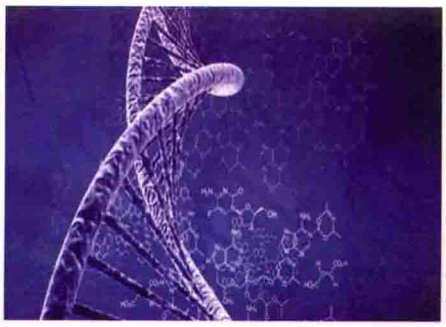
环保公益性行业科研专项经费项目系列丛书

R and Its Applications
in Environmental Epidemiology



R 软件

及其环境流行病学应用

彭晓武　王家春　余松林　著

中国环境出版社

环保公益性行业科研专项经费项目系列丛书
课题：珠江三角洲地区灰霾天气对人群健康影响评估影响（200709004）

R软件及其环境流行病学应用

彭晓武　王家春　余松林　著

中国环境出版社·北京

图书在版编目（CIP）数据

R 软件及其环境流行病学应用/彭晓武，王家春，余松林著．—北京：中国环境出版社，2013.9
ISBN 978-7-5111-1470-9

Ⅰ．①R… Ⅱ．①彭… ②王…③余… Ⅲ．①统计分析—应用软件 Ⅳ．①C819

中国版本图书馆 CIP 数据核字（2013）第 107103 号

出 版 人 王新程
责任编辑 李卫民
责任校对 唐丽虹
封面设计 刘丹妮

出版发行 中国环境出版社
（100062 北京市东城区广渠门内大街 16 号）
网 址：http：//www.cesp.com.cn
电子邮箱：bjgl@cesp.com.cn
联系电话：010-67112765（编辑管理部）
发行热线：010-67125803，010-67113405（传真）
印 刷 北京中科印刷有限公司
经 销 各地新华书店
版 次 2013 年 9 月第 1 版
印 次 2013 年 9 月第 1 次印刷
开 本 787×1092 1/16
印 张 34.5
字 数 832 千字
定 价 80.00 元（附光盘）

《环保公益性行业科研专项经费项目系列丛书》

编著委员会

“十一五”环保公益性行业科研专项经费项目系列丛书

序　言

我国作为一个发展中的人口大国，资源环境问题是长期制约经济社会可持续发展的重大问题。党中央、国务院高度重视环境保护工作，提出了建设生态文明、建设资源节约型与环境友好型社会、推进环境保护历史性转变、让江河湖泊休养生息、节能减排是转方式调结构的重要抓手、环境保护是重大民生问题、探索中国环保新道路等一系列新理念新举措。在科学发展观的指导下，“十一五”环境保护工作成效显著，在经济增长超过预期的情况下，主要污染物减排任务超额完成，环境质量持续改善。

随着当前经济的高速增长，资源环境约束进一步强化，环境保护正处于负重爬坡的艰难阶段。治污减排的压力有增无减，环境质量改善的压力不断加大，防范环境风险的压力持续增加，确保核与辐射安全的压力继续加大，应对全球环境问题的压力急剧加大。要破解发展经济与保护环境的难点，解决影响可持续发展和群众健康的突出环境问题，确保环保工作不断上台阶出亮点，必须充分依靠科技创新和科技进步，构建强大坚实的科技支撑体系。

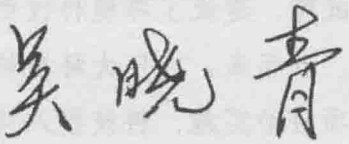
2006年，我国发布了《国家中长期科学和技术发展规划纲要（2006—2020年）》（以下简称《规划纲要》），提出了建设创新型国家战略，科技事业进入了发展的快车道，环保科技也迎来了蓬勃发展的春天。为适应环境保护历史性转变和创新型国家建设的要求，原国家环境保护总局于2006年召开了第一次全国环保科技大会，出台了《关于增强环境科技创新能力的若干意见》，确立了科技兴环保战略，建设了环境科技创新体系、环境标准体系、环境技术管理体系三大工程。五年来，在广大环境科技工作者的努力下，水体污染控制与治理科技重大专项启动实施，科技投入持续增加，科技创新能力显著增强；发布了502项新标准，现行国家标准达1 263项，环境标准体系建设实现了跨越式发展；完成了100余项环保技术文件的制修订工作，初步建成以重点行业污染防治技术政策、技术指南和工程技术规范为主要内容的国家环境技术管理体系。环境

科技为全面完成“十一五”环保规划的各项任务起到了重要的引领和支撑作用。

为优化中央财政科技投入结构，支持市场机制不能有效配置资源的社会公益研究活动，“十一五”期间国家设立了公益性行业科研专项经费。根据财政部、科技部的总体部署，环保公益性行业科研专项紧密围绕《规划纲要》和《国家环境保护“十一五”科技发展规划》确定的重点领域和优先主题，立足环境管理中的科技需求，积极开展应急性、培育性、基础性科学研究。“十一五”期间，环境保护部组织实施了公益性行业科研专项项目234项，涉及大气、水、生态、土壤、固废、核与辐射等领域，共有包括中央级科研院所、高等院校、地方环保科研单位和企业等几百家单位参与，逐步形成了优势互补、团结协作、良性竞争、共同发展的环保科技“统一战线”。目前，专项取得了重要研究成果，提出了一系列控制污染和改善环境质量技术方案，形成一批环境监测预警和监督管理技术体系，研发出一批与生态环境保护、国际履约、核与辐射安全相关的关键技术，提出了一系列环境标准、指南和技术规范建议，为解决我国环境保护和环境管理中急需的成套技术和政策制定提供了重要的科技支撑。

为广泛共享“十一五”期间环保公益性行业科研专项项目研究成果，及时总结项目组织管理经验，环境保护部科技标准司组织出版“十一五”环保公益性行业科研专项经费项目系列丛书。该丛书汇集了一批专项研究的代表性成果，具有较强的学术性和实用性，可以说是环境领域不可多得的资料文献。丛书的组织出版，在科技管理上也是一次很好的尝试，我们希望通过这一尝试，能够进一步活跃环保科技的学术氛围，促进科技成果的转化与应用，为探索中国环保新道路提供有力的科技支撑。

中华人民共和国环境保护部副部长



2011年10月

序

目前国际上有许多很好的统计软件，如 SAS（Statistical Analysis System），SPSS（Statistical Package for the Social Sciences，现更名为 Statistical Product and Service System）等。虽然它们都具有强大的统计计算功能，并得到国际公认，但都属于商业软件，对于经济欠发达国家的用户来说难以承受，而且这些软件中模块的更新较慢。

R 软件是一种免费软件，具有强大的统计计算和绘图功能，最初由新西兰奥克兰大学统计学系的 Robert Gentleman 和 Ross Ihaka 编写。其源代码于 1995 年公布于众。由于该软件具有自编程序的功能，许多统计学者不断添加自编的新程序，使其功能日益完善，受到全世界统计工作者的青睐。本书的目的在于介绍该统计软件，为使用者提供一种新的软件选择。

随着经济的快速发展，环境保护越来越受到社会的重视，环境与健康关系的流行病学研究日益深入。本书向广大的环境流行病学工作者推荐应用 R 软件作为计算工具，书中内容包括处理独立观察资料的常用统计方法和处理非独立观察资料的时间序列方法。

本书的内容按由浅入深、循序渐进的顺序编排。全书内容包括三部分：第一部分为 R 基础，第二部分为常用统计方法，第三部分为时间序列分析方法。各部分的内容安排为：R 基础包含 R 软件的安装、窗口的介绍和 R 的基本工作原理，R 的数据运算和数据集操作，R 的绘图和编程等。常用统计方法包括定量变量的描述性统计，分类变量的描述性统计，区间估计与假设检验（t 检验，χ^2 检验，单向和双向方差分析等），二项分布与泊松分布，生存率的计算和比较方法，线性回归与相关，Logistic 回归，Cox 比例风险模型等。时间序列分析方法包括时间序列的经典分析方法，平稳时间序列分析方法，非平稳时间序列分析方法，季节非平稳时间序列模型，带输入变量的时间序列模型和泊松分布广义加性模型，以及经济效应分析等。

初学者在开始接触 R 软件时往往会感到不知所措，为了给大家的学习提供方便，我们编写了各部分的示例程序（R 代码），大家可以在其导引下，更容易地学习和理解书中的内容。由于对一个新软件的学习与掌握是一个从不熟悉到熟悉的过程，建议读者仔细阅读 R 基础，并在计算机上进行大量的操作练习。为了顺利地学习和掌握第三部分的时间序列分析方法，读者应具有一定的统计学基础。虽然从学习统计学的角度来说，已具备坚实的基本统计知识的读者可以跳过第二部分而直接开始学第三部分。但是从学习 R 软件的角度来说，我们建议从长远考虑，最好还是全书完整地学习一遍，以便于系统地掌握 R 语言的基础知识。

本书为环境流行病学工作者介绍 R 软件，使之应用于环境污染与健康关系的研究。鉴于本书内容的系统性和叙述的详尽性，并有配套的案例分析，亦可作为一般统计工作者的参考书以及高等院校的统计学软件教材。

在本书的编写过程中，得到华中科技大学同济医学院公共卫生学院领导和环境保护部华南环境科学研究所领导的关心和支持，得到了我们的很多同事、朋友以及家人的关心、鼓励和帮助，谨此致以谢意。崔伊薇教授还对本书的时间序列分析部分做了仔细阅读和认真修改，特此一并致谢。还要特别感谢李卫民编辑，她为本书的编审工作付出了辛勤劳动。她不仅对书稿进行了认真细致的逐字逐句审核，还帮助作者发现了许多原书稿中的错误和不足之处。本书的出版与她的努力是分不开的。

当今统计学和相应软件的发展一日千里。限于作者的学术水平，本书难免存在许多错误和不足之处，诚望广大读者给予批评指正，以便再版时予以修正。

本书中大部分例子引自所附参考文献，特别是引自余松林主编的《医学统计学》（人民卫生出版社，2002 年）。谨向相关编委王洪源教授、王增珍教授、宇传华教授、张菊英教授、周燕荣教授、骆福添教授和曹素华教授致以谢意。

本书得到环境保护部环保公益性行业科研专项经费项目“珠江三角洲地区灰霾天气对人群健康影响评估研究”（200709004）的资助。

彭晓武　王家春　余松林

2013 年 3 月 2 日

目 录

第一篇 R基础

第二篇 常用统计方法

第三篇 时间序列分析方法

第一篇　R 基础

第1章 绪 论

1.1 R的起源和发展

R是一种统计计算和制图的语言或环境，同时又是一款统计软件。R最初是20世纪80年代问世的S统计计算语言系统的一个分支，因此可以把R看成S语言的一种补充。所以很多用S语言编写的程序代码可以直接在R环境中运行。R语言最先由新西兰奥克兰大学统计系的Robert Gentleman和Ross Ihaka合作创建，其源代码于1995年公布于众。R语言的基本参考书是由Richard A. Becker，John M. Chamber和Allan R. Wilks所著的*The New S Language: A Programming Environment for Data Analysis and Graphic*。与S语言一样，R语言在设计理念上与其他统计软件系统存在着许多不同。虽然R出自S，但R与S也还是有一些不同之处，读者不要混淆：

①在R中，统计分析通常按一系列步骤来完成，并将部分统计分析结果直接显示在屏幕上，某些中间结果（如P值、回归系数、残差等）保存在所谓的“对象”中。②做模型拟合时，SAS软件和SPSS软件会直接给出可以复制的输出结果，而R则将其计算结果保存在一个“拟合对象”中，需要用另外的R语句（称作“函数")来访问，直接输出的结果很少。③SAS软件和SPSS软件对英文字母的大小写是不区分的，而R则是区分字母的大小写的，即如果指定了大写字母，而你用了小写字母的话，程序将停止运行并提示错误。

R是在GNU协议（General Public Licence）下免费发行的。它的开发及维护工作由R开发核心小组（R Development Core Team）具体负责。由科学家维护，为科学家服务。R自公布以来，很快赢得了很多用户，并不断有贡献者加入新的程序包，以改善和扩展R的功能。软件及其程序包的版本更新也是很快的，几个月内就有新的版本出现。

1.2 R的功能和特点

R内部包含了许多实用的统计分析函数和作图函数。作图函数能够在一个独立的窗口中展示它所产生的图形。用户还可以将这些图形保存为（甚至直接输出进）各种格式的图形文件（如bmp、emf、jpg、png、ps、pdf、tif等）。统计分析结果也能够直接显示出来，一些中间结果（如P值，回归系数和残差等）既可以用作进一步分析的对象，也可以保存到指定的文件。

在R语言中，用户既可以使用循环语句来连续分析多个数据集，也可以将多个不同的统计函数放在同一个语句中，执行更为复杂的分析。用户还可以借鉴网上提供的用S语言编写的大量的程序，而且大多数都能被R直接调用。

用户在初学阶段可能会觉得R比较复杂多变，令人琢磨不定。其实，R的一个突出的优点正是它的灵活性。一般的统计软件通常会直接展示其分析结果，而R则是把这些结果

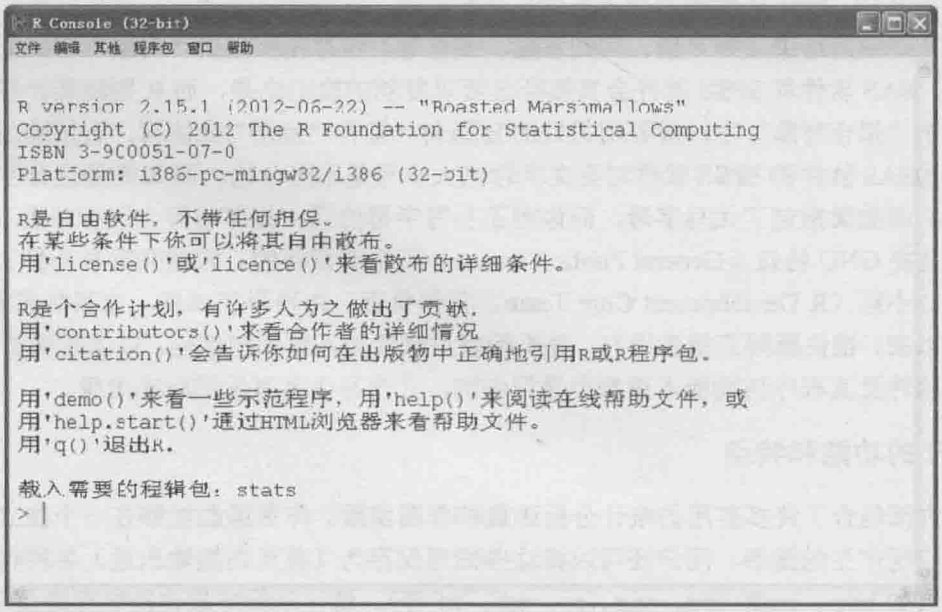
保存在对象(object)中，所以常常在分析执行结束后并不或很少显示任何结果，初学者可能会对此感到困惑。其实这个特点是很有用的，因为用户随时可以从大量的结果中选择性地提取出感兴趣的部分。例如，我们要运行 20 个回归分析而只想比较其回归系数，在 R 中就可以选择只显示所有分析所得出的回归系数，这样，结果就只占了一排，而用别的软件就可能要一下打开 20 个窗口。在下面的章节中，我们会看到更多 R 相对于传统软件更为灵活和优越的例子。

1.3 R 软件的获取与安装

R 由“综合 R 档案馆网站”（Comprehensive R Archive Network，CRAN）发布。下载网址为 http://cran.r-project.org/mirrors.html。到本书完成初稿时为止，R 的 Windows 版本为 R 2.15.1，大小约为 47MB。

R 的安装过程很简单，登录网站 http://cran.r-project.org → download R → 选择一个靠近你的地区（如 China 下的 http://mirrors.ustc.edu.cn/CRAN/）→ Download R for Windows → base → Download R 2.15.1 for Windows → 指定一个文件夹以保存所下载的 R 安装文件 → 双击该文件以开始安装 R 软件 → 遵循自动安装中的提示指引 → 完成 R 的安装。

安装完成后，双击桌面上 R 软件图标即可启动 R。显示控制台（R Console）窗口如下。



在以上窗口中，最底下的一行为红色的“>", 后边还有一个红色的“|”在闪动，这个“>”就是 R 提示符，它的右边如果没有任何显示内容，则表明 R 正在等待你的命令（这时候可以键入命令，然后按回车键以执行此命令）。如果窗口中最底下的一行不是提示符，则表明R正在执行命令(这时候需要你等待R,直到窗口中最底下的一行出现提示符为止)。

从以上过程不难看出，我们使用 R 的过程就是我们与 R 对话的过程（就像使用 QQ 那样），这就是 R 的交互式操作。

当控制台窗口中的内容不再需要了，我们可以按 Ctrl-L 键，或用右键单击窗口内部的