文章编号:1008-5564(2013)02-0035-05

时空地理加权回归模型的统计诊断

刘美玲,王　博,张雪敏
(西安建筑科技大学 理学院,西安 710055)

摘　要:时空地理加权回归模型作为一类能简单有效的解决数据时空特性问题的数据分析方法已经得到了广泛的应用.主要研究该模型的统计诊断与影响分析.首先,基于数据删除模型定义了 Cook 统计量;其次,基于均值漂移模型讨论了异常点的检验问题,并构造了相应的检验统计量.

关键词:时空地理加权回归模型; Cook 距离;均值漂移模型
中图分类号:O212　　**文献标志码**:A

Statistical Diagnostics for Spatial - Temporal Geographical Weighted Regression Model

LIU Mei-ling, WANG Bo, ZHANG Xue-min
(College of Sciences, Xi'an University of Architecture & Technology, Xi'an 710055, China)

Abstract: As a spatial data analysis method, Spatial - Temporal Geographical Weighted Regression Model is effective and simple to address the spatial and temporal features of data and therefore has a wide application. In this paper, we mainly discuss the statistical diagnostics and influence analysis on this model. Firstly, Cook statistic based on data deletion model is introduced; secondly, based on the mean shift model, the testing of a given observation is discussed and the corresponding test statistic is constructed.

Key words: spatial - temporal geographical weighted regression model; Cook's distance; mean shift model

时空地理加权回归模型 GTWR 是 Huang 等人[1]在传统的地理加权回归模型 GWR[2,3]基础上扩展提出的.该模型通过假定回归系数是地理位置和观测时刻的任意函数,将数据的时空特性嵌入到回归模型中,为分析回归系数的时空特性以及解决时空非平稳性提供了有效的可行性.最近,在经济学、统计学、房地产分析等领域内得到广泛的应用[4].

GTWR 模型记为如下形式:

$$y_i=\beta_0(u_i,v_i,t_i)+\sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik}+\varepsilon_i,i=1,2,\cdots,n \tag{1}$$

其中$\{y_i;x_{i1},x_{i2},\cdots,x_{id}\}_{i=1}^{n}$,是在观测点$(u_i,v_i,t_i)$处因变量 Y 和自变量 $X_1,X_2,\cdots,X_d$ 的 n 组观测值($i=$

收稿日期:2013-01-20
作者简介:刘美玲(1986—),女,陕西渭南人,西安建筑科技大学理学院硕士研究生,主要从事应用数学研究;
通讯作者:苏变萍(1963—),教授,硕导,主要从事数量经济与金融数学研究.

$1,2,\cdots,n)$. $\beta_k(u_i,v_i,t_i)(k=1,2,\cdots,d)$是第$i$个观测点$(u_i,v_i,t_i)$处的未知参数,其中各元素是$(u_i,v_i,t_i)$的任意函数,$\varepsilon_1,\varepsilon_1,\cdots,\varepsilon_n$为独立同分布的误差项,通常假定均值为零,方差为$\sigma^2$.

1 地理加权回归模型的估计

XUAN Hai-yan[5]等人基于加权最小二乘原理,给出了模型(1)的参数估计,以及与之相关的权函数选取原则,在此对该方法做一简单介绍.

鉴于模型(1)在观测点(u_i,v_i,t_i)处的观测值$(y_i;x_{i1},x_{i2},\cdots,x_{id})$,$i=1,2,\cdots,n$相对于$(u_0,v_0,t_0)$处的参数值来说的"重要程度"不一样,因此在给定研究区域的任一观测点(u_0,v_0,t_0)处指定一组权,记为$w_1(u_0,v_0,t_0),w_2(u_0,v_0,t_0),\cdots,w_n(u_0,v_0,t_0)$来表示这种"重要程度",其中$w_i(u_0,v_0,t_0)$对应于第$i$组观测值$(y_i;x_{i1},x_{i2},\cdots,x_{id})$,$i=1,2,\cdots,n$. 相应的距离$(u_0,v_0,t_0)$点较近的观测赋予较大的权值,距离$(u_0,v_0,t_0)$点较远的观测赋予较小的权值. 根据加权最小二乘原理[6],对时空空间的任一点(u_0,v_0,t_0)处的未知参数$\beta(u_0,v_0,t_0)$,可通过使目标函数

$$\sum_{i=1}^{n}\left[y_i-\beta(u_0,v_0,t_0)-\sum_{k=1}^{d}\beta_k(u_0,v_0,t_0)x_{ik}\right]^2 w_i(u_0,v_0,t_0) \tag{2}$$

达到最小来予以估计,其中

$$\beta(u_0,v_0,t_0)=(\beta_0(u_0,v_0,t_0),\beta_1(u_0,v_0,t_0),\cdots,\beta_d(u_0,v_0,t_0))^{\mathrm{T}}.$$

下面令 $$\hat{\beta}(u_0,v_0,t_0)=(\hat{\beta}_0(u_0,v_0,t_0),\hat{\beta}_1(u_0,v_0,t_0),\cdots,\hat{\beta}_d(u_0,v_0,t_0))^{\mathrm{T}}$$

$$W(u_0,v_0,t_0)=diag\left[w_1(u_0,v_0,t_0),w_2(u_0,v_0,t_0),\cdots,w_n(u_0,v_0,t_0)\right]$$

$$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},X=\begin{pmatrix}1&x_{11}&\cdots&x_{1d}\\1&x_{21}&\cdots&x_{2d}\\\vdots&\vdots&\ddots&\vdots\\1&x_{n1}&\cdots&x_{nd}\end{pmatrix}$$

得到(u_0,v_0,t_0)处的参数估计为

$$\hat{\beta}(u_0,v_0,t_0)=\left[X^{\mathrm{T}}W(u_0,v_0,t_0)X\right]^{-1}X^{\mathrm{T}}W(u_0,v_0,t_0)Y$$

因变量Y在(u_0,v_0,t_0)处的拟合值为 $$\hat{Y}=SY$$

得残差向量 $$\hat{e}=(\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n)^{\mathrm{T}}=Y-\hat{Y}=Y-SY=(I-S)Y$$

拟合的残差平方和为 $$RSS=e^{\mathrm{T}}e=(Y-\hat{Y})^{\mathrm{T}}(Y-\hat{Y})=Y^{\mathrm{T}}(I-S)^{\mathrm{T}}(I-S)Y$$

其中

$$S=\begin{bmatrix}X_1^{\mathrm{T}}[X^{\mathrm{T}}W_1(u_0,v_0,t_0)X]^{-1}X^{\mathrm{T}}W_1(u_0,v_0,t_0)\\X_2^{\mathrm{T}}[X^{\mathrm{T}}W_2(u_0,v_0,t_0)X]^{-1}X^{\mathrm{T}}W_2(u_0,v_0,t_0)\\\vdots\\X_n^{\mathrm{T}}[X^{\mathrm{T}}W_1(u_0,v_0,t_0)X]^{-1}X^{\mathrm{T}}W_n(u_0,v_0,t_0)\end{bmatrix}$$

由以上得到的结果,构造σ^2的估计

$\hat{\sigma}^2=RRS/tr[(I-S)^{\mathrm{T}}(I-S)]$,其中$tr(A)$表示矩阵$A$的迹.

在时空数据分析中,观测点离(u_0,v_0,t_0)越近,对点(u_0,v_0,t_0)处的参数估计影响越大,以d_i来度量点(u_0,v_0,t_0)与观测点(u_i,v_i,t_i),$i=1,2,\cdots,n$间的时空距离,距离大的赋予较大的权值,反之则赋予较小的权值. 权函数记为如下形式:

$$w_i(u_0,v_0,t_0)=\exp\left\{-\frac{(u_0-u_i)^2+(v_0-v_i)^2}{h_1^2}\right\}\times\exp\left\{-\frac{(t_0-t_i)^2}{h_2^2}\right\}$$

其中$h_1=\sqrt{h^2/\lambda}$和$h_2=\sqrt{h^2/\mu}$分别为空间和时间窗宽参数. 并用交叉确认法[7]对h_1和h_2予以确定. 即令

$$CV(h_1,h_2)=\sum_{i=1}^{n}[y_i-\hat{y}_i(h_1,h_2]^2$$

然后选择 h_{10} 和 h_{20} 使得 $CV(h_1,h_2)$ 达到最小.

2　基于数据删除模型的影响分析

为了研究 GTWR 模型中数据集与模型的拟合问题,类似线性回归模型,我们将基于数据删除模型来研究各组观测点的影响大小,考虑删除第 i 组观测点 $(y_i;x_{i1},x_{i2},\cdots,x_{id})$, $i=1,2,\cdots,n$ 以后的模型及参数的估计问题,则模型可以表示为:

$$y_i=\beta_0(u_j,v_j,t_j)+\beta_1(u_j,v_j,t_j)x_{j1}+\cdots+\beta_d(u_j,v_j,t_j)x_{jd}+\varepsilon_j,j=1,2,\cdots,n,j\neq i \tag{3}$$

这种模型称之为数据删除模型(CDM),依照第二节介绍的时空地理加权回归方法,我们可得到该模型在点(u,v,t)处的参数$\beta(u,v,t)$的估计值$\hat{\beta}_{(i)}(u,v,t)$和 RSS 的局部加权最小二乘估计 $RSS(i)$. 则就会有下面的结论:

定理 1　模型(2)和模型(3)中相应的参数估计有如下关系

$$\hat{\beta}_{(i)}(u,v,t)=\hat{\beta}(u,v,t)-\frac{\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_iW_i(u,v,t)\hat{e}_i}{1-h_{ii}(u,v,t)} \tag{4}$$

其中 $\hat{e}_i=y_i-x_i^{\mathrm{T}}\hat{\beta}(u,v,t)$, $h_{ii}=x_i^{\mathrm{T}}\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_iW_i(u,v,t)$

证明　由局部加权最小二乘估计可得数据删除模型(3)中参数的估计为

$$\hat{\beta}_{(i)}(u,v,t)=\left[X_{(i)}^{\mathrm{T}}W_{(i)}(u,v,t)X_{(i)}\right]^{-1}X_{(i)}^{\mathrm{T}}W_{(i)}(u,v,t)Y_{(i)} \tag{5}$$

首先,由于

$$X^{\mathrm{T}}W(u,v,t)X=X_{(i)}{}^{\mathrm{T}}W_{(i)}(u,v,t)X_{(i)}+x_iW_i(u,v,t)x_i^{\mathrm{T}} \tag{6}$$

利用恒等式$(A-ug^{\mathrm{T}})^{-1}=A^{-1}+\frac{A^{-1}ug^{\mathrm{T}}A^{-1}}{1-u^{\mathrm{T}}A^{-1}g}$,其中 A 为 $n\times n$ 可逆阵.

则有

$$\left[X_{(i)}^{\mathrm{T}}W_{(i)}(u,v,t)X_{(i)}\right]^{-1}=$$

$$\left[X^{\mathrm{T}}W(u,v,t)X-x_iW_i(u,v,t)x_i^{\mathrm{T}}\right]^{-1}=$$

$$\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}+\frac{\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_iW_i(u,v,t)x_i^{\mathrm{T}}\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}}{1-h_{ii}(u,v,t)} \tag{7}$$

其中

$$h_{ii}(u,v,t)=\left[x_iW_i(u,v,t)\right]^{\mathrm{T}}\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_i=x_i^{\mathrm{T}}\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_iW_i(u,v,t)$$

其次有

$$X^{\mathrm{T}}W(u,v,t)Y=X_{(i)}{}^{\mathrm{T}}W_{(i)}(u,v,t)Y_{(i)}+x_iW_i(u,v,t)y_i \tag{8}$$

然后将(7)(8)两式代入式(5)中,得到

$$\hat{\beta}_{(i)}(u,v,t)=\hat{\beta}(u,v,t)-\frac{\left[X^{\mathrm{T}}W(u,v,t)X\right]^{-1}x_iW_i(u,v,t)\hat{e}_i}{1-h_{ii}(u,v,t)}$$

差值$\hat{\beta}-\hat{\beta}_{(i)}$是 i 组观测点的影响大小的一种度量,差值的绝对值越大,影响也就越大,但是它是一个向量,在实际应用中不方便比较,所以为了解决这个问题,我们就必须先定义一个合适的数量或者距离,以便于比较影响的大小. 因此,根据线性回归分析中度量其影响常用的 Cook 距离,即

$$D_i=\frac{1}{\sigma^2q}(\hat{\beta}-\hat{\beta}_{(i)})^{\mathrm{T}}[X^{\mathrm{T}}X]^{-1}(\hat{\beta}-\hat{\beta}_{(i)}) \tag{9}$$

将以上的结论推广到多组观测点,可以定义对应于第 i 组观测点的 Cook 距离为

$$D_i = \frac{\sum_{j=1}^{n}\left\{[\hat{\beta}_{(i)}(u_j,v_j,t_j)-\hat{\beta}(u_j,v_j,t_j)]^{\mathrm{T}}x_i x_i^{\mathrm{T}}[\hat{\beta}_{(i)}(u_j,v_j,t_j)-\hat{\beta}(u_j,v_j,t_j)]\right\}}{tr(s)\hat{\sigma}^2}$$

$$= \frac{\| Y - Y_{(i)} \|^2}{tr(s)\hat{\sigma}^2} \tag{10}$$

其中 $Y_{(i)} = [x_1^{\mathrm{T}}\hat{\beta}_{(i)}(u_1,v_1,t_1),\cdots,x_n^{\mathrm{T}}\hat{\beta}_{(i)}(u_n,v_n,t_n)]^{\mathrm{T}}$

3 均值漂移模型与异常点的检验

数据删除模型是构造有效诊断统计量的基础,在实践中也是一种重要的诊断模型,另一种重要的常用诊断模型是均值漂移模型(MSOM).针对模型(1),考虑对第 i 组观测点增加一个扰动 γ,则关于第 i 组观测点的均值漂移模型可记为

$$\begin{cases} y_j = \beta_0(u_j,v_j,t_j) + \sum_{k=1}^{d}\beta_k(u_j,v_j,t_j)x_{jk} + \varepsilon_j, j \neq i \\ y_i = \beta_0(u_i,v_i,t_i) + \sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik} + \gamma + \varepsilon_i \end{cases} \tag{11}$$

进一步将此模型记为如下形式:

$$y_i = \beta_0(u_0,v_0,t_0) + \sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik} + \gamma d_i + \varepsilon_i, i = 1,2,\cdots,n \tag{12}$$

其中 γ 为扰动值,是一未知的待估参数,记虚拟变量 d 为 $d_i = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$,我们记 $d_i = (d_1,d_2,\cdots,d_n)^{\mathrm{T}}$

那么,此模型变为了混合地理加权回归模型.

数据删除模型与均值漂移模型对于一般的线性回归模型具有等价性,即两个模型的估计量是相同的.对于其他如空间变系数模型[8]、混合地理加权回归模型[9]等证明了二者仍具有等价性.我们把这种结论推广到 GTWR 这一模型中.

基于上面的均值漂移模型,检验第 i 组观测值是否为异常点,等价于下面的假设检验问题:

H_0: $\gamma = 0$ 对应 H_1: $\gamma \neq 0$

首先,如果假设 H_1 成立,即模型(12)中参数 γ 已知,则模型可写为

$$y_i - \gamma d_i = \beta_0(u_0,v_0,t_0) + \sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik} + \varepsilon_i, i = 1,2,\cdots,n \tag{13}$$

则该模型就变成了空间变系数回归模型.则可用两步估计方法为基础估计该模型中的变系数部分与参数部分.

首先,基于局部加权最小二乘法得到系数的最初估计

$$\hat{\beta}(u_i,v_i,t_i) = \left[X^{\mathrm{T}}W(u_i,v_i,t_i)X\right]^{-1}X^{\mathrm{T}}W(u_i,v_i,t_i)(Y-\gamma d_i), i = 1,2,\cdots,n \tag{14}$$

将(14)式代入到(12)式中得

$$y_i - \sum_{k=1}^{d}\hat{\beta}_k(u_i,v_i,t_i)x_{ik} = \gamma d_i + \varepsilon_i, i = 1,2,\cdots,n$$

此模型就变为了一般的线性回归模型,因此我们用一般最小二乘法估计常系数 γ,

即使

$$L(\beta) = \sum_{i=1}^{n}\left\{y_i - \sum_{k=1}^{d}\hat{\beta}_k(u_i,v_i,t_i)x_{ik} - \gamma d_i\right\}^2$$

$$= \| Y - \gamma d_i - S(Y-\gamma d_i) \|^2$$

$$= \| (I-S)Y - (I-S)\gamma d_i \|^2$$

达到最小,得 γ 的估计为

$$\hat{\gamma} = \left[d_i^{\mathrm{T}}(I-S)^{\mathrm{T}}(I-S)d_i\right]^{-1}d_i^{\mathrm{T}}(I-S)^{\mathrm{T}}(I-S)Y \tag{15}$$

将 $\hat{\gamma}$ 代入(14)式得到变系数部分的最终估计为

$$\hat{\beta}(u_i,v_i,t_i)=\left[X^{\mathrm{T}}W(u_i,v_i,t_i)X\right]^{-1}X^{\mathrm{T}}W(u_i,v_i,t_i)(Y-\gamma d_i),i=1,2,\cdots,n \tag{16}$$

基于上面的拟合方法，得模型(12)中因变量的拟合值为

$$\hat{Y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n)^{\mathrm{T}}=\hat{\gamma}d_i+S(Y-\gamma d_i)=L_1Y$$

其中 $L_1=S+(I-S)d_i\left[d_i^{\mathrm{T}}(I-S)^{\mathrm{T}}(I-S)d_i\right]^{-1}d_i^{\mathrm{T}}(I-S)^{\mathrm{T}}(I-S)$

得残差平方和为

$$RSS_1=\|Y-\hat{Y}\|^2=Y^{\mathrm{T}}(I-L_1)^{\mathrm{T}}(I-L_1)Y \tag{17}$$

再次，如果原假设 H_0 成立，则模型(12)变为

$$y_i=\beta_0(u_0,v_0,t_0)+\sum_{k=1}^{d}\beta_k(u_i,v_i,t_i)x_{ik}+\varepsilon_i,i=1,2,\cdots,n$$

那么，模型拟合的残差平方和为

$$RSS_0=Y^{\mathrm{T}}(I-L)^{\mathrm{T}}(I-L)Y \tag{18}$$

对于假设检验 H_0，类似于传统的 F 检验，构造检验统计量如下：

$$T=\frac{(RSS_0-RSS_1)/\lambda_1}{RSS_1/\delta_1}=\frac{Y^{\mathrm{T}}(H_0-H_1)Y/\lambda_1}{Y^{\mathrm{T}}H_1Y/\delta_1} \tag{19}$$

其中 $H_0=(I-L)^{\mathrm{T}}(I-L)$，$H_1=(I-L_1)^{\mathrm{T}}(I-L_1)$，$\lambda_1=tr(H_0-H_1)$，$\delta_1=tr(H_1)$. RSS_0-RSS_1 反映了原假设与备择假设下模型的拟合效果差异. 若二者有显著差异，则倾向于拒绝原假设 H_0，即可认为第 i 组观测值是异常点. 由于模型拟合的复杂性，统计量 T 在原假设下一般不服从 F 分布，但在模型误差为正态分布的假定下可用 F 分布逼近法求其检验 p 值. 令 t 为 T 的观测值，则对于 p_0 有如下结果

$$p_0=P_{H_0}(T>t)\approx P(F(r_1,r_2)>t) \tag{20}$$

其中 $r_1=\lambda_1^2/\lambda_2$，$r_2=\delta_1^2/\delta_2$，$\lambda_2=tr[(H_0-H_1)^2]$，$\delta_2=tr(H_1^2)$. 对于给定的显著性水平 α，若 $p_0<\alpha$，则拒绝原假设 H_0

本文将统计学的相关理论引入到时空地理加权回归模型这一常用模型中，研究了该模型对应的数据删除模型和均值漂移模型，并提出了相应 Cook 统计量和构造了用于检验异常点的检验统计量.

[参 考 文 献]

[1] HUANG B, WU B, BARRY M. Geographically and temporally weighted regression for modeling spatio - temporal variation in house prices [J]. International Journal of Geographically Information Science, 2010, 24(3): 383 - 401.

[2] BRUNSDON C, FOTHERINGHAM A S, CHARLTON M. Geographically weighted regression: A method for exploring spatial nonstationarity[J]. Geographical Analysis, 1996, 28(4): 281 - 299.

[3] BRUNSDOM C, FOTHERINGHAM A S, CHARITON M. Some notes on parametric significance test for geographically weighted regression[J]. Journal of Regional Science, 1999, 39(1): 321 - 332.

[4] 吴波，刘彪，詹锡兰. 应用改进的时空地理加权回归模型分析城市住宅价格变化[J]. 自然资源学报，2006，21(6)：55 - 63.

[5] 玄海燕，李帅峰. 时空地理加权回归模型及其拟合[J]. 甘肃科学学报，2011，23(4)：119 - 121.

[6] 童恒庆. 理论计量经济学[M]. 北京：科学出版社，2005.

[7] HASTIE T J, TIBSHIRANI R J. Generalized additive models[M]. London: Chapman and Hall, 1990.

[8] 魏传华，吴喜之. 空间变系数模型的统计推断[J]. 数理统计与管理，2007，26(6)：1027 - 1033.

[9] 魏传华，吴喜之. 混合地理加权回归模型的统计推断[J]. 统计与信息论坛，2009，24(1)：9 - 13.

[责任编辑　王新奇]