回归诊断在城市空气质量预报中的应用研究

郑选军　王国强

（浙江省绍兴市气象局，312000）

提　要

城市空气质量回归预报模型的残差分布存在着不对称现象，它是由高杠杆点引起。这些高杠杆试验点的残差存在着统计天气预报意义上的不合理性，导致回归系数 Ls 估计的误差，从而引起预报的误差。针对这些问题提出了城市空气质量的回归诊断预报模型。实例计算说明，回归诊断预报模型要优于常规回归预报模型。进一步分析指出，城市空气质量回归预报模型的不合理性并非个别例子的特殊性所造成，而是由模型的数学特点所决定，因此城市空气质量的回归诊断预报模型具有普遍意义。

关键词： 空气质量预报模型　回归诊断　残差分布　高杠杆点

引　言

城市空气质量与一定范围内污染源的分布和排放有关，与大气运动对空气中污染物的稀释、扩散、清除和聚集的强度有关。前者可用当地环境监测站的实测空气质量记录来反映，并认为污染源在短期内具有相对稳定性，而预报主要从天气过程与污染物关系出发进行研究[1,2]。在空气质量预报的统计学[3,4]和动力统计方法[5,6]中广泛地应用多元回归预报模型，因为多元回归模型具有优良的统计学性质。但回归预报模型的优良性质是有前提的，一般有高斯-马可（Gauss-Markov)假设[7]：

$$\mathrm{Cov}(y) = \sigma^2 I_n \qquad (1)$$

公式(1)左边是应变量 y 的协方差阵，σ 为未知参数，I_n 为 n 阶单位矩阵，公式(1)的含义是各试验点的应变量互不相关且有等方差。还有假设

$$\mathrm{E}(e) = 0 \qquad (2)$$

即回归预报模型的随机误差 e 的均值为零。如果假设基本成立，可以认为回归预报模型的一些统计学性质成立。反之如果假设不成立，则这些性质也不能成立，回归预报模型的预报误差就可能因此增加。对于具体的城市空气质量预报模型，先要测定式(1)和式(2)是否成立，若不成立，再找出具体的原因，并设法予以解决，这就是回归诊断[8]的思路。至于如何分析原因和如何寻找解决办法，回归诊断理论没能提供确定的方法，而要根据城市空气质量预报的内容和专业特点进行具体分析和设计。

绍兴市城市空气质量预报系统是投入业务应用的一项科研成果，目的是预报绍兴市区第二天的二氧化硫(SO_2)、二氧化氮(NO_2)和可吸入颗粒物的级别。因子经过天气学和统计学方面的加工处理[9]。本文以此预报模型中可吸入颗粒物级别预报为例提出回归诊断在城市空气质量预报中的应用方法，目的是减少回归系数估计的误差，提高模型预报的精度。预报方程由 8 因子组成，样本容量为 92。因子的具体内容为：x_1 为 MM5 输出的 12 小时雨量等级；x_2 为前一时次可吸入颗粒物指数实测值；x_3 为 T213 的 850hPa 温度露点差；x_4 为 T213 的 700hPa 垂直速度；x_5 为 T213 的 850hPa 温度梯度；x_6 为 T213 的海平面气压梯度；x_7 为 T213 的 1000hPa

风速风向组合指数；x_8 为58453站24小时气温变幅。

1 城市空气质量预报模型的残差分析

在第 i 次预报中，预报量为 y_i，预报量估计为 $\hat{y}_i$，则普通残差为 $\delta_i = y_i - \hat{y}_i$。为了分析不同单位和数量级别的预报量 y_i 残差分布，可以使用学生化残差 r_i，

$$r_i = \delta_i / \sigma(1 - h_i)^{1/2} \tag{3}$$

式中 σ 为预报量 y_i 的标准差，h_i 为一定意义下的距离（见第2节）。图1是城市空气质量预报模型的学生化残差分布示意图。横坐标为预报量估计 $\hat{y}_i$，纵坐标为学生化残差 r_i。图中由 $r_i = 0$，$\hat{y}_i = y_{\max}$，$\hat{y}_i = y_{\min}$（$y_{\max}$ 和 $y_{\min}$ 为样本中 y_i 的最大值和最小值）三条直线分割成6个区域，其中D1、D2、D3和D4是残差分布范围，可见残差出现不对称分布。城市空气质量预报模型的残差分布不对称现象并非偶然，并不是由具体的预报个例造成，而是模型的数学特点所决定：当 $\hat{y}_i < y_{\min}$ 时不等式 $\delta_i = y_i - \hat{y}_i > 0$ 成立，由于 r_i 与 δ_i 同号，故 $r_i > 0$ 成立。也就是说，此时只有 $r_i > 0$ 的情况存在，而没有 $r_i < 0$ 的情况存在。这是图1 D1区域中的残差。同样D4中的情况可以类推。这是造成残差分布不对称的两种情况。D1和D4中试验点的条件可写为：

$$[(y_i = y_{\max}) \wedge (\hat{y}_i > y_{\max})] \vee [(y_i = y_{\min}) \wedge (\hat{y}_i < y_{\min})] \tag{4}$$

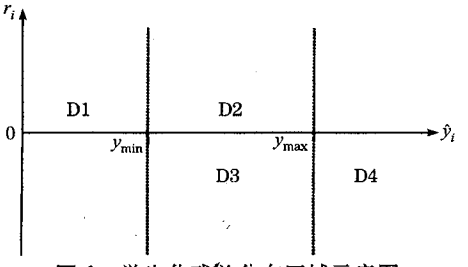

图1 学生化残差分布区域示意图

图1所示的不对称现象表示了城市空气质量预报的回归残差分布的方差非齐性（Heterogeneity of variances）[10]，表示了回归模型的Gauss-Markov假设已难以成立，因而可以推断城市空气质量预报模型存在着不合理性。

2 城市空气质量预报模型的回归诊断

为了进一步分析导致模型不合理性的原因，本文计算了城市空气质量预报中各因子对于预报量的诊断统计量，表1仅列出第1个因子的情况。表中第5～9列分别为普通残差、学生化残差、距离统计量、距离统计量的检验统计量和影响函数，它们的详细解释见文献[8]。其中

$$h_i = 1/n + (x_i - \bar{x})^{\tau}(X^{*\tau}X^{*})^{-1}(x_i - \bar{x}) \tag{5}$$

式中 n 为样本容量，$X = \begin{pmatrix} x_1^{\tau} \\ x_2^{\tau} \\ \vdots \\ x_n^{\tau} \end{pmatrix}$，$X^{*} = \begin{pmatrix} (x_1 - \bar{x})^{\tau} \\ (x_2 - \bar{x})^{\tau} \\ \vdots \\ (x_n - \bar{x})^{\tau} \end{pmatrix}$，$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$。式(5)第二项为马氏（Mahalanobis）距离，它的意义是在自变量空间中试验点 x 到试验中心的距离，因此 h_i 是描述距离的统计量[10]。另外

$$F_i = [(n - p - 1)/p] \times [(h_i - 1/n)/(1 - h_i)] \tag{6}$$

F_i 为 h_i 的检验统计量，p 为自变量维数，当 $F_i > F_{p,n-p-1}(\alpha_0)$，对应的试验点 x_i 可以判定为高杠杆点（High Leverage Case）。所谓高杠杆点就是远离试验中心 $\bar{x}$ 的试验点。本文取信度 $\alpha_0 = 0.05$，样本长度为92，则 $f_{1,90}(0.05) = 3.95$。此外

$$d_i = \frac{1}{p+1} r_i^2 \frac{h_i}{1 - h_i} \tag{7}$$

d_i 为强影响点（Strong Influence Case）的判别函数。所谓强影响点就是对回归系数有较大影响的试验点。另外学生化残差 r_i 的表达式见公式(3)，式中 h_i 即式(5)中的距离统计量 h_i。

表1 x_1 的回归诊断量

i	x_i	y_i	$\hat{y}_i$	δ_i	γ_i	h_i	f_i	d_i
1	−3	1	1.718	−0.718	−0.097	0.015	0.373	0.000072
2	−1	3	1.823	1.177	0.159	0.012	0.117	0.000155
3	2	3	1.980	1.020	0.137	0.011	0.004	0.000104
4	3	3	2.033	0.967	0.130	0.011	0.038	0.000097
5	−22	1	0.722	0.278	0.040	0.120	11.212	0.000108
6	4	1	2.085	−1.085	−0.146	0.012	0.109	0.000131
7	6	3	2.190	0.810	0.109	0.015	0.359	0.000090
8	9	1	2.347	−1.347	−0.183	0.022	1.011	0.000373
9	26	3	3.238	−0.238	−0.034	0.129	12.218	0.000087
10	0	2	1.876	0.124	0.017	0.011	0.043	0.000002
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表2是预报因子 x_1 的有关统计。表中 x_1 有7个不对称残差点和8个高杠杆点，前者是图1不对称区域D1和D4中的点。进一步统计可知，8个因子合计判定高杠杆点61次，其中56次为不合理残差点，重合率为91.8%，同时61个高杠杆点概括了所有的不合理残差点。结果表明城市空气质量回归预报模型的残差分布不对称现象是由一些高杠杆点引起，而这些高杠杆点造成了回归系数 Ls 估计的误差，从而导致城市空气质量预报的误差。

表2 不对称残差点与高杠杆点的关系(x_1)

序号									比例	
不对称残差点	5	9	21	26	37	55	89		高杠杆点为不对称残差点比例	7/8
高杠杆点	5	9	21	24	26	37	55	89	不对称残差点为高杠杆点比例	7/7

3 残差不对称性的统计预报意义

城市空气质量回归预报中，应变量 y 的元素集合为可数点可集[11] $A = \{y_{\min}, y_{\min}+1, \cdots, y_{\max}\}$，应变量的估计 $\hat{y}$ 既是 y 的拟合值，又表示了预测可靠程度。例如当应变量估计分别为 $\hat{y}_1 = 1.4$ 和 $\hat{y}_2 = 1.1$，据最近距离法则，它们的预测结论都是空气质量为1级，但后者比前者的预测可靠性大。预测可靠性可用 $\hat{y}_i$ 与 A 集的两个元素的距离之差 $\Delta H = |H_1 - H_2|$ 表示，H_1 和 H_2 分别为 $\hat{y}_i$ 与相近两个元素的距离。ΔH 为可靠性指数，其定义域为[0,1]，其值越大表示预测可靠性越大。对于 $\hat{y}_1 = 1.4$ 和 $\hat{y}_2 = 1.1$，它们的可靠性指数分别为 $\Delta H_1 = 0.2$ 和 $\Delta H_2 = 0.8$，以 $\hat{y}_2$ 的可靠性较大。

表3列出3个试验点的情况。第1和3点预测成功，可靠性指数均达到最大值1.0。但它们的残差却不同，第3点的残差反而与预测失败的第2点残差(绝对值)相等，可见第3点的残差是不合理的。分布于图1非对称区域D1和D4的所有试验点均属这种残差不合理的情况[8]。

表3 D1集和D4集试验点的残差不合理性分析

NO.	y	$\hat{y}$	预报结论	评定	ΔH	δ
1	1	1.0	1	对	1.0	0.0
2	1	1.6	2	错	0.2	−0.6
3	1	0.4	1	对	1.0	0.6

4 城市空气质量回归诊断预报模型

城市空气质量回归预报模型的残差非对称分布表明，模型的前提——Gauss-Markov等假设不能成立；回归诊断表明，这种非对称分布由一些高杠杆点引起；统计预报意义分析表明，造成残差非对称分布的试验点又存在着残差的不合理性。那么从回归诊断的观点看，为了使一些假设能成立，为了减少回归系数 Ls 估计的误差，必须探讨试验点的正确性问题，或者探讨统计模型的修改问题。具体方法如下：

(1)分别计算各因子的一元回归方程 $\hat{y}_i$

$= c_0 + c_1 x_i$，并剔除符合式(4)条件的试验点。显然这些就是D1和D4集合中的试验点。高杠杆点中可能包括了反映天气异常的个例，这些点自然不宜剔除。而这里剔除的高杠杆点均属于“最易预报”的个例(如表3中NO.3)，这种剔除只是计算订正值的需要，在以后步骤中这些点同样参加计算。

(2)原样本长度为 n，剔除了 u 个试验点以后新样本长度为$(n-u)$，新样本的回归方程为：$\hat{y}_i = c_0(u) + c_1(u)x_i(u)$。式中 $c_0(u)$ 和 $c_1(u)$ 表示剔除 u 个试验点后的回归系数 Ls 估计。

(3)对原因子作订正

$$x'_i = \begin{cases} \max\left[\dfrac{y_{max}-c_0(u)}{c_1(u)}, \dfrac{y_{min}-c_0(u)}{c_1(u)}\right] & \text{当 } x_i > \max\left[\dfrac{y_{max}-c_0}{c_1}, \dfrac{y_{min}-c_0}{c_1}\right] \\ \min\left[\dfrac{y_{max}-c_0(u)}{c_1(u)}, \dfrac{y_{min}-c_0(u)}{c_1(u)}\right] & \text{当 } x_i < \min\left[\dfrac{y_{max}-c_0}{c_1}, \dfrac{y_{min}-c_0}{c_1}\right] \\ x_i & \text{当 } \min\left[\dfrac{y_{max}-c_0}{c_1}, \dfrac{y_{min}-c_0}{c_1}\right] \leqslant x_i \leqslant \max\left[\dfrac{y_{max}-c_0}{c_1}, \dfrac{y_{min}-c_0}{c_1}\right] \end{cases} \quad (8)$$

这里订正的试验点是D1和D4中所有的试验点，这些点的原预报量估计 $\hat{y}$ 均大于 y_{max} 或小于 y_{min}。不难看出，如果应用了订正公式，类似于表3中第3试验点的残差将大大减小，残差不合理性得到改善，进而减小回归系数 Ls 估计的误差。

(4)用订正后的新因子(样本长度仍为n)建立多元城市空气质量回归预报方程，方法与普通回归法一致。两种模型的基本情况见表4。

表4 回归诊断预报模型中自变量的映射

	原一元回归方程的回归系数		诊断后一元回归方程的回归系数		自变量订正	
	c_0	c_1	c_0	c_1		
x_1	2.094	−0.0847	2.145	−0.097	当 $x_1 > 12.9$, $x'_1 = 12.9$	当 $x_1 \leqslant -10.7$, $x'_1 = -10.7$
x_2	1.875	0.0524	1.817	0.103	当 $x_2 > 21.5$, $x'_2 = 21.5$	当 $x_2 \leqslant -16.7$, $x'_2 = -16.7$
x_3	1.590	0.0540	1.314	0.0973	当 $x_3 > 26.1$, $x'_3 = 26.1$	当 $x_3 \leqslant -10.9$, $x'_3 = -10.9$
x_4	1.945	0.029	1.962	0.059	当 $x_4 > 36.4$, $x'_4 = 36.4$	当 $x_4 \leqslant -32.6$, $x'_4 = -32.6$
x_5	1.950	0.0483	1.960	0.083	当 $x_5 > 21.7$, $x'_5 = 21.7$	当 $x_5 \leqslant -19.7$, $x'_5 = -19.7$
x_6	1.936	0.092	1.931	0.162	当 $x_6 > 11.6$, $x'_6 = 11.6$	当 $x_6 \leqslant -10.2$, $x'_6 = -10.2$
x_7	1.878	0.068	1.813	0.107	当 $x_7 > 16.5$, $x'_7 = 16.5$	当 $x_7 \leqslant -12.9$, $x'_7 = -12.9$
x_8	2.205	−0.0148	2.389	−0.022	当 $x_8 > 81.4$, $x'_8 = 81.4$	当 $x_8 \leqslant -53.7$, $x'_8 = -53.7$

5 两种空气质量回归预报模型的效果比较

从表5可见，城市空气质量预报各因子在回归诊断预报模型中的相关系数均大于在常规回归预报模型中的相关系数。从多元回归方程的复相关系数和残差平方和看，也以前者为好，前者比后者减少残差平方和14.5%。另外对2003年9月中下旬20天的独立样本作对比试验，回归诊断模型和常规回归模型的残差平方和分别为 $Q_1 = 11.33$ 和 $Q_2 = 13.37$，前者比后者减少15.3%。由于绍兴大部分日期的空气质量为优或良，因此对比试验又对轻度污染以上的事件进行技术评分，它们的成功界限指数分别为 $CSI_1 = 0.50$ 和 $CSI_2 = 0.44$。以上两项评分均以回归诊断模型为好。

令回归诊断预报模型的各因子与预报量的相关系数为 R'，常规回归模型的相关系数为 R，可以证明恒有不等式$|R'| \geqslant |R|$成立(证明从略)，至于差值$|R'-R|$的大小，则取决于样本试验点在样本空间中的分布。如果在图1所示不对称区域D1和D4的试验点越多，试验点与试验中心的距离越远，则差值$|R'-R|$越大，反之越小；如果这类试验点不出现，则差值$(R'-R)=0$。

表5 两种回归模型的效果比较

	回归诊断模型	常规回归模型	评定
x_1	0.457	0.417	提高
x_2	0.448	0.445	略提高
x_3	0.460	0.425	略提高
x_4	0.471	0.426	提高
x_5	0.505	0.464	提高
x_6	0.480	0.443	提高
x_7	0.401	0.372	提高
x_8	0.438	0.409	提高
复相关系数	0.609	0.514	提高
Q	50.260	58.765	提高

6 讨 论

当预报量为离散型随机变量时，预报量与因子的数学关系是阶梯函数，严格地说线性回归模型不能用于离散型预报量的回归课题。为方便起见，预报工作中近似地作了引用，自然导致回归系数估计的误差。本文提出的回归诊断预报模型从回归诊断的角度部分地解决了这一问题。

本文对的城市空气质量回归预报模型所存在的不合理性问题的分析，以及针对此问题提出的城市空气质量回归诊断预报模型，对于离散型应变量的预报问题带有普遍适用的意义。

参考文献

1 孙明华，徐大海，朱蓉等. 城市空气臭氧污染业务预报方案研究. 气象，2002，28(4)：3～8.

2 徐大海，朱蓉. 大气平流扩散的箱格预报模型与污染潜势指数预报. 应用气象学报，2000，11(1)：1～12.

3 周斌斌. 论雾与污染的关系. 气象，1994，22(9)：19～24.

4 王伟平，杨军，蔡菊珍. 空气污染气象条件预报的试验研究. 浙江气象科技，2001，22(3).

5 刘实，王宁，朱其文等. 长春市空气污染潜势预报的统计模型研究. 气象，2002，28(1)：8～12.

6 徐祥德，杨绪，徐大海等. 城市化环境气象学引论. 北京：气象出版社，2001：201～205.

7 王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990：195～216.

8 王松桂. 线性回归诊断. 数理统计与管理. 1985，4(6)，38～49，1986，5(1)：40～47.

9 王国强. 近邻估计-线性回归预报模型及其台风暴雨预报中的试验. 气象科技，1999，27(4)：25～29.

10 陈希孺，王松桂. 近代回归分析. 合肥：安徽教育出版社，1987：91～146.

11 左孝陵，李为，刘永才. 离散数学. 上海：上海科学文献出版社，1982：147～170.

An Application of Regression Diagnosis to City Air Quality Forecasting

Zheng Xuanjun Wang Guoqiang

(Shaoxing Meteorological Office, Zhejiang Province 312000)

Abstract

There exists asymmetry characteristic of residual distribution in the regression model of city air quality forecast. It is caused by some high leverage cases. The residues of these high leverage cases have no rationality in the sense of statistical weather forecasting, and the errors of Least Square Estimation (LS) of the regression coefficient are occurred. So the errors of city air quality forecast are occurred. Thus the regression diagnosis prediction model for city air quality forecasting is proposed. The mathematical proving and the calculation of the examples show that this new model is superior to the general regression prediction model. It is illustrated by means of further analysis that rationality of regression model of city air quality is not caused by some examples, but by mathematical characteristics of the model.

Key Words: air quality forecasting regression diagnosis residual distribution high leverage cases