GenStat®

GenStat
统计方法与数据分析

R.W. Payne D.B. Baird M. Cherry A.R. Gilmour S.A. Harding
A.F. Kane P.W. Lane D.A. Murray D.M. Soutar R. Thompson
A.D. Todd G. Tunnicliffe Wilson R.Webster S.J. Welham.

郑可锋 张浩 祝利莉 等 译
王 磊 Xiang-ming Xu 校

中国农业科学技术出版社

GenStat®

GenStat
统计方法与数据分析

R.W. Payne D.B. Baird M. Cherry A.R. Gilmour S.A. Harding
A.F. Kane P.W. Lane D.A. Murray D.M. Soutar R. Thompson
A.D. Todd G. Tunnicliffe Wilson R.Webster S.J. Welham.

郑可锋 张 浩 祝利莉 等 译
王 磊 Xiang-ming Xu 校

中国农业科学技术出版社

图书在版编目（CIP）数据

GenStat 统计方法与数据分析／（英）佩恩（Payne，R.）等著；郑可锋等译．—北京：中国农业科学技术出版社，2009.12
ISBN 978-7-5116-0080-6

Ⅰ.①G… Ⅱ.①佩…②郑… Ⅲ.①统计分析-应用软件，GenStat Ⅳ.①C819

中国版本图书馆 CIP 数据核字（2009）第 225047 号

责任编辑 梅 红
责任校对 贾晓红

出 版 者 中国农业科学技术出版社
北京市中关村南大街 12 号 邮编：100081
电 话 (010)82109704(发行部)(010)82106630(编辑室)
(010)82109703(读者服务部)
传 真 (010)82106636
网 址 http://www.castp.cn
经 销 者 新华书店北京发行所
印 刷 者 北京富泰印刷有限责任公司
开 本 889 mm×1 194 mm 1/16
印 张 13
字 数 330 千字
版 次 2010 年 4 月第 1 版 2010 年 4 月第 1 次印刷
定 价 32.00 元

内容简介

本书是英国 VSNi 公司开发的 GenStat 统计软件的配套教材。基本内容包括线性回归、非线性回归、广义线性模型、方差分析、混合模型分析、RMEL 综合分析及空间分析等，其中包括近年来的一些较新进展。大部分统计方法都给出了 GenStat 软件的操作过程及输出结果的解读，便于广大科研人员的教学、自学和应用。

本书可作为农业科研机构科研人员、高等农业院校或综合性院校生物类各专业本科生、研究生的教材，也可供各领域需要进行数据分析处理的实际工作者自学参考。

声　明

序

欢迎使用 GenStat 统计方法与数据分析。和世界上很多从事生物科学研究的科学家一样，你也选择了使用 GenStat 统计软件用于分析科研数据。

英国洛桑自 1843 年 John Bennet Lawes 开展小麦试验时开始农业研究，到目前为止，洛桑试验站是世界上运行时间最长的试验站。同时，洛桑试验站也率先在生物科学研究中应用统计方法。1919 年 Ronald Fisher 在研究小麦前期累积试验结果和后续试验结果时，他意识到在整个农业和生物研究中需要提高统计技术，故他和他的同事在 19 世纪 20 年代和 30 年代奠定了现代应用统计的基础。

当 Fisher 的继任者 Frank Yates 将一台 Elliot 401 计算机用于自己的统计工作时，洛桑试验站开始了统计计算工作。Fisher 将传统的统计研究延伸到现实中的生物统计，也促进了洛桑试验站在实践中更有效地应用现代统计软件。1968 年 John Nelder 接替 Yates 成为统计方面的负责人后，洛桑试验站开始开发 GenStat 统计软件。1985 年 Nelder 退休后，Roger Payne 继续领导 GenStat 的开发工作。

19 世纪 70 年代，GenStat 开始走出洛桑试验站。1979 年，世界上最古老的计算技术公司之一 Numerical Algorithms Group（NAG）开始负责 GenStat 的经营。近年来，VSNi 公司负责 GenStat 的开发和销售。VSNi 公司组建于 2000 年，作为洛桑试验站和 NAG 旗下的子公司，其整合了洛桑试验站的开发团队和 NAG 的商业经营团队，促进了 GenStat 的开发和市场经营。然而，开发团队作为洛桑试验站和 VSNi 公司连接者，仍然与研究者保持着紧密的联系。所以，用户不仅能从公司严格的质量控制中获益，而且能感受到浓厚的研究氛围。

GenStat 的一个重要特点是，其开发团队（包括合作团队）的成员本身就从事广泛的应用统计工作，如方差分析、试验设计、线性模型、典型变量分析及最近开发的混合模型分析等。因此，GenStat 不仅能运行现有的统计模型，而且还不断地发展和应用新的模型。所以我们相信，GenStat 不仅能满足生物统计方面的需求，而且是解决其他应用统计问题的理想工具。

Roger Payne

Professor Roger Payne
Technical Director
November, 2009.

前　言

随着现代信息技术的发展，在农业和生物科学研究中利用统计软件进行统计分析越来越普遍。统计软件不仅为相关科研工作者消除了大量数据处理的烦恼，同时可以促进其对统计理论和方法的深入理解，不断提高人们应用统计学的能力，统计软件已成为农业和生物科学研究工作中最有力且不可缺少的工具。GenStat 统计软件是一款强大而灵活的统计软件，它属于完全交互式系统，具有先进的图形化工具和友好的图形化用户界面，还有强大的统计学程序编制功能。而且，GenStat 具有悠久的成功历史，经过不断更新发展，它始终能活跃在生物统计学技术的最前沿。

GenStat 统计软件有比较详细的原版使用手册，但对于大多数中国读者而言，阅读这些英文材料无疑会耗费较多的时间。因此，如何让中国读者快速熟悉和使用 GenStat 统计软件，并能解决农业和生物科学研究中实际问题，是翻译《GenStat 统计方法与数据分析》一书的初衷。因译者水平有限，本书的翻译出版只是一次初步的尝试，但愿它能起到抛砖引玉之作用。

本书以 GenStat 7.0 统计软件包作为实现复杂统计计算的工具，从而省去了大量的篇幅着重介绍各种试验设

计方法、各种统计分析方法及其适用条件、结合具体问题正确选用统计方法的技术以及对计算结果的正确解释和应用。在一切从实际出发的思想指导下，经过合理调整结构及书写形式，貌似复杂的统计问题被化繁为简，更实用更方便。

全书分为3篇共15章。第1篇由第1章至第3章组成，主要介绍回归模型，包括线性回归、非线性回归和广义线性回归。第2篇由第4章至第11章组成，着重介绍GenStat在T检验、区组结构、处理结构、假设检验、误差项设计等方面的应用。第3篇由第12章至第15章组成，分别介绍了GenStat在线性混合模型分析、RMEL综合分析、空间分析等方面的应用。

整体编译工作安排如下：第1章至第3章由郑可锋、张浩、祝利莉同志编译，第4章至第11章由郑可锋、张浩、张小斌和姚旭国同志编译，第12章至第15章由胡为群和叶少挺同志编译，郑可锋同志负责全书编译工作的统筹和安排。

本书可用作生物类研究生、本科生、大中专生的统计学教材，亦可作为涉农高等院校和科研机构的教师、学者、科技人员、管理工作者等学习和应用统计方法的参考书，还可作为用GenStat软件解决统计问题的实用手册。

本书的问世，与英国VSNi公司和浙江省农业科学院在资金上给予的大力扶持是息息相关的，与中国农业科学技术出版社的热情关心和帮助是分不开的。中国水稻研究所的王磊研究员对本书内容提出了很多宝贵的建议。英国洛桑试验站的Xiang-ming Xu博士为本书的初稿作了大量认真而又细致的校对工作。在此，我们一并表示衷心的感谢！

最后，由于我们水平有限，编译时间仓促，难免有错漏之处，敬请读者批评指正，以便我们不断改进。

译者

二〇一〇年三月

目　　录

第一篇

第二篇

第三篇

第 1 章　线性回归

本章将学习以下内容：
— 如何拟合单个自变量的回归模型
— 输出的含义是什么
— 如何绘制拟合模型
— 分析有哪些假设，如何验证
— 拟合、显示、评估线性模型所需命令★
— 如何用排列检验评价回归★
— 如何保存 GenStat 数据结构的输出以便将来使用★
— 如何根据回归分析进行预测
— 如何拟合多元线性回归（多个自变量）
— 存在多个自变量时如何探索其他模型
— 如何用全子集回归评估和汇总所有模型★
— 既有说明因子又有自变量时，如何拟合平行和非平行回归线
备注：标出★的内容阅读时可选。

1.1　简单线性回归

线性回归是描述一个变量和其他一个或多个变量之间关系的方法：
— 响应变量（又称 y - 变量或因（应）变量）是被描述的变量；
— 自变量（又称 x - 变量或自变量）是用于描述响应变量的变量。
“简单线性回归”只有一个自变量，即 x。因此，我们希望用下列模型描述应变量 y：

$$y = b \times x + c$$

模型中参数为：
b：回归系数
c：常数。

简单线性回归中，因为 c 是 x 为 0 时的 y 值，所以常数 c 常被称为截距（intercept）。一般模型都带常数，后面我们将讨论如何拟合无常数的模型。回归系数 b 一般被称为回归线的斜率（slope）。

上述模型表示了对 y 假设的理论值，但实际情况不太可能与观测值相同：可能存在随机变异，或者模型可能只是真实情况的一种近似。假设我们已对 x 和 y 进行 n 次观察，观察次数以下标 i 表示。我们可以定义一个描述观察的统计模型：

$$y_i = b \times x_i + c + \varepsilon_i \qquad i = 1 \cdots n$$

其中：ε_i 为第 i 次观察的残差（residual），代表对第 i 次观察模型预测的理论值和实际观测值 y_i 的差异。

模型预测的理论值称为拟合值：

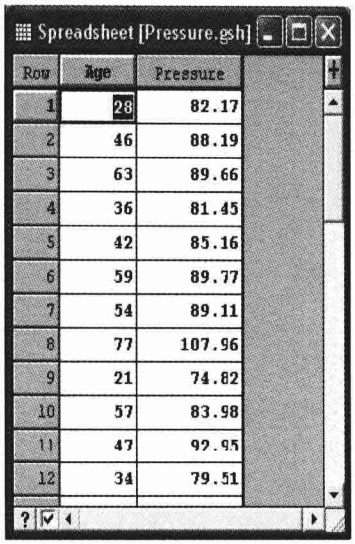
$$f_i = b \times x_i + c \qquad i = 1 \cdots n$$

一般的线性回归中，残差 ε_i 假设来自独立的正态分布，其方差相同。1.3 节将介绍如何验证此假设。第 3 章将介绍如何拟合来自其他分布数据的模型。

用最小平方（least squares）估算参数值，即取使残差平方和最小的参数值：

$$\sum{}_i \varepsilon_i^{\,2} = \sum{}_i (y_i - b \times x_i - c)^2$$

如果残差确实为正态分布，我们得到的估计为极大似然（maximum likelihood）估计（即参数值使得观察到的数据有最大概率）。本节后面在介绍统计检验时还会用到正态分布假设。此处不对数学统计理论作深入讨论，更多信息可参考《应用回归分析》（Draper & Smith，Applied Regression Analysis，1981，Wiley，NewYork）等标准统计学教材。



Row	Age	Pressure
1	28	82.17
2	46	88.19
3	63	89.66
4	36	81.45
5	42	85.16
6	59	89.77
7	54	89.11
8	77	107.96
9	21	74.82
10	57	83.98
11	47	99.95
12	34	79.51

图 1.1

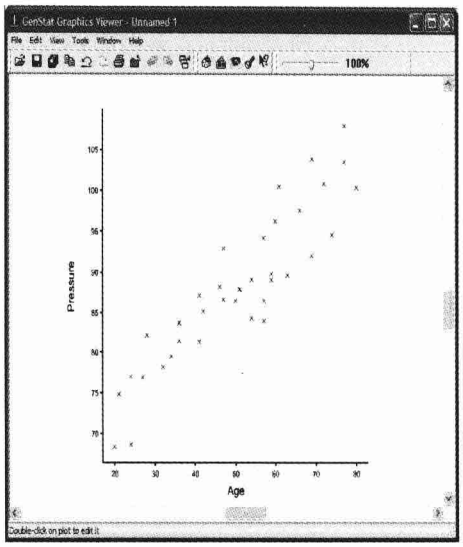


图 1.2

表格文件 Pressure. gsh（图 1.1）为 38 位 20 ~ 80 岁妇女的血压样本记录。选择 Graphics 菜单的 Point Plot 可绘制血压（Pressure）和年龄（Age）关系图（图 1.2），它显示血压和年龄之间有一定的线性关系，因此适用线性回归。图 1.3 表示的就是回归线，以及从数据点到回归线上所对应拟合值的垂直距离表示的残差。

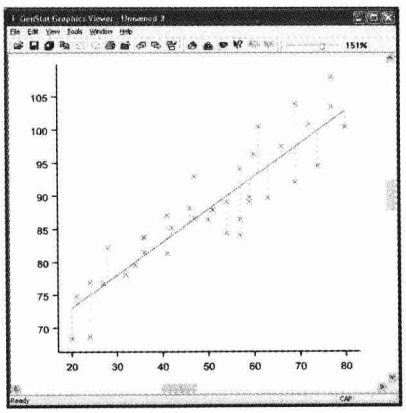


图 1.3

为了利用 GenStat 拟合回归，选择菜单栏中 Stats 下拉菜单的 Regression Analysis 选项，点击 Linear Models（线性模型）子选项（图 1.4）。

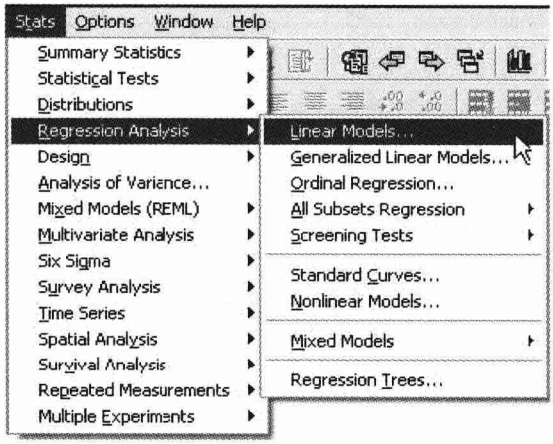


图 1.4

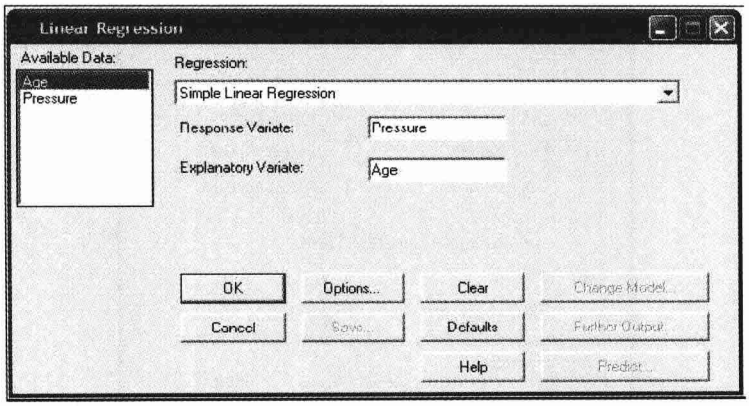


图 1.5

这时弹出 Linear Regression（线性回归）对话框（图 1.5）。如果选择对话框中顶部的 Regression 下拉列表的 Simple Linear Regression（简单线性回归）选项，我们只需在窗口中指定应变量（Response Variate）和自变量（Explanatory Variate）即可。点击“OK”，生成下列输出（Output）：

***** Regression analysis *****

Response variate: Pressure

Fitted terms: Constant, Age

*** Summary of analysis ***

Source	d. f.	s. s.	m. s.	v. r.	F pr.
Regression	1	2 647.7	2 647.69	169.73	<0.001
Residual	36	561.6	15.60		
Total	37	3 209.3	86.74		

续表

Percentage variance accounted for 82.0				
Standard error of observations is estimated to be 3.95				
*** Estimates of parameters ***				
Parameter	estimate	s. e.	t (36)	t pr.
Constant	63.04	2.02	31.27	<0.001
Age	0.498 3	0.038 2	13.03	<0.001

点击 Linear Regression 窗口中的 Options 按钮即可打开 Linear Regression Options（线性回归选项）对话框（图 1.6）来控制输出。默认输出的开始部分为模型的描述、应变量、拟合项（fitted terms，它们为常数和自变量）。输出的默认设置中包含常数（constant），如果希望省略，请不要勾选 Estimate Constant Term 框。它会强制拟合直线通过原点（即自变量值为 0 时应变量必须为 0），但分析仍然基于这样的假设：在整个数据范围内沿着直线的变异是固定的，而且线性关系保持至原点。因此这样的设置有时可能是不合理的，尤其当原点附近有观测值时。

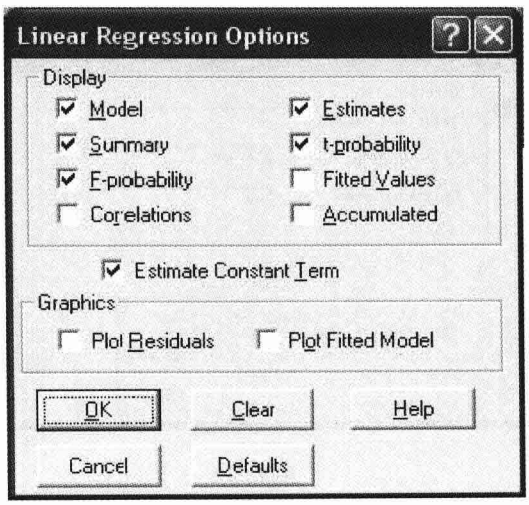


图 1.6

输出的下一部分内容（summary of analysis）为帮助评价模型的方差分析。“s. s.”栏的“Residual”（残差）行对应的数值为残差平方和，被视为随机变异。“Total”（总和）行包括了仅含常数的模型的残差平方和。在此模型中这一常数是用应变量值的均值（即总体均值）来估计的，因此，这一行包括：

$$\sum_i (y - \mu)^2$$

其中：μ 为总体均值：

$$\mu = \sum_i y/n$$

更精确地说，它是基于总体均值校正的平方总和。

线性回归中，我们希望发现应变量和自变量存在线性关系的证据，因此我们要比较仅含常数的模型和含常数和自变量的模型。两个模型的残差平方和差值显示在“Regression”（回归）那一行：

$$2\ 647.7 = 3\ 209.3 - 561.6$$

它被称为“回归引起的”平方和，表示含有自变量的模型所对应的回归系数能“解释”（即从残差中去掉的）的变异量。

“d. f.”（自由度）栏记录构成各平方和的独立参数数目，“Total”行中对应的自由度为37（样本数减1，因为已拟合了常数项），“Residual”行中为36（样本数减2，因为已拟合常数项和自变量的回归系数），“Regression”行中为1（因为这一行代表在模型再增加1个参数的效应）。

“m. s.”（均方）栏为平方和除以自由度的数值，从而将平方和转变为方差。“v. r.”（方差比）栏为回归均方除以残差均方的比值。假设无效假设为真，即应变量和自变量无线性关系，则方差比服从F分布，它的自由度对应在回归（regression）和残差（residual）行中给出的自由度。例如，在前面例子中输出结果的F值为169.73，自由度为1和36。“F pr.”栏给出相应的概率。此值小于0.001（<0.001），因此在0.1%显著性水平关系显著。

需要记住的是，F分布是基于残差服从独立正态分布、方差相同这一假设，1.3节将介绍如何对此进行验证。

方差解释的百分比（percentage variance accounted for）总结了拟合模型已经解释了多少样本值的变异。它是用总均方与残差均方之差占总均方的百分比来表示的。如果此值以比例而非百分比表示，统计上称为校正的决定系数R^2（Adjusted R^2），它不同于相关系数的平方（R^2）。校正的R^2考虑了模型参数数目与观察数的比较。

输出的最后一部分为模型参数估计值（estimates of parameters）。年龄（Age）的回归系数为0.4983，标准误为0.0382。因此模型预测血压每年增加0.4983个单位。对应的t统计量的值较大，为13.03（自由度为36），再次说明血压和年龄之间存在显著线性关系。实际上，回归模型仅有1个自由度，估计表中的t值为方差分析中F值的平方根。因此，它们其实为同一检验。t分布的应用也基于回归假设成立。

点击Linear Regression窗口的Further Output（更多输出）可以获得更多输出。有多个不同回归模型时，如果希望省略某些默认输出，可在Linear Regression Options窗口（图1.6）的相关复选框反勾选，直到决定哪个模型最佳为止。

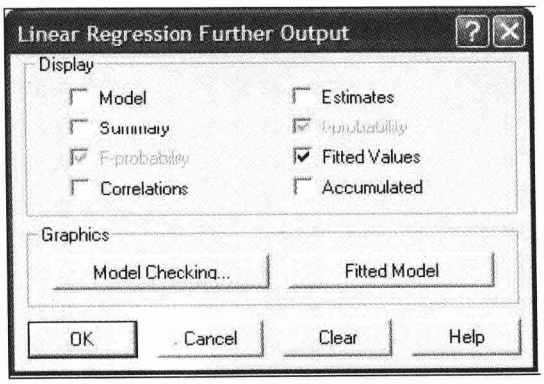


图1.7

图1.7为Linear Regression Further Output窗口。例如，勾选复选框中的Fitted Values（拟合值），点击“OK”，得到下列输出：

***** Regression analysis *****

*** Fitted values and residuals ***

Unit	Response	Fitted value	Standardized residual	Leverage
1	82.17	77.00	1.36	0.072
2	88.19	85.97	0.57	0.028
3	89.66	94.44	-1.24	0.042
4	81.45	80.98	0.12	0.045
5	85.16	83.97	0.31	0.032
6	89.77	92.44	-0.69	0.034
7	89.11	89.95	-0.22	0.028
8	107.96	101.41	1.74	0.095
9	74.82	73.51	0.35	0.105
10	83.98	91.45	-1.92	0.031
11	92.95	86.46	1.67	0.027
12	79.51	79.99	-0.12	0.050
13	87.86	88.46	-0.15	0.026
14	76.85	76.50	0.09	0.076
15	76.93	75.00	0.51	0.090
16	87.09	83.47	0.93	0.034
17	97.55	95.93	0.42	0.050
18	92.04	97.43	-1.41	0.060
19	100.85	98.92	0.51	0.072
20	96.30	92.94	0.87	0.036
21	86.42	87.96	-0.39	0.026
22	94.16	91.45	0.70	0.031
23	78.12	78.99	-0.23	0.057
24	89.06	92.44	-0.87	0.034
25	94.58	99.92	-1.41	0.080
26	103.48	101.41	0.55	0.095
27	81.30	83.47	-0.56	0.034
28	83.71	80.98	0.71	0.045
29	68.38	73.01	-1.24	0.111
30	86.64	86.46	0.05	0.027
31	87.91	88.46	-0.14	0.026
32	86.42	91.45	-1.29	0.031
33	103.87	97.43	1.68	0.060
34	83.76	80.98	0.72	0.045
35	84.35	89.95	-1.44	0.028
36	68.64	75.00	-1.69	0.090
37	100.50	93.44	1.82	0.038
38	100.42	102.91	-0.67	0.111
Mean	87.95	87.95	0.00	0.053