right)},$$

$$d = \inf_{\beta \in \mathbf{R}^m} \sup_{z \in [0,1]^r} \left| g(z) - \sum_{j=1}^m \beta_j g_j(z) \right|.$$

不妨设 $\frac{n-m-2}{2} = c$, 由 Stirling 公式知

$$A = \frac{c^c}{2c!} e^{-c} = \frac{c^c e^{-c}}{c^c e^{-c} 2 \sqrt{2\pi c} e^{\theta_c}} = \frac{1}{2 \sqrt{2\pi c}} e^{-\theta_c},$$

其中, $0 < \theta_c < \frac{1}{12c}$, 因此 A 有界。由于 $\lim_{n \rightarrow \infty}$

$\left\{ \left[\Gamma \left(\frac{n-m+1}{2} \right) \right] / \left[\sqrt{n} \Gamma \left(\frac{n-m}{2} \right) \right] \right\} = 1/\sqrt{2}$, 因此

$B_{\sigma,d}$ 中 $\left[\sqrt{n} \Gamma \left(\frac{n-m-1}{2} \right) \right] / \Gamma \left(\frac{n-m}{2} \right) (n \rightarrow \infty)$ 有界。

3 数值模拟

分别考虑回归函数自变量的维数 $q=1$ 和 $q=2$ 的情况。首先, 把回归函数用已知函数的线性组合来近似, 从而得到模型(7)。事实上, 有很多方法可以对非参数函数 $g(z)$ 进行这种线性近似, 这里考虑3种方法, 即用已知函数的分段线性组合对 $g(z)$ 近似、对 $g(z)$ 进行局部线性化以及取 $g(z)$ 的局部多项式。

在模拟中取显著性水平 $\alpha=0.05, 0.1$; 误差标准差 $\sigma=1, 10, 100, 1000$ 。对每组不同的 σ 和 $g(z)$

的选取,取10000组样本,每组样本的容量 n 分别取为30,50,100。

3.1 $q=1$ 时的模拟结果

回归函数 $g(z)$ 分别取为 $g(z)=e^z, (z-1/2)^2, \sin(2\pi z), \sin(4\pi z)$ 且 $z \in U[0,1]$ 。分别就回归函数的3种线性近似方法来进行模拟。

3.1.1 用已知函数的分段线性组合来近似 $g(z)$

取 $[0,1]$ 的一个均匀分割 $0=z^{(1)} < \dots < z^{(m)} = 1$,满足 $|z^{(j)} - z^{(j-1)}| \leq Hn^{-k}$,其中 H, k 为常数, $j=2, \dots, m$ 。定义 $l_j(z) = l(z^{(j-1)} \leq z \leq z^{(j)})$, $2 \leq j \leq m-1$; $l_m(z) = l(z^{(m-1)} \leq z \leq z^{(m)})$,则 $g(z)$ 可以用线性函数组合式(6)来近似,其中

$$\begin{cases} g_1(z) = -\frac{z-z^{(2)}}{z^{(2)}-z^{(1)}}l_2(z), \\ g_{j-1}(z) = \frac{z-z^{(j-2)}}{z^{(j-1)}-z^{(j-2)}}l_{j-1}(z) - \frac{z-z^{(j)}}{z^{(j)}-z^{(j-1)}}l_j(z), \quad 3 \leq j \leq m, \\ g_m(z) = -\frac{z-z^{(m-1)}}{z^{(m)}-z^{(m-1)}}l_m(z). \end{cases} \quad (11)$$

分别考虑节点 $m=6, 11$ 的情况,可以按照以下模拟步骤计算区间估计的真实覆盖率:

①以 z 的分布 $U[0,1]$ 产生容量为 n 的随机样本,代入式(11),得式(7)中的回归系数 X 。

②以模型(1)中误差项 ε 的分布 $N(0, \sigma^2)$ 产生容量为 n 的随机样本,由式(1)得 Y 的观察值 Y_{obs} 。再代入式(4)和式(9),得区间估计 $D(Y_{obs})$ 。

③重复步骤②10000次,可得 $D_1(Y_{obs}), \dots, D_{10000}(Y_{obs})$ 。由此可以计算基于 Y_{obs} 观察值和 s 基础上真实覆盖率 $1-\alpha$ 的估计 $\hat{p} = \#\{\sigma^2 \in D_i(Y_{obs}), i=1, \dots, 10000\}/10000$ (表1)。

3.1.2 对 $g(z)$ 进行局部线性化

取定 z_0 ,把 $g(z)$ 在 z_0 处局部线性展开,即

$$g(z) = g(z_0) + g'(z_0)(z - z_0) = \beta_1 + \beta_2(z - z_0) = \beta_1 g_1(z) + \beta_2 g_2(z),$$

其中 $g_1(z) \equiv 1, g_2(z) = z - z_0$ 。因此, $g(z)$ 可用式(6)来近似。类似于上述模拟步骤,计算得出此种近似下 $D_{1-\alpha}(Y)$ 的真实覆盖率,见表2。

表1 局部线性化近似下不同 α 时 $D_{1-\alpha}(Y)$ 的真实覆盖率

α	$g(z)$	n	$m=6$				$m=11$			
			$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$	$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$
0.05	e^z	30	0.9478	0.9470	0.9530	0.9513	0.9525	0.9507	0.9505	0.9505
	e^z	50	0.9547	0.9493	0.9491	0.9493	0.9483	0.9519	0.9501	0.9492
	e^z	100	0.9507	0.9505	0.9477	0.9524	0.9504	0.9504	0.9530	0.9505
	$(z-1/2)^2$	30	0.9478	0.9474	0.9507	0.9498	0.9507	0.9536	0.9501	0.9518
	$(z-1/2)^2$	50	0.9501	0.9499	0.9480	0.9545	0.9530	0.9489	0.9507	0.9483
	$(z-1/2)^2$	100	0.9508	0.9497	0.9465	0.9531	0.9507	0.9501	0.9533	0.9544
	$\sin(2\pi z)$	30	0.9499	0.9497	0.9505	0.9506	0.9500	0.9516	0.9528	0.9476
	$\sin(2\pi z)$	50	0.9503	0.9501	0.9533	0.9512	0.9501	0.9509	0.9504	0.9461
	$\sin(2\pi z)$	100	0.9493	0.9473	0.9545	0.9532	0.9528	0.9470	0.9506	0.9472
	$\sin(4\pi z)$	30	0.9480	0.9486	0.9518	0.9485	0.9545	0.9476	0.9509	0.9499
	$\sin(4\pi z)$	50	0.9370	0.9475	0.9477	0.9514	0.9488	0.9502	0.9536	0.9469
	$\sin(4\pi z)$	100	0.9296	0.9490	0.9503	0.9521	0.9489	0.9472	0.9521	0.9500
0.1	e^z	30	0.9018	0.9006	0.8997	0.9036	0.8974	0.9012	0.9009	0.9000
	e^z	50	0.8992	0.9034	0.9013	0.9005	0.9028	0.8967	0.8961	0.9008
	e^z	100	0.8986	0.9014	0.8997	0.9003	0.8991	0.9005	0.9038	0.9017
	$(z-1/2)^2$	30	0.9020	0.9003	0.9018	0.9020	0.9031	0.9008	0.9000	0.9000
	$(z-1/2)^2$	50	0.8975	0.9004	0.8988	0.8958	0.8977	0.9002	0.9017	0.8998
	$(z-1/2)^2$	100	0.8966	0.8988	0.9001	0.9033	0.8980	0.9004	0.9056	0.8971
	$\sin(2\pi z)$	30	0.8960	0.8968	0.9036	0.9006	0.8995	0.8985	0.9004	0.8991
	$\sin(2\pi z)$	50	0.8865	0.9018	0.9049	0.8939	0.8969	0.8979	0.9039	0.8961
	$\sin(2\pi z)$	100	0.8734	0.8990	0.9064	0.9014	0.8946	0.8978	0.9049	0.9020
	$\sin(4\pi z)$	30	0.9018	0.9006	0.8997	0.9036	0.8974	0.9012	0.9009	0.9000
	$\sin(4\pi z)$	50	0.8992	0.9034	0.9013	0.9005	0.9028	0.8967	0.8961	0.9008
	$\sin(4\pi z)$	100	0.8986	0.9014	0.8997	0.9003	0.8991	0.9005	0.9038	0.9017

表 2 $z_0=1/3, \alpha$ 不同时 $D_{1-\alpha}(Y)$ 的真实覆盖率

α	$g(z)$	n	$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$	α	$g(z)$	n	$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$
0.05	e^z	30	0.9484	0.9480	0.9511	0.9520	0.1	e^z	30	0.8955	0.8978	0.9044	0.9023
	e^z	50	0.9510	0.9481	0.9524	0.9514		e^z	50	0.9001	0.8988	0.9045	0.8986
	e^z	100	0.9477	0.9492	0.9476	0.9436		e^z	100	0.9018	0.9050	0.9010	0.9032
	$(z-1/2)^2$	30	0.9504	0.9484	0.9473	0.9505		$(z-1/2)^2$	30	0.9023	0.9000	0.8991	0.8958
	$(z-1/2)^2$	50	0.9525	0.9510	0.9478	0.9481		$(z-1/2)^2$	50	0.8988	0.8968	0.8996	0.9002
	$(z-1/2)^2$	100	0.9500	0.9520	0.9500	0.9516		$(z-1/2)^2$	100	0.8967	0.8985	0.9026	0.9026
	$\sin(2\pi z)$	30	0.8628	0.9532	0.9511	0.9496		$\sin(2\pi z)$	30	0.8348	0.9003	0.9012	0.9008
	$\sin(2\pi z)$	50	0.8572	0.9514	0.9493	0.9450		$\sin(2\pi z)$	50	0.7087	0.9018	0.9024	0.9044
	$\sin(2\pi z)$	100	0.7861	0.9525	0.9501	0.9508		$\sin(2\pi z)$	100	0.6614	0.9019	0.8991	0.8942
	$\sin(4\pi z)$	30	0.9505	0.9528	0.9488	0.9496		$\sin(4\pi z)$	30	0.6316	0.8979	0.9037	0.9026
0.1	$\sin(4\pi z)$	50	0.4003	0.9494	0.9456	0.9504	0.1	$\sin(4\pi z)$	50	0.4269	0.9015	0.8976	0.9013
	$\sin(4\pi z)$	100	0.1405	0.9473	0.9511	0.9546		$\sin(4\pi z)$	100	0.2669	0.8963	0.9023	0.9023

3.1.3 用 $g(z)$ 的局部多项式近似

取定 z_0 , 把 $g(z)$ 在 z_0 处 m 次泰勒展开, 即

$$g(z) = g(z_0) + g'(z_0)(z-z_0) + \frac{g''(z_0)}{2!}(z-z_0)^2 + \dots + \frac{g^{(m)}(z_0)}{m!}(z-z_0)^m + \dots$$
$$\beta_m(z-z_0)^m/m! = \beta_1 g_1(z) + \beta_2 g_2(z) + \dots + \beta_m g_m(z)$$

(z),

其中 $g_1(z) = 1, g_2(z) = (z-z_0), \dots, g_m(z) = (z-z_0)^m/m!$ 。因此, $g(z)$ 可用式(6)近似。分别考虑泰勒展开次数 $m=3, 5$ 的情况。类似于上述模拟步骤, 计算得出此种近似下 $D_{1-\alpha}(Y)$ 的真实覆盖率(表 3)。

表 3 局部多项式近似下不同 α 时 $D_{1-\alpha}(Y)$ 的真实覆盖率

α	$g(z)$	n	$m=3$				$m=5$			
			$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$	$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$
0.05	e^z	30	0.9513	0.9501	0.9457	0.9528	0.9491	0.9479	0.9500	0.9515
	e^z	50	0.9530	0.9535	0.9495	0.9534	0.9503	0.9509	0.9514	0.9547
	e^z	100	0.9489	0.9483	0.9500	0.9485	0.9487	0.9472	0.9526	0.9496
	$(z-1/2)^2$	30	0.9454	0.9524	0.9504	0.9505	0.9516	0.9490	0.9511	0.9483
	$(z-1/2)^2$	50	0.9505	0.9514	0.9510	0.9508	0.9526	0.9532	0.9504	0.9524
	$(z-1/2)^2$	100	0.9467	0.9528	0.9526	0.9479	0.9487	0.9535	0.9507	0.9506
	$\sin(2\pi z)$	30	0.9449	0.9477	0.9499	0.9503	0.9525	0.9498	0.9502	0.9451
	$\sin(2\pi z)$	50	0.9507	0.9529	0.9477	0.9506	0.9511	0.9503	0.9490	0.9453
	$\sin(2\pi z)$	100	0.9506	0.9502	0.9547	0.9473	0.9502	0.9526	0.9510	0.9498
	$\sin(4\pi z)$	30	0.7715	0.9513	0.9482	0.9498	0.9515	0.9489	0.9505	0.9483
0.1	$\sin(4\pi z)$	50	0.6667	0.9532	0.9481	0.9498	0.9415	0.9509	0.9518	0.9527
	$\sin(4\pi z)$	100	0.4100	0.9516	0.9461	0.9478	0.9355	0.9489	0.9465	0.9485
	e^z	30	0.9033	0.8991	0.8997	0.8999	0.9004	0.8999	0.8947	0.8970
	e^z	50	0.9006	0.8983	0.8997	0.8982	0.9001	0.9019	0.9011	0.9049
	e^z	100	0.9023	0.8968	0.9017	0.9041	0.9037	0.9027	0.8981	0.8975
	$(z-1/2)^2$	30	0.9045	0.8964	0.9026	0.8989	0.8940	0.8954	0.8990	0.9030
	$(z-1/2)^2$	50	0.8966	0.8986	0.9032	0.8978	0.9006	0.8969	0.9005	0.8973
	$(z-1/2)^2$	100	0.8969	0.8953	0.8923	0.9028	0.9007	0.8977	0.9032	0.8909
	$\sin(2\pi z)$	30	0.8979	0.8979	0.8994	0.8986	0.9026	0.8998	0.9045	0.9014
	$\sin(2\pi z)$	50	0.9000	0.8954	0.8969	0.9023	0.8997	0.9445	0.8982	0.9032

3.2 $q=2$ 时的模拟结果

对于自变量为 2 维的回归函数, 考虑用局部多项式来进行近似。回归函数 $g(z^{(1)}, z^{(2)})$ 分别取

$$g_a(z^{(1)}, z^{(2)}) = e^{z^{(1)} + z^{(1)}z^{(2)} + (z^{(2)})^2},$$
$$g_b(z^{(1)}, z^{(2)}) = \sin(2\pi(z^{(1)} + z^{(2)})),$$

且 $z^{(1)}, z^{(2)}$ 独立同服从 $U[0, 1]$ 分布。对给定 $(z_0^{(1)}, z_0^{(2)})$, 把 $g(z^{(1)}, z^{(2)})$ 在 $(z_0^{(1)}, z_0^{(2)})$ 处 m 次泰勒展开, 即

即

$$g(z^{(1)}, z^{(2)}) = g(z_0^{(1)}, z_0^{(2)}) + \left[(z^{(1)} - z_0^{(1)}) \frac{\partial}{\partial z^{(1)}} + (z^{(2)} - z_0^{(2)}) \frac{\partial}{\partial z^{(2)}} \right] g(z_0^{(1)}, z_0^{(2)}) + \frac{1}{2!} \left[(z^{(1)} - z_0^{(1)}) \times \right.$$

$$\frac{\partial}{\partial z^{(1)}} + (z^{(2)} - z_0^{(2)}) \frac{\partial}{\partial z^{(2)}} \Big]^2 g(z_0^{(1)}, z_0^{(2)}) + \dots +$$

$$\frac{1}{m!} \left[(z^{(1)} - z_0^{(1)}) \frac{\partial}{\partial z^{(1)}} + (z^{(2)} - z_0^{(2)}) \frac{\partial}{\partial z^{(2)}} \right]^m \times$$

$$g(z_0^{(1)}, z_0^{(2)}) = \beta_1 g_1(z^{(1)}, z^{(2)}) + \beta_2 g_2(z^{(1)}, z^{(2)}) + \dots +$$

$$\beta_{m(\frac{n+1}{2})} g_{m(\frac{n+1}{2})}(z^{(1)}, z^{(2)}).$$

其中 $g_1(z^{(1)}, z^{(2)}) = 1, g_2(z^{(1)}, z^{(2)}) = z^{(1)} - z_0^{(1)},$
 $g_3(z^{(1)}, z^{(2)}) = z^{(2)} - z_0^{(2)}, \dots$

因此, $g(z^{(1)}, z^{(2)})$ 可用式(6)近似。分别考虑

表4 $z_0^{(1)} = 1/2, z_0^{(2)} = 1/3$, 不同 α 时 $D_{1-\alpha}(Y)$ 的真实覆盖率

α	$g(z)$	n	$m=3$				$m=4$			
			$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$	$\sigma=1$	$\sigma=10$	$\sigma=100$	$\sigma=1000$
0.05	e^z	30	0.9481	0.9485	0.9498	0.9493	0.9481	0.9491	0.9524	0.9501
	e^z	50	0.9487	0.9499	0.9517	0.9465	0.9504	0.9507	0.9512	0.9496
	e^z	100	0.9416	0.9478	0.9511	0.9516	0.9505	0.9526	0.9473	0.9483
	$(z-1/2)^2$	30	0.9383	0.9498	0.9481	0.9531	0.9497	0.9497	0.9469	0.9481
	$(z-1/2)^2$	50	0.8844	0.9523	0.9520	0.9488	0.9305	0.9508	0.9514	0.9543
	$(z-1/2)^2$	100	0.7645	0.9502	0.9496	0.9473	0.9588	0.9506	0.9514	0.9548
0.1	e^z	30	0.9035	0.8974	0.8991	0.8999	0.9040	0.9032	0.9017	0.8993
	e^z	50	0.9029	0.8963	0.8959	0.9044	0.9003	0.8977	0.8986	0.8993
	e^z	100	0.8963	0.8984	0.9009	0.8982	0.9032	0.8958	0.8966	0.8989
	$(z-1/2)^2$	30	0.8997	0.9036	0.9013	0.8966	0.8874	0.9000	0.8988	0.9021
	$(z-1/2)^2$	50	0.7801	0.9046	0.8963	0.8955	0.8745	0.8993	0.9054	0.8989
	$(z-1/2)^2$	100	0.5861	0.9005	0.8990	0.9035	0.6609	0.8925	0.8960	0.8997

4 定理的证明

4.1 引理1的证明

对任意 $\delta > 0$,

$$G(x_0) = P(\eta \leq x_0) = P\left(\xi \leq x_0 \frac{\xi}{\eta}\right) =$$

$$P\left(\xi \leq x_0 \frac{\xi}{\eta}, \frac{\xi}{\eta} < 1 + \delta\right) + P\left(\xi \leq x_0 \frac{\xi}{\eta}, \frac{\xi}{\eta} \geq 1 + \delta\right) \leq$$

$$P(\xi < x_0(1 + \delta)) + P\left(\frac{\xi}{\eta} \geq 1 + \delta\right) = P(\xi < x_0(1 +$$

$$\delta)) - P(\xi < x_0) + P(\xi < x_0) + P\left(\frac{\eta}{\xi} \leq \frac{1}{1 + \delta}\right) =$$

$$P(x_0 < x_0(1 + \delta)) + P(\xi < x_0) + P\left(\frac{\eta}{\xi} - 1 \leq -\frac{\delta}{1 + \delta}\right) =$$

$$P(x_0 < x_0(1 + \delta)) + P(\xi < x_0) + P\left(\left|\frac{\eta}{\xi} - 1\right| \geq \frac{\delta}{1 + \delta}\right) =$$

$$M\delta x_0 + P(\xi < x_0) + \frac{1 + \delta}{\delta} E\left|\frac{\eta}{\xi} - 1\right|.$$

另一方面,

$$P(\eta \leq x_0) \geq P(\eta \leq x_0, \xi < x_0) =$$

$$P\left(\xi \leq \frac{\xi}{\eta} x_0, \xi < x_0\right) = P(\xi < x_0) - P\left(\xi > \frac{\xi}{\eta} x_0, \xi < x_0\right).$$

又

$$P\left(\xi > \frac{\xi}{\eta} x_0, \xi < x_0\right) = P\left(\xi > \frac{\xi}{\eta} x_0, \xi < x_0, \frac{\xi}{\eta} \geq 1 - \delta\right) +$$

泰勒展开次数 $m=3, 4$ 的情况。用已知函数的分段线性组合来近似 $g(z)$ 计算在此种近似下 $D_{1-\alpha}(Y)$ 的真实覆盖率(表4)。

从模拟结果可看出,由式(9)所给出的 σ^2 的区间估计 $D_{1-\alpha}(Y)$ 对于小样本情况也是适用的。此外,在 σ 一定大的情况下, $D_{1-\alpha}(Y)$ 的真实覆盖率接近于名义覆盖率。当采用分段线性及局部多项式来对回归函数进行近似时,随着节点个数和泰勒展开次数的增加,真实覆盖率更接近于名义覆盖率。

$$P\left(\xi > \frac{\xi}{\eta} x_0, \xi < x_0, \frac{\xi}{\eta} < 1 - \delta\right) \leq P(\xi \geq x_0(1 - \delta), \xi <$$

$$x_0) + P\left(\frac{\xi}{\eta} < 1 - \delta\right) \leq Mx_0\delta + P\left(\frac{\eta}{\xi} > \frac{1}{1 - \delta}\right) \leq Mx_0\delta +$$

$$P\left(\left|\frac{\eta}{\xi} - 1\right| > \frac{\delta}{1 - \delta}\right) \leq Mx_0\delta + \frac{1 - \delta}{\delta} E\left|\frac{\eta}{\xi} - 1\right| \leq Mx_0\delta +$$

$$\frac{1 + \delta}{\delta} E\left|\frac{\eta}{\xi} - 1\right|.$$

综合两方面结果知

$$|G(x_0) - F(x_0)| \leq Mx_0\delta + \frac{1 + \delta}{\delta} E\left|\frac{\eta}{\xi} - 1\right|.$$

上式右边作为 δ 的函数,其最小值点为

$$\delta = \frac{\sqrt{E\left|\frac{\eta}{\xi} - 1\right|}}{\sqrt{Mx_0}},$$

再代入上式,可知引理1结论成立。

4.2 定理1的证明

记

$$\Delta = (g(z_1) - \sum_{j=1}^m g_j(z_1)\beta_j, \dots, g(z_n) - \sum_{j=1}^m g_j(z_n)\beta_j)^T,$$

$$\xi = V^T(I - P_X)V \sim \chi^2(n - m).$$

由式(4)知,

$$S^2 = Y^T(I - P_X)Y = (G + \sigma V)^T(I - P_X)(G + \sigma V) =$$

$$(\Delta + \sigma V)^T(I - P_X)(\Delta + \sigma V),$$

代入式(9),因此

$$P\{\sigma^2 \in D_{1-\alpha}(Y)\} =$$

$$P\left(\chi_{\frac{n-m}{2}}^2 \leq \left(\frac{\Delta}{\sigma} + V\right)^T (I - P_X) \left(\frac{\Delta}{\sigma} + V\right) \leq \chi_{1-\frac{\alpha}{2}}^2(n-m)\right).$$

令 $\eta = \left(\frac{\Delta}{\sigma} + V\right)^T (I - P_X) \left(\frac{\Delta}{\sigma} + V\right)$, 由引理1知,

$$\begin{aligned} & |P(\sigma^2 \in D_{1-\alpha}(Y)) - (1-\alpha)| \leq \\ & \left| P\left(\eta \leq \chi_{\frac{n-m}{2}}^2(n-m)\right) - P\left(\xi \leq \chi_{1-\frac{\alpha}{2}}^2(n-m)\right) \right| + \\ & \left| P\left(\eta \leq \chi_{\frac{n-m}{2}}^2(n-m)\right) - P\left(\xi \leq \chi_{\frac{n-m}{2}}^2(n-m)\right) \right| \leq \\ & 2\sqrt{A\chi_{1-\frac{\alpha}{2}}^2(n-m)E\left|\frac{\eta}{\xi} - 1\right|} + \\ & 2\sqrt{A\chi_{\frac{n-m}{2}}^2(n-m)E\left|\frac{\eta}{\xi} - 1\right|} + 2E\left|\frac{\eta}{\xi} - 1\right|, \end{aligned} \quad (12)$$

其中 A 是 $\chi^2(n-m)$ 分布密度函数的上界, 由式(10)给出.

因此只需证 $E\left|\frac{\eta}{\xi} - 1\right|$ 有界.

$$\begin{aligned} & \left|\frac{\eta}{\xi} - 1\right| = \\ & \left|\frac{\frac{\Delta^T}{\sigma}(I - P_X)\frac{\Delta}{\sigma} + 2\frac{\Delta^T}{\sigma}(I - P_X)V + V^T(I - P_X)V}{V^T(I - P_X)V} - 1\right| = \\ & \left|\frac{\frac{\Delta^T}{\sigma}(I - P_X)\frac{\Delta}{\sigma} + 2\frac{\Delta^T}{\sigma}(I - P_X)V}{V^T(I - P_X)V} - 1\right| \leq \\ & \frac{\frac{\Delta^T}{\sigma}(I - P_X)\frac{\Delta}{\sigma}}{V^T(I - P_X)V} + 2\sqrt{\frac{\frac{\Delta^T}{\sigma}(I - P_X)\frac{\Delta}{\sigma}}{V^T(I - P_X)V}}, \end{aligned} \quad (13)$$

又由于 $V^T(I - P_X)V \sim \chi^2(n-m)$, 因此

$$\begin{aligned} E\frac{1}{V^T(I - P_X)V} &= \frac{1}{n-m-2}, \\ E\frac{1}{\sqrt{V^T(I - P_X)V}} &= \frac{\Gamma\left(\frac{n-m-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{n-m}{2}\right)}, \end{aligned}$$

将其代入式(13), 有

$$\begin{aligned} E\left|\frac{\eta}{\xi} - 1\right| &\leq \frac{1}{\sigma^2} \frac{n}{n-m-2} d^2 + \frac{\sqrt{2}}{\sigma} \frac{\sqrt{n}\Gamma\left(\frac{n-m-1}{2}\right)}{\Gamma\left(\frac{n-m}{2}\right)} d \\ &= \frac{d}{\sigma} B_{\sigma,d}, \end{aligned}$$

将其代入式(12), 可知定理1结论成立.

参考文献:

- [1] SPECKMAN P. Kernel smoothing in partial linear models [J]. Journal of the Royal Statistical Society, Ser B, 1988, 50:413-436.
- [2] CHEN H. Convergence rates for parametric components in partly linear model [J]. The Annals of Statistics, 1988, 16:136-146.
- [3] HARDLE H, LIANG H, GAO J. Partially linear models [M]. New York: Physica Verlag, 2000.
- [4] GHOSH D. Efficiency considerations in the additive hazards model with current status data [J]. Statistica Neerlandica, 2001, 55: 367-376.
- [5] HUANG J, ROSSINI A J. Sieve estimation for the proportional-odds failure time regression model with interval censoring [J]. Journal of the American Statistical Association, 1997, 92:960-967.
- [6] SHEN X. On methods of sieves and penalization [J]. The Annals of Statistics, 1997, 25:2555-2591.
- [7] SHEN X, WONG W H. Convergence rate of sieve estimates [J]. The Annals of Statistics, 1994, 22:413-436.
- [8] FAN J Q, GJEBELS I. Variable bandwidth and local linear regression smoothers [J]. The Annals of Statistics, 1992, 20:2008-2036.
- [9] FAN J Q. Local polynomial modelling and its applications [M]. London: Chapman and Hall, 1996.
- [10] GASSER T, SROKA L, JENNEN S C. Residual variance and residual pattern in nonlinear [J]. Biometrika, 1986, 73:625-633.
- [11] HALL P, KAY J W, TITTERINGTON D M. Asymptotically optimal difference-based estimation of variance in nonparametric regression [J]. Biometrika, 1990, 77(3):521-528.
- [12] HALL P, MARRON S. On variance estimation in nonparametric regression [J]. Biometrika, 1990, 77(2):415-419.
- [13] DAWID A P, STONE M. The functional model basis of fiducial inference [J]. The Annals of Statistics, 1982, 10:1054-1067.
- [14] XU X, LI G. Fiducial inference in the pivotal family of distributions [J]. Science in China: Ser A, Mathematics, 2006, 49:410-432.

(编辑 修荣荣)