文章编号:1000-5862(2013)05-0471-04

相依误差下非参数回归模型的方差变点 Ratio 检验

孙耀东,徐　宝,赵志文
(吉林师范大学数学学院,吉林 四平　136000)

摘要:运用 Ratio 检验方法研究相依误差下非参数回归模型的方差变点检验. 首先利用核估计方法估计模型中的回归函数得到残差序列,其次利用残差序列构造 Ratio 检验统计量,并推导检验统计量的极限分布,最后通过数值模拟验证检验方法的有效性.

关键词:Ratio 检验;非参数回归模型;方差变点
中图分类号:O 212.1　　**文献标志码**:A

0　引言

考虑时间序列模型 $AR(p)$:

$$x_i = \sum_{k=1}^{p} \varphi_k x_{i-k} + e_i, i = 0, \pm 1, \pm 2, \cdots, \quad (1)$$

其中 e_i 是随机误差. 对 $AR(p)$ 模型性质的讨论依赖于观测数据 $X = (x_1, \cdots, x_n)'$,然而在很多实际问题中 X 常常是观测不到的,观测到数据往往包含趋势项,对此常用的方法是将趋势项假定为某一具体的参数模型,在分析数据时剔除趋势项[1],这种方法的缺点是对趋势项的参数假定常常是主观的,为此很多学者采用非参数的方法[2],考虑如下非参数回归模型:

$$y_i = f(u_i) + x_i, i = 1, 2, \cdots, n, \quad (2)$$

其中 y_i 是实际观测的数据,$f(\cdot)$ 是回归函数,表示趋势项,$u_i = i/n$ 是固定设计点,x_i 同(1)式. 已有许多学者对于模型(2)进行了研究[3-6].

本文考虑模型(2)的方差变点检测问题,具体如下:

$H_0: E(e_i^2) = \sigma^2$,对所有的 $i = 1, 2, \cdots, n$,

$H_1: H_0$ 不成立.

常用的变点检验方法是 CUSUM 检验,如 S. Lee 等[7]研究了固定设计下非参数回归模型方差的变点检测问题,赵文芝等[8]研究了随机设计下非参数回归模型方差的变点检验问题. 由于 CUSUM 检验在讨论检验统计量渐近性质时需要对模型尺度参数进行估计,而当数据相依时,该估计很难获得,针对这个问题,H. Lajos 等[9]在 CUSUM 检验的基础上提出 Ratio 检验,并讨论了时间序列数据均值变点检验问题,赵文芝等[10]研究了随机设计下非参数回归模型方差的变点 Ratio 检验问题,郭小芳等[11]研究了基于 PCA 的时间序列异常检测方法.

本文意在运用 Ratio 检验方法研究非参数回归模型(2)的方差变点检测问题,利用核估计方法估计回归函数,用残差序列构造 Ratio 检验统计量,并在一定条件下推导 Ratio 检验统计量的极限分布,最后通过数值模拟验证方法的有效性.

1　主要结果

应用 M. B. Priestley 等[12]提出的核估计方法估计模型(2)中的趋势函数 $f(\cdot)$,估计量为

$$\hat{f}(x) = f_n(x) = \frac{1}{n}\sum_{t=1}^{n} y_t K_h(x - u_t), 0 \leqslant x \leqslant 1,$$

其中 $K_h(x) = K(x/h)/h$,h 是窗宽,$K(\cdot) \geqslant 0$ 是核函数. 残差序列为 $\hat{x}_i = y_i - f_n(u_i)$. 下面分别记 $x_i^2, \hat{x}_i^2$ 的部分和为 $S_k, \hat{S}_k$,即 $S_k = \sum_{i=1}^{k} x_i^2, \hat{S}_k = \sum_{i=1}^{k} \hat{x}_i^2, k = 1, 2, \cdots, n$. 类似于文献[10],定义 Ratio 检验统计量为

$$T_n = \max_{n\delta \leqslant k \leqslant n - n\delta} \frac{\max\limits_{1 \leqslant i \leqslant k} \left| \sum\limits_{j=1}^{i} \hat{x}_j^2 - \frac{i}{k} \sum\limits_{j=1}^{k} \hat{x}_i^2 \right|}{\max\limits_{k < i \leqslant n} \left| \sum\limits_{j=i+1}^{n} \hat{x}_i^2 - \frac{n-i}{n-k} \sum\limits_{j=k+1}^{n} \hat{x}_i^2 \right|},$$

其中 $0 < \delta < 1/2$,并记 $W(t)$ $(0 \leqslant t < \infty)$ 是维纳过程,定义如下随机过程:

$$\eta_1(t) = \max_{0 < s \leqslant t} |W(s) - (s/t) W(t)|,$$

收稿日期:2013-06-10
基金项目:吉林省教育厅"十一五"科学技术研究(2010350)资助项目.
作者简介:孙耀东(1982-),男,吉林德惠人,讲师,主要从事数理统计方面的研究.

$$\eta_2(t) = \max_{t<s\leqslant 1} |W^*(s) - (1-s)/(1-t)\cdot W^*(t)|, W^*(t) = 1 - W(t).$$

为得到 Ratio 检验统计量 T_n 的极限性质,需要以下假设:

(C1) $\{e_i\}$ 独立同分布,$E(e_i) = 0$,$\exists r > 0$ 使得 $E(|e_i^2 - \sigma^2|^r) < \infty$.

(C2) $\{x_i\}$ 是 α 强混合序列,$\exists C > 0, \rho > 0$ 使得混合系数 $\alpha_k \leqslant Ce^{-\rho k}$,存在序列 $\{\psi_j\}$ 满足

$$\sum_{j=0}^{\infty} |\psi_j| < \infty, \sum_{j=0}^{\infty} \psi_j \neq 0,$$

使得 $x_i = \sum_{j=0}^{\infty} \psi_j e_{i-j}$.

(C3) 趋势函数 $f(\cdot)$ 是 Lipschitz 连续的,即 $\exists K_1 > 0$,对于 $0 \leqslant x, y \leqslant 1$ 使得

$$|f(x) - f(y)| < K_1 |x - y|.$$

(C4) 核函数 $K(\cdot)$ 在区间 $[-1,1]$ 内取值,是 Lipschitz 连续的,即 $\exists K_2 > 0$,对于 $-1 \leqslant x, y \leqslant 1$,有 $|K(x) - K(y)| < K_2 |x - y|$,$\int_{-1}^{1} K(x)\mathrm{d}x = 1$.

(C5) 当 $n \to \infty$ 时,窗宽 $h = h_n$ 满足 $nh^2 \to \infty$,$nh^4 \to 0$.

注 1 (ⅰ) 很多过程满足条件(C2);当 e_i 的分布函数连续时可逆 $AR(p)$ 模型满足(C2).

(ⅱ)(C3)~(C5) 是非参数分析中的一般性假设.

引理 1 [13] 设条件(C1)和(C2)成立,当 H_0 为真时,$\frac{1}{n}\sum_{1\leqslant i\leqslant [nt]} x_i \xrightarrow{D[0,1]} \sigma\left|\sum_{j=0}^{\infty}\psi_j\right| W(t)$($\xrightarrow{D[0,1]}$ 表示在 $D[0,1]$ 上弱收敛).

引理 2 设条件(C1)~(C5)成立,当 H_0 为真时,

$$\max_{1\leqslant k<n} \frac{1}{\sqrt{n}} |\hat{S}_k - S_k| \xrightarrow{P} 0.$$

证 $\frac{1}{\sqrt{n}}(\hat{S}_k - S_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^{k}(\hat{x}_i^2 - x_i^2) = \frac{1}{\sqrt{n}}\cdot \sum_{i=1}^{k}[\hat{f}(u_i) - f(u_i)]^2 + \frac{2}{\sqrt{n}}\sum_{i=1}^{k}[f(u_i) - \hat{f}(u_i)]x_i = J_1 + J_2$.

首先考虑 J_1.

$$J_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{k}(\hat{f}(u_i) - f(u_i))^2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{k}(f_n(u_i) - f(u_i))^2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{t=1}^{n} x_t K_h(u_i - u_t) + \frac{1}{n}\sum_{t=1}^{n} f(u_t)\cdot K_h(u_i - u_t) - f(u_i)\right)^2 \leqslant 3\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{k}\frac{1}{n}\sum_{t=1}^{n} x_t K_h(u_i - u_t)\right]^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^{k}\left[\frac{1}{n}\sum_{t=1}^{n}(f(u_t) - f(u_i))K_h(u_i - u_t)\right]^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^{k}\left[f(u_i)\left(\frac{1}{n}\sum_{t=1}^{n} K_h(u_i - u_t) - 1\right)^2\right]. \quad (3)$$

由(C1),C(2) 和 C(4) 知,

$$E\left[\frac{1}{n}\sum_{t=1}^{n} x_t K_h(u_i - u_t)\right]^2 = O(1/(nh)),$$

从而(3) 式的第 1 项为 $O_p(1/(\sqrt{n}h))$,又由 C(3) 和 C(4) 知,

$$\frac{1}{n}\sum_{t=1}^{n}[f(u_t) - f(u_i)]K_h(u_i - u_t) = O(h),$$

$$f(u_i)\left[\frac{1}{n}\sum_{t=1}^{n} K_h(u_i - u_t) - 1\right] = O(1/(nh)),$$

从而(3) 式的第 2 项、第 3 项分别为 $O_p(1/(\sqrt{n}h^2))$,$O_p(1/(\sqrt{n^3}h^2))$,进而由 C(5),得 $\max_{1\leqslant k<n}|J_1| \xrightarrow{P} 0$.

接下来考虑 J_2,

$$J_2 = \frac{2}{\sqrt{n}}\sum_{i=1}^{k}(f(u_i) - f_n(u_i))x_i = \frac{2}{\sqrt{n}}\sum_{i=1}^{k}\left\{f(u_i)\cdot\left[1 - \frac{1}{n}\sum_{t=1}^{n} K_h(u_i - u_t)\right]x_i\right\} + \frac{2}{\sqrt{n}}\sum_{i=1}^{k}\left\{\left[\frac{1}{n}\sum_{t=1}^{n}(f(u_t) - f(u_t))K_h(u_i - u_t)\right]x_i\right\} - \frac{2}{\sqrt{n}}\sum_{i=1}^{k}\left\{\left[\frac{1}{n}\sum_{t=1}^{n} x_t K_h(u_i - u_t)\right]x_i\right\} = A_1 + A_2 - A_3.$$

$$|A_1| \leqslant \frac{2}{\sqrt{n}}\sum_{i=1}^{k}\left|f(u_i)\left[1 - \frac{1}{n}\sum_{t=1}^{n} K_h(u_i - u_t)\right]\right| |x_i|,$$

由 $f(u_i)\left[\frac{1}{n}\sum_{t=1}^{n} K_h(u_i - u_t) - 1\right] = O(1/(nh))$,从而 $\max_{1\leqslant k<n}|A_1| = O_p(1/(\sqrt{n}h))$. 由文献[7]的引理 2 的证明得 $\max_{1\leqslant k<n}|A_2| = O_p(1)$,$\max_{1\leqslant k<n}|A_3| = O_p(1)$. 综上所述,命题 2 得证.

定理 1 设条件(C1)~C(5)成立,当 H_0 为真时,

$$T_n \xrightarrow{D} \max_{\delta\leqslant t<1-\delta}\frac{\eta_1(t)}{\eta_2(t)}. \quad (4)$$

证 记 $Z_{n,1}(t) = \frac{1}{n}\sum_{1\leqslant i\leqslant [nt]}\left(x_i^2 - \sigma^2\sum_{j=0}^{\infty}\psi_j^2\right)$,$Z_{n,2}(t) = \frac{1}{n}\sum_{[nt]\leqslant i\leqslant n}\left(x_i^2 - \sigma^2\sum_{j=0}^{\infty}\psi_j^2\right)$,其中 $0 \leqslant t \leqslant 1$.

由引理 1 及文献[9]的定理 1.1 有

$$(Z_{n,1}(t), Z_{n,2}(t)) \xrightarrow{D^2[0,1]} E\left(x_i^2 - \sigma^2\sum_{j=0}^{\infty}\psi_j^2\right)^2 (W(t), W^*(t)). \quad (5)$$

因为

$$\frac{1}{\sqrt{n}}\max_{0<i\leqslant [nt]}\left|\sum_{j=1}^{i}\hat{x}_j^2 - \frac{i}{[nt]}\sum_{j=1}^{[nt]}\hat{x}_j^2\right| = \max_{0<i\leqslant [nt]}\left|Z_{n,1}\left(\frac{i}{n}\right) - \frac{i}{[nt]}Z_{n,1}(t) + \frac{1}{\sqrt{n}}\sum_{j=1}^{i}(\hat{x}_j^2 - x_j^2) -

$$\frac{1}{\sqrt{n}}\left|\frac{i}{[nt]}\sum_{j=1}^{[nt]}(\hat{x}_j^2-x_j^2)\right|,$$

$$\frac{1}{\sqrt{n}}\max_{[nt]<i\leqslant n}\left|\sum_{j=i+1}^{n}\hat{x}_j^2-\frac{n-i}{n-[nt]}\sum_{j=[nt]+1}^{n}\hat{x}_j^2\right|=\max_{[nt]<i\leqslant n}\left|Z_{n,2}\left(\frac{i}{n}\right)-\frac{n-i}{n-[nt]}Z_{n,2}(t)+\frac{1}{\sqrt{n}}\sum_{j=i+1}^{n}(\hat{x}_j^2-x_j^2)-\frac{1}{\sqrt{n}}\frac{n-i}{n-[nt]}\sum_{j=[nt]+1}^{nt}(\hat{x}_j^2-x_j^2)\right|,$$

从而对 $0<\delta<1/2$,由(5)式及引理 2,有

$$\frac{1}{\sqrt{n}}\left(\max_{1\leqslant i\leqslant[nt]}\left|\sum_{j=1}^{i}\hat{x}_j^2-\frac{i}{[nt]}\sum_{j=1}^{[nt]}\hat{x}_j^2\right|,\max_{[nt]<i\leqslant n}\left|\sum_{j=i+1}^{n}\hat{x}_j^2-\frac{n-i}{n-[nt]}\sum_{j=[nt]+1}^{n}\hat{x}_j^2\right|\right)\xrightarrow{D[\delta,1-\delta]}$$

$$E\left(x_i^2-\sigma^2\sum_{j=0}^{\infty}\psi_j^2\right)^2\left(\max_{0<s\leqslant t}|W(s)-(s/t)W(t)|,\max_{t<s\leqslant 1}|W^*(s)-(1-s)/(1-t)W^*(t)|\right),$$

故由连续映射定理得(4)式,定理 1 得证.

2　数值模拟

下面通过数值算例验证本文方法的有效性. 由于 $\max_{\delta\leqslant t<1-\delta}\eta_1(t)/\eta_2(t)$ 分布的具体形式很难得到,采用文献[2]的由样本分位数求得的临界值,对于 $\delta=0.2$,在显著性水平分别取值为 0.01,0.05 和 0.10 下,临界值分别为 6.540 6,4.803 1 和 4.273 3.

考虑模型

$$y_i=f(u_i)+x_i,$$

$$x_i=\varphi x_{i-1}+e_i,u_i=i/n,i=1,2,\cdots,n,$$

这里 $f(x)=25x^3-45x^2+24x-3.6$,e_i 是均值为 0,方差为 σ^2 的独立同分布正态随机变量. 为估计趋势函数,选用 Epanechnikov 核函数:

$$K(x)=\frac{3}{4}(1-x^2)I_{[-1,1]}(x),$$

其中 $I(\cdot)$ 是示性函数,窗宽 $h=h_n=0.4n^{-0.3}$. 假定方差变点发生的时刻为 $t_0=[n\tau_0]$,考虑如下检验问题:

$$H_0:\sigma^2=1,\text{对所有的 } i=1,2,\cdots,n,$$

$$H_1:\sigma^2 \text{ 在 } t_0=[n\tau_0] \text{ 之前为 } 1,\text{之后为 } \alpha^2,$$

其中 α^2 取值为 2,4,9,τ_0 取值为 0.25,0.50,0.60 且 φ 取值为 0,0.3,0.5,0.8,样本容量 $n=200,300,500$. 实验重复 500 次,在显著性水平为 0.05 下计算拒绝 H_0 的频率. 模拟结果如表 1 ~ 表 3 所示.

表 1　当 $\tau_0=0.25$ 时拒绝 H_0 的频率

φ		0			0.3			0.5			0.8		
α^2		2	4	9	2	4	9	2	4	9	2	4	9
n	200	0.530	0.680	0.745	0.485	0.660	0.745	0.555	0.600	0.615	0.480	0.530	0.550
	300	0.610	0.880	0.885	0.565	0.745	0.845	0.550	0.705	0.700	0.500	0.575	0.585
	500	0.870	0.950	0.960	0.800	0.920	0.950	0.690	0.880	0.850	0.525	0.590	0.645

表 2　当 $\tau_0=0.5$ 时拒绝 H_0 的频率

φ		0			0.3			0.5			0.8		
α^2		2	4	9	2	4	9	2	4	9	2	4	9
n	200	0.530	0.770	0.620	0.550	0.725	0.735	0.560	0.655	0.685	0.490	0.520	0.540
	300	0.750	0.855	0.875	0.565	0.820	0.845	0.645	0.690	0.755	0.515	0.575	0.590
	500	0.865	0.980	0.975	0.815	0.935	0.945	0.775	0.880	0.885	0.545	0.620	0.700

表 3　当 $\tau_0=0.6$ 时拒绝 H_0 的频率

φ		0			0.3			0.5			0.8		
α^2		2	4	9	2	4	9	2	4	9	2	4	9
n	200	0.360	0.550	0.625	0.355	0.555	0.580	0.390	0.515	0.525	0.350	0.395	0.492
	300	0.585	0.645	0.710	0.565	0.630	0.685	0.420	0.610	0.695	0.372	0.560	0.562
	500	0.755	0.865	0.890	0.695	0.875	0.920	0.620	0.755	0.850	0.392	0.610	0.695

对比表1,表2和表3所得到的模拟结果可以看出,与本文的Ratio检验是一致的,样本容量 n 越大,检验效果越好,并且当 n 足够大时,检验的势函数值一定为1. 变点之后的方差 α^2 越大,检验效果越好. 检验效果与系数 φ 有关,随着 φ 越接近1,模型趋近非平稳,检验效果变差. 当变点位置位于观测值中间位置时检验效果最好,其次是当变点位于观测值前段时,而当变点位于观测值后段时检验效果相对较差.

3 参考文献

[1] Fuller W A. Introduction to statistical time series [M]. New York:John Wiley Sons,1996.

[2] 吴小腊,刘万荣,李泽华. 变系数模型变窗宽局部M-估计的渐近正态性 [J]. 重庆师范大学学报:自然科学版,2008,25(1):50-53.

[3] Truong Y K. Nonparametric curve estimation with time series errors [J]. Journal of Statistical Planning and Inference,1991,28(2):167-183.

[4] Altman N S. Estimating error correlation in nonparametric regression [J]. Statistics Probability Letters,1993,18(3):213-218.

[5] Shao Qin,Yang Lijian. Autoregressive coefficient estimation in nonparametric analysis [J]. Journal of Time Series Analysis,2011,32(6):587-597.

[6] 甘登文. 半相依回归模型中的回归系数新估计的效率 [J]. 江西师范大学学报:自然科学版,1993,17(2):127-130.

[7] Lee S,Na O,Na S. On the CUSUM of square test for variance change in nonstationary and nonparametric time series models [J]. Ann Inst Statist Math,2003,55(3):467-485.

[8] 赵文芝,夏志明,刘勤社,等. 随机设计下非参数回归模型方差变点检验 [J]. 西北大学学报:自然科学版,2011,41(1):15-18.

[9] Horváth L,Horváth Z,Hušková M. Ratio tests for change point detection [J]. Institute of Mathematical Statistics Collections,2008,1(1):293-304.

[10] 赵文芝,夏志明,贺兴时. 随机设计下非参数回归模型方差变点Ratio检验 [J]. 数学的实践与认识,2012,42(16):224-229.

[11] 郭小芳,李锋,宋晓宁. 一种基于PCA的时间序列异常检测方法 [J]. 江西师范大学学报:自然科学版,2012,36(3):280-283.

[12] Priestley M B,Chao M T. Nonparametric function fitting [J]. Journal of the Royal Statistical Society Series B,1972,34(3):385-392.

[13] Wang Qiying,Lin Yanxia,Gulati C M. The invariance principle for linear process with applications [J]. Econometric Theory,2002,18(1):119-139.

The Ratio Test for Change Point Detection in Nonparametric Regression Model with Dependent Errors

SUN Yao-dong,XU Bao,ZHAO Zhi-wen

(College of Mathematics,Jilin Normal University,Siping Jilin 136000,China)

Abstract: The detection problem of change point in nonparametric regression model with dependent errors is considered by Ratio test. The residual sequence is obtained when the regression function is estimated by kernel smoothers and the ratio test statistics is established based on the residual sequence. Asymptotic properties of the test statistics are derived. The performance of the method is illustrated by simulation studies.

Key words: Ratio test;nonparametric regression model;variance change point

(责任编辑:曾剑锋)