第 27 卷第 5 期　　四川理工学院学报(自然科学版)　　Vol. 27　No. 5
2014 年 10 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Oct. 2014
文章编号:1673-1549(2014)05-0096-05　　DOI:10.11863/j.suse.2014.05.22

我国省际 SO_2 排放量的特异性研究

李晓琳, 惠　洁, 刘　倩

(成都理工大学管理科学学院, 成都　610059)

摘　要:采用《中国统计年鉴》的数据,利用 SPSS 软件在主成分分析的基础上构造了多元回归模型。根据模型产生的杠杆值、剔除学生化残差、库克距离、DFFITS 统计量对我国各省的 SO_2 排放量进行了特异数据挖掘。此外还探讨了特异值产生原因。

关键词:SO_2 排放量;主成分分析;多元线性回归;特异值

中图分类号:O212.4　　**文献标志码**:A

引　言

SO_2 是典型的大气污染物,主要来自含硫能源(煤炭、石油)的燃烧、机动车辆的尾气排放、金属熔炼等过程。SO_2 易溶于人体血液和其它粘稠液对人体危害大,SO_2 会对植物产生急性危害破坏生态系统多样性,同时,SO_2 是主要酸沉降(酸雨)前体物,它会使自然和人造材料遭到腐蚀、使水质酸化[1]。

南方根据石家庄市的大气质量特征,确定了石家庄市的环境方案和指标分配方案[2]。黄青等应用 MM5/CMAX 耦合模型对北京市一次 SO_2 污染过程进行了模拟,采用 SO_2 贡献来源识别技术对北京市 SO_2 来源进行了识别分析,为筛选 SO_2 贡献重点地区和重点行业、制定城市和区域大气污染控制方案[1]。张昭利利用投入产出模型,讨论对外贸易对我国 SO_2 污染的影响,并利用结构分解分析方法将贸易含污量的变动分解为各种效应,分析各种效应的大小及变化趋势,并从贸易政策的角度提出相关的意见建议[3]。

综上所述有关 SO_2 的研究大多是单因素的定量研究,本文从多角度讨论了 SO_2 的影响因素,选取了社会指标、能源控制指标、大气污染控制指标三个一级指标。

Hawkins 认为,"一个离群点是一个观察点,它偏离其它观察点如此之大,以致引起怀疑是由不同机制生成的"[4]。传统的数据挖掘算法主要划分为五类:基于密度的方法,基于距离的方法,基于统计的方法,基于偏离的方法和基于聚类的挖掘算法[5-6]。传统的统计方法不适用于高维空间,本文首先采用主成分分析方法进行了降维,然后构造了多元回归模型,最后对 2011 年全国 30 个省自治区(西藏地区缺少数据)SO_2 排放量进行了特异性研究[7]。

1　研究方法

1.1　指标选择及数据来源

在描述我国省际 SO_2 排放量时,所选取的指标主要分为三类:社会指标、能源消耗指标、大气污染控制指标。本文选取 x_1 工业增加值(亿元)、x_2 年末常住人口(万人)、x_3 煤炭消费量(万吨)、x_4 焦炭消费量(万吨)、x_5 柴油消费量(万吨)、x_6 燃料油消费量(万吨)、x_7 建城区绿化覆盖率(%)、x_8 生活无害化处理能力(吨/日)、x_9 森林覆盖率(%)、x_{10} 民用汽车拥有量(万辆)、x_{11} 居民消费水平(元)作为研究对象,其中我国各省的 SO_2 排放量(万吨)作为因变量 y, 其余 11 个自变量的选取是反映个省的综合情况的指标。

本文研究地区包括中国大陆 30 个省、直辖市和自

收稿日期:2014-05-07

基金项目:四川省应用基础计划项目(2012JY0033)

作者简介:李晓琳(1989-),女,河北新乐人,硕士生,主要从事统计分析模型及数据分析方面的研究,(E-mail)li18328503467@163.com

治区,其中西藏自治区因缺少数据暂不做研究,且不包括香港、澳门特别行政区和台湾省。采用的数据来自《中国统计年鉴—2011》。

1.2 研究思路

先对数据进行标准化,通过相关性分析得出与 SO_2 排放量显著相关的因素。由于选取的各变量间还有显著的相关性,因而建立回归模型时各个观察值必须是独立的,这就会影响建模质量,因此先进行主成分分析[8]。它在对原有变量进行信息重构时不仅使变量的数量远少于原有指标的数量,而且变量间不存在线性关系,能让模型的质量得到很大提高。接着由得到的主成分构建回归模型,根据回归模型得出的杠杆值、剔除学生化残差、库克距离、DFFITS 统计量对各省的 SO_2 排放量进行特异性挖掘[9-10],检验出异常值,进而对回归模型进行重建。

2 异常值诊断原理

本文中异常值的讨论是针对一般多元般线性回归模型:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i$$

用矩阵形式表示为 $Y = X\beta + \varepsilon$, 多元回归型的参数估计同一元线性回归方程一样也是在要求误差平方和最小的前提下,用最小二乘法求解参数。就是求解估计向量 $\hat{\beta}$。该最小二乘估计为:$\hat{\beta} = \beta = (X'X)^{-1}X'Y, \hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$。

主要从 4 个角度进特异性数据挖掘研究:

(1)杠杆水平原理

由公式 $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$ 可知矩阵的 $X(X'X)^{-1}X'$ 作用是把因变量 Y 变为拟合值 $\hat{Y}$, 形象地称矩阵 $X(X'X)^{-1}X'$ 为帽子矩阵,记为 H[11]。h_{ii} 为帽子矩阵 H 对角线上的的元素,称 h_{ii} 为第 i 个观察点的杠杆水平(leverage)。可以证明[12]:

(i) $0 \leqslant h_{ii} \leqslant 1$ 且 $h_{ii} = 1$ 时, $h_{ij} = 0, j \neq i$

(ii) $\sum_{i=1}^{n} h_{ii} = p + 1$

(iii) $h_{ii} = \frac{1}{n} + (x_i - \bar{x})'(X^{*\prime}X^{*})(x_i - \bar{x})$

这里

$$X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix}, X^{*} = \begin{pmatrix} (x'_1 - \bar{x})' \\ (x'_2 - \bar{x})' \\ \vdots \\ (x'_n - \bar{x})' \end{pmatrix}, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

从(iii)可知, h_{ii} 表示第对 i 个观察点的样本数据与 x_i 样本与自变量平均值之间的距离远近。杠杆值越大说明与自变量平均值之间的距离越远。又由于 $\hat{Y} = H \cdot Y$,可以看出 h_{ii} 实际上是观测值 $\hat{Y}$ 在拟合值 $\hat{Y}$ 中的权重。h_{ii} 还表示当观测值 y_i 有微小变化时相应的拟合值 $\hat{y}_i$ 的变化大小,因此 h_{ii} 越大,$\hat{y}_i$ 对 y_i 的变化越敏感。

由(ii)可得,杠杆值 h_{ii} 的平均值为 $\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p+1}{n}$,这样一个杠杆值 h_{ii} 大于2 倍或3 倍 $\bar{h}$,就认为是高杠杆点。本文用 SPSS 软件计算杠杆值时,计算出的是中心化的杠杆值 ch_{ii},由参考文献[11]可知 $ch_{ii} = h_{ii} - \frac{1}{n}$, 因此 $\sum_{i=1}^{n} ch_{ii} = p$, 中心化杠杆值 ch_{ii} 的均值为 $\bar{ch} = \frac{1}{n}\sum_{i=1}^{n} ch_{ii} = \frac{p}{n}$。

(2)剔除学生化残差[12]

$$r_i^{*} = \frac{\delta_i}{\delta(i)\sqrt{i - h_{ii}}}, i = 1, 2, \cdots, n$$

其中,

$$\delta = Y - \hat{Y} = Y - X\hat{\beta} = (1 - X(X'X)X')Y = (1 - H)Y$$

是回归模型估计的普通残差向量

$$\hat{\delta}^2(i) = Y_{(i)}[I - X_{(i)}(X_{(i)}'X_{(i)})X_{(i)}']Y/(n - p - 2)$$

是剔除第 i 组数据之后对于剩余 $n - 1$ 组数据得到的回归模型估计导出的误差方差 δ^2 的估计, $X_{(i)}, Y_{(i)}$ 是剔除第 i 个观测值剩下的数据。

其优越之处是:①在误差 $e: N(0, \delta^2, I)$ 条件下 $\hat{\delta}^2(i)$ 和 δ_i 相互独立,且 r_i^{*} 服从自由度 $(n-1) - p - 1$ 的 t 分布。②如果目的是检查异常值,那么 r_i^{*} 比其它残差都为有效。在 0.20 显著性水平下,有 $(n-1) - p - 1$ 个自由度的 t 分布的双侧分位数 $t_{0.10}$, $|r_i^{*}| > t_{0.10}((n-1) - p - 1)$ 时,则第 i 次观测值是一个异常值。

(3) *Cook's* 距离(库克距离)[12]

第 i 个点的数据对应的库克距离定义为:

$$D_i = \frac{\|\hat{Y} - \hat{Y}_{(i)}\|^2}{P\hat{\delta}^2} = p^{-1}r_{(i)}^2\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

其中, r_i 为学生化残差, D_i 是第 i 个学生化残差 r_i 的平方与 h_{ii} 的一个单调函数的乘积,因此,统计量 D_i 是从残差和观测点所处位置两方面反映第 i 点对参数估计的影响程度。如果 p 固定, D_i 由 r_i 和 h_{ii} 两个不同的方面决定,当 $D_i > \frac{4}{(n - k - 1)}$, 可认为第 i 个观测点为影响点,其中 k 是预测变量数。

(4) $DFFITS_i$ 统计量[12]

$DFFITS_i$ 统计量法是通过测量一个观测值排除是否对其预测值的影响,判断该观测值是否为强影响点的一种方法。$DFFITS_i$ 统计量定义为:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{S_{(i)}\sqrt{1 - x_i(X'X)^{-1}x_i'}}$$

其中,$s_{(i)}$ 为排除第 i 个观测值后拟和回归的均方误差。一般建议,当 $|DFFITS_i| > 2\sqrt{(p+1)/n}$ 时,该观测值对回归有较大影响,应加以关注。

$DFFITS_i$ 统计量与 *Cook's* 距离法的主要区别在于它观察的是第 i 个观测值的预测值,而不是所有点的预测。

3 实证分析

3.1 我国省际 SO_2 排放量与选取因素的相关性分析

以我国大陆 30 个省、直辖市和自治区,以上述选择的指标为数据样本,用 SPSS 软件先对数据进行标准化,然后计算出 SO_2 排放量与其它因素的 *Person* 相关系数 r 和 p 值。设定 $p < 0.1$ 表示相关性显著,将 $p > 0.1$ 的指标删除,就可以得到 zx_1、zx_2、zx_3、zx_4、zx_5、zx_6、zx_8、zx_9、zx_{10}、zx_{11} 与 SO_2 排放量显著相关。

3.2 基于主成分(PCA)分析回归模型的建立[13-14]

由主成分分析得到三个主成分 F_1、F_2、F_3,且方差占全部方差比例的 84.247%,说明基本上保留了原有指标的所有信息,同时根据成分得分系数矩阵得到 F_1、F_2、F_3 的线性组合:

$$\begin{aligned}F_1 &= 0.169x_1 + 0.157x_2 + 0.155x_3 + 0.118x_4 + 0.169x_5 + 0.121x_6 + 0.147x_8 - 0.006x_9 + 0.167x_{10} + 0.026x_{11}\\ F_2 &= 0.139x_1 + 0.010x_2 - 0.233x_3 - 0.270x_4 + 0.116x_5 + 0.259x_6 - 0.289x_8 + 0.203x_9 + 0.143x_{10} + 0.437x_{11}\\ F_3 &= 0.052x_1 + 0.281x_2 - 0.095x_3 - 0.124x_4 + 0.072x_5 - 0.229x_6 - 0.019x_8 + 0.697x_9 + 0.052x_{10} - 0.387x_{11}\end{aligned}$$

然后对由原始标量得出的三个因子 F_1、F_2、F_3 进行回归分析,建立 y 与 F_1、F_2、F_3 的回归模型。得到相关系数 $R = 0.947$,判定系数 $R^2 = 0.896$,调整的判定系数 $\bar{R}^2 = 0.884$,回归估计的标准误差 $S = 0.3399$,说明样本回归代表性强。

由回归方程的方差分析表:统计量 $F = 75.001$,显著性为 $0.000 < 0.001$ 表示回归模型的整体解释变异量达到显著性水平。最后得出多元线性方程为:

$$\hat{y} = -1.851E-16 + 0.796F_1 - 0.513F_2 - 0.024F_3$$

将 F_1、F_2、F_3 代入整理得:

$$\begin{aligned}\hat{y} = {} & 0.062x_1 + 0.113x_2 + 0.245x_3 + 0.235x_4 + 0.073x_5 - \\ & 0.031x_6 + 0.266x_8 - 0.126x_9 + 0.058x_{10} - 0.221x_{11}\end{aligned}$$

4 模型的异常检测

从模型的各统计量观察,发现了几个异常点,表 1 给出了各异常统计量的检测值。

表 1 模型异常残差分析表

序号	地区	标准化残差	剔除学生化残差	库克距离	杠杆值	DFFITS
3	河北省	-2.2124	-2.9292	0.5933	0.2298	-0.2687
5	内蒙古自治区	1.6088	1.7666	0.08277	0.0696	0.0628
9	上海市	0.5117	0.6929	0.10672	0.4323	0.1515
15	山东省	0.8107	1.0336	0.16539	0.3497	0.1711
19	广东省	-0.5807	-0.7084	0.06616	0.3076	-0.1021
24	贵州省	2.3805	2.7718	0.12006	0.0395	0.0636
	临界值	2,-2	+1.7081	0.1428	0.2000	0.7302

由表 1 可以看出河北省的剔除学生化残差,库克距离都高于临界值,且 DFFITS 统计量远高于其它值,比较异常。内蒙古自治区,贵州省,只有学生化残差值偏高,应具体研究。上海市和广东省只有杠杆值偏高,是高杠杆点。山东省的杠杆值,库克距离都偏高,为强影响点。

各统计值的分布结果如图 1 ~ 图 4 所示,各图中把超过临界值的样本作了标记,这样可以清楚地看到异常情况。

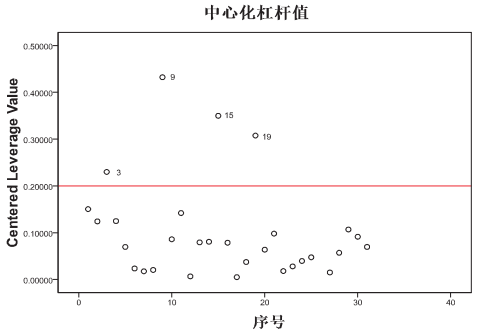

图 1 中心化的杠杆值分布图

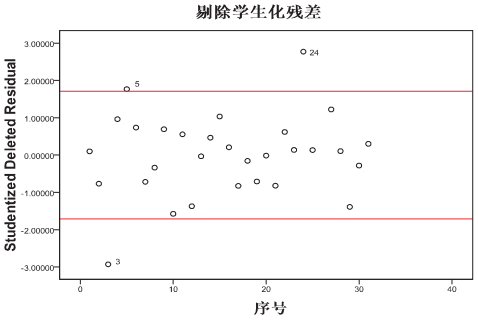

图 2 剔除学生化残差分布

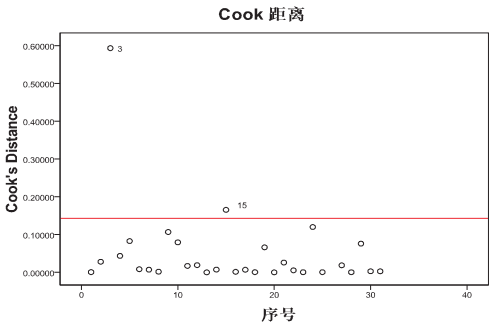

图 3　Cook 距离分布

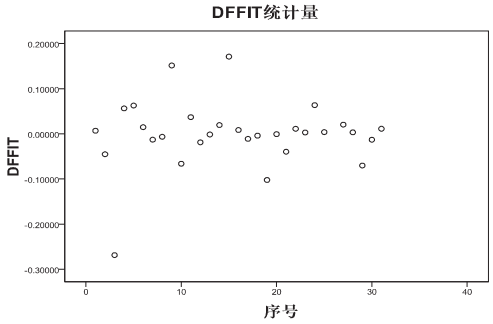

图 4　DFFIT 统计量

5　模型的异常检测的结果分析

从实际情况分析河北省的 SO_2 排放量居于全国第二。河北省包含京津唐工业基地的两个中心城市唐山和秦皇岛,所以它的煤炭、焦炭和柴油的耗费都高于其它地区,同时建成区绿化覆盖率、生活垃圾无害化处理能力都很好,但居民生活水平、森林覆盖率偏低。因此,河北省是因为居民生活水平和森林覆盖率偏低引起的异常。

内蒙古自治区的 SO_2 的排放量居于第三,它的煤炭、焦炭和柴油的用量位于前五位,生活无害化处理能力和居民名生活水平偏高,工业增加值和年末常住人口排名十八位。因此,内蒙古自治区是由于工业增加值和年末常住人口引起的异常。

山东省的 SO_2 排放居于全国第一,其它指标也处在第四名左右,所以是强影响点。

上海市和广东省都是高杠杆点,但引起的原因却不同。上海市的 SO_2 排放量位于倒数第五位,它的居民的生活水平排第一位,燃料油的使用排在第二位,年末常住人口位于最后几位,其它指标差不多在十四名左右,因此,上海市是居民生活水平、燃料油的消费量、年末常住人口引起的异常。

广东省的 SO_2 排放量位于第十位,但它的工业增加值、年末常住人口和民用汽车拥有量都位于第一位,且远高于其它样本,柴油使用量、建成区绿化率和森林覆盖率也名列前茅,其它指标差不多处于中等水平,因此,广东省是工业增加值、年末常住人口和民用汽车拥有量引起的异常。

贵州省的 SO_2 排放量居于第八位,污染主要来自以煤炭和柴油燃油的燃烧,除了煤炭消费量较高外其它指标都位于后几位,因此,贵州省是由于煤炭消费量引起的异常。

6　模型重建

将河北省、上海市和广东省这些异常点剔除,观察模型的变化情况。

剔除这些观测值后建立回归模型,得到相关系数 $R = 0.959$,判定系数 $R^2 = 0.919$,调整的判定系数 $\bar{R}^2 = 0.909$,回归估计的标准误差 $S = 0.2984$。与前面模型相比,R,R^2 修正值都略有提高,残差有所减少。说明模型拟合情况较好,并且提高了模型的拟合优度。得到的回归方程为:

$$\hat{y} = 0.063x_1 + 0.137x_2 + 0.277x_3 + 0.267x_4 + 0.077x_5 - 0.059x_6 + 0.306x_8 - 0.115x_9 + 0.058x_{10} - 0.255x_{11}$$

从系数分析,x_3,x_4,x_8,x_{11} 的系数变化相对较大,其余变量的系数也有略微变化。说明剔除三个观测量对模型的影响还是很大。

如果在数据中剔除其它异常值,模型的拟合优度没有上述情况好,所以没有一一列举。

7　对策及建议

在经济较发达的地域,其工业生产总值较高,SO_2 排放并没有随之增大。这说明工业生产中重视脱硫技术发挥了作用。这对其它地区遏制 SO_2 的排放起到了一定的借鉴作用。但这些地域的人口较密集、汽车拥有量较多,这是解决 SO_2 污染面对的新问题。

在经济欠发达的地域,工业增加值较低、能源耗费较高、绿化程度较低。因此,首先应加大环境的美化程度,增大绿化率,使环境的自净能力增强。其次应因地制宜减少能源消耗,根据具体情况确定脱硫技术,并且应用新型方法实现脱硫的高效化、综合化和经济化。最后,转变经济增长方式,大力促进服务业发展,走新型工业化道路。

参 考 文 献:

[1] 黄 青,程水源,陈东升,等.北京市一次 SO_2 污染过程来源分析[J].环境科学与技术,2010,33(1):89-93.

[2] 南 方.区域二氧化硫排放总量控制指标确定方法研究[D].北京:中国地质大学,2009.

[3] 张昭利.中国二氧化硫污染的经济分析[D].上海:上海交通大学,2012.

[4] 张英杰,冷伏海.基于离群点的前沿趋势探测方法研究高技术通讯[J].2011,21(11):1109-1114.

[5] Yang Ling, Wang Yuhrau, Pai Suzanne. Statistical and economic analyses of all EWMA-based snthesised control scheme for monitoring processes withoutliers[J]. International Journal of Systems Science, 2012, 43(2): 285-295.

[6] Su Weixing, Zhu Yunlong, Liu Fang. On-line outlier and change point detection for time series[J]. J. Cent. South Univ, 2013, 20:114-122.

[7] 王怀亮.回归诊断在统计数据异常值检测中的应用[J].黑龙江对外经贸,2011,200(2):118-119.

[8] 何晓群.现代统计分析方法与应用[M].北京:中国人民大学出版社,2007.

[9] 王寅琮.回归分析中异常值与共线性的诊断[D].秦皇岛:燕山大学,2012.

[10] 魏凤荣.回归分析中离群点的作用与搜寻[J].中央民族大学学报:自然科学版,2000,9(2):108-111.

[11] 何晓群,刘文卿.应用回归分析[M].北京:中国人民大学出版社,2007.

[12] 张洪阳.重庆市人口普查数据分析与模型诊断[D].重庆:重庆大学,2003.

[13] 宋志刚,谢蕾蕾,何旭洪.spss16.0 实用教程[M].北京:人民邮电出版社,2008.

[14] 冯 力,刘文卿.回归分析方法原理及 spss 实际操作[M].北京:中国金融出版社,2004.

Study on the Specificity of Provincial SO_2 Emissions in China

LI Xiaolin, *HUI Jie*, *LIU Qian*

(College of Management Sciences, Chengdu University of Technology, Chengdu 610059, China)

Abstract: According to the data of *China Statistical Yearbook*, the multiple regression model is constructed by using SPSS software which is based on the principal component analysis. Then, through the leverage values, eliminate studentized residuals, *Cook's* distance and *DFFITS* statistics generated by the model, the specific data of SO_2 emissions of each province in China is mined. In addition, the cause for the production of specific value is discussed.

Key words: SO_2 emissions; principal component analysis; multiple linear regression; outlier