第50卷 第2期　　气 象 学 报　　Vol. 50, No. 2
1992年5月　　ACTA METEOROLOGICA SINICA　　May, 1992

离散型应变量的回归诊断和离散回归法*

王国强　陈晓平　诸晓明

（浙江省绍兴市气象台）

本文从大量天气预报实例的普查结果，归纳了离散型应变量的回归残差分布模式。回归诊断表明模式的单向性是由一些高杠杆点引起。对这些高杠杆点的进一步分析可以看到，它们的残差存在着统计预报意义上的不合理性，不合理性又造成回归系数*LS*估计的误差。针对这些问题，本文提出了离散回归法。

1. 离散型应变量的回归残差分布模式

线性回归模型具有优良的统计性能，被广泛地应用于天气预报业务。这种模型不但常用于连续型变量的预测问题，而且也常用于离散型变量的预测问题。然而对于离散型变量的预测，线性回归方法存在着一些不合理性，以致影响了回归方程的质量。

在线性回归中，设自变量为 X_i，应变量为 Y_i，应变量的拟合值 $\hat{Y}_i=\alpha+\beta^\tau X_i$，普通残差 $\delta_i=Y_i-(\alpha+\beta^\tau X_i)$。式中列向量 $\beta=(\beta_1,\beta_2,\beta_3\cdots\beta_p)^\tau$，列向量 $X_i=(x_{1i},x_{2i},x_{3i}\cdots x_{pi})^\tau$，$p$ 为自变量向量的维数，τ 为转置符号。普查预报业务中大量离散型应变量的逐步回归方程发现，它们的残差分布存在着一种特殊现象，这种现象可归纳成如图1的残差分布模式。图中的阴影部分表示残差的分布区域。这种模式可表述如下：应变量 Y 是定义在区间[M，N]的离散型随机变量，当回归拟合值 $\hat{Y}_i>N$ 时，有残差 $\delta_i<0$ 而不出现 $\delta_i\geqslant 0$；当 $\hat{Y}_i<M$ 时，有残差 $\delta_i>0$ 而不出现 $\delta_i\leqslant 0$。残差分布呈现单向性。

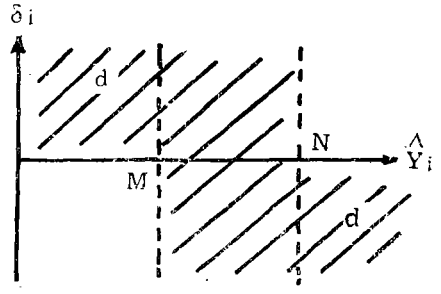

图 1　离散型变量的回归残差分布模式

普查说明这种现象是连续型应变量的回归方程所不具备的，而是线性回归模型用于离散型应变量的特殊结果。显然这种单向性表示了回归残差分布的方差非齐性[1]，表示了与线性回归模型的 Gauss-Markov 假设不符合。由此可见线性回归模型在用于离散型随机变量的预测问题时存在着不合理性。

2. 残差分布单向性的回归诊断

为了研究造成这种残差分布的单向性现象的原因，现以绍兴市短期客观预报系统中梅雨期逐日0—24小时雨量量级预报为例进行分析。该预报模型为一个五元逐步回归方程，方程中的自变量均是由天气形势图(含数值预告图)中格点资料组合而成的天气学因子。本文计算了例子中所有个例的一些回归诊断统计量，表1列出了部分诊断统计量 h 和 F[2]的计算值。其中

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

* 本文于1990年3月2日收到，1990年8月28日收到修改稿。

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

表 1 一元回归的诊断统计量（部份数据）

i	x_1		x_2		x_3		x_4		x_5	
	h	F	h	F	h	F	h	F	h	F
1	0.016	2×10^{-3}	0.023	0.422	0.017	0.051	0.024	0.513	0.018	0.147
2	0.057	2.694	0.102	5.836	0.018	0.118	0.033	1.053	0.078	4.114
3	0.057	2.694	0.063	3.097	0.018	0.118	0.019	0.168	0.030	0.887
4	0.030	0.910	0.016	3×10^{-3}	0.018	0.118	0.033	1.052	0.033	1.087
5	0.019	0.206	0.020	0.287	0.026	0.636	0.016	0.011	0.047	2.024
6	0.028	0.759	0.023	0.422	0.026	0.636	0.026	0.655	0.016	0.027
7	0.057	2.694	0.016	3×10^{-3}	0.016	3×10^{-3}	0.024	0.513	0.019	0.177
8	0.017	0.083	0.020	0.287	0.029	0.841	0.033	1.052	0.067	3.342
9	0.022	0.384	0.016	3×10^{-3}	0.017	0.051	0.048	2.062	0.027	0.714
10	0.022	0.384	0.020	0.287	0.029	0.841	0.048	2.062	0.016	8×10^{-3}
⋮		⋮				⋮				⋮

$$F_i=\frac{n-p-1}{p}\cdot\frac{h_i-\frac{1}{n}}{1-h_i}$$

F_i 为 h_i 值的检验统计量，当 $F_i>F_{p,n-p-1}(\alpha_0)$ 时，对应的试验点 (x_i,y_i) 可判定为高杠杆点 (High Leverage Case)。取 F 检验的信度 $\alpha_0=0.10$，则 $F_{1-61}(0.10)=279$。表 2 列出了各自变量一元回归的高杠杆点，凡被各自变量判定为高杠杆点的所有试验点编号均归纳在表中末行。

表 2 高 杠 杆 点 的 判 定

		试 验 点 序 号															
d,d' 中的试验点		2	3	8	22		27	34	35		39	40	41				60	61
$F>F(\alpha_0)$	x_1																	
	x_2	2	3		22	23		34					41					
	x_3								35	36	39			42				
	x_4						27										60	61
	x_5	2		8				34							49	51		
综 合		2	3	8	22	23	27	34	35	36	39		41	42	49	51	60	61

本例子样本长度 $n=63$，样本中属于 d 和 d'(见图 1)的共有 12 个例，几乎全部$\left(\frac{11}{12}\right)$为高杠杆点，剩下 1 个例的 F 值也十分接近高杠杆点的标准。可见残差分布的单向性是由一些高杠杆点所引起的。虽然这些高杠杆点不全是强影响点，但是它们的联合[2]确实对回归系数的 LS(最小二乘)估计产生很大的影响，后面第五节的计算说明了这个问题。为了探求解决这一问题的途径，下面对这些点再从天气

预报的角度进行分析。

3．残差分布单向性的统计预报意义分析

在前面例子即绍兴市梅雨期逐日0—24小时雨量量级预报中，Y_i的集合为可数点集[8]$A=\{1, 2, 3\}$，$\hat{Y}_i=\alpha+\beta^T X_i$既是$Y_i$的拟合值，同时$\hat{Y}_i$与点集$A$中元素的距离也表示了该试验点预测的可靠性大小。例如对于$\hat{Y}_A=2.9$和$\hat{Y}_B=2.6$，根据最近距离法则，预测结论都应是3，但是它们的预测可靠性不一样。这种可靠性可以用Y_i与点集$A=\{1,2,3\}$中二个最近元素的距离差$\Delta H=H_1-H_0$表示，H_0和H_1分别为最近和次近距离，ΔH为可靠性指数，可靠性指数的取值范围为[0,1]。如对于$\hat{Y}_A=2.9$可靠指数$\Delta H_A=0.8$，而对于$\hat{Y}_B=2.6$，则$\Delta H_B=0.2$。据此可见虽然它们的预测结论均为3，但$\hat{Y}_A$较$\hat{Y}_B$的预测可靠性要大。

表3　不合理点的残差

i	y_i	$\hat{y}_i$	预报结论	正确与否	ΔH_i	δ_i
1	3	3.0	3	对	1.0	0
2	3	2.4	2	错	0.2	0.6
3	3	3.6	3	对	1.0	−0.6

根据最近距离法则，现在比较表3的三种预测情况：第一次预测效果很好；第二次预测错误；第三次预测效果也很好，预报结论十分明确，但第三次预测的残差却与第一次的不同，而与第二次的相等（指绝对值）。可见第三次的残差是不合理的。图1中点集d与d'的所有个例都属于这类不合理情况。

4．离散回归法

既然离散型应变量的回归残差分布模式表明，残差分布单向性使线性回归模型的前提——Gauss-Markov假设不再成立，既然回归诊断表明，残差分布单向性是由一些高杠杆点所引起，既然造成单向性的个例在预报意义上又存在着残差的不合理性，那么从回归诊断的观点看，我们必须或者探讨对样本数据的处理(含剔除)问题，或者探讨对统计模型的修改问题，以便使Gauss-Markov假设在此能基本成立，以便较好地减少回归系数LS估计中的误差。而在残差分布单向性的天气预报意义讨论中，对于残差的不合理性分析实际上已提示了这种处理的着眼点。具体处理可按以下步骤进行：

1．分别计算各自变量对于应变量的一元回归方程$\hat{Y}_i=a+bx_i$，并剔除符合条件：

$\{[(Y_i=N)\wedge(\hat{Y}_i>N)]\vee[(Y_i=M)\wedge(\hat{Y}_i<M)]\}$的个例，$Y_i$的定义域为$[M,N]$。

2．如果原样本长度为n，剔除的个例数为u，那么新样本的长度为$(n-u)$。新样本的一元回归方程$\hat{Y}_i(u)=a(u)+b(u)x_i(u)$，$a(u)$与$b(u)$表示剔除$u$个个例后计算的回归系数$LS$估计。

3．订正原自变量资料(包括剔除的个例)：

$$x_i'\begin{cases}=\dfrac{N-a(u)}{b(u)} & \text{当 } \hat{Y}_i>N\\ =\dfrac{M-a(u)}{b(u)} & \text{当 } \hat{Y}_i<M\\ =x & \text{当 } M\leqslant\hat{Y}_i\leqslant N\end{cases}$$

前面第二节已分析到，图1表示的残差分布的单向性是由一些高杠杆点所引起，而上式订正的$\hat{Y}_i>N$与$\hat{Y}_i<M$的个例就是图1中d与d'的个例。订正的目的是使这些个例的拟合值得到控制，从而对应的残差也得到控制。这里不难看出，如果应用了订正公式，表3中第三次预测的残差将大大减小，残差的不合理现象将得到改善。也可作如下订正：

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

这两种变换是等价的，可根据资料处理的情况而采用其中之一，机器处理一般以后者为方便。

4．用订正后的新自变量资料 X'（样本长度仍为 n）与应变量资料 Y 作相关计算，建立逐步回归方程，方法与常规的线性回归一致。

5．在用样本资料作拟合或用独立样本资料作预测时，资料也需相应订正，方法如步骤 3 中的第二种方法。

5．两种回归法的效果比较

可以证明自变量 X 与应变量 Y 在离散回归法中的相关关系比在常规回归法中的关系要密切（证明从略）。现再用前面的预报实例对两种不同回归方法的效果进行比较。

先利用绍兴市客观短期预报系统建立一个常规五元逐步回归方程，接着以此方程中的五个自变量和应变量资料建立五元离散回归方程。两个对应方程的有关统计量列于表 4。由表可见虽然两种方法的原始资料相同，但是回归系数 LS 估计不同，从回归方程的复相关系数 R 看，以离散回归法为好。另外以最近距离法则用回归方程来预测雨量量级，常规回归法的预测相关概率通过信度为 0.05 的显著性检验，离散回归法则可通过 0.01 的检验(见表 4)，可见相关概率的提高是显著的。

表 4 两种回归法的效果比较

	常规回归法	离散回归法
回归系数	$a=1.209$ $\beta^T=(0.0312,0.190,-0.0674,-0.0523,0.0951)$	$a=1.094$ $\beta^T=(0.0220,0.322,-0.0695,-0.0442,0.106)$
剩余方差	0.312	0.262
复相关系数	0.793	0.832
相关概率	0.667	0.778
显著性检验	通过信度0.05	通过信度0.01

如果说剩余方差和复相关系数表示了回归方程的拟合精度，相关概率表示了回归方程对离散变量的分辨能力，从这两方面看离散回归法都优于常规回归法。

参 考 文 献

[1] 王松桂，线性回归诊断，数理统计与管理，**6**，38—49，1985，1，40—47，1986。
[2] 陈希孺等，近代回归分析，安徽教育出版社，1987。
[3] 左孝凌等，离散数学，上海科学技术文献出版社，1982。

THE REGRESSION DIAGNOSIS FOR THE DEPENDENT VARIABLE OF DISCRETE TYPE AND THE METHOD OF DISCRETE REGRESSION

Wang Guoqiang　Chen Xiaoping Zhu Xiaoming

(Shaoxing Meteorological Observatory, Zhejiang Province)

Abstract

In this paper, the distribution model of the regression residue of discrete dependent variables is obtained by general survey of a vast amount of the examples of the weather predictions. The regression diagnosis shows that unidirectional characteristic of the model is caused by some high leverage cases. In the sense of statistical forecasting, the residues of these high leverage cases have no rationality, and the errors of least square (LS) estimation of the regression coefficient are occured. Therefore, the method of discrete regression is proposed in this paper. The results show that the method of discrete regression is superior to general regression method by mathematical proving and the calculation of the examples.