# 回归诊断在事件概率回归预报中的应用

王国强

（浙江省绍兴市气象台，312000）

**提　要**

针对事件概率回归模型的特点，用残差分析和统计量诊断的方法归纳了回归残差的非对称分布现象，揭示了这种现象是由高杠杆点所引起，探讨了概率回归模型的残差不合理性的统计天气预报意义，从而提出适用于概率预报问题的事件概率回归改进模型。分析指出，事件概率回归模型的不合理性并不是个别例子的特殊性所造成，而是由该模型的数学特点所决定。概率回归改进模型要优于普通的概率回归模型。

**关键词：** 回归诊断　残差分布　高杠杆点事件概率

## 引言

在天气预报业务中，天气事件的概率预报经常应用线性回归模型——事件概率回归模型。线性回归模型加LS(最小二乘)方法具有优良的统计学性质，但是这是有前提的，一般有 Gauss-Markov 假设：

$$cov(y) = \sigma^2 I_n \qquad (1)$$

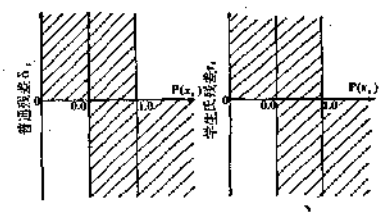
式中 $y$ 为应变量。其含义第一是各试验点的应变量互不相关，第二是它们有等方差。还有假设：

$$E(e) = 0 \qquad (2)$$

$e$ 为模型的随机误差。有时还需要另外一些假设[1]。而某些假设在概率回归模型中并不满足，因此线性回归模型加 LS 方法的一些性质在此不再成立。

## 1　事件概率回归模型的残差分布

为了研究概率回归模型的残差分布规律，现以绍兴市气象台冬季 120 小时降水概率预报为例进行分析。此概率回归方程为六元逐步回归方程，自变量是由数值天气预报产品的格点资料组合而成的天气学因子，均为连续型变量，其中部分自变量已作非线性处理，样本长度 $n=93$。图 1 是该预报方程的 2 张概率回归残差分布图，其中 $\hat{P}(X_i)=\alpha+\beta^T X_i$ 为天气事件的概率回归估计，$\delta_i$ 为普通残差，$r_i=\delta_i/\sigma\sqrt{1-h_i}$ 为学生氏残差；其中列向量 $X=(x_{1i},x_{2i},\cdots,x_{mi})^T$ 为自变量，列向量 $\beta=(\beta_1,\beta_2,\cdots,\beta_m)^T$ 为回归系数，$T$ 为转置号，$\sigma$ 为标准差，$m$ 为自变量维数，$h_i$ 为某种意义下的距离。附图中阴影区表示残差分布范围。由图可见该预报方程的普通残差和学生氏残差均出现非对称分布。




**附图**　事件概率回归的残差分布图

事件概率回归模型出现残差非对称分布的现象并非偶然，并非由具体预报例子的个别情况造成，而是模型的数学特点的反映：当 $\hat{p}(x_i)>1$ 时，由于恒有 $p_i\leqslant 1$，所以必然有不等式 $\delta_i=p_i-\hat{p}(x_i)<0$ 成立；当 $\hat{p}(x_i)<0$ 时，由于恒有 $p_i\geqslant 0$，也必然有不等式 $\delta_i=p_i-\hat{p}(x_i)>0$ 成立。残差的这些特点造成了如

图 1 的残差非对称分布现象。

显然概率回归的残差非对称分布表示了回归残差分布的方差非齐性[2](Heterogeneity of variances)，揭示了事件概率回归模型的前提——Gauss-Markov 等假设已经不能成立，由此可推断事件概率回归模型存在着不合理性。

## 2 概率回归模型的回归诊断

为了进一步分析导致概率回归模型不合理的具体原因，本文计算了例子中各自变量对于应变量的诊断统计量 $h_i$、$F_i$ 和 $D_i$(见表 1)。其中

$$h_i = \frac{1}{n} + (x_i - \bar{x})^{\mathrm{T}} (X^{*\mathrm{T}} X^*)^{-1} (x_i - \bar{x}) \tag{3}$$

表 1 高杠杆点的判定

| | $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | …… | 93 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | $h$ | 0.058 | 0.041 | 0.038 | 0.027 | 0.019 | 0.014 | 0.012 | 0.021 | 0.012 | 0.021 | …… | 0.021 |
| | $F$ | 4.558 | 2.828 | 2.553 | 1.502 | 0.808 | 0.316 | 0.112 | 0.959 | 0.094 | 0.961 | …… | 0.130 |
| | $D$ | 0.028 | 0.028 | 0.028 | 0.025 | 0.022 | 0.020 | 0.004 | 0.000 | 0.001 | 0.013 | …… | 0.027 |
| | $F>2.76$ | √ | √ | | | | | | | | | | |
| $x_2$ | $h$ | 0.012 | 0.023 | 0.024 | 0.018 | 0.057 | 0.082 | 0.021 | 0.011 | 0.011 | 0.012 | …… | 0.014 |
| | $F$ | 0.129 | 1.138 | 1.252 | 0.665 | 4.452 | 7.035 | 0.930 | 0.058 | 0.001 | 0.134 | …… | 0.284 |
| | $D$ | 0.022 | 0.022 | 0.022 | 0.022 | 0.013 | 0.004 | 0.021 | 0.002 | 0.003 | 0.001 | …… | 0.046 |
| | $F>2.76$ | | | | | √ | √ | | | | | | |
| $x_3$ | $h$ | 0.059 | 0.076 | 0.036 | 0.015 | 0.011 | 0.012 | 0.019 | 0.056 | 0.027 | 0.011 | …… | 0.012 |
| | $F$ | 4.687 | 6.470 | 2.415 | 0.373 | 0.007 | 0.099 | 0.802 | 4.391 | 1.559 | 0.022 | …… | 0.069 |
| | $D$ | 0.031 | 0.030 | 0.029 | 0.020 | 0.020 | 0.025 | 0.001 | 0.002 | 0.000 | 0.003 | …… | 0.018 |
| | $F>2.76$ | √ | √ | | | | | | √ | | | | |
| $x_4$ | $h$ | 0.012 | 0.019 | 0.016 | 0.019 | 0.015 | 0.011 | 0.015 | 0.029 | 0.020 | 0.022 | …… | 0.013 |
| | $F$ | 0.117 | 0.785 | 0.452 | 0.756 | 0.435 | 0.001 | 0.435 | 1.743 | 0.852 | 1.004 | …… | 0.242 |
| | $D$ | 0.019 | 0.022 | 0.021 | 0.022 | 0.041 | 0.021 | 0.001 | 0.000 | 0.000 | 0.000 | …… | 0.020 |
| | $F>2.76$ | | | | | | | | | | | | |
| $x_5$ | $h$ | 0.050 | 0.083 | 0.029 | 0.011 | 0.017 | 0.038 | 0.013 | 0.011 | 0.038 | 0.011 | …… | 0.015 |
| | $F$ | 3.762 | 7.142 | 1.696 | 0.005 | 0.542 | 2.590 | 0.178 | 0.036 | 2.592 | 0.040 | …… | 0.413 |
| | $D$ | 0.029 | 0.025 | 0.026 | 0.021 | 0.013 | 0.028 | 0.004 | 0.002 | 0.038 | 0.003 | …… | 0.038 |
| | $F>2.76$ | √ | √ | | | | | | | | | | |
| $x_6$ | $h$ | 0.018 | 0.019 | 0.021 | 0.065 | 0.062 | 0.033 | 0.016 | 0.012 | 0.074 | 0.015 | …… | 0.012 |
| | $F$ | 0.652 | 0.749 | 0.925 | 5.242 | 4.940 | 2.049 | 0.527 | 0.118 | 6.234 | 0.377 | …… | 0.158 |
| | $D$ | 0.052 | 0.022 | 0.023 | 0.022 | 0.023 | 0.025 | 0.009 | 0.004 | 0.160 | 0.001 | …… | 0.030 |
| | $F>2.76$ | | | | √ | √ | | | | √ | | | |

式中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$，$X = \begin{bmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{bmatrix}$，$X^* = \begin{bmatrix} (x_1 - \bar{x})^{\mathrm{T}} \\ (x_2 - \bar{x})^{\mathrm{T}} \\ \vdots \\ (x_n - \bar{x})^{\mathrm{T}} \end{bmatrix}$。等式右边第二项为马氏(Mahalanobis)距离，$h_i$ 在几何上是试验点 $x_i$ 在自变量空间中离试验中心 $\bar{x}$ 的远近的量度。另外

$$F_i = \frac{n-m-1}{m} \frac{h_i - \frac{1}{n}}{1 - h_i} \tag{4}$$

$F_i$ 为 $h_i$ 的检验统计量，当 $F_i > F_{m,n-m-1}(\alpha_0)$ 时，对应的试验点 $x_i$ 可认为是高杠杆点(High Leverage Lase)。此处取信度 $\alpha_0 = 0.05$，样本长度 $n=93$，$m=1$，故 $F_{m,n-m-1}(\alpha_0) = 2.76$。此外

$$D_i = \frac{1}{m+1} r^2 \frac{h_i}{1 - h_i} \tag{5}$$

$D_i$ 为影响函数，它是强影响点(Strong Influ-

ence Case)的判别函数。表1还列出各自变量的$F_i>2.76$的试验点，这些点被认为是高杠杆点。

令图1的非对称区域$d_1$和$d_2$中试验点集合为$D$。对于表1中各自变量，它们属于$D$集的试验点全部或几乎全部被判为高杠杆点，个别的试验点也接近于高杠杆点的判别标准。例如$x_1$有16个试验点属于$D$集，全部为高杠杆点。因此说概率回归模型的残差非对称分布由一些高杠杆点引起。虽然这些高杠杆点不全是强影响点，但是它们的联合[3]确实对回归系数的LS估计产生明显的影响。为了探求解决问题的途径，我们对$D$集的试验点再从统计天气预报的角度作分析。

## 3 残差不合理性的统计预报意义分析

天气事件的出现与不出现为互逆事件，它们构成了完备群[4]，因此条件概率估计$\hat{p}(x_i)$表示了预测结论是事件出现还是不出现，同时也表示这种预测的可靠程度。例如条件概率估计分别为$\hat{p}(x_A)=0.6$和$\hat{p}(x_B)=0.9$，据最近距离法则，它们的预测结论都是事件出现，但后者比前者的预测可靠性大。在样本资料中$P_i$集合的元素为$c=\{0,1\}$，预测可靠性可用$\hat{p}(x_i)$与$c$集的两个元素的距离之差$\Delta H=|H_1-H_0|$表示，$H_1$和$H_0$分别为$\hat{p}(x_i)$与元素1和元素0的距离。$\Delta H$称可靠性指数，其定义域为$[0,1]$，其值越大，表示预测可靠性越大。对于$\hat{p}(x_A)$和$\hat{p}(x_B)$，它们的可靠性指数分别为$\Delta H_A=0.2$和$\Delta H_B=0.8$，以$\hat{p}(x_B)$的可靠性较大。

表2 残差不合理性举例

| 序号 | $P_i$ | $\hat{p}(x_i)$ | 结论 | 评定 | $\Delta H_i$ | $\delta_i$* |
|---|---|---|---|---|---|---|
| 1 | 1 | 1.0 | 1 | 对 | 1.0 | 0.0 |
| 2 | 1 | 0.4 | 0 | 错 | 0.2 | 0.6 |
| 3 | 1 | 1.6 | 1 | 对 | 1.0 | −0.6 |

* $\delta_i=P_i-\hat{p}(x_i)$

表2列出了本例中3个试验点情况。第1和第3点预测成功，可靠性指数均达到最大值1.0。但是它们的残差却不同，第3点的残差反而与预测失败的第2点残差相等(指绝对值)，可见第3点的残差是不合理的。在图1中分布于非对称区域$d_1$和$d_2$的所有试验点，均属于这种残差不合理的情况。

## 4 事件概率回归改进模型

事件概率回归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，必须或者探讨试验点的增删问题，或者探讨统计模型的修改问题。对此回归诊断理论不能提供确定的方法，而需要根据天气预报中的内容和特点，作具体的分析和设计。前面从统计预报意义对残差不合理性的讨论，实际上已提示了处理的着眼点，具体方法如下：

### 4.1 分别计算各自变量的一元回归方程$\hat{p}(x_i)=a+bx_i$，并剔除符合条件

$$\{[(P_i=1)\wedge(\hat{p}(x_i)>1)]\vee[(p_i=0)\wedge(\hat{p}(x_i)<0)]\} \quad (6)$$

的试验点。显然这些就是$D$集中的试验点，如对于$x_1$，是表1中$F>2.76$的16个试验点。

高杠杆点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“最易预测”的个例，它们的预测可靠性指数均达到最大值1.0，不管用普通的回归模型还是用本文的改进模型，它们均被准确地“预测”。这种剔除处理只是为了计算订正值的需要，在以后步骤中这些剔除点同样参加方程的计算。

### 4.2 原样本长度为$n$，剔除了$u$个试验点以后新样本长度为$(n-u)$，新样本的概率回归方程为$\hat{p}_i(u)=a(u)+b(u)x_i(u)$。$a(u)$和$b(u)$表示剔除$u$个试验点后的回归系数LS估计。

4.3 订正原自变量

$$x'_i=\begin{cases}\max\left[\dfrac{1-a(u)}{b(u)},\dfrac{-a(u)}{b(u)}\right] & \text{当 } x_i>\max\left[\dfrac{1-a}{b},\dfrac{-a}{b}\right]\\ \min\left[\dfrac{1-a(u)}{b(u)},\dfrac{-a(u)}{b(u)}\right] & \text{当 } x_i<\min\left[\dfrac{1-a}{b},\dfrac{-a}{b}\right]\\ x_i & \text{当 } \min\left[\dfrac{1-a}{b},\dfrac{-a}{b}\right]\leqslant x_i\leqslant\max\left[\dfrac{1-a}{b},\dfrac{-a}{b}\right]\end{cases}\tag{7}$$

这里订正的试验点是 $D$ 集中所有的试验点，这些点的原概率回归估计 $\hat{p}(x_i)=a+bx_i$ 均大于 1 或小于 0，订正的目的是通过控制它们的估计值，而控制对应的残差。不难看出，如果应用了订正公式，类似于表 2 中第 3 试验点的残差将大大减小，残差的不合理性得到改善，进而减小回归系数 LS 估计的误差。

4.4 用订正后的新自变量 $x'_i$（样本长度仍为 $n$）建立多元的概率回归方程，方法与普通的回归方法一致。

4.5 用样本资料作拟合和用独立样本资料作预测时，资料需要作同样的订正。

## 5 两种概率回归模型的效果比较

可以证明概率回归改进模型中各自变量的相关系数 $R'$ 大于等于普通的概率回归模型中各自变量的相关系数 $R$，即 $R'-R\geqslant 0$（证明从略）。至于差值$(R'-R)$的大小，则取决于样本试验点在$(m+1)$维（自变量为 $m$ 维，应变量为一维）的样本空间中的分布。如果分布在图 1 不对称区域 $d_1$ 和 $d_2$ 中的试验点越多，则差值$(R'-R)$越大，反之差值越小。现在仍用前面的例子对两种模型的效果作比较。

首先利用绍兴市客观预报系统自动组合因子，自动初选因子，再用逐步回归方法筛选因子建立六元的 120 小时降水概率预报回归方程。接着对此 6 个预报因子，用概率回归改进模型建立新的降水概率预报方程。两个方程对应的统计量列于表 3。它们所用的样本资料完全相同，但是回归系数的 LS 估计不同，预报方程的质量也不同。从表 3 的几个有关统计量看，概率回归改进模型均优于普通的概率回归模型。当然采用改进模型的效果在具体预报实例中各不相同，它取决于样本在空间的分布，具体地说它取决于在事件概率回归模型中样本回归残差的不合理程度。

表 3 两种概率回归模型的效果比较

| 统计量 | 概率回归模型 | 概率回归改进模型 |
|---|---|---|
| 复相关系数 | 0.580 | 0.666 |
| 均方差 | 0.350 | 0.320 |
| 相关概率 | 0.848 | 0.893 |

## 6 几点讨论

6.1 在绍兴市气象台的预报业务中，从初选因子开始到逐步回归建立预报方程，都采用概率回归改进模型中的订正处理，因此效果比第 5 节所述的情况要更好一些。

6.2 令普通一元概率回归方程系数的 LS 估计为 $a$（截距）和 $b$（斜率），一元概率回归改进方程系数的 LS 估计为 $a'$ 和 $b'$，如果 $b$ 与 $b'$ 的正负符号不一致，说明该自变量与应变量的相关关系不够稳定，这种自变量应予剔除。在具体的预报业务中，通过初选的自变量的质量都比较好，一般不易出现这种现象。

6.3 本文分析的事件概率回归模型不合理性，并不是由个别预报例子的特殊性造成，而是由该模型的数学特点所决定，因此本文讨论的问题和提出的事件概率回归改进模型具有普遍的意义。不管样本的具体情况如何，该模型总是优于或等于普通的事件概率回归模型，因此在预报业务中如果用前者代替后者将是可行的。

## 参考文献

1 王学仁，王松桂. 实用多元统计分析. 上海技术出版社，1990.

2 王松桂. 线性回归诊断. 数理统计与管理，6(1985)，1(1986).

（下转 49 页）

（上接第 46 页）

3 陈希孺等. 近代回归分析. 安徽教育出版社，1987.
4 左孝陵等. 离散数学. 上海科学技术文献出版社，1982.

# The Application of Regression Diagnosis Method to Weather Event Probability Regression Prediction

Wang Guoqiang
(Shaoxing Meteorological Observatory, Zhejiang Province 312000)

**Abstract**

The regression diagnosis for the event probability regression model was carried out. It is showed that probability regression model have no rationality in the sense of statistical forecasting by the distribution of the probability residua and diagnosis amounts. Thus an advancement of event probability regression model is proposed and the results show that it is superior to general regression model by mathematical proving and calculation of the examples.

**Key Words**: regression diagnosis event probability distribution of residua advancement of regression model