现代外国统计学优秀著作译丛

应用线性回归

APPLIED LINEAR REGRESSION

[美] S. Weisberg 著

王静龙

梁小筠 译

李宝慧

柴根象 校

中国统计出版社

现代外国统计学优秀著作译丛

应用线性回归

（第　二　版）

[美]　S·Weisberg　著

王静龙　梁小筠　李宝慧　译

柴根象　校

中国统计出版社

（京）新登字 041 号

图书在版编目（CIP）数据

应用线性回归/（美）S. 韦斯伯格著（Weisberg，S.）；
王静龙等译. —北京：中国统计出版社，1998. 3
（现代外国统计学优秀著作译丛）
书名原文：Applied Linear Regression
ISBN 7-5037-1966-4

Ⅰ. 应…
Ⅱ. ①韦… ②王…
Ⅲ. 线性回归—回归分析
Ⅳ. O212・1

中国版本图书馆 CIP 数据核字（97）第 00139 号

著作权合同登记：图字 01－97－0561 号

中国统计出版社出版
（北京三里河月坛南街 75 号 100826）
新 华 书 店 经 销
科伦克三莱印务（北京）有限公司印刷
*
850×1168 毫米 32 开本 11.5 印张 29 万字
1998 年 3 月第 1 版 1998 年 3 月北京第 1 次印刷
印数：1－4000 册
*
定价： 26.00 元

版权公告：

Copyright notice：

应用线性回归（第二版）

APPLIED LINEAR REGRESSION

(Second Edition)

[美] S. Weisberg

现代外国统计学优秀著作译丛
专家委员会

出 版 说 明

为了加强对国外统计理论与实践的研究和了解，全面反映国外统计科研和教学的发展，促进我国统计教学改革和教材内容更新，在国家统计局领导的大力支持下，全国统计教材编审委员会组织翻译出版了这套“现代外国统计学优秀著作译丛”。

随着我国社会主义市场经济体系的逐步建立，统计教育正面临着十分严峻的挑战。一方面，在社会主义市场经济条件下，不论国家的宏观经济调控还是企业的生产经营管理，都要求准确地把握市场运行的态势，科学地分析经济中各种错综复杂的关系，因而，对统计信息的需求越来越大，对统计人才的业务素质提出了更高的要求；另一方面，我国过去的统计教育模式是按为高度集中的计划经济管理体制服务的要求建立的，培养的统计人才的知识结构比较单一，难以适应经济体制、统计体制改革的需要。为使统计人才的培养适应建立社会主义市场经济体制的需要，满足二十一世纪现代化建设的要求，缩小与国际先进水平的差距，基础在教育，关键在教材。在继续组织有关专家、学者编写一批反映国内统计科学和统计实践发展的新教材的同时，必须尽快引进并翻译出版一批外国先进统计教材。这是学习外国先进统计知识的一种直接而且十分有效的方式，对于推动国内统计教材内

容更新和教学改革，造就一大批具有渊博知识和多方面业务技能的复合型人才，具有十分重要的意义。

为了做好这套丛书的翻译出版工作，全国统计教材编审委员会成立了现代外国统计学优秀著作译丛专家委员会，对国外统计著作的出版和使用情况进行了调查研究，分析了国内对外国统计教材的需求，在此基础上制定了翻译著作选题规划。在这套丛书的翻译出版过程中，我们得到了国内外有关专家、有关院校统计系和国外有关出版公司的大力帮助和支持，在此表示衷心的谢意。

全国统计教材编审委员会

1995 年 7 月

译 者 序

回归分析是探求变量之间关系的一门科学。它理论丰富、应用广泛，因而是统计工作者不可缺少的工具，也是数理统计领域内最活跃的分支之一。

S·韦斯伯格的“应用线性回归”系著名的Wiley系列的专著。它深入浅出地阐述了回归分析的理论和方法，特别介绍了七、八十年代的许多新思想、新方法。这本书没有把重点放在数学公式的推导上，而是着重于统计思想的分析和讨论。难能可贵的是书中附有大量生动的实例，作者从问题的实际背景出发，介绍了统计方法的应用。这些实例涉及的面很广，大都有原始数据，作者对统计推断的结果进行了周密、细致的分析，给读者留下了很好的范例。统计方法在实际中的应用离不开计算机，这本书在介绍回归分析的各种方法时，也对使用计算机的读者进行了指导，并介绍了有关的统计软件包。

这本书可以作为高等院校统计等专业的师生和广大实际工作者的参考书，具有概率统计和线性代数基础知识的读者不难看懂其中的内容。

国家统计局李宝慧翻译了第二、三章，华东师范大学统计系梁小筠翻译了其余各章，华东师范大学统计系王静龙作了综合修改并定稿。

同济大学应用数学系柴根象教授认真、仔细地审阅了译稿，指出了翻译中不少疏漏之处，并提出许多中肯的意见，对提高翻译

的质量起了很大的作用，在此表示衷心的感谢。我们还要特别感谢国家统计局统计教育中心对我们翻译工作的大力支持。

由于我们水平有限，并且新的统计术语的翻译国内尚未统一，译文中难免有不少缺点和错误，恳请专家和广大读者指正。

译　　者

1996 年夏

第二版前言

自从完成本书第一版以来的五年中，在线性回归领域内的研究发展迅猛。提出了许多新的方法，以改变实际工作者进行回归分析的途径。这次修改意在纳入参考书目中引入的1980年以后的60多篇文献所反映的许多新思想。

大部分正文是重新写过的，以分清已有过的讨论及引入的新思想。增加了几个新的家庭作业的问题、实例和图。主要增加的是新的总结性的一章，它给出了对非线性、*logistic* 和广义线性回归模型的介绍。不过，我在引入新的论题的同时，删去了一部分比较陈旧的，不大重要的材料，由此达到某种平衡，以保持第一版的简洁的特性。

和第一版一样，对在这一工作中给我鼓励的我的许多同事和朋友表示感谢。Michael Lavine 阅读了最后的手稿并纠正了其中的不少错误。Stephen Stigler 特别对第一版中的许多评注给予帮助，由此引导了第二版的修改。我还要感谢Dennis Cook 关于我们在回归模型方面连续的对话。

SANFORD WEISBERG
St. Paul，明尼苏达
1985年4月

第一版前言

线性回归分析由一组探求变量之间关系的技术组成。我们对它感兴趣，在理论上，是因为其基本理论的优美，而从应用的角度看，是由于已经出现并且每天继续出现的各色各样的回归的应用。本书中讨论了用于拟合一个响应变量作为一个或多个自变量的函数的模型的回归方法，以供想要将它们应用于数据的读者使用。中心主题是建立模型，评价拟合及可靠性，并作出结论。如果用作教科书，本书适宜作为统计的第二或第三门课程。仅有的明确的准备知识是熟悉显著性检验，p-值，置信区间，随机变量，参数估计的思想，也要熟悉正态分布及由它导出的分布，如学生 t、F 和 χ^2 分布。当然，附加的统计方法或线性代数知识也将是有用的。

本书分为 12 章。第 1 章和第 2 章分别给出了在简单回归和多元回归中最小二乘估计的相当标准的结论。第 3 章称为“下结论”，是对前两章中的方法的结论作解释。还给出了对未完全测量的自变量的影响的讨论。第 4 章给出了关于最小二乘估计的附加结论。第 5 章和第 6 章涉及研究一个模型拟合失真的方法，检验假设的失败，并评估一个拟合的模型的可靠性。在第 5 章，给出了必需的统计量的理论上的结果，因为许多读者对此可能并不熟悉，第 6 章给出了基于这些统计量的图形和其它方法，以及对它们暴露出来的问题的可能的补救方法。在第 7 章，所论述的主题与建立模型的问题有关，包括虚拟变量、多项式回归和主成分。然

后，第8章提供了基于变量的一个子集选择模型的方法。在第9章，讨论了回归方法用于进行预测时要特别考虑的地方。在上述每一章中，所讨论的回归方法通过使用实际数据的例子加以说明。

接下去的两章比前面的短。第10章给出对部分测量的或不完全数据的分析的指导。最后，在第11章，讨论了除最小二乘估计以外的其它方法。

有几章有相关的附录被收集在正文后面，它们的编号与章相应。例如：附录1A.2是第1章的第二个附录。章是按照线性回归一个学期或三个月的课程安排的，第1章至第8章组成了严格的一个学期的课程。

前面九章中的每一章都给出了家庭作业。理论问题只打算给那些具备必要的统计背景的学生。要求数据分析的问题是给每个人的。某些问题的要求并没有规定清楚，可以根据学生的兴趣进行变化。大部分问题使用真实数据，并且可以用多种方法求解。

计算机　回归方法的应用发展可以直接追溯到计算机的广泛的利用。尽管本书并不打算成为任何特定计算机程序的手册，它指导希望使用计算机的读者应用学到的技术。我们可以得到回归计算的高质量的软件，介绍其来源的必要的参考书目列在正文、家庭作业问题以及附录中。

感谢　在此，对给本书的初稿作出评论，提供例子或通过讨论在被涉及的论题中澄清我的见解的人们表示感谢。这些人中包括Christopher Bingham，Morton Brown，Cathy Campbell，Dennis Cook，Stephen Fienberg，James Frane，Seymour Geisser，John Hartigan，David Hinkley，Alan Izenman，Soren Johansen，Kenneth Kcehler，Darid Lane，Kinley Larntz，John Rice，Donald Rubin，Wei-Chung Shih，G. W. Stewart，Douglas Tiffany，Carol Weisberg，Howard Weisberg和一位不具名的读者。另外，我要感谢明尼苏达大学成果组的成员，Naomi Miner，Sue Hangge，Therese Therrien，特别是Marianne O'Brien，他的熟练的帮助使本书的完

稿变为现实。

在本书写作过程中，我得到了美国国家医学科学总研究所拨款的部分资助。另外，计算机方面，明尼苏达大学计算中心提供了帮助。

SANFORD WEISBERG

St. Paul，明尼苏达

1980 年 2 月

1

简单线性回归

回归被用于研究可以测量的变量之间的关系。线性回归则被用于研究一类特殊的关系，即可用直线或多维时的直线的推广描述的关系。这一技术被用于几乎是所有的研究领域，包括社会科学、物理、生物、商业、科技和人文科学。正如本书例子所示，用线性回归模型拟合的原因因应用而异，而最通常的原因是描述关系并对未来值进行预测。

一般地，回归分析由许多步骤组成。为研究一组变量之间的关系，要收集这些变量在一组单元或案例中的每一个数据。这里研究的回归模型，一个变量起着响应的作用，称为响应变量，而所有其它变量看成是响应的预报因子，称为自变量。我们可以方便地，而且也常是准确地认为，自变量有数据收集者所得的数据值，而把响应变量看作这些自变量的一个函数。除了若干未知参数，对于给定值的自变量，假设模型详细说明了响应变量的行为。模型通常还会指出，由于假设误差项而不能给出准确拟合的某些特征。然后，数据被用于得到未知参数的估计值。尽管存在着多种估计方法，本书中大部分研究的是最小二乘法。这种分析方法称为综合分析，因为其主要目的是将数据聚集在一起，并综合出数据的一个拟合模型。接着，同样重要的回归分析的下一个阶段

是案例分析。这里数据被用于检验拟合模型对被研究的关系是否合适、有用。其结果有可能导致对原先指定的拟合模型的修改。对数据或假设修改以后，回复至综合分析。

本章讨论简单回归，它具有一个自变量和一个响应变量。重点是对一个合适模型的描述，假设的讨论，最小二乘估计，置信区间及检验等过程。

例 1.1　Forbes 数据

在十九世纪四、五十年代，苏格兰物理学家 James D. Forbes，试图通过水的沸点来估计海拔高度。他知道通过气压计测得的大气压可用于得到海拔高度，高度越高，气压越低。在这里讨论的实验中，他研究了气压和沸点之间的关系。由于在 40 年代运输精密的气压计相当困难，这引起了他研究此问题的兴趣。测量沸点将给旅行者提供一个快速估计高度的方法。

Forbes 在阿尔卑斯山及苏格兰收集数据。选定地点后，他装起仪器，测量气压及沸点。气压单位采用水银柱高度，并根据测量时周围气温与标准气温之间的差异校准气压。沸点用华氏温度表示。我们从他 1857 年的论文中选取了 $n=17$ 个地方的数据，见表 1.1（Forbes，1857）。在研究这些数据时，有若干可能引起兴趣的问题，气压及沸点是如何联系的？这种关系是强是弱？我们能否根据温度预测气压？如果能，有效性如何？

Forbes 的理论认为，在观测值范围内，沸点和气压值的对数成一直线。由此，我们取 10 作为对数的底数。事实上统计分析和对数的底数是没有关系的。由于气压的对数值变化不大，最小的为 1.318，而最大的为 1.478，我们将所有气压的对数值乘以 100，如表 1.1 中第 5 列所示。这将在不改变分析的主要性质的同时，避免研究非常小的数字。

着手进行回归分析的一个有效途径是，画一个变量对另一个变量的图。这图称为散点图，它既能用于提示某种关系，也能用于说明这种关系可能是不适当的。散点图可手工在一般作图纸上绘制。X 轴即水平轴，通常留作用于自变量。在 Forbes 的数据中为沸点。Y 轴即垂直轴，通常被用于表示响应变量。在本例中，Y 轴的值为 $100\times\log$（气压）。对 n 对 (x, y) 数据中的每一对，在图上作一个点。大多数回归分析的计算机程序可以作这个图。

Forbes 数据的散点图的总的印象是，这些点基本上，但并不精确地，落在一条直线上。图 1.1 所画的直线将在后面讨论。它指出两个变量之间的关系至少可以初步近似地用一条直线的方程来描述。

在我们学习这一章的过程中，学到的方法都将被用来分析这批数据。

表 1.1　在阿尔卑斯山及苏格兰的 17 个地方沸点(°F)及大气压（英寸汞柱）的 Forbes 数据

案例号	沸　点 (°F)	气　压 (英寸汞柱)	log (气压)	100×log (气压)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

1.1　建立简单回归模型

在简单回归中，两个量，如 X 和 Y 之间的关系将被研究。首先，我们希望这一关系可以用一条直线来描述。为使之合理，我们可能需要变换 X 和（或者）Y 的尺度，如我们在 Forbes 数据中所做的那样，将气压变换成 log（气压）。在本章中，X 和 Y 的观测值将用带下标的小写字母（x_i，y_i）表示，指 X 和 Y 在研究中

的第 i 个案例。这里给出的是简单回归模型的主要特征。更正式的讨论参见附录 1A.1。

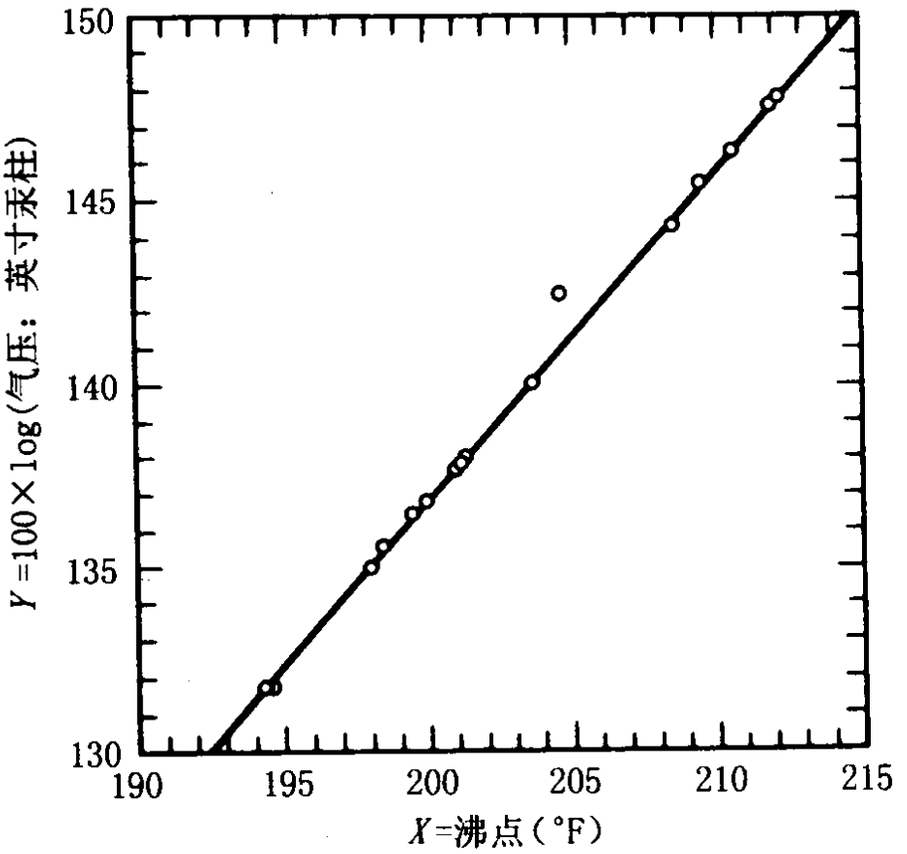

图 1.1　Forbes 数据的散点图

直线方程　关于两个量 X 和 Y 的直线可用方程

$$Y = \beta_0 + \beta_1 X \tag{1.1}$$

表示。在方程（1.1）中，β_0 是截距，它是 X 取 0 时 Y 的值。斜率 β_1 为 X 变化一个单位时 Y 的变化率，见图 1.2。β_0 和 β_1 称为参数，并且因为它们取遍所有可能的值，它们给出所有可能的直线。在大多数统计模型应用中，参数是未知的，并且要通过数据进行估计。

误差　实际数据几乎是从来不会准确地落在一条直线上的。测得的响应变量的值与模型给出的值的差（对简单回归，即 Y 的观测值减去（$\beta_0+\beta_1 \cdot X$）的差）称为统计误差。这一术语不能同通常所用的近义词“错误”相混淆。模型不能给出一个精确的拟