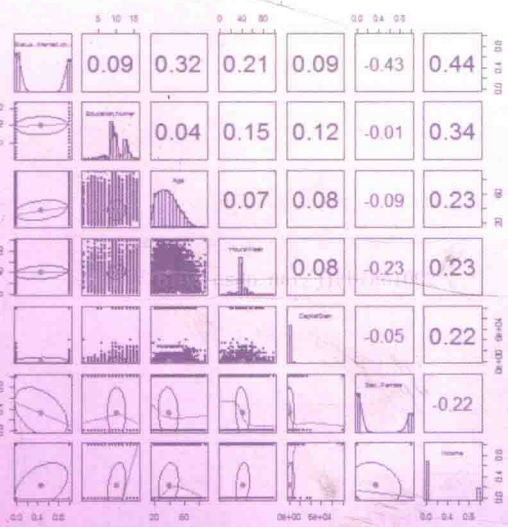
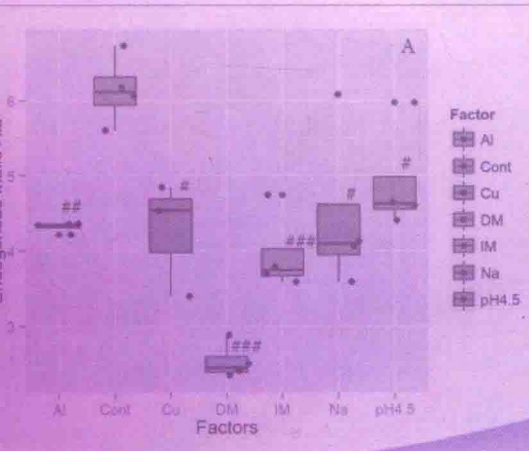
国家林业局普通高等教育“十三五”规划教材

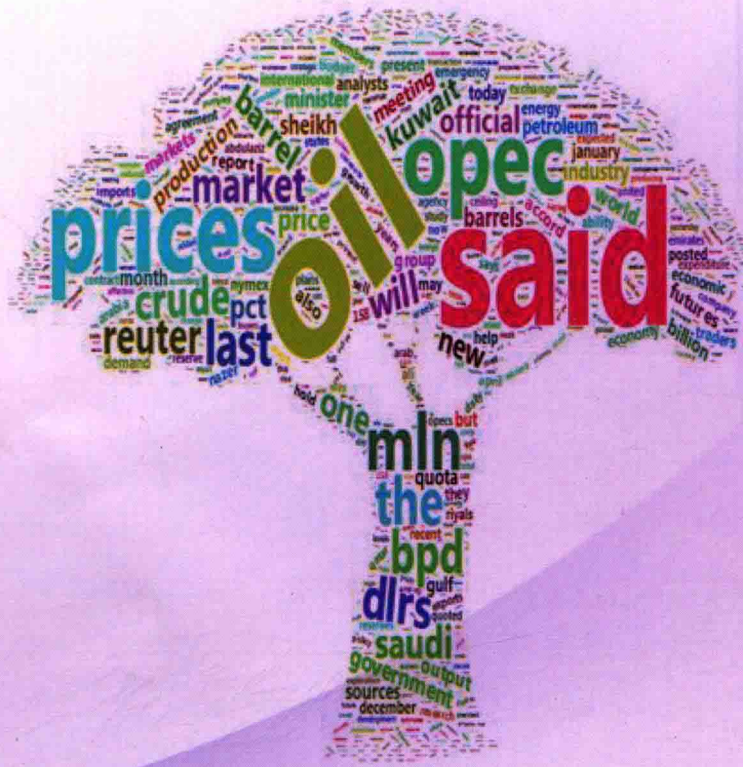
R与ASReml-R统计学

R AND ASREML-R STATISTICS

林元震 ■ 主编
张卫华 郭 海 ■ 副主编







中国林业出版社

国家林业局普通高等教育“十三五”规划教材

R 与 ASReml-R 统计学

林元震　主编
张卫华　郭　海　副主编

中国林业出版社

内 容 简 介

R 语言近年来成为统计分析的最受欢迎软件之一，已广泛用于生态、金融、统计、互联网、医疗和农林牧渔等行业，并涉及大数据、生物信息学以及人工智能等领域。本书主要面向农林业试验数据，系统介绍了 R 与 ASReml - R 的统计应用，全书共分 11 章，具体包括 R 语言简介、基础语法、数据创建、数据管理、基础统计、高级统计、试验设计、基础绘图、高级绘图、遗传评估和程序包开发。本书内容新颖，覆盖面广，应用性强，而且章节合理、结构清晰、行文规范，适用于林学类、植物生产类、生物科学类、草学类、医学类等专业本科生的统计分析教材，也可供相关专业的研究生和科研工作者参考使用。

图书在版编目（CIP）数据

R 与 ASReml-R 统计学/林元震主编. —北京：中国林业出版社，2016.12
国家林业局普通高等教育“十三五”规划教材
ISBN 978-7-5038-8869-4

Ⅰ.①R… Ⅱ.①林… Ⅲ.①统计分析 - 统计程序 - 高等学校 - 教材 Ⅳ.①C819

中国版本图书馆 CIP 数据核字（2016）第 305347 号

国家林业局生态文明教材及林业高校教材建设项目

中国林业出版社·教育出版分社

策划编辑： 肖基浒　　**责任编辑：** 高兴荣　肖基浒

电　　话：（010）83143555　　**传　　真：**（010）83143561

出版发行　中国林业出版社（100009　北京市西城区德内大街刘海胡同 7 号）
E-mail：jiaocaipublic@163.com　电话：（010）83143500
http：//lycb.forestry.gov.cn

经　　销　新华书店
印　　刷　北京昌平百善印刷厂
版　　次　2017 年 1 月第 1 版
印　　次　2017 年 1 月第 1 次印刷
开　　本　850mm × 1168mm　1/16
印　　张　38.5
字　　数　920 千字
定　　价　68.00 元

《R 与 ASReml-R 统计学》编写人员

主　编　林元震（华南农业大学）

副主编　张卫华（广东省林业科学研究院）

郭　海（高原圣果沙棘制品有限公司）

编　委（按姓氏笔画排序）

王　锋（广东省龙眼洞林场）

朱航勇（哈尔滨林业科学研究院）

苏　艳（华南农业大学）

吴元奇（四川农业大学）

陆钊华（中国林科院热带林业研究所）

林露湘（中科院西双版纳热带植物园）

欧阳昆唏（华南农业大学）

罗昊澍（中国农业大学）

周　玮（华南农业大学）

赵曦阳（东北林业大学）

钮世辉（北京林业大学）

骈瑞琪（华南农业大学）

Preface

I have beena SAS software user since 1992. I have used it for research as well as for teaching. Although it is great software for data manipulation and statistical analyses, I have been using R and ASReml more and more in recent years. There are several reasons for this.

First, the algorithm used by SAS procedures, such as MIXED and GLIMMIX are not efficient (slow) to analyse large-scale genetic data. They are not as flexible to fit complex variance-covariance structures in mixed models. Second, molecular data have become readily available for my research. We expect the amount of molecular data increase substantially in coming years. Companies producing SAS and SPSS have difficulty meeting the growing demand for different algorithms for large-scale genetic data. Third, SAS and some other commercially available software are typically not affordable for many individual scientists and professionals. Students and professionals are expected to use freely available software instead of paying large annual license fees.

ASReml software has become an industry standard to fit mixed models. It is very powerful and flexible. Its algorithm can solve large number of mixed models in a fraction of time what SASprocedures requires. ASReml is not a free software but the cost is not steep as SAS. I have been using ASReml for the last 15 years for research and teaching and I recommend it.

In recent years we have witnessed a growing interest in freely available software, specifically R. The software is open source. This is a great way to invite thousands of volunteers to contribute to the program. As of June 2016, R had more than seven thousand packages and this number is growing every year. The packages in R environment are sometimes developed for very specific needs, which is important for research. I have included R in my teaching in recent years. However, it has been the primary software for genetic data analyses in my research.

I have to admit that my understanding of Mandarin language is limitedto a few words. When Yuanzhen Lin asked me to write a preface, I hesitated. However, after scanning though the book, I was pleasantly surprised with the in depth coverage of R and ASRemlR. The book includes large number of screen shots of R scripts and ASReml scripts which makes it easier to follow. It looks like the scripts and output are interpreted in detail. It is obviously a product of hard work. I am sure readers will greatly benefit from using this book to improve their R and ASReml skills.

July 2016

Fikret Isik, Professor
North Carolina State University, Raleigh, US

序

从1992年起我就是SAS用户，不仅用于科研，也用于教学。虽然SAS是数据管理和统计分析方面的先锋软件，但近年来我逐渐使用R和ASReml软件，原因如下：

首先，SAS模块（比如MIXED、GLIMMIX）在处理大规模遗传数据时效率低，而且在拟合混合模型的复杂方差协方差结构时不够灵活。其次，分子数据已用到我的研究中，并且预计未来分子的数据量将大幅增加。而开发SAS和SPSS的公司难以满足针对大规模遗传数据的不同算法的不断增长需求。第三，SAS和其他一些商业软件对许多科学家和专业人员来说，通常并不实惠。学生和专业人员更倾向于使用免费的软件，而非每年支付高额的软件使用费。

ASReml软件已成为拟合混合模型的行业标准。ASReml非常强大和灵活。在求解大规模的混合模型时，ASReml的运算时间远少于SAS程序所需要的时间。ASReml不是免费的软件，但其费用远低于SAS。这15年的研究和教学中，我一直使用ASReml，因此我也推荐使用ASReml。

近年来，我们已目睹了人们对免费软件日益增长的兴趣，尤其是R。R是开源软件，有数千名志愿者为R作贡献，这是一个非常棒的方式。截至2016年6月，R有七千多个程序包，而且每年还在增加。R中的程序包有时是专门为一些特殊需求开发的，这对于科学研究来说非常重要。我最近几年也将R用到教学中，R一直是我分析遗传数据的主要软件。

我得承认，自己对中文的理解比较有限。当林元震邀请我为此书做序时，我有点犹豫。然而浏览此书后，我对本书在R和ASReml-R方面的深入覆盖面，感到惊喜。书中含有大量R程序和ASReml程序，这便于读者学习。同时，该书对程序和程序结果做了详细的解答，可见作者颇费功夫。我相信读者通过这本书来提高R和ASReml技能时，将受益颇丰。

2016年7月

菲克里特·艾斯克教授
北卡罗来纳北卡罗莱纳州立大学

附注：此序为Isik所写preface的中文版，由林元震博士翻译，如有不当之处，敬请读者批评指正。

前　言

近些年国内 R 语言会议的参加人数变化即可看出 R 语言在国内日趋热门，2012 年约为 400 人，2013 年约为 600 人，2014 年约为 1400 人，2015 年达到 4200 人，据统计 2016 年参会人员将突破 1 万人，俨然是国内规模较大的专题会议之一。R 语言会议地点也从最早的北京，到上海、深圳、广州，慢慢拓展到各省会城市。R 语言现已渗透在国内的生态、金融、统计、互联网、医疗和农林牧渔等行业，且在大数据、生物信息学以及人工智能等领域大展身手。正如笔者在《R 与 ASReml-R 统计分析教程》前言中所写的“R 语言在数据挖掘和可视化应用领域的快速崛起意味着 R 语言已经为大数据时代做好准备”，从 R 语言在国内的应用领域来看，已然得到佐证。

大约 3 年以前，笔者组织编写了农林领域第一部有关 R 与 ASReml-R 软件的“十二五”规划教材——《R 与 ASReml-R 统计分析教程》，该教材在业界内获得一定的好评。但正如 R 语言的迅猛发展一样，该教材的匹配的章节不够齐全、部分内容亟需更新，加之近年来比较热门的基因组选择，上述原因正是编写本书的动力所在。

与笔者编写的第一部农林领域教材一样，对阅读本书的读者，没有统计编程或 R 语言背景的要求，当然读者如有 R 语言基础知识将会更好地理解、掌握本书的知识点。本书结构已完全不同于《R 与 ASReml-R 统计分析教程》教材，在本书中，总共包含 11 章，且每章都附有思考题。

本书的第 1 ~ 3 章介绍 R 语言、基础语法和数据创建，让读者对 R 语言有一些直观的概念，了解 R 及其语法的特点，熟悉 R 中数据类型及其创建，这些对于后续的数据管理、统计分析以及图形绘制等操作是必需的。

第 4 章介绍了数据管理的各种操作，包括数据转换、排序、合并、重构、分段、汇总、查重以及子集提前，重点介绍了数据综合处理包 dplyr 包和 data. table 包的用法。熟练掌握数据管理的各种操作对于统计分析和图形绘制非常重要。

第 5、6 章较全面介绍了 R 的基础统计和高级统计，其中基础统计包括描述性统计、频数表分析、方差分析、协方差分析、t 检验、卡方检验、线性回归、相关分析和通径分析，高级统计包括广义线性模型、生长模型、生存分析、主成分分析、因子分析、聚类分析、判别分析、功效分析、重抽样和综合评价分析。

第 7 章专门介绍了 R 的试验设计和数据分析，设计类型包括完全随机设计、随机区组设计、平衡不完全区组设计、拉丁方设计、正交设计、裂区设计、巢式设计、析因设计、循环设计、格子设计、α 设计和条区设计，并介绍了各种设计的基本概念、R 出设计表以

及数据分析的过程。

第 8 章介绍了 R 的基础绘图，包括条形图、直方图、散点图、热图、散点图矩阵等常见图形，并介绍了绘图参数的设置，以及数学公式、文本的添加。此外，还展示了交互图形的绘制。

第 9 章重点演示了 R 包 lattcie 和 ggplot2 的高级绘图，其中 lattcie 包绘图包括基础语法、单变量绘图、双变量绘图、多变量绘图以及高级绘图参数的设置，ggplot2 包绘图包括基础语法、各种图形绘制以及高级绘图参数的设置。本章节是 R 绘图优势和强大功能的展现。

第 10 章介绍了 R 包在遗传评估上的应用，重点介绍了 MCMCglmm 包和 ASReml-R 包。尤其是 ASReml-R 包，作为商业软件包，已广泛应用于农林牧渔、生态等各行业。在本章节中，特别演示了 ASReml-R 包在单性状模型、双性状模型、模型比较、阈性状模型、泊松分布型模型、协变量模型以及批量分析的基础用法，也拓展了遗传参数评估（遗传力、育种值、遗传相关与遗传增益）的各种类型，包括子代测定、无性系测定、空间分析（规则与不规则）、多地点 G × E 分析、多年份分析、多交配分析、多世代分析以及基因组选择。本章节对于动植物遗传试验的数据分析具有较重要的参考价值。

第 11 章介绍了 windows 系统下的 R 包开发，包括所需软件、函数编写及 R 包制作，并专门演示了笔者自编程序包 AAfun 的一些功能。本章的目的是让读者了解 R 包的开发流程，希望有更多的 R 读者加入到程序包的开发中，更好、更快地促进 R 在各领域中的应用。

附录部分给出了索引、网络资源，便于读者进一步查询或学习 R 语言的相关知识。与之前那部教材一样，本书继续秉着 R 开源免费的精神，将本书中所有的数据、代码和彩图存放于网盘 http：//yzhlin-asreml. ys168. com/，供读者免费下载、自由使用。

最后，笔者要衷心感谢美国北卡罗来纳州立大学的 Fikret Isik 教授，Isik 教授是国际知名的遗传统计学家，感谢他百忙之中欣然为本书作序。此外，也要特别感谢瑞典农业大学的合作导师 Harry Wu 教授以及 ASReml 的软件开发者 Arthur Gilmour，他们对于我在 R 与 ASReml-R 的学习路程上起着不可磨灭的推动作用。

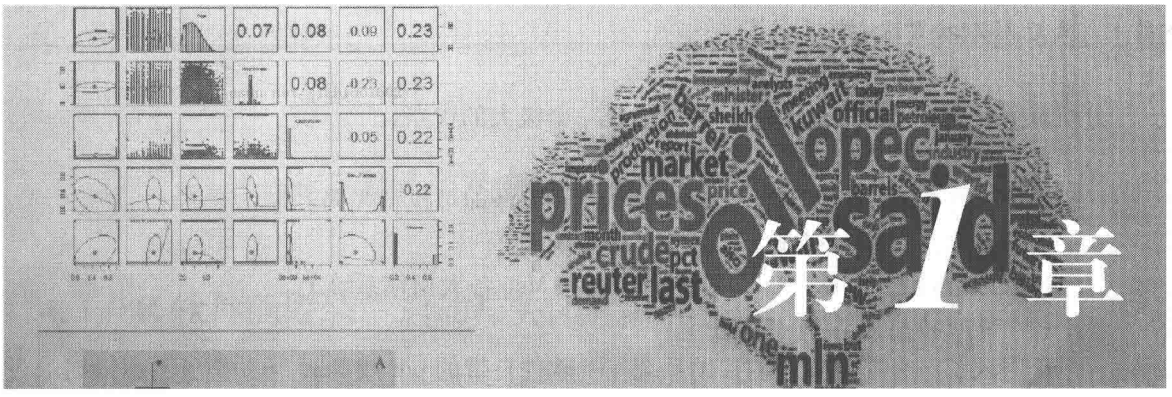
本书由广东省高水平大学经费（4400—216202）资助出版，特此谢忱！

由于编者的知识水平有限，书中难免会有疏漏和不足，恳请广大读者批评指正。如对本书有任何建议或意见，请发送邮件到 yzhlinscau@ 163. com。

林元震

2016 年 6 月

目　录



R 简介

1.1 R 语言

R 既是软件，也是语言，在 GNU 协议(General Public License)下免费发行，是 1995 年由新西兰奥克兰大学统计系的 Ross Ihaka 和 Robert Gentleman 基于 S 语言基础上共同开发的一种统计软件。现在由 R 开发核心小组(R Development Core Team)负责维护与更新，并将全球优秀的统计应用程序包免费提供给大家使用、共享。

R 是一套由数据操作、计算和图形展示功能整合而成的软件系统，包括：

①有效的数据存储和处理功能；

②一套完整的数组(特别是矩阵)计算操作模块；

③拥有完整体系的数据分析工具；

④为数据分析和显示提供强大的图形功能；

⑤一套完善、简单、有效的编程语言(包括条件、循环、自定义函数和输入输出功能等)。

1.2 R 的特点

现在越来越多的人开始学习和使用 R，因为 R 有以下几个显著的优点。

(1)免费且开源

目前市面上有许多流行的统计和制图软件，如 Microsoft Excel、SAS、SPSS、MiniTab 以及 Original 等，但这些多属商业软件，并且费用昂贵。但是 R 是一个免费且开源的统计