半参数空间变系数回归模型的 Back-Fitting估计

魏传华[1], 梅长林[2]

(1. 中国人民大学统计学院, 北京 100872)
(2. 西安交通大学理学院, 西安 710049)

摘要: 针对半参数空间变系数回归模型给出了一种估计方法 - 后向拟合估计,该方法可得到模型中常值系数估计量的精确解析表达式, 广泛的数值模拟表明所提出的估计方法对估计常值系数具有满意的精度和稳定性,最后,利用该方法分析了一个实际的例子.

关键词: 半参数空间变系数回归模型;地理加权回归方法;后向拟合法;广义交叉证实法

1 引 言

在空间数据分析中,虽然一般线性回归模型作为一种最常用的方法,也可用来确定和分析变量之间的关系,且有完备的理论体系和统计推断方法[1].然而,此模型要求回归系数在所研究的空间区域内具有一致性(即为常数),没有考虑空间数据的最典型特征—— 空间非平稳性(spatial nonstationarity)[2,3],因而其分析结果不能全面反映空间数据的真实特征,尤其是数据随空间区域的变化规律.

近年来,在这方面已作了许多有益的改进.其中受到人们普遍重视的一个模型是下面的空间变系数回归模型(spatially varying-coefficient regression model)[4,5]:

$$y_i = U_0(\__i,\nu_i) + U_1(\__i,\nu_i)x_{i1} + \cdots + U_p(\__i,\nu_i)x_{ip} + X_i, \quad i = 1,2,\cdots,n. \tag{1}$$

这里 $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$ $(i = 1,2,\cdots,n)$是因变量 Y和自变量 $X_1, X_2, \cdots, X_n$的 n组观测值, $X_i, X_2, \cdots, X_n$为独立同分布的随机误差项,均值为零,方差为 e^2. $(__i,\nu_i)$是对应于第 i组观测 $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$的地理位置的坐标(如经度和纬度), $U(\__i,\nu_i) = (U_0(\__i,\nu_i), U_1(\__i,\nu_i),\cdots,U_p(\__i,\nu_i))^T$ $(i = 1,2,\cdots,n)$是未知的回归系数向量,其中各元素是空间位置 $(__i,\nu_i)$的函数. 此模型将数据的空间位置嵌入到回归系数之中,故其既能描述因变量和自变量的关系,又能反映数据的空间变化特征,在对来自地理,经济,环境,地质等领域的数据的分析中有广泛的应用.

关于模型(1)的拟合, Brunsdon等提出了一种称之为地理加权回归(geographically weighted regression)的技术[4,5], 该方法利用局部加权最小二乘法拟合模型,其中的权取为观测点之间距离的某一非增函数.为使本文更具可读性,这里对此方法作简要介绍(详细叙述可参见[4- 9]等).

对研究区域内任一位置 $(\_,\nu)$, 指定一组权 $w_1(_,\nu), w_2(_,\nu),\cdots,w_n(_,\nu)$来表示不同点处的观测值对拟合该点处 Y值的贡献,其中第 i个权值对应于第 i组观测 $(y_i; x_{i1}, x_{i2},\cdots,x_{ip})$. 令

收稿日期: 2004-03-18
基金项目: 国家自然科学基金(10431010)资助

$$\boldsymbol{W}(\_,\nu)=\operatorname{diag}[w_1(\_,\nu),w_2(\_,\nu),\cdots,w_n(\_,\nu)],$$

和

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix},\quad \boldsymbol{Y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

若 $(\_,\nu)$ 处的自变量的观测值为 $(x_1,x_2,\cdots,x_p)$,则由加权最小二乘法可得 $(\_,\nu)$ 处的因变量 Y 的拟合值为

$$\hat{Y}(\_,\nu)=(1,x_1,x_2,\cdots,x_p)[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\_,\nu)\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\_,\nu)\boldsymbol{Y}. \tag{2}$$

特别对 n 个设计点 $(\__i,\nu_i)$, $i=1,2,\cdots,n$, 以 $\hat{\boldsymbol{Y}}_L=(\hat{y_1},\hat{y_2},\cdots,\hat{y_n})^{\mathrm{T}}$ 记因变量在各设计点处的拟合值所组成的向量,则有 $\hat{\boldsymbol{Y}}=\boldsymbol{L}\boldsymbol{Y}$. 其中

$$\boldsymbol{L}=\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__1,\nu_1)\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__1,\nu_1)\\ \boldsymbol{x}_2^{\mathrm{T}}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__2,\nu_2)\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__2,\nu_2)\\ \vdots\\ \boldsymbol{x}_n^{\mathrm{T}}[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__n,\nu_n)\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\__n,\nu_n)\end{bmatrix},\quad \boldsymbol{x}_i^{\mathrm{T}}=(1,x_{i1},\cdots,x_{ip}),\quad i=1,2,\cdots,n.$$

由上可知,我们需要在每个估计点位置 $(\_,\nu)$ 上指定一组权 $w_1(\_,\nu),w_2(\_,\nu),\cdots,w_n(\_,\nu)$,并且这组权会随 $(\_,\nu)$ 的变化而变化. 根据地理学 Tobler第一基本定理及一般非参数光滑的思想,距离位置 $(\_,\nu)$ 较近的观测值对估计 $(\_,\nu)$ 处的因变量的值所起的作用较大,而距离 $(\_,\nu)$ 较远的观测值的影响较小. 因此 $(\__i,\nu_i)$ 处的权可取为 Gauss函数

$$w_j(\__i,\nu_i)=\exp(-\theta d_{ij}^2),\quad j=1,2,\cdots,n, \tag{3}$$

其中 d_{ij} 为空间位置点 $(\__i,\nu_i)$ 与 $(\__j,\nu_j)$ 之间的距离. $\theta>0$为待定参数,它可用交叉证实法确定[10,11],即令

$$\mathrm{CV}(\theta)=\sum_{i=1}^{n}[y_i-\hat{y}_{(i)}(\theta)]^2$$

其中 $\hat{y}_{(i)}(\theta)$ 是在给定的 θ 值下,去掉第 i 组观测数据 $(y_i;\ x_{i1},\ x_{i2},\cdots,\ x_{ip})$ 后,利用前述方法所求得的 y_i 的预测值. 选择 θ_0, 使得

$$\mathrm{CV}(\theta_0)=\min\mathrm{CV}(\theta),\quad (\theta>0),$$

则以 θ_0 作为 θ 的值,实际计算中,可在 θ 的定义域内选择一系列的值,分别计算 $\mathrm{CV}(\theta)$ 的值,以最小的 $\mathrm{CV}(\theta)$ 的值所对应的 θ 值作为 θ_0.

2 半参数空间变系数回归模型及估计

对于空间变系数回归模型的应用,有一个重要的问题需要解决,就是给定了因变量 Y 与 p 个回归变量 $x_1,x_2,\cdots,x_n$ 的 n 组观测 $(Y_i;\ x_{i1},\ x_{i2},\cdots,\ x_{ip})$, $i=1,2,\cdots,n$, 及其在空间的相应“位置信息”后,我们并不能肯定哪些回归变量前的系数是常值,哪些是随观测点的变化而改变的.因此,必须基于观测数据对此给出一个合理的估计,确定哪些系数可以认为是常值.

对于这个问题,我们可以应用文[6-8]中检验空间变系数回归模型(varying-coefficientmodel)中各回归变量的系数变化是否显著来予以解决,即在一定的显著水平下,如果某些系数的变化是不显著的,则可认为这些系数可取为常数,否则按变系数对待.这样,经适

当调整回归变量的次序,可以得到如下的半参数空间变系数回归模型.

$$y_i = U_0 + \sum_{k=1}^{q} U_k x_{ik} + \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{ik} + X_i, \quad i = 1, 2, \cdots, n; \tag{4}$$

或者

$$y_i = \sum_{k=1}^{q} U_k x_{ik} + U_0(\__i, \nu_i) + \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{ik} + X_i, \quad i = 1, 2, \cdots, n, \tag{5}$$

其中各项的意义与模型(1)相同,我们称模型(4)或者(5)为半参数空间变系数回归模型.

需要指出的是, Brunsdon等在研究空间变系数回归模型时,也提及了以上模型,并且给出了该模型拟合的一种迭代算法[12]. 此迭代算法,不但计算很耗时,而且也只能得到常值系数的近似估计. 实际上,对模型(4)或(5),基于通常的最小二乘估计和前述的地理加权回归技术,采用[13]中的后向拟合法(Back-Fitting procedure),在一定条件下,可以得到常值系数估计以及因变量拟合值的明确表达式. 为明确起见,我们仅对模型(4)进行讨论.

从模型构成上看,除去误差项,模型(4)由两个可加部分构成,一部分为常系数线性部分,而另外一部分为变系数部分.且它们的各自的拟合都是线性拟合,则我们利用 back-fitting方法进行估计.

仍以 $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$ $(i = 1, 2, \cdots, n)$ 表示因变量 Y 和自变量 $X_1, X_2, \cdots, X_n$ 的 n 组观测值, $(\__i, \nu_i)$, $i = 1, 2, \cdots, n$ 表示观测点空间位置. 令

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X}_1 = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1q} \\ 1 & x_{21} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nq} \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} x_{1,q+1} & x_{1,q+2} & \cdots & x_{1p} \\ x_{2,q+1} & x_{2,q+2} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,q+1} & x_{n,q+2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{f}_2 = \begin{bmatrix} \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{1k} \\ \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{2k} \\ \vdots \\ \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{nk} \end{bmatrix},$$

继续用

$$\boldsymbol{W}(\__i, \nu_i) = \mathrm{diag}[w_1(\__i, \nu_i), w_2(\__i, \nu_i), \cdots, w_n(\__i, \nu_i)]$$

表示 $(\__i, \nu_i)$ 处的权矩阵. 以 $U_1 = (U_0, U_1, \cdots, U_q)^{\mathrm{T}}$, $U_2(\__i, \nu_i) = (U_{q+1}(\__i, \nu_i), U_{q+2}(\__i, \nu_i), \cdots, U_p(\__i, \nu_i))^{\mathrm{T}}$ 分别记常值系数向量和 $(\__i, \nu_i)$ 处的变系数向量.

由 Back-Fitting方法,若假设模型(4)中的变系数部分已知,则模型变为如下的线性回归模型

$$y_i - \sum_{k=q+1}^{p} U_k(\__i, \nu_i) x_{ik} = U_0 + \sum_{k=1}^{q} U_k x_{ik} + X_i, \quad i = 1, 2, \cdots, n,$$

得 U_1 的最小二乘估计为

$$\hat{U}_1 = (\hat{U}_0, \hat{U}_1, \cdots, \hat{U}_q)^T = [X_1^T X_1]^{-1} X_1^T (Y - f_2),$$

从而线性部分拟合值为

$$\hat{f}_1 = X_1 \hat{U}_1 = X_1 [X_1^T X_1]^{-1} X_1^T Y. \tag{6}$$

若假设常值系数已知,则模型变为第一部分讨论的变系数模型,

$$y_i - U_0 - \sum_{k=1}^{q} U_k x_{ik} = \sum_{k=q+1}^{p} U_k(u_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

由地理加权回归技术可得变系数部分拟合值为

$$\hat{f}_2 = L_2 (Y - X_1 U_1). \tag{7}$$

其中

$$L_2 = \begin{bmatrix} x_1^T [X_2^T W(u_1, v_1) X_2]^{-1} X_2^T W(u_1, v_1) \\ x_2^T [X_2^T W(u_2, v_2) X_2]^{-1} X_2^T W(u_2, v_2) \\ \vdots \\ x_n^T [X_2^T W(u_n, v_n) X_2]^{-1} X_2^T W(u_n, v_n) \end{bmatrix},$$

而 $x_i^T = (x_{i,q+1}, x_{i,q+2}, \cdots, x_{ip})$, $i = 1, 2, \cdots, n$.

根据 Back-Fitting原理(详见 [13], P118-120),得到下面的估计方程

$$\begin{cases} X_1 U_1 = X_1 [X_1^T X_1]^{-1} X_1^T (Y - f_2); \\ f_2 = L_2 (Y - X_1 U_1). \end{cases} \tag{8}$$

将第二式代入第一式,并假定矩阵 $X_1^T (I - L_2) X_1$ 可逆,则可求得

$$\begin{cases} \hat{U}_1 = [X_1^T (I - L_2) X_1]^{-1} X_1^T (I - L_2) Y; \\ \hat{f}_2 = L_2 (Y - X_1 \hat{U}_1). \end{cases} \tag{9}$$

从而可以得到 $Y = (y_1, y_2, \cdots, y_n)^T$ 的拟合值为

$$\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^T = X_1 \hat{U}_1 + \hat{f}_2 = SY, \tag{10}$$

其中

$$S = L_2 + (I - L_2) X_1 [X_1^T (I - L_2) X_1]^{-1} X_1^T (I - L_2) \tag{11}$$

对于模型(5),只需将上面的 X_1 的第一列作为 X_2 的第一列,其余均不变,上述结论仍然成立.

需要注意的是,由非参数回归的性质可知 $(I - L_2)$ 一般情况下不是幂等对称阵,所以上面得到的 Back-Fitting估计不同于文 [14]中的两步估计方法.

在模型(4)的估计中,权函数同样可选择为(3)式的形式. 而其中的光滑参数可由下述计算量较小的广义交叉证实法(见 [13]第 3章)来确定.设 $\theta > 0$为任一给定的参数值,为明确拟合值与 θ 的关系,记 $\hat{Y}(\theta) = (\hat{y}_1(\theta), \hat{y}_2(\theta), \cdots, \hat{y}_n(\theta))^T = S(\theta) Y$, 其中 $S(\theta) = (S_{ij}(\theta))$ 如(12)式所示.令

$$\mathrm{GCV}(\theta) = \sum_{i=1}^{n} \left[\frac{y_i - \hat{y}_i(\theta)}{1 - S_{ii}(\theta)} \right]^2 = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i(\theta)}{[1 - S_{ii}(\theta)]^2}, \tag{12}$$

其中 $\hat{\varepsilon}_i(\theta) = y_i - \hat{y}_i(\theta)$ $(i = 1, 2, \cdots, n)$ 为利用所有观测拟合模型(5)所得残差.由于 $S(\theta)$ 有明确的表达式,因此其主对角线上的各元素 $S_{ii}(\theta)$ 是容易求得的.广义交叉证实法即选择 θ_0 使得 $\mathrm{GCV}(\theta)$ 达到最小.

3 模拟试验

和半参数(部分线性)回归模型一样,常值系数的有效估计是此类模型所关心的主要内容,对了解各自变量对因变量的影响关于空间位置的变化的规律具有重要作用,下面我们将通过一系列的模拟试验来考察第二部分所给予的关于模型(4)和(5)中常值系数估计的精确性和该方法的稳定性情况.

3.1 试验设计

在模拟中,我们取所研究的空间区域是边长为 $m-1$ 个距离单位的正方形,观测位置在 $m\times m$ 个格子点上,各点之间的水平与垂直距离均为 1个长度单位.以 $\_,\nu$ 分别表示观测点的横坐标和纵坐标,其中规定观测点的顺序按由左至右,由下至上的顺序排列.即对于每一个 $\_=0,1,\cdots,m-1$,令 ν 分别取 $0,1,\cdots,m-1$ 而依次排列.则第 i 个观测点位置的坐标 $(\__i,\nu_i)$ 为 $\__i=\mathrm{mod}(i-1,m)$,$\nu_i=\left[\frac{i-1}{m}\right]$ $(i=1,2,\cdots,m^2)$. 其中的 mod(·) 为取余函数,[·]为取整函数.正如文[15]所指出的,这样的正规格子点区域有着广泛的应用背景,例如在地理分析中占有十分重要地位的遥感数据的空间位置便呈上述格子点形式.

下面我们就如下几种混合空间变系数回归模型进行模拟,其中误差项均服从正态分布 $N(0,\ \sigma^2)$. 为进一步考察噪声方差对估计值的影响,我们分别取 $\sigma=0.2,\ 0.6,\ 1$.

1) $y_i=U_0+U_1(\__i,\nu_i)x_i+X_i$

M1: $U_0=5$; $U_1(\__i,\nu_i)=\sin\__i$;

M2: $U_0=5$; $U_1(\__i,\nu_i)=\__i+\nu_i$.

2) $y_i=U_0(\__i,\nu_i)+U_1x_i+X_i$

M3: $U_0(\__i,\nu_i)=2\__i$; $U_1=5$;

M4: $U_0(\__i,\nu_i)=2\ln\frac{1+\__i}{4}$; $U_1=10$.

3) $y_i=U_0+U_1x_{i1}+U_2(\__i,\nu_i)x_{i2}+X_i$

M5: $U_0=5$; $U_1=15$; $U_2(\__i,\nu_i)=2\__i$.

4) $y_i=U_0+U_1(\__i,\nu_i)x_{i1}+U_2(\__i,\nu_i)x_{i2}+X_i$

M6: $U_0=6$; $U_1(\__i,\nu_i)=\lg\frac{5+\__i}{5}$; $U_2(\__i,\nu_i)=\cos\__i$.

权函数选择(3)式的形式,即在每个 $(\__i,\nu_i)$ 点取权函数为

$$w_j(\__i,\ \nu_i)=\exp\{-\theta[(\__i-\__j)^2+(\nu_i-\nu_j)^2]\},\quad i,j=1,2,\cdots,m^2,$$

其中参数 θ 由 2.1节的广义交叉证实法确定.各模型中自变量的值是独立产生的服从区间(01)上均匀分布的随即数. X_i, $i=1,2,\cdots,m^2$, 是服从 $N(0,\sigma^2)$ 的随机数.对每一种模型(σ 分别取 0.2, 0.6, 1)都取 $m=6,\ 7,\ 8,\ 9,\ 10$ 五种情况进行模拟. 对每个给定的 m,σ 和模型,我们在每次只改变误差向量 X 的情况下,重复计算 500次,从而对于每个常值系数,比如 U_k,将得到 500个估计值, $\hat{U}_{1k},\hat{U}_{2k},\cdots,\hat{U}_{500k}$, 我们以这 500个估计值的平均值作为该常值系数的最终估计值,记为 $\hat{U}_k$. 以其样本标准差 $s_n(k)=\sqrt{\frac{1}{500}\sum_{i=1}^{500}\left(\hat{U}_{ki}-\frac{1}{500}\sum_{i=1}^{500}\hat{U}_{ki}\right)^2}$ 衡量估计方法的稳定性.

3.2 模拟实验结果

模拟试验结果见表 1. 从模拟结果可以得知

表 1 500次重复下各常值系数估计的平均值和标准差

模型常系数	e	$\hat{\theta}$及 $s_n(k)$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
M1 ($\theta_0=5$)	0.2	$\hat{\theta}_0$	4.9513	5.0035	4.9670	5.0383	4.9554
		$s_n(0)$	0.0663	0.0580	0.0554	0.0511	0.0437
	0.6	$\hat{\theta}_0$	4.8516	4.9944	5.0201	4.9955	5.0427
		$s_n(0)$	0.2208	0.1715	0.1830	0.1505	0.1186
	1	$\hat{\theta}_0$	5.0347	5.0874	5.0105	5.1427	5.0672
		$s_n(0)$	0.3207	0.2703	0.2713	0.2455	0.2153
M2 ($\theta_0=5$)	0.2	$\hat{\theta}_0$	4.9832	5.0565	4.9917	5.0247	5.0076
		$s_n(0)$	0.0642	0.0564	0.0514	0.0481	0.0391
	0.6	$\hat{\theta}_0$	4.9653	4.9198	4.9966	5.0143	5.0153
		$s_n(0)$	0.1932	0.2131	0.1692	0.1324	0.1325
	1	$\hat{\theta}_0$	5.0473	4.9346	5.0857	5.0427	5.0965
		$s_n(0)$	0.3526	0.2867	0.3015	0.2485	0.1969
M3 ($\theta_1=5$)	0.2	$\hat{\theta}_1$	5.0781	4.7139	5.0191	4.9551	4.9446
		$s_n(1)$	0.1324	0.1194	0.1092	0.0948	0.0789
	0.6	$\hat{\theta}_0$	5.0781	5.4401	4.9769	4.9659	5.1303
		$s_n(0)$	0.1324	0.1297	0.1041	0.0814	0.0860
	1	$\hat{\theta}_0$	5.0781	5.4401	4.9769	4.9659	5.1303
		$s_n(0)$	0.1324	0.1297	0.1041	0.0814	0.0860
M4 ($\theta_1=10$)	0.2	$\hat{\theta}_1$	10.0960	9.9037	9.9846	10.0152	10.0061
		$s_n(0)$	0.1323	0.1194	0.1092	0.0948	0.0788
	0.6	$\hat{\theta}_0$	10.0580	10.1661	9.9151	9.8763	9.9484
		$s_n(0)$	0.3782	0.3816	0.2923	0.2323	0.2381
	1	$\hat{\theta}_0$	10.0854	10.2345	9.8884	9.8491	9.9269
		$s_n(0)$	0.6308	0.6404	0.4807	0.3897	0.3928
M5 ($\theta_0=5$)	0.2	$\hat{\theta}_0$	4.9114	4.8702	4.8833	4.9632	5.0687
		$s_n(0)$	0.1200	0.1066	0.0824	0.0790	0.0592
	0.6	$\hat{\theta}_0$	4.9008	4.9095	5.0476	4.9397	5.0819
		$s_n(0)$	0.3202	0.2849	0.2667	0.1965	0.1767
	1	$\hat{\theta}_0$	4.8975	4.8685	4.9728	4.9023	5.1281
		$s_n(0)$	0.5186	0.4641	0.4346	0.3210	0.2881
M5 ($\theta_1=15$)	0.2	$\hat{\theta}_1$	15.2491	15.2077	15.1899	15.1077	14.9993
		$s_n(1)$	0.1828	0.1387	0.1129	0.1054	0.0791
	0.6	$\hat{\theta}_0$	15.2929	15.2358	14.8286	15.0980	14.9857
		$s_n(0)$	0.4581	0.3459	0.3657	0.2763	0.2348
	1	$\hat{\theta}_0$	15.2919	15.3051	14.8983	15.1648	14.8522
		$s_n(0)$	0.7191	0.5492	0.5840	0.4431	0.3738
M6 ($\theta_0=6$)	0.2	$\hat{\theta}_0$	6.0667	6.0743	6.2242	6.1857	6.1760
		$s_n(0)$	0.0364	0.0311	0.0307	0.0295	0.0219
	0.6	$\hat{\theta}_0$	6.0720	6.1813	6.2117	6.1729	6.1496
		$s_n(0)$	0.1081	0.0925	0.0827	0.0709	0.0637
	1	$\hat{\theta}_0$	6.0375	6.0926	6.1266	6.2096	6.1687
		$s_n(0)$	0.1828	0.1535	0.1403	0.1216	0.1072

1) 对于不同 m, e下的各个模型,通过所提出的后向拟合法得到的常值系数的估计值都非常接近其真实值,并且该估计方法较为稳定.

2) 随着 m 的增大,即观测点的增多,对于以上模型来说,虽然估计的精度(估计值逼近精确值的程度)提高不大,但其稳定性有明显的提高.

3) 随着 σ^2的增大,即噪声方差变大,对模型的干扰增强,估计的精度变化不大,但稳定性有明显的降低.

4) 当模型中的变系数项增多时,比如模型 M5中含有两个变系数项,估计的精度有所降低.

4 实例分析

极端温度分析是极端气候事件研究的一部分,也是全球变暖背景下越来越引起人们关注的的问题.基于我国 110个观测站点 1960至 2000年温度逐日观测资料,包括每日的平均气温,最高温度,最低温度,我们来研究近 40年来每个站点的平均气温的变化与极端高温,极端低温的变化的关系 (关于观测点的详细分布信息可见 [16]).由于我国幅员辽阔,气温地区差异特别大.我们采用如下的空间变系数回归模型

$$y(v_i) = U_0(v_i) + U_1(v_i)x_1(v_i) + U_2(v_i)x_2(v_i) + X(v_i), \quad i = 1, 2, \cdots, 110 \qquad (13)$$

其中 v_i 表示第 i个观测点的位置 (包括经度,纬度,和海拔三个指标), $y(v_i)$,$x_1(v_i)$,$x_2(v_i)$ 分别是站点 v_i 对应的平均气温,极端高温,极端低温 40年来的变化量.根据文 [17],平均气温,极端高温和极端低温近 40年来的变化量可用相对应的最后三年的平均值减去最初三年的平均值来度量.即上述变化量由下面的公式计算得到, $z = \frac{t_{2000} + t_{1999} + t_{1998}}{3} - \frac{t_{1962} + t_{1961} + t_{1960}}{3}$.

我们感兴趣的是极端高温和极端低温的变化对于平均气温的变化是否具有空间差异性,即对下面的三个原假设进行检验 $H_0(i)$: $U_i(v_1) = U_i(v_2) = \cdots = U_i(v_{110})$, $i = 0, 1, 2$. 根据文 [6- 8]中的方法,最后所得检验 p值分别为 0.0004647, 0.2422481, 0. 即极端高温的变化对平均气温的变化没有空间差异性,而极端低温的变化对平均气温的变化的影响有空间差异性.

由以上的检验结果,空间变系数模型 (13)转化为如下的半参数空间变系数回归模型

$$y(v_i) = U_0(v_i) + U_1x_1(v_i) + U_2(v_i)x_2(v_i) + X(v_i), \quad i = 1, 2, \cdots, 110$$

由第二部分的后向拟合估计方法,即可得到 $U_0(v_i)$,$U_2(v_i)$,$i = 1, 2, \cdots, 110$和 U_1的估计值.由于篇幅所限,具体结果没有在本文列出.

5 小　　结

至此,我们给出了半参数空间变系数回归模型的一种新的拟合方法,该方法计算简单并给出了常值系数的精确估计表达式,并且通过大量的 数值模拟验证了该方法的合理性和稳定性.最后我们还给出了一个实际的例子.

由于局部拟合的复杂性,要分析常值系数估计的偏差及方差,以及与之相应的统计推断问题,有一定的研究难度. 但如果利用其渐进性质或者利用 Bootstrap方法有望解决,这些问题有待进一步研究.

参考文献:

[1] Anselin L. Spatial Econometrics: Methods and Models[M]. Kluwer Academic, Dordrecht, 1988.

[2] Fotheringham A S, Charlton M, Brunsdon C. The geography of parameter space: an investigation into spatial

non-stationarity[J]. International Journal of Geographical Information Systems, 1996, (10): 605-627.
[3] Fotheringham A S. Trends in quantitative methods I: stressing the local[J]. Progress in Human Geography 1997, 21: 88-96.
[4] Brunsdon C, Fotheringham A S, Charlton M. Geographically weighted regression: a method for exploring spatial nonstationarity[J]. Geographical Analysis, 1996, 28: 281-298.
[5] Fotheringham A S, Charlton M, Brunsdon C. Measuring spatial variation in relationships with geographically weighted regression[J]. In Recent Developments in Spatial Analysis, Edited by M M Fischer and A Getis, Springer-Verlag, London, 1997. 60-82.
[6] Mei Changlin, Zhang wenxiu, Leung Yee. Statistical inferences for varying-cofficient models based on locally weighted regression technique[J]. Acta Mathematicae Applicatae Sinica English Series, 2001, 17(3): 407-417.
[7] Leung Yee, Mei Changlin, Zhang Wenxiu. Statistical tests for spatial nonstationarity based on the geographically weighted regression model[J]. Environment and Planning A, 2000, 32: 9-32.
[8] Mei Changlin, He Shuyuan, Fang Kaitai. A note on the mixed geographically weighted regression model[J]. Journal of Regional Science A, 2004, 44: 143-157.
[9] Mei Changlin, Wang Ning. Functional-coefficient regression and its estimation[J]. Appl Math J Chinese Univ Ser B, 2001, 16(3): 304-314.
[10] Bowman A W. An alternative method of cross-validation for the smoothing of density estimate[J]. Biometrika, 1984, 71: 353-360.
[11] Cleveland W S. Robust locally weighted regression and smoothing scatterplots[J]. Journal of the American Statistical Association, 1979, 74: 829-836.
[12] Brunsdon C, Fotheringham A S, Charlton M. Some notes on parametric significance test for geographically weighted regression[J]. Journal of Regional Science, 1999, 39: 497-524.
[13] Hastie T J, Tibshirani R J. Generalized Additive Models[M]. Chapman and Hall, London, 1990.
[14] 魏传华, 梅长林. 半参数空间变系数回归模型的两步估计方法及其数值模拟[J]. 统计与信息论坛, 2005, 1: 16-19.
[15] Anselin L, Rey S. Properties of tests for spatial dependence in linear regression models[J]. Geographical Analysis, 1991, 23: 112-131.
[16] 续秋霞. 非参数回归方法研究及其在我国极端温度分析中的应用[D].西安交通大学硕士论文, 2004.
[17] 陈宝风, 王羽翔. 南阳近 45年气候变化分析[J]. 河南气象, 1998, 6: 29-29.

Back-Fitting Procedure for Semiparametric Spatially Varying-Coefficient Regression Model

WEI Chuan-hua[1], MEI Chang-lin[2]

(1. School of Statistics, Renmin University of China, Beijing 100872, China)
(2. School of Science, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract This paper proposes a novel procedure for fitting the semiparametric spatially varying-coefficient regression model, by which an explicit expression for estimators of the constant coefficients in the model can be obtained. Extensive simulations are then conducted to examine the performance of the proposed fitting procedure and the results demonstrate that the estimators for the constant coefficients are quite accurate and stable, finally, we analysis a real data.

Keywords semiparametric spatially varying-coefficient regression model; geographically weighted regression procedure; back-fitting procedure; generalized cross-validation method