统计诊断引论

韦博成　鲁国斌　史建清　著

东南大学出版社

统计诊断引论

韦博成　鲁国斌　史建清　著

（国家自然科学基金资助项目）

东南大学出版社

内 容 简 介

本书系统介绍统计诊断的原理、方法和应用。内容包括：线性回归的异常点分析、线性回归的残差分析、线性回归的影响分析、数据变换、数据变换的诊断、广义影响分析基础、多元线性回归模型的统计诊断、多元分析中的影响评价、其他模型的统计诊断、回归诊断的 Bayes 方法等。书中少数标有"*"号的段落供有兴趣的读者参考。

本书可作为应用统计、管理科学、计量经济等方向的研究生教材。但在取材与叙述上力求深入浅出，其中大部分内容可供工程、农林、生物、医学等领域从事实际应用的大学生、研究生、教师、科技人员和统计工作者参考。

责任编辑 徐步政

统计诊断引论

韦博成 鲁国斌 史建清 著

东南大学出版社出版
（南京四牌楼2号）
江苏省新华书店发行 东南大学印刷厂印刷
开本850×1165毫米1/32 印张 20.125 字数523千字
1991年6月第1版 1991年6月第1次印刷
印数：1—2000册

ISBN 7-81023-424-2

O·43 定价：5.25元

序　言

统计诊断是晚近20多年迅速发展起来的一门统计学新分支。它以强烈的应用背景、新颖的统计思想、广泛的研究内容和丰富的实际成果在广大统计工作者面前展现出一个理论与应用紧密结合的崭新领域。顾名思义，统计诊断就是对实际问题中得到的数据和提炼出的模型以及相应的统计推断方法的合理性进行研究；检查数据和模型及其推断方法中可能存在的"毛病"，并提出相应的"治疗"措施。大量的理论研究和应用实践使人们对统计诊断的意义和价值有了肯定而明确的认识，现今已广泛应用于各种统计问题和许多统计模型；并被编入许多通用软件包，成为统计学使用过程中不可缺少的一个重要步骤。

本书系统介绍统计诊断的原理、方法和应用。鉴于统计诊断的应用范围十分广泛，为了尽可能兼顾多方面读者的需要，本书在详细阐述基本原理的同时，重点突出实用诊断方法，并列举来自工程、农林、经济、医学、教育、心理等领域的大量实例（近100个），还辅之以各种诊断图表，力求做到学以致用。全书分为两大部份：前六章介绍线性回归模型的诊断，这是理论上较成熟、应用上最成功的部分；后五章介绍其他统计模型的诊断以及统计诊断的新方法，其中不少章节都包含作者近年来的研究成果。这几章内容较新，许多问题仍在继续研究和探索之中。本书内容大体分三个层次，前六章的主要内容只要求读者具有工科类型大学微积分、线性代数和概率统计的知识；后五章的大部分内容则要求读者具有理科类型概率统计的基本知识；少数比较专门的内容则标以"*"号，供有兴趣的读者参考。作者期望：对于从事实际应用的读者，通过阅读本书能够较快的掌握统计诊断的基本方法，并将其用于本专业的实际问题；对于侧重基础理论的读

者，能够较快的了解统计诊断的基本原理和发展方向，并能逐步接触近代文献和有关的研究工作。

本书初稿为我校研究生讲义。1986年以来，其主要内容曾在我校进行过多次讲授；部分内容也在全国一些短训班作过介绍。在成书过程中，我们对原有讲义作了全面的充实和加工改写。本书采用集体讨论、分头执笔的方式完成初稿（其中韦博成负责第二、三、四章以及第一、七、十章部分内容；鲁国斌负责第八、九章和第一、七、十章部分内容以及部分审定工作；史建清负责第五、六、十一章以及大部份数值计算工作；徐亮负责§10.1的编写工作及部分计算工作）。在初稿的基础上，由韦博成负责综合修改、审稿、定稿。由于我们水平有限，难免有不妥和谬误之处，恳请同行专家和广大读者提出批评和建议。

作　者

1990年6月于东南大学

目　录

第一章　引论 …… 1

§1.1　统计诊断概述 …… 2

1.1.1　统计诊断的内容和意义 …… 2

1.1.2　统计诊断的两个基本概念 …… 9

§1.2　预备知识 …… 16

1.2.1　矩阵代数 …… 17

1.2.2　多元正态分布及其他有关分布 …… 23

1.2.3　线性模型的参数估计和假设检验 …… 28

第二章　线性回归的异常点分析 …… 38

§2.1　异常点的检验 …… 39

2.1.1　数据删除模型 …… 39

2.1.2　均值漂移模型与异常点的检验 …… 43

§2.2　多个异常点的检验 …… 53

2.2.1　数据删除模型 …… 54

2.2.2　均值漂移模型 …… 58

§2.3　方差扩大模型 …… 61

2.3.1　Score 检验 …… 63

2.3.2　最大似然估计 …… 71

第三章　线性回归的残差分析 …… 78

§3.1　学生化残差 …… 80

§3.2　残差图 …… 85

3.2.1　标准残差图 …… 86

3.2.2　附加变量残差图 …… 94

3.2.3　正态图、半正态图和包络图 …… 104

§3.3　其他形式的残差 …… 110

3.3.1　预测残差 …… 110

3.3.2* 不相关残差和BLUS …… 112
3.3.3* 递推残差 …… 117

第四章 线性回归的影响分析 …… 122
§4.1 度量影响的基本统计量 …… 124
4.1.1 Cook 统计量 …… 125
4.1.2 广义Cook 距离和 W-K 统计量 …… 129
4.1.3 子集参数的 Cook 距离和 W-K 统计量 …… 135
4.1.4 多个数据点的影响度量 …… 142
§4.2 度量影响的其他统计量 …… 143
4.2.1 协方差比统计量 …… 144
4.2.2 AP 统计量 …… 145
4.2.3 置信域的体积比 …… 147
4.2.4 信息比统计量 …… 148
§4.3 似然距离 …… 156
4.3.1 线性模型的似然距离 …… 158
4.3.2 似然距离的近似计算 …… 164

第五章 数据变换 …… 167
§5.1 方差稳定化变换和线性化变换 …… 168
§5.2 Box-Cox变换 …… 181
5.2.1 Box-Cox 变换 …… 182
5.2.2 变换参数的最大似然估计 …… 184
5.2.3 变换参数的 Atkinson 估计 …… 193
5.2.4 Box-Cox变换的大样本性质 …… 196
§5.3 推广的幂变换族 …… 199
5.3.1 带有漂移参数的幂变换族 …… 199
5.3.2 折叠的幂变换族 …… 204
5.3.3 模变换族 …… 205
§5.4 自变量的 Box-Cox 变换 …… 206
§5.5* 双边 Box-Cox 变换 …… 210
§5.6 数据变换的假设检验 …… 214

5.6.1 检验统计量 …… 215
5.6.2 各种检验方法的比较 …… 222

第六章 数据变换的诊断 …… 231
§6.1 异常点的图诊断法 …… 232
§6.2 数据变换诊断的常用方法 …… 241
6.2.1 诊断模型分析 …… 241
6.2.2 似然距离诊断法 …… 244
6.2.3 Atkinson 诊断法 …… 255
6.2.4 子集参数的 Cook 统计量诊断法 …… 260
§6.3 自变量变换的诊断 …… 265
6.3.1 Atkinson 诊断法 …… 266
6.3.2 子集参数的 Cook 统计量诊断法 …… 269
6.3.3 似然距离诊断法 …… 275
§6.4* 双边变换的诊断 …… 279

第七章 广义影响分析基础 …… 285
§7.1 影响函数 …… 286
7.1.1 影响函数的定义及其样本形式 …… 286
7.1.2 线性回归模型中的影响函数 …… 295
§7.2 局部影响分析 …… 302
7.2.1 影响图 …… 303
7.2.2 局部影响的曲率度量 …… 307
7.2.3 子集参数的局部影响 …… 314
§7.3 线性回归的局部影响分析 …… 319
7.3.1 方差扰动模型 …… 319
7.3.2 自变量扰动模型 …… 324
§7.4* 数据变换模型的局部影响分析 …… 331
7.4.1 方差扰动模型 …… 332
7.4.2 自变量扰动模型 …… 335
7.4.3 自变量变换的局部影响 …… 337
7.4.4 变换数据扰动的影响 …… 341

第八章　多元线性回归模型的统计诊断 …… 347
§8.1　多元线性回归模型 …… 347
§8.2　均值漂移模型异常点的识别 …… 352
§8.3　影响分析 …… 357
8.3.1　多元广义 Cook 距离 …… 357
8.3.2　其他影响度量 …… 363
8.3.3　实例 …… 371
§9.4　似然距离 …… 377
第九章　多元分析中的影响评价 …… 383
§9.1　基本扰动展开式 …… 385
§9.2　线性判别中的影响分析 …… 392
9.2.1　理论影响函数 …… 392
9.2.2　影响函数的逼近及其样本形式 …… 398
§9.3　主成分分析中的影响评价 …… 403
9.3.1　特征根和特征向量的影响函数 …… 403
9.3.2　样本场合下影响函数的计算 …… 406
9.3.3　三种样本形式的进一步分析 …… 409
§9.4* 典型相关分析中的影响评价 …… 416
9.4.1　典型相关度量的影响函数 …… 416
9.4.2　相关系数影响函数的分布 …… 422
9.4.3　样本场合下的影响度量 …… 425
9.4.4　强影响点的检测 …… 426
第十章　其他模型的统计诊断 …… 442
§10.1　广义线性模型的诊断 …… 442
10.1.1　广义线性模型 …… 442
10.1.2　广义线性模型的诊断 …… 448
10.1.3　诊断图示法 …… 465
10.1.4* 局部影响分析 …… 472
§10.2　非线性回归模型的统计诊断 …… 480
10.2.1　非线性回归模型 …… 481

10.2.2 基于线性近似的诊断方法…… 483
10.2.3 诊断模型分析…… 488
10.2.4* 均值漂移模型的曲率度量…… 492
10.2.5* 基于统计曲率的诊断统计量…… 496
§10.3 时间序列的统计诊断 …… 503
10.3.1 时间序列中异常点的类型 …… 504
10.3.2 基于删除数据点的识别方法 …… 512
10.3.3 识别异常点的 Score 检验方法 …… 523
10.3.4* ARMA模型的强影响点诊断…… 535
第十一章 回归诊断的 Bayes 方法 …… 548
§11.1 异常点诊断的 Box-Tiao 方法 …… 549
11.1.1 Box-Tiao 方法概述 …… 549
11.1.2 方差加权模型的 Bayes 诊断 …… 553
11.1.3 均值漂移模型的 Bayes 诊断…… 563
§11.2 异常点诊断的 Chaloner-Brant 方法…… 567
§11.3 影响分析的 Kullback-Leibler 距离方法 …… 577
11.3.1 Kullback-Leibler 距离方法…… 577
11.3.2 Bayes 估计的影响度量 …… 580
11.3.3 Bayes 预测的影响度量 …… 586
§11.4 异常点诊断的条件预测方法…… 590
§11.5* Bayes 局部影响分析…… 594
11.5.1 Bayes 局部影响 …… 594
11.5.2 线性模型 Bayes 估计的局部影响分析 …… 599
11.5.3 线性模型 Bayes 预测的局部影响分析 …… 606
附录 1 Score 检验统计量 …… 610
附录 2 多元 t 分布 …… 615
参考文献 …… 618

第一章 引 论

统计诊断是70年代中期发展起来的一门统计学新分支。70年代正是统计学在高速计算机有力支持下大放异彩的年代，是二次大战后统计学蓬勃发展浪潮中的一个新的高峰期。以往大多数经典的统计方法，诸如参数估计、假设检验、线性回归、多元分析等等，大多是在计算条件慢速而昂贵的限制下发展起来的。某些需要较大计算量的方法，如随机模拟、Jackknife 等方法，尽管提出来了，但由于受到当时计算条件的限制，往往得不到人们足够的重视。70年代高速计算机的迅速发展和普及给统计学注入了新的活力，也引起了统计学观念上的更新。许多新的方法，诸如随机模拟、投影追踪法（Projection Pursuit）、刀切-自助法、(Jackknife-Bootstrap)、统计图形法（Statistical Graphics）、统计诊断（Statistical Diagnostics）等等，也应运而生，并迅速地发展成为统计学的一些新的活跃的分支。与经典方法相比，这些方法的一个鲜明的特点就是广泛应用已变得十分快速而又便宜的电子计算机，代表着统计学与计算机密切结合的新潮流。

顾名思义，统计诊断就是对从实际问题中收集起来的数据和提炼出来的模型以及由此出发所作的推断方法的合理性进行深入细致的分析，并通过一些诊断统计量来检查数据、模型及推断方法中可能存在的“毛病”，进而提出“治疗”方案。也就是说对统计方法解决问题的全过程进行诊断。10多年来的理论研究和应用实践，使人们对统计诊断的必要性有了肯定而明确的认识。今天它已成为统计学使用过程中不可缺少的一个分析步骤，已经编入许多统计软件包，并正在受到统计界和广大实际工作者越来越多的重视。

本章引论分两部分：第1节简要介绍统计诊断的内容、意义及其在统计中的地位和作用；第2节介绍本书各章所需的若干基本知识。

§1.1 统计诊断概述

1.1.1 统计诊断的内容和意义

统计学研究的出发点是一个数据集D，该数据集往往是根据在实际工作中逐步积累起来的历史资料或围绕某一特定目标收集起来的数据经初步加工整理而成。为了通过数据集D研究实际问题，通常的做法是把它纳入某一方便有效的统计模型M进行研究。但是，任何统计模型都只能是对客观复杂过程的一种近似描述，它不可避免地要包含某些假定，甚至模型本身也就是一种假定。人们自然有理由要问：我们选择的模型M究竟能不能大体上反映所要研究的实际问题？它是否与数据集D中绝大多数的数据相一致？我们所得到的数据集D中会不会有个别数据由于收集或整理过程中的疏忽和失误或其它种种原因而出现较大的误差？这些错误数据会不会严重干扰我们对问题所作的结论？另外，数据集D中各个数据点对我们进行统计推断的影响是否大致相仿，会不会有某些点的影响特别大？等等。在使用统计方法解决具体问题的过程中，人们必须慎重地回答上述种种问题，才能作出更加符合客观实际的结论。这一点，在以往的统计分析中常常被忽视，从而有可能得到与实际情况严重不符合的分析结果。

统计诊断就是针对上述种种问题而发展起来的一种分析方法。为了克服既定模型（postulated model）与客观实际之间可能存在的不一致性，通常有两种途径可循：第一，寻找一种统计方法使之当模型有微小变动或扰动（perturbation）时统计推断不受太大的影响，亦即这种统计方法对模型的扰动具有某种稳

健性，这就是所谓稳健统计。第二，寻找一种诊断方法，判断实际数据是否与既定模型有较大偏离并采取相应对策，这就是统计诊断的主要内容。通过统计诊断，可以找出严重偏离既定模型的数据点，即所谓异常点（outlier）；也可以区分出对于统计推断影响特别大的点，即所谓强影响点（influential point）；还可以找出那些远离数据主体的点，即所谓高杠杆点(high leverage point)。此外，还可研究模型中若干具体因素对于统计推断的影响。对数据进行这些初步诊断后，还需要尽可能研究“治疗”方案。如果实际数据中仅有个别点与既定模型偏离较大，这时我们往往肯定模型，而对这些个别点再作进一步考察。如果实际数据中许多点都与既定模型偏差比较大，则需要采取“更有力”的治疗措施。在多数情况下，仍然希望保留方便有效的既定模型。为此，可对数据集进行合适的数据变换，使得变换后的数据符合既定模型，从而进行必要的统计分析。如果数据变换后统计分析的效果仍然不够理想，那就要进行“大手术”，寻找较为复杂但更加有效的模型。显然，这是比较复杂的问题。

今结合最常用的线性回归模型进一步说明回归诊断的内容和意义。

线性回归模型可表示为

$$y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_{p-1}x_{i\,p-1}+\varepsilon_i,\ i=1,\ 2,\cdots,n\ .$$

其中 y_i 为因变量，$x_{i1},\cdots,x_{i\,p-1}$ 为自变量，ε_i 为随机误差。其第 i 组观察值为 $(y_i;\ x_{i1},\cdots,x_{i\,p-1})$。通常可表示为矩阵形式如下

$$Y=X\beta+\varepsilon\ . \tag{1.1.1}$$

其中 $Y=(y_1,\cdots,y_n)^T$，$\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)^T$，$\beta=(\beta_0,\beta_1,\cdots,\beta_{p-1})^T$。$X$为 $n\times p$ 阶列满秩矩阵，其第 i 行为 $(1,\ x_{i1},\cdots,x_{i\,p-1})$。本书前几章主要讨论这一模型。对于随机误差项 ε，通常假定其分量 $\varepsilon_1,\cdots,\varepsilon_n$ 互相独立，数学期望为零，方差具有

齐性，即

$$E(\varepsilon)=0,\ \mathrm{var}(\varepsilon)=\sigma^2 I\ .$$

其中 σ^2 为未知常数，I 为 n 阶单位矩阵。这时记为

$$\varepsilon\sim(0,\ \sigma^2 I)\ . \tag{1.1.2}$$

在多数情况下还假定 ε 服从标准正态分布，即

$$\varepsilon\sim N(0,\ \sigma^2 I)\ . \tag{1.1.3}$$

通常的线性回归，大多采用了这些假设。这里有一个值得注意的重要问题，即给定的数据集$(y_i;x_{i1},\cdots,x_{ip-1}),i=1,\cdots,n$，是否确实符合关于模型的假设(1.1.1)、(1.1.2)或(1.1.3)式？这就是回归诊断要研究的主要问题。一般可能只有少量数据点不符合关于模型的假设条件，这种点就是所谓异常点。识别、判定和检验异常点是回归诊断的重要内容。如果数据集与既定模型有很大的或系统性的偏离，则可设法对模型的假设进行修正。但更多的是保留模型，对数据进行变换，使得变换后的数据适合假设条件(1.1.1)、(1.1.2)或(1.1.3)式。这显然是因为只有在这些假设下，对于数据的回归分析才能行之有效。回归诊断的另一重要内容就是所谓影响分析。很显然，每组数据$(y_i;x_{i1},\cdots,x_{ip-1})$对回归模型的统计推断（如估计、检验、预测等等）都有一定的影响，但并非每组数据的影响都一样。通过统计量定量地刻画数据点影响的大小，从而找出强影响数据点，这也是回归诊断的重要内容。另外，残差分析和残差图是研究既定模型是否能够很好拟合给定数据的行之有效的综合方法，它与异常点分析、影响分析以及数据变换都有密切关系，也是回归诊断的重要内容。我们通过一个例子进一步说明回归诊断的意义。

例 1.1 Anscombe 数据。

表 1.1 是Anscombe 于1973年给出的有名的数据(见[10])。这四组人造数据集强有力地说明了回归诊断的必要性。

表 1.1 Anscombe 数据

No.	x_1	y_1	y_2	y_3	x_2	y_4
1	10.00	8.04	9.14	7.46	8.00	6.58
2	8.00	6.95	8.14	6.77	8.00	5.76
3	13.00	7.58	8.74	12.70	8.00	7.71
4	9.00	8.81	8.77	7.11	8.00	8.84
5	11.00	8.33	9.26	7.81	8.00	8.47
6	14.00	9.96	8.10	8.84	8.00	7.04
7	6.00	7.24	6.13	6.08	8.00	5.25
8	4.00	4.26	3.10	5.39	19.00	12.05
9	12.00	10.08	9.13	8.15	8.00	5.56
10	7.00	4.82	7.26	6.42	8.00	7.91
11	5.00	5.68	4.74	5.73	8.00	6.89

这四组数据分别记为 $D_1(x_1, y_1)$，$D_2(x_1, y_2)$，$D_3(x_1, y_3)$ 和 $D_4(x_2, y_4)$。今对这四组数据分别做一元线性回归，即取模型

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,\ 2,\ \cdots,\ 11\ .$$

经简单计算可知，数据集 D_1—D_4 的基本统计量都相同，即

$$\hat{\beta}_0=3.0\ ,\quad \hat{\beta}_1=0.5\ ,\ \hat{\sigma}^2=1.531,\ R^2=0.667$$

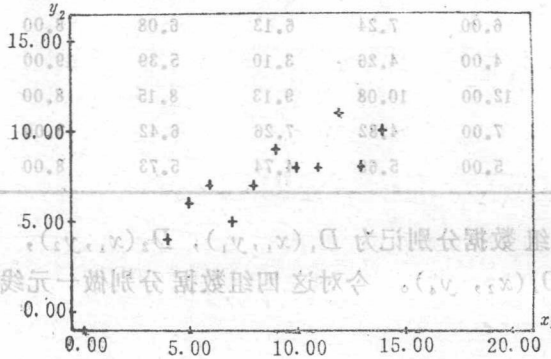
其中 σ^2 为 y_i 的方差，R 为复相关系数（见（1.2.57）式）。因此，若按经典回归分析观点考虑，回归方程 $y=3+0.5x$ 拟合这四组数据集效果完全相同。但是，我们看一下散点图 1.1 即可发现，这四组数据集的情况相差极大！

图 (a) 表明，数据集 D_1 用上述模型拟合基本合适。图 (b) 表明，散点图与回归直线有很大不同，也许用二次曲线 $y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\varepsilon_i$ 对数据集 D_2 进行拟合效果会更好。图(c)表明，回归直线对数据集 D_3 的大多数点都合适。但是第 3 个数据点（13.0，12.74)偏离回归线很远。这是一个很明显的异常点，

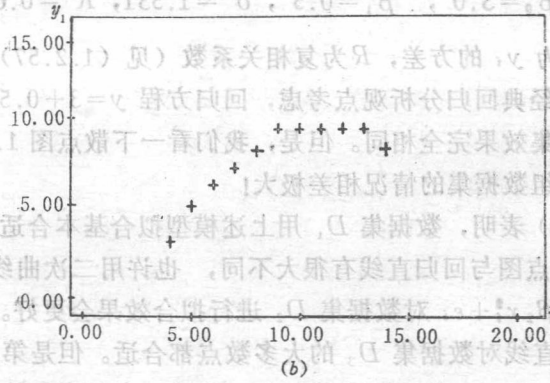
由于它的存在，使得回归线斜率有所增加，截距有所减小。如果把这个点删除再进行回归，可能效果更好。这时相应的基本统计量为

$$\hat{\beta}_0=4\ ,\quad \hat{\beta}_1=0.346\ ,\quad \hat{\sigma}^2=9.497\times10^{-6},\quad R^2=1\ .$$

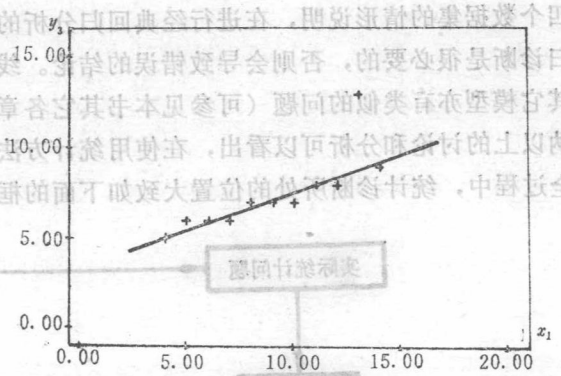
这时回归方程 $y=4+0.346x$ 与数据集 D_3 拟合得更好（不考虑



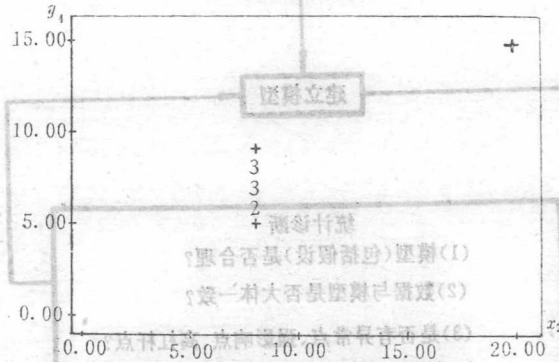
(a)



(b)



(c)



(d)

图1.1 Anscombe 数据的散点图(*)

第3点，见图(c)中直线)。图(d)表明，所求的回归线与数据集 D_5 的散点图完全不合适。其原因是数据集提供的信息太少，不足以进行模型拟合。这是因为自变量 x 的11个观察值中有10个集中在 $x=8$，而其它点，除了 $x=19$ 外，没有进行任何观察。显然，自变量 x 的观察值必须多取一些不同的点，才能提供更多的信息，以便进行合理的回归分析。

(*) 散点图中有一个"+"代表一个点，而数字则表明所在位置有多个点。如(d)中的"3"表示所在位置有3个点，以后的散点图均按此法表示。