时空地理加权回归模型的统计诊断

高丽群

(东北林业大学)

【摘 要】时空地理加权回归模型是一类有效的空间数据分析方法,该文将介绍时空地理加权回归模型的统计诊断问题以及基于数据删除模型和均值漂移模型的影响分析,并且利用扰动进行异常点检验.

【关键词】时空地理加权回归模型;数据删除模型;均值漂移模型;扰动;异常点检验

中图分类号:O212 文献标识码:A 文章编号:1000-5617(2015)06-0050-03

0 引言

数据删除模型[1]和均值漂移模型[2]是构造诊断统计量的两种常用模型,数据删除模型比较直观,均值漂移模型便于解释,虽然各有优劣,但是在有关估计的统计性质方面具有等价性. 该文将基于这两种模型对时空地理加权回归模型进行分析.

定义 时空地理加权回归模型的一般形式为:

$$y_i = \sum_{k=1}^{p} \beta_k(u_i,v_i,t_i)x_{ik} + \varepsilon_i, i = 1,2,\cdots,n \tag{1}$$

其中,$(y_i,x_{i1},x_{i2},\cdots,x_{ip})$是第$i$个观测点$(u_i,v_i,t_i)$的观测值,$i=1,2,\cdots,n$. $\beta_k(u_i,v_i,t_i)$是在第i个观测点位置的未知函数. $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$服从均值为0,方差为$\sigma^2$的独立同分布的随机误差项.

1 时空地理加权回归模型基于数据删除模型的参数估计

为了研究数据与模型的拟合情况,一个重要的方法就是考虑每个点对回归分析的影响. 将上述模型(1)删除第j组观测数据$(y_j;x_{j1};x_{j2},\cdots,x_{jp})$,模型可以表示为

$$y_{i(j)} = \sum_{k=1}^{p} \beta_{k(j)}(u_i,v_i,t_i)x_{ik(j)} + \varepsilon_{i(j)}, i = 1,2,\cdots,n, i \neq j \tag{2}$$

定理1 记在删除第j组观测点后,观测点(u_0,v_0,t_0)处系数函数向量的估计为$\hat{\beta}_{(j)}(u_0,v_0,t_0)$. 那么模型(1)的系数函数的估计$\hat{\beta}(u_0,v_0,t_0)$与模型(2)的系数函数的估计$\hat{\beta}_{(j)}(u_0,v_0,t_0)$有如下的关系

$$\hat{\beta}_{(j)}(u_0,v_0,t_0) = \hat{\beta}(u_0,v_0,t_0) - \frac{[X^TW(u_0,v_0,t_0)X]^{-1}x_j(y_j - x_j^T\hat{\beta}(u_0,v_0,t_0))w_j(u_0,v_0,t_0)}{1-x_j^T[X^TW(u_0,v_0,t_0)X]^{-1}x_jw_j(u_0,v_0,t_0)}$$

证明略.

定理2 令删除第j组观测值后方差表示为$\sigma^2_{(j)}$,则$\hat{\sigma}^2_{(j)}$不是$\sigma^2_{(j)}$的无偏估计.

证明 对于时空地理加权回归模型(1),由于

$$\hat{y}(u_0,v_0,t_0) = x_0^T[X^TW(u_0,v_0,t_0)X]^{-1}X^TW(u_0,v_0,t_0)Y,$$

故有$E(\hat{y}(u_0,v_0,t_0)) = E(y)$,即$E(\hat{Y}) = E(Y)$,那么时空地理加权回归模型的残差向量的估计值$\hat{\varepsilon} = (I-L)\varepsilon$,其中

收稿日期:2015-08-22

$$L=\begin{pmatrix} x_1^T[X^TW(u_1,v_1,t_1)X]^{-1}X^TW(u_1,v_1,t_1) \\ x_2^T[X^TW(u_2,v_2,t_2)X]^{-1}X^TW(u_2,v_2,t_2) \\ \vdots \\ x_n^T[X^TW(u_n,v_n,t_n)X]^{-1}X^TW(u_n,v_n,t_n) \end{pmatrix}$$

则方差估计值的数学期望 $E(\hat{\sigma}^2)=\sigma^2$[3]. 对于删除一组观测值之后的时空地理加权回归模型(2),因变量 Y 在观测点 (u_0,v_0,t_0) 处的拟合值为

$$\hat{y}_{(j)}(u_0,v_0,t_0)=x_0^T\hat{\beta}_{(j)}(u_0,v_0,t_0)$$
$$=x_0^T\hat{\beta}(u_0,v_0,t_0)-$$
$$\frac{x_0^T[X^TW(u_0,v_0,t_0)X]^{-1}x_j(y_j-x_j^T\hat{\beta}(u_0,v_0,t_0))w_j(u_0,v_0,t_0)}{1-x_j^T[X^TW(u_0,v_0,t_0)X]^{-1}x_jw_j(u_0,v_0,t_0)}$$

则

$$E(\hat{y}_{(j)}(u_0,v_0,t_0))=\sum_{k=1}^{p}E(\hat{\beta}_{(j)}(u_0,v_0,t_0))x_{ik}=$$
$$\sum_{k=1}^{p}(\beta(u_0,v_0,t_0)-$$
$$\frac{[X^TW(u_0,v_0,t_0)X]^{-1}x_j(x_j\beta_j-x_j^T\beta(u_0,v_0,t_0))w_j(u_0,v_0,t_0)}{1-x_j^T[X^TW(u_0,v_0,t_0)X]^{-1}x_jw_j(u_0,v_0,t_0)})x_{ik}$$

由于其拟合值 $E(\hat{y}_{(j)}(u_0,v_0,t_0))\neq y_{(j)}(u_0,v_0,t_0)$,故其残差向量估计值 $\hat{\varepsilon}\neq(I-L)\varepsilon$,则必有 $E(\hat{\sigma}_{(j)}^2)\neq\sigma_{(j)}^2$,即 $\hat{\sigma}_{(j)}^2$ 不是 $\sigma_{(j)}^2$ 的无偏估计.

2　时空地理加权回归模型基于均值漂移模型的异常值检验

数据删除模型是建立诊断统计量的最基本的模型,由于它非常直观且计算简单因而被广泛应用于实际问题,另一种常见的统计诊断模型是均值漂移模型,考虑对第 i 组数据点增加一个扰动 η[4],它是一未知待估参数,则式(1)关于第 j 组观测的均值漂移模型可记为

$$y_i=\sum_{k=1}^{p}\beta_k(u_i,v_i,t_i)x_{ik}+\eta\lambda_i+\varepsilon_i,\quad i=1,2,\cdots,n \tag{3}$$

对于虚拟变量 λ,有 $\lambda_i=\begin{cases}0,i\neq j\\1,i=j\end{cases}$,并且,$\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)^T$,称此模型为均值漂移模型.

利用均值漂移模型进行异常点的检验就是提出假设检验问题,即扰 η 是否为0的问题,从均值漂移模型来看,如果 η 非零,则意味着数据点 (u_i,v_i,t_i) 不服从原来的模型,则为异常点,这时就需要对模型进行全面的分析. 下面就利用均值漂移模型来进行异常点的检验.

假设检验:

$H_0:\eta=0$

$H_1:\eta\neq0$

首先看 H_1 成立的条件下,利用两步估计法[5] 对该估计方法做介绍.

假定模型中参数 η 已知,则模型(3)可以记为

$$y_i-\eta\lambda_i=\sum_{k=1}^{p}\beta_k(u_i,v_i,t_i)x_{ik}+\varepsilon_i \tag{4}$$

这时,该模型变为空间变系数模型,用地理加权回归方法拟合,得变系数在空间位置点 (u_i,v_i,t_i) 的初次估计表达式[6] 为

$$\hat{\beta}(u_i,v_i,t_i)=\begin{pmatrix}\hat{\beta}_1(u_i,v_i,t_i)\\ \hat{\beta}_2(u_i,v_i,t_i)\\ \vdots\\ \hat{\beta}_p(u_i,v_i,t_i)\end{pmatrix}$$
$$=[X^TW(u_i,v_i,t_i)X]^{-1}X^TW(u_i,v_i,t_i)(Y-\eta Z) \tag{5}$$

将(5)带入到(3)中,认为变系数部分已知,可得

$$y_i-\sum_{k=1}^{p}\hat{\beta}_k(u_i,v_i,t_i)x_{ik}=\eta\lambda_i+\varepsilon_i\quad i=1,2,\cdots,p \tag{6}$$

该模型为一般的线性回归模型. 对于常系数 η 的估计可以用一般最小二乘法得

$$\hat{\eta}=[\lambda^T(I-L)^T(I-L)\lambda]^{-1}\lambda^T(I-L)^T(I-L)Y \tag{7}$$

将(7)带入(5),可得变系数的最终估计为

$$\hat{\beta}(u_i,v_i,t_i)=\begin{pmatrix}\hat{\beta}_1(u_i,v_i,t_i)\\ \hat{\beta}_2(u_i,v_i,t_i)\\ \vdots\\ \hat{\beta}_p(u_i,v_i,t_i)\end{pmatrix}$$
$$=[X^TW(u_i,v_i,t_i)X]^{-1}X^TW(u_i,v_i,t_i)(Y-\hat{\eta}Z) \tag{8}$$

由(7)和(8)可得因变量的拟合值为

$$\hat{Y}=\hat{\eta}\lambda+L(Y-\hat{\eta}\lambda)=SY \tag{9}$$

其中

$$S=L+(I-L)\lambda[\lambda^T(I-L)^T(I-L)\lambda]^{-1}\lambda^T(I-L)^T(I-L)$$

由此可得(3)的残差平方和为

$$RSS_1=\|Y-\hat{Y}\|^2=Y^T(I-S)^T(I-S)Y$$

若 H_1 成立,则由前面描述可知,残差平方和为

$$RSS_0=Y^T(I-L)^T(I-L)Y$$

针对假设 H_0 构造统计量 F 如下

$$F = \frac{(RSS_0 - RSS_1)/m}{RSS_1/n} = \frac{Y^T(K_1 - K_2)Y}{Y^T K_2 Y} \tag{10}$$

其中，$K_1 = (I-L)^T(I-L)$，$K_2 = (I-S)^T(I-S)$，$m = \mathrm{tr}(K_1 - K_2)$，$n = \mathrm{tr}K_2$.

当拒绝 H_0 时，F 有偏大的趋势，否则不然. 故检验 p 值为

$$p_0 = P_{H_0}(F > f)$$

f 由(10) 求得.

对于原假设 H_0，在其为真时，一般不服从 F 分布，故可以在模型误差为正态分布的假定下用 F 逼近法求其检验 p 值.

3 结束语

该文将数据删除模型应用在时空地理加权回归模型之中，给出了系数函数的估计值，并且分析了删除一组观测点之后方差估计的无偏性. 经分析可知，当删除一个观测点后，对距离其近的观测点影响大，反之，对距离其远的观测点影响小，并且删除观测点后，对模型整体还是有很大影响的.

其次，该文基于均值漂移模型将扰动值 η 插入到时空地理加权回归模型之中，构造了 F 统计量检验异常点.

时空地理加权回归模型是一种全面分析空间特性的模型，由于其参数较多，因此模型整体需要统计诊断来进行影响分析. 需要说明的是，该文在数据删除模型中只讨论了删除一组观测点的情况，在该文的基础上可以进一步延伸，探讨删除多个观测点的情况.

参 考 文 献

[1] Fung W K, Zhu Z, We i B. And He X. Influence diagnostics and outlier tests for semiparametric mixed models[J]. J Roy Statist Soc Ser B64, 2002: 565 - 579.

[2] 魏传华，梅长林. 半参数空间变系数回归模型的 Back - Fitting 估计[J]. 数学的实践与认识，2006，3(19).

[3] 梅长林，王林. 近代回归分析方法[M]. 北京：科学出版社，2012.

[4] 朱忠义，韦博成. 半参数非线性模型的统计诊断与影响分析[J]. 应用数学学报，2001，24(4)：568 - 581.

[5] 曾林蕊，朱忠义，茆诗松. 半参数广义线性模型的影响分析与异常点检验[J]. 高校应用数学学报，2004，19(3)：323 - 332.

[6] 吴玉鸣，徐建华. 中国区域经济增长集聚的空间统计分析[J]. 地理科学，2004(6)：654 - 659.

Statistical Diagnosis of Geographically and Temporally Weighted Regression

Gao Liqun

(Northeast Forestry University)

Abstract: Geographically and Temporally Weighted Regression is an effective spatial data analysis method, in this paper, the problems of statistical diagnosis of the geographically and temporally weighted regression are studied and analyzed based on the data deleted model and mean shift model. The abnormal points test is made by using disturbance.

Keywords: Geographically and Temporally Weighted Regression; Data Deleted Model; Mean Shift Model; Disturbance; Abnormal points test

(责任编辑：于达)