

» 数据分析与模拟丛书

Song S. Qian 著

曾思育 译

Environmental and Ecological
Statistics with R

环境与生态统计

——R 语言的应用

» 数据分析与模拟丛书

Song S. Qian 著

曾思育 译

Environmental and Ecological
Statistics with R

环境与生态统计

——R 语言的应用

HUANJING YU SHENGTAI TONGJI — R YUYAN DE YINGYONG

图字：01-2010-5363 号

Environmental and Ecological Statistics with R

© 2010 by Taylor & Francis Group, LLC

All Rights Reserved.

Authorized translation from English language edition published by

CRC Press, part of Taylor & Francis Group LLC.

Copies of this book sold without a Taylor & Francis sticker on the cover are unauthorized and illegal.

图书在版编目 (CIP) 数据

环境与生态统计：R 语言的应用/钱松著，曾思育译。—北京：高等教育出版社，2011.7

书名原文：Environmental and Ecological Statistics with R

ISBN 978-7-04-031893-7

I. ①环… II. ①钱… ②曾… III. ①环境生态学：环境统计学-统计模型 IV. ①X171.1 ②X11

中国版本图书馆 CIP 数据核字 (2011) 第 121828 号

策划编辑 陈正雄

责任编辑 柳丽丽

封面设计 张楠

版式设计 马敬茹

插图绘制 尹莉

责任校对 杨凤玲

责任印制 张福涛

出版发行 高等教育出版社
社址 北京市西城区德外大街4号
邮政编码 100120
印刷 北京天来印务有限公司
开本 787 mm × 1092 mm 1/16
印张 25.25
字数 470千字
购书热线 010-58581118
咨询电话 400-810-0598

网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>

版 次 2011年7月第1版
印 次 2011年7月第1次印刷
定 价 49.00元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物料号 31893-00

译者序

2010年春节过后，有幸拜读了美国杜克大学钱松（Song S. Qian）教授的专著《Environmental and Ecological Statistics with R》。后来受高等教育出版社的委托，我开始着手翻译这本书。之所以接手这项工作，是因为原著对我有两方面的吸引力，希望能亲自将它翻译出版，让更多的人受益。一方面，因为循序渐进、深入浅出的原则始终贯穿在各个章节中，所以原著能够把原本让人人都发憷的统计学方法讲得有声有色，非常有利于读者的学习和掌握。原著从基本概念讲起，对长期以来很容易被大家混淆和误解的一些概念则花费了更多笔墨，让读者知其然又知其所以然。在介绍每种方法时所选择的案例都非常贴近环境与生态科学领域的研究实践，且具有足够的典型性，容易引起读者兴趣，在不知不觉中得到提高。我现在正给清华大学的本科生讲授“环境数据处理与数学模型”课程，涉及的不少统计分析方法都可以在书中找到极好的例子，完全可以把它用作教学参考书。另一方面，每讲到一种模型，原著都配套描述了如何用R语言来实现。R语言的应用可以让我们摆脱昂贵的商业统计软件，为相关的教学和科研工作提供更多的灵活性。利用R语言来完成环境与生态科学研究中的数据分析和统计建模任务，是该书非常突出的特色。对R语言的讲解详细而实用，使之成为国内目前不可多得的材料。

正是这样一本极具特色和优势的好原著，激励我用了大致一年的时间来完成艰苦的翻译和校对工作，为的是能给读者奉献一本好译著。不过，由于时间和水平有限，翻译过程中难免有疏漏，可能还存在这样或那样的问题，敬请读者批评指正！

曾思育

2011年5月16日

于清华园

前 言

统计学是全世界高等教育机构中几乎所有环境和生态学研究院系和培养项目都开设的课程之一。统计学还常被认为是最不讨人喜欢、教学效果差的科目，尤其是对于数学/统计学领域之外的学生更是如此。学习统计学过程中普遍存在的问题是，统计学常被认定是数学的一个分支领域。因此，我们的期望是学到一套规则并能够将统计学应用到我们的工作中。但是，应用统计学并不是数学。本书致力于在典型的应用统计学和环境与生态学领域对统计学的需求之间架起一座桥梁，重点则放在统计学思考过程中的归纳特征上。内容上避开很多数学和理论背景，通过案例进行概念的介绍和方法的阐释。如同 R. A. Fisher 将统计学引入应用科学那样，本书将统计学当做一种推动科学思考的工具加以介绍。

本书所采用的方法遵循了 R. A. Fisher 关于统计建模的一般步骤，即模型定义、参数估计以及模型评估。这些步骤与研究人员在科研项目中所采取的步骤是相似的。但是，正如很多人讨论的那样，统计学往往不是科学和工程领域的学生们所喜欢的话题 (Berthouex 和 Brown, 1994)，生态学家们也常常会在这上面犯错误 (Peters, 1991)。这是由于在典型的应用统计学课程/教材和典型的科学问题之间缺乏联系。求解一个科学问题时，我们先从潜在机理的假设开始，以之作为收集数据的依据。所提出的假设提供了用公式表达模型的基础，公式中则常常带有未知参数。实验和其他数据收集工作则为估计这些未知参数提供数据。一旦估计出这些参数的值，科研人员就可以通过比较模型的预测值和新的观测值来评价模型。在对科学问题求解过程的简单概括中，第一步(形成假设)往往是最困难的部分，需要研究人员既要有经验又要有创造性。因为一个错误的模型绝不可能把我们带向成功，所以模型/假设的公式表达是这一过程中最重要的步骤。在应用统计学中，正如 R. A. Fisher 所描述的那样，我们所遵循的典型步骤与求解科学问题的步骤是相似的。对于一个具体问题，我们首先必须考察数据，然后提出统计模型来描述我们感兴趣的变量的分布。统计学模型参数化靠的是那些需要利用数据进行估值的未知参数。当完成参数估值后，我们必须通过检验估计出的参数值的抽样分布来评估模型的不确定性。然而，对于研究人员而言，因为困难在于从科学假设到统计学模型的转换，所以科学问题求解和统计学模型开发过程中的相似性并没有解决统计学学

II 前言

习中的困难。典型的应用统计学课程/教材将这一科目当做不同类型统计学模型的方法汇总来讲授,或多或少地忽视了模型公式表达中的问题。因为模型的公式表达必然是一个科学问题,所以这种处理是不可避免的。应用统计学教材或者课程集中于讨论参数估值和模型评估中的问题。不同类型的模型往往需要不同的数学求解。这样对待统计学通常会导致大家错误地去认识什么是统计学和我们为什么要学习统计学。

本书的灵感来自于统计学思考和科学方法之间潜在的联系。本书以统计学模型为基础来组织。但是,通过贯穿全书的案例来讨论每种类型的统计学模型,其中一些例子覆盖了几种模型。这些案例的重点是模型的公式化,背后的数学/统计学理论则大部分被略去了,代之以介绍如何用 R 语言来实现这些模型。本书的基础是我在杜克大学环境学院积累的教学资料。本书划分为 3 个单元。

第 1 章到第 5 章曾被用于研究生水平的应用数据分析课程,可以作为高级统计建模的预备知识来阅读。这些章节的目的是打好基础,以便读者能够开展简单的数据分析工作,如探索性数据分析和拟合线性回归模型等。

第 6 章到第 8 章曾被用于统计建模的后续课程。本单元中的这 3 章之间基本上是相互独立的,可以分别阅读。第 8 章中的 3 个主题(第 8.1—8.4 节、第 8.5 节和第 8.6 节)也是一样的情况。

第 9 章和第 10 章曾被用于博士水平的独立研究课程。第 9 章讨论了如何使用模拟手段进行模型检验,为开发出的模型提供了评估鉴定的工具。模拟方法在参数估值和不确定性评估中是会普遍使用的。虽然在文献中对使用模拟方法来检验模型的讨论并不多,但它是模型开发与评估的重要方面。第 10 章讨论了多层回归模型的应用,这是一类可以对环境和生态学数据分析产生广泛影响的模型。

本书中使用的数据和 R 脚本可以在线获取: <http://www.duke.edu/~song/eeswithr.htm>。

很多人对本书的撰写提供了帮助。Kenneth H. Reckhow、Curtis J. Richardson 和 Michael Lavine 是我的导师和长期合作者。本书反映了他们对我走进环境与生态统计学领域的影响。与 Yandong Pan 的合作加强了我对生态学问题和生态学问题求解过程的认识。Craig A. Stow 不断地给我提供有意思的想法和论文,非常感谢他在分析鱼体内 PCB 数据时所做的工作。Olli Malve、George B. Arhonditsis 和 Andrew D. Gronewold 花了无数小时帮我理清思路和概念。Thomas F. Cuffney 和 Gerard McMahon 给了我 EUSE 的案例,并花了大量时间与我讨论第 10 章中的案例。沈泽昊于 2007 年夏天在北京大学接待了我的访问,并提供了很多有趣的案例。Richard L. Smith 阅读了本书的草稿并提出了批

评意见，从而大大改善了本书的表达，对一些关键概念的讨论也使这些概念变得更为清晰。Meg Mobley、Ibrahim Alameddine、Itai Shelem、Kristen Marine、Emily Sharp、Erin Gray 和 Wyatt Hartman 发现了多处错误，并提出了改进建议。

Song S. Qian

于美国北卡罗来纳州 Durham

2009 年 3 月

目 录

表清单

图清单

第 I 部分 基本概念

第 1 章 引言	3
1.1 美国佛罗里达 Everglades 湿地案例	5
1.2 统计学问题	7
1.3 参考文献说明	10
第 2 章 R 语言	11
2.1 什么是 R 语言?	11
2.2 开始使用 R 语言	11
2.2.1 R 提示符与赋值	12
2.2.2 数据类型	13
2.2.3 R 的函数	15
2.3 R Commander	17
第 3 章 统计假设	24
3.1 正态性假设	24
3.2 独立性假设	28
3.3 等方差假设	29
3.4 探索性数据分析	30
3.4.1 展示分布的图形	30
3.4.2 比较分布的图形	32
3.4.3 识别变量间依存关系的图形	34
3.5 从图形到统计学思维	41
3.6 参考文献说明	43
第 4 章 统计推断	44
4.1 总体均值和置信区间的估计	45
4.1.1 估计标准误的自举法	51
4.2 假设检验	55
4.2.1 t 检验	56

II 目录

4.2.2	双侧备择	62
4.2.3	用置信区间进行假设检验	63
4.3	一般过程	64
4.4	假设检验的非参数方法	65
4.4.1	秩变换	66
4.4.2	Wilcoxon 符号秩检验	66
4.4.3	Wilcoxon 秩和检验	68
4.4.4	关于分布无关检验方法的讨论	69
4.5	置信水平 α 、统计功效 $1-\beta$ 和 p 值	73
4.6	单因素方差分析	80
4.6.1	方差分析	81
4.6.2	统计推断	82
4.6.3	多重比较	85
4.7	案例	89
4.7.1	美国佛罗里达 Everglades 湿地案例	90
4.7.2	Kemp 的鳞龟	91
4.7.3	水质达标评价	96
4.7.4	红树林和海绵体之间的相互作用	99
4.8	参考文献说明	104

第 II 部分 统计建模

第 5 章	线性模型	107
5.1	作为线性模型的 ANOVA	110
5.2	简单和多元线性回归模型	113
5.2.1	最小二乘法	113
5.2.2	鱼样本中的 PCBs	114
5.2.3	用一个预测变量来回归	116
5.2.4	多元回归	118
5.2.5	相互作用	119
5.2.6	残差和模型评估	121
5.2.7	类型预测变量	128
5.2.8	芬兰湖泊案例和共线性	132
5.3	构建预测性模型的一般考虑	140
5.4	模型预测的不确定性	144
5.5	双因素 ANOVA	146
5.5.1	相互作用	151
5.6	参考文献说明	153

第 6 章 非线性模型	154
6.1 非线性回归	154
6.1.1 分段线性模型	162
6.1.2 案例：美国丁香花初次开花的日期	168
6.2 平滑	171
6.2.1 散点图平滑	171
6.2.2 拟合局部回归模型	173
6.3 平滑和加性模型	174
6.3.1 加性模型	175
6.3.2 加性模型的拟合	177
6.3.3 北美湿地数据库	179
6.3.4 讨论：科学中非参数回归模型的作用	182
6.3.5 时间序列的季节分解	186
6.4 参考文献说明	194
第 7 章 分类和回归树	196
7.1 美国俄勒冈 Willamette 河案例	196
7.2 统计学方法	199
7.2.1 种植和修剪一棵回归树	201
7.2.2 种植和修剪一棵分类树	208
7.2.3 绘图选项	212
7.3 讨论	214
7.3.1 将 CART 用做建模工具	214
7.3.2 离差平方和与概率假设	217
7.3.3 CART 和生态阈值	218
7.4 参考文献说明	219
第 8 章 广义线性模型	221
8.1 逻辑斯蒂回归	222
8.1.1 案例：评估将紫外线作为饮用水消毒剂的有效性	223
8.1.2 统计学问题	223
8.1.3 在 R 中拟合模型	224
8.2 模型解释	227
8.2.1 逻辑特变换	227
8.2.2 截距	227
8.2.3 斜率	228
8.2.4 其他的预测变量	228
8.2.5 相互作用	230
8.2.6 对隐孢子虫案例的讨论	231

IV 目录

8.3	诊断学	232
8.3.1	箱式残差图	232
8.3.2	偏大离差	233
8.4	啮齿动物食用种子：逻辑斯蒂回归的第二个案例	235
8.5	泊松回归模型	248
8.5.1	中国台湾西南部的种数据	248
8.5.2	泊松回归	250
8.5.3	暴露和偏移	254
8.5.4	偏大离差	255
8.5.5	相互作用	258
8.5.6	泊松回归与逻辑斯蒂回归	265
8.5.7	负二项分布	267
8.6	广义加性模型	269
8.6.1	案例：西南极半岛的鲸	271
8.7	参考文献说明	280

第 III 部分 高级统计建模

第 9 章	用于模型检验和统计推断的模拟	285
9.1	模拟	285
9.2	用模拟来概括线性和非线性回归	287
9.2.1	一个入门案例	287
9.2.2	概括线性回归模型	290
9.2.3	用于模型评估的模拟	295
9.3	基于重采样的模拟	300
9.3.1	自举聚合	301
9.3.2	案例：基于 CART 的阈值的置信区间	302
9.4	参考文献说明	305
第 10 章	多层回归	306
10.1	多层结构和可交换性	306
10.2	多层 ANOVA	309
10.2.1	食用潮间海藻的动物	310
10.2.2	农田的 N_2O 背景释放量	314
10.2.3	何时使用多层模型？	318
10.2.4	双因素 ANOVA	319
10.3	多层线性回归	326
10.3.1	非嵌套分组	337
10.3.2	多元回归问题	342

10.4 广义多层模型	351
10.4.1 利物浦飞蛾——一个逻辑斯蒂回归案例	351
10.4.2 美国饮用水中的隐孢子虫——一个泊松回归案例	356
10.4.3 采用模拟手段来检验模型	360
10.5 参考文献说明	366
参考文献	367
索引	374

表 清 单

表 3.1	基于模型的百分点和基于数据的百分点	27
表 4.1	ANOVA 表	82
表 4.2	Everglades 湿地数据的样本容量	91
表 5.1	线性模型的 ANOVA 表	123
表 5.2	具有两个类型预测变量的线性模型系数	149
表 6.1	用图 6.11 中数据估计出的分段线性模型系数（及其标准误）	170
表 8.1	种子食用模型中 24 个时间-地形组合的截距	242
表 8.2	饮用水中砷的案例数据	249
表 8.3	砷标准对癌症死亡率的影响	255
表 8.4	性别和癌症类型之间的相互作用	259
表 10.1	芬兰湖泊类型的定义	345

图 清 单

图 3.1	标准正态分布	25
图 3.2	湿地 TP 背景浓度分布	26
图 3.3	用 S-L 图比较标准差	29
图 3.4	湿地 TP 浓度的直方图	30
图 3.5	分位数图的一个例子	31
图 3.6	箱图的解释	32
图 3.7	Q-Q 图中可加的偏移和可乘的偏移	33
图 3.8	双变量散点图	34
图 3.9	散点图矩阵	35
图 3.10	蝴蝶花数据	36
图 3.11	北美湿地数据库的散点图	37
图 3.12	幂变换后的正态性	38
图 3.13	美国巴尔的摩市的 PM2.5 每日浓度	39
图 3.14	美国巴尔的摩市 PM2.5 每日浓度的季节性模式	39
图 3.15	空气质量的条件图	40
图 4.1	模拟中心极限定理	48
图 4.2	样本标准差的分布	50
图 4.3	Everglades 湿地 TP 背景浓度的 75 百分点的分布	50
图 4.4	t 分布	57
图 4.5	α 、 β 和 p 值之间的关系	58
图 4.6	一次双侧检验	63
图 4.7	影响统计功效的因素	74
图 4.8	来自 ANOVA 模型的残差	83
图 4.9	ANOVA 模型残差的 S-L 图	84
图 4.10	ANOVA 残差	84
图 4.11	ANOVA 残差的正态分位数图	85
图 4.12	Everglades 国家公园的年降雨量	90
图 4.13	Everglades 湿地中 TP 浓度的年际变化	91
图 4.14	统计功效是样本容量的函数	98

II 图清单

图 4.15	红树林-海绵体相互影响数据的箱图	100
图 4.16	红树林-海绵体相互影响数据的正态 Q-Q 图	100
图 4.17	红树林-海绵体数据的两两比较	101
图 5.1	鱼体组织中 PCB 浓度的时间演变趋势	115
图 5.2	鱼体组织内 PCB 浓度与鱼的长度	116
图 5.3	PCB 例子的简单回归模型	117
图 5.4	PCB 例子的多元线性回归模型	119
图 5.5	PCB 模型残差的正态 Q-Q 图	125
图 5.6	PCB 模型残差与拟合值	126
图 5.7	PCB 模型残差的 S-L 图	126
图 5.8	PCB 模型的 Cook 距离	127
图 5.9	PCB 模型的 rfs 图	128
图 5.10	修正后的 PCB 模型与拟合值	131
图 5.11	芬兰湖泊案例	133
图 5.12	条件图: 以 TN 为条件, 对叶绿素 a 和 TP 作图 (没有相互作用)	135
图 5.13	条件图: 以 TP 为条件, 对叶绿素 a 和 TN 作图 (没有相互作用)	135
图 5.14	芬兰湖泊案例: 相互作用图 (没有相互作用)	137
图 5.15	条件图: 以 TN 为条件, 对叶绿素 a 和 TP 作图 (正的相互作用)	138
图 5.16	条件图: 以 TP 为条件, 对叶绿素 a 和 TN 作图 (正的相互作用)	138
图 5.17	芬兰湖泊案例: 相互作用图 (正的相互作用)	139
图 5.18	芬兰湖泊案例: 相互作用图 (负的相互作用)	140
图 5.19	响应变量变换的 Box-Cox 图	143
图 6.1	非线性 PCB 模型	156
图 6.2	非线性模型残差的正态 Q-Q 图	156
图 6.3	非线性 PCB 模型残差与拟合出的 PCB	157
图 6.4	非线性模型残差的 S-L 图	157
图 6.5	非线性 PCB 模型的残差分布	157
图 6.6	4 个非线性 PCB 模型	161
图 6.7	模拟出的 2000—2007 年 PCB 减少的百分比	161
图 6.8	曲棍球球棍模型	164
图 6.9	分段线性回归模型	165

图 6.10	为指定年份估计的分段线性回归模型	167
图 6.11	北美丁香花首次开花日期	169
图 6.12	北美丁香花首次开花日期的所有数据	170
图 6.13	移动平均平滑器	173
图 6.14	loess 平滑器	174
图 6.15	多元线性回归模型的图形表达	175
图 6.16	对数变换后的多元线性回归模型的图形表达	176
图 6.17	对数变换后的多元线性回归模型的图形表达	176
图 6.18	鱼体内 PCB 的加性模型	177
图 6.19	平滑参数的影响	179
图 6.20	北美湿地数据库	180
图 6.21	出水浓度-负荷率关系	181
图 6.22	用 mgcv 默认值拟合出的加性模型	182
图 6.23	用 gam 拟合出的双变量平滑器的等值线图	184
图 6.24	用 gam 拟合出的双变量平滑器的三维透视图	184
图 6.25	1 克规则模型	185
图 6.26	利用用户选定的平滑参数拟合出的加性模型	186
图 6.27	来自美国夏威夷 Mauno Loa 的 CO ₂ 时间序列	187
图 6.28	Neuse 河的粪大肠杆菌时间序列	191
图 6.29	Neuse 河粪大肠杆菌时间序列的 STL 模型	192
图 6.30	Neuse 河总磷时间序列的 STL 模型	193
图 6.31	Neuse 河 TKN 的长期趋势	194
图 7.1	蝴蝶花数据的分类树	198
图 7.2	蝴蝶花数据的分类规则	199
图 7.3	Willamette 流域内敌草隆的浓度	201
图 7.4	第一个敌草隆 CART 模型	202
图 7.5	敌草隆 CART 模型的 CP 图	205
图 7.6	修剪后的敌草隆 CART 模型	206
图 7.7	修剪后的敌草隆 CART 模型	207
图 7.8	敌草隆数据的分位数图	209
图 7.9	第一个敌草隆 CART 分类模型	210
图 7.10	敌草隆分类模型的 CP 图	211
图 7.11	修剪后的敌草隆分类模型	211
图 7.12	CART 图选项 1	212
图 7.13	CART 图选项 2	213

IV 图清单

图 7.14	CART 图选项 3	214
图 7.15	4 个敌草隆分类模型	216
图 8.1	剂量-响应曲线	226
图 8.2	逻辑特变换	227
图 8.3	鼠感染数据	229
图 8.4	逻辑斯蒂回归残差	233
图 8.5	箱式残差图	233
图 8.6	种子被食用与种子重量	237
图 8.7	随时间变化的种子食用情况	239
图 8.8	随时间变化的食用率	240
图 8.9	不同时间和种子重量条件下的食用概率	241
图 8.10	种子被食用的概率是种子重量的函数	244
图 8.11	种子重量和地形分类的相互作用	246
图 8.12	种子食用模型的箱式残差图	247
图 8.13	饮用水中砷浓度数据 1	252
图 8.14	饮用水中砷浓度数据 2	252
图 8.15	饮用水中砷浓度数据 3	253
图 8.16	饮用水中砷浓度数据 4	253
图 8.17	加性泊松模型的原始残差和标准化残差	256
图 8.18	拟合出的考虑了偏大离差的泊松模型	261
图 8.19	将年龄同时作为变量时拟合出的偏大离差泊松模型	264
图 8.20	泊松模型的残差	265
图 8.21	南极鲸调查地点	272
图 8.22	南极鲸调查数据散点图	273
图 8.23	南极鲸调查 CART 模型的 CP 图	274
图 8.24	南极鲸调查 CART (回归) 模型	274
图 8.25	南极鲸调查 CART (分类) 模型	275
图 8.26	南极鲸调查的泊松 GAM	277
图 8.27	带有偏大离差的 GAM 的残差	278
图 8.28	南极鲸调查的逻辑斯蒂 GAM	279
图 9.1	2000—2007 年鱼组织内 PCB 的降低预测	294
图 9.2	鱼的尺寸与年份	294
图 9.3	残差作为拟合优度的度量	296
图 9.4	用模拟手段进行模型评估	296
图 9.5	鱼体内 PCB 案例的尾部面积	297