高等院校研究生用书

TEXTBOOKS FOR GRADUATES

复杂数据统计方法
——基于R的应用

吴喜之 编著

中国人民大学出版社

高等院校研究生用书
TEXTBOOKS FOR

复杂数据统计方法
——基于R的应用

吴喜之　编著

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

复杂数据统计方法：基于 R 的应用 / 吴喜之编著. — 北京：中国人民大学出版社，2012. 9

高等院校研究生用书

ISBN 978-7-300-16399-4

Ⅰ. ①复… Ⅱ. ①吴… Ⅲ. ①统计分析－应用软件－研究生－教材 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2012）第 220059 号

高等院校研究生用书

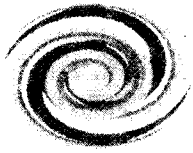
复杂数据统计方法

——基于 R 的应用

吴喜之 编著

Fuza Shuju Tongjifangfa

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	邮政编码	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京民族印务有限责任公司		
规　　格	170 mm × 228 mm　16 开本	版　　次	2012 年 10 月第 1 版
印　　张	15　插页 1	印　　次	2012 年 10 月第 1 次印刷
字　　数	248 000	定　　价	33.00 元



前　言

什么是复杂数据? 没有人能够确切定义. 本书将通常统计基本教科书中的例子所代表的数据称为简单数据, 例如通常最小二乘线性回归所能够完满处理的独立同正态分布数据、用标准多元分析方法能够处理的服从多元正态分布的数据等. 其他本科教科书中能够相对完满处理的数据应该不算复杂数据. 显然, 现实世界中遇到的绝大多数数据都不是标准教科书中所介绍的方法能够完满处理的, 因此都应该被认为是复杂数据. 按照这个含义, 绝大多数真实数据是复杂数据.

对于一个实际工作者来说, 拿到一个真实数据以后, 很可能需要查阅不少文献来寻找适合这个数据的几种可能模型 (假定知道用什么模型可能解决问题), 再翻阅若干种软件手册来查阅这些文献所使用软件的计算方法 (假定购买了这些软件). 造成这种情况的原因是, 多数统计教科书是以模型或方法为导向的, 内容也多是按照数学思维展开的.

以模型或方法为导向的教科书通常以介绍某种数学模型和方法为主, 同时说明这种模型适用于满足某些数学假定的数据, 最后说明该模型对于这些满足假定的数据拟合的优越性. 实际上, 任何一种真实数据是否满足某种数学假定几乎无法证明, 每一类数据都可能有不止一种现成的统计方法来处理, 还有无数的未知方法等待人们去开发. 以模型或方法为主导的方式往往让读者忽略了其他有关的方法, 而那些被忽略的方法在某种意义下很可能更有效, 或者更优越.

笔者认为, 现在需要一本具有以下特点的书:

- 用实际数据做案例.
  - 介绍的数据种类尽可能广泛;
  - 这些数据必须是真实的;
  - 这些数据必须不是简单平凡的教科书例子;
  - 每个数据都有理论及应用方面的背景;
  - 所有数据都能从网上下载.
- 对每种数据都介绍可能的方法.
  - 这些方法尽可能新;
  - 对各种方法进行比较;
  - 所有方法必须有计算支持.
- 全书使用一种软件.
  - 该软件必须是免费的, 可以从网上下载的;
  - 该软件必须能够包含尽可能多的最新统计方法;
  - 该软件必须不断更新;
  - 书中所有结论都可以通过运行该软件程序而得出, 并给出所有代码.
- 篇幅不能太大.
- 必须由浅入深, 对经典知识和模型进行必要的回顾.
- 不能有太多数学公式, 但至少必须让读者能直观理解各种方法的含义.
- 其宗旨是训练动手的能力, 而不是面面俱到地告诉人们所有细节.
- 不仅提供各种方法, 而且提醒人们使用各种方法存在的风险.

本书以数据形式为导向, 对应不同的数据形式介绍可能使用的一些方法. 首先引入某些感兴趣类型的数据, 再介绍并且对比可能适合这些数据的一些统计方法. 这些统计方法可能属于许多不同的模型, 属于不同的统计方向, 但只要适用于同一类数据, 我们就尽量将它们都予以介绍. 笔者觉得这种以数据为主导的学习方式有助于理解统计作为数据科学的本质, 有助于实际工作者通过数据学习多种统计方法的应用. 我们列举了可能用于同类数据的若干方法, 希望对创造新的数据分析方法有所启发并促使进一步探索, 同时也让读者免受查阅大量不同文献之苦. 本书不可能介绍所有的方法, 大量新方法在你阅读本书的时候正在诞生.

本书所有的分析都通过免费的自由软件 R 来实现. 读者可以毫不困难地

重复本书所有的计算. R 网站[①]拥有世界各地统计学家贡献的大量最新软件包(package), 这些软件包以飞快的速度增加和更新, 已从 2009 年年底的大约 1000 个增加到 2012 年 8 月底的 4009 个, 仅 2012 年 8 月就增加了 449 个. 它们代表了统计学家创造的崭新的统计方法. 这些软件包的代码都是公开的[②]. 与此相对比, 所有商业软件远没有如此多的资源, 也不会更新得如此之快, 而且商业软件的代码都是保密的昂贵 "黑匣子". 在发达国家, 不能想象一个统计专业的研究生不会使用 R 软件. 那里很多学校都开设了 R 软件的课程. 今天, 任何一个统计学家想要介绍和推广其创造的统计方法, 都必须提供相应的计算程序, 而发表该程序的最佳地点就是 R 网站. 由于方法和代码是公开的, 这些方法很容易引起有关学者的关注, 这些关注对研究相应方法形成群体效应, 推动其发展. 不会编程的统计学家在今天是很难生存的.

在学校讲授任何一种商业软件都是为该公司做义务广告, 如果没有相关软件公司的资助, 就没有学校愿意花钱讲授商业软件. 在教学中使用盗版软件是违法行为, 绝对不应该或明或暗地鼓励师生使用盗版商业软件.

对 R 软件编程的熟悉还有助于学习其他快速计算的语言, 比如 C++ 和 FORTRAN, 这对于应对因快速处理庞大的数据集而面临的巨大的计算量有所裨益.

本书首先通过一些简单的统计和数学内容介绍 R 软件的基本知识, 然后介绍数据分析的一些基本逻辑和常识. 本书的主体则是根据不同数据形式介绍相应的方法. 本书以数据为主导, 各章都是完全独立的. 有一些统计基本知识的读者可以选读本书的任何一个完整的部分. 虽然本书介绍的方法涉及应用统计的各个方面, 但不可能介绍所有的数学和统计细节, 否则将会是一部巨型的百科全书. 笔者尽量用文字和少量数学公式对各种方法的原理予以直观介绍, 并且引导读者做进一步的阅读.

由于本书没有按照数学模型的分类来编排, 因此对各种方法的介绍不可能满足数学上的系统、整洁和完美的要求, 但这正是对现实数据和现实世界的反映. 如果现实数据都像标准教科书例子那样 "规范", 统计就没有存在和发展的必要了. 本书试图让读者理解世界是复杂的, 数据形式是多种多样的. 必须有超越书本、超越所谓权威的智慧和勇气, 才能充满自信地面对世界上出现的各

① 网址: http://www.r-project.org/.

② 除了极个别并非秘密的子程序之外, 因为它们很费时间, 用机器代码实行.

种挑战.

由于统计正以前所未有的速度发展, R 网站及其各个软件包也在不断更新, 因此, 笔者希望读者通过对本书的学习, 学会如何通过 R 不断学习新的知识和方法. "授人以鱼不如授之以渔", 成功的教师不是像百科全书那样告诉学生一些现成的知识, 而是让学生产生疑问和兴趣, 以促进其做进一步的探索.

本书所有的数据例子都可以在网上找到并且下载. 这些例子背后都有一些理论和应用的故事. 笔者并没有刻意挑选例子所在的领域, 这没有关系. 你学会了一加一等于二, 也就学会了一个苹果加一个苹果等于两个苹果, 或一个梨加一个梨等于两个梨这样的计算. 那个把作为科学的统计按照工种来划分 (诸如工业统计, 农业统计, 劳动统计) 的时代早已一去不复返了. 统计是为各个领域服务的, 我们想要得到的是到任何领域都能施展的能力, 而不是有限的行业培训. 如果你能够处理具有挑战性的数据, 那么无论该数据来自何领域, 你的感觉都会很好.

虽然本书冠以 "复杂数据统计方法" 之名, 但对处理 "非复杂" 数据的方法都有较完整的回顾, 并给出了相应的运算程序, 只不过没有像标准教科书那样详细地解释细节而已.

本书的适用范围很广, 其内容曾经在中国人民大学、首都经贸大学、中央财经大学、西南财经大学、云南财经大学、四川大学、哈尔滨理工大学、新疆财经大学、中山大学讲授过, 对象包括数学、应用数学、统计、精算、经济、旅游、环境等专业的本科生以及数学、应用数学、统计、计量经济学、生物医学、经济学等专业的硕士和博士研究生. 作为成绩评定, 给每个学生分配两个国外网站上的实际数据, 并且要求他们在学期末将他们分析处理这些数据的结果形成报告. 这些数据如何处理, 没有标准答案, 甚至有些必要的方法还超出了授课的范围, 需要学生做进一步探索和学习. 笔者认为, 应用统计硕士所学的内容应该包括本书的大部分内容. 希望本书对于各个领域的教师以及实际工作者都有参考价值.

在任何国家及任何制度下都能够生存和发展的知识和能力, 就是科学, 是人们在生命的历程中应该获得的.

吴喜之

目　录

第 1 章

引　言

§1.1　作为科学的统计

统计是科学 (science), 而科学的基本特征是其方法论: **对世界的认识源于观测或实验的信息 (或者数据), 总结信息时会形成模型 (亦称假说或理论), 模型会指导进一步的探索, 直到遇到这些模型无法解释的现象, 这就导致对这些模型的更新和替代**. 这就是科学的方法. 只有用科学的方法进行的探索才能被称为科学.

科学的理论完全依赖于实际, 统计方法则完全依赖于来自实际的数据. 统计可以定义为 "收集、分析、展示和解释数据的科学," 或者称为**数据科学** (science of data). 统计几乎应用于所有领域.

统计的思维方式是归纳 (induction), 也就是从数据所反映的现实得到比较一般的模型, 希望以此解释数据所代表的那部分世界. 这和以演绎 (deduction) 为主的数学思维方式相反, 演绎是在一些人为的**假定**(比如一个公理系统) 之下, 推导出各种结论.

在统计科学发展的前期, 由于没有计算机, 不可能应付庞大的数据量[①], 只能在对少量数据的背景分布做出诸如独立同正态分布之类的数学假定后, 建立一些假定的数学模型, 进行手工计算, 并推导出一些由这些模型所得结果的性质, 诸如置信区间, 假设检验的 p 值, 无偏性及相合性等. 在数据与数学假定相

① 请想象一下用纸和笔得到简单线性回归所必须计算的预测矩阵 $X(X^TX)^{-1}X^T$, 假定 X 为 30×5 的数值矩阵.

差较远的情况下, 人们又利用诸如中心极限定理或大样本定理等得到当样本量趋于无穷时的一些类似的性质. 统计的这种发展方式, 给统计打上了很深的数学烙印.

统计发展的历史痕迹体现在很多方面, 特别是流行 "模型驱动" 的研究及教学模式. 各统计院系的课程大都以数学模型作为课程的名称和主要内容, 一些数理统计杂志也喜欢发表没有数据背景的关于数学模型的文章. 很多学生毕业后只会推导一些课本上的公式, 却不会处理真实数据. 一些人对于有穷样本, 也假装认为是大样本, 并且堂而皇之地用大样本性质来描述从有穷样本中得到的结论. 至于数据是否满足大样本定理的条件, 数据样本是不是 "大样本" 等关键问题则尽量不谈或少谈. 按照模型驱动的研究方式, 一些学者不从数据出发, 而是想象出一些他们感觉很好的数学模型, 然后在世界上到处寻求 "适合" 他们模型的数据来 "证明" 自己的模型的确有意义. 这种自欺欺人的做法绝对是不科学的.

以模型而不是数据为主导的研究方式导致统计在某种程度上成为自我封闭、自我欣赏及自我评价的系统. 固步自封的后果是, 三十多年来, 统计丢掉了许多属于数据科学的领域, 也失去了许多人才. 在现成数学模型无法处理大量的复杂数据的情况下, 计算机领域的研究人员和部分概率论及统计学家开发了许多计算方法, 处理了传统统计无法解决的大量问题. 诸如人工神经网络、决策树、boosting、随机森林、支持向量机等大量算法模型的相继出现宣告了传统数学模型主导 (如果不是垄断的话) 数据分析时代的终结. 这些研究最初根本无法刊登在传统统计杂志上, 因此大都出现在计算机及各应用领域的杂志中.

模型驱动的研究方法在前计算机时代有其合理性. 但是在计算机快速发展的今天, 仍然固守这种研究模式就不会有前途了. 人们在处理数据时, 首先寻求现有的方法, 当现有方法不能满足他们的要求时, 往往会根据数据的特征创造出新的可以计算的方法来满足实际需要. 这就是统计科学近年来飞速发展的历程. 创造模型的目的是适应现实数据. 统计研究应该是由问题或者数据驱动的, 而不是由模型驱动的.

随着时代的进步, 各个统计院系现在也开始设置诸如数据挖掘、机器学习等课程, 统计杂志也开始逐渐重视这些研究. 这些算法模型大都不是用封闭的数学公式来描述, 而是体现在计算机算法或程序上. 对于结果的风险也不是用假定的分布 (或渐近分布) 所得到的 p 值来描述, 而是用没有参加建模训练的

测试集的误差来描述. 这些方法发展很快, 不仅因为它们能够解决问题, 而且因为那些不懂统计或概率论的人也能够完全理解结果 (这也是某些有 "知识垄断欲" 的传统统计学家不易接受的现实). 现在, 无论承认与否, 多数统计学家都明白, 如果不会计算机编程, 也不与编程人员合作, 则不会产生任何有意义的成果.

§1.2 数据分析的实践

数据收集 首先要根据实际目的收集数据. 有些数据是需要人工收集的, 通过普查、调查、实验、观察等手段得到; 另一些则来自各种与计算机联系的数据源, 比如遥感数据、网络数据、商务数据、远程上报数据等. 确定哪些变量的数据需要收集是非常关键的, 这个决策不是基于数学或统计的知识, 而是基于对相应领域的了解和经验. 有数据不一定能够得到需要的结论, 我们需要的是与所关心问题**充分有关**的变量的数据.

数据预处理 原始数据往往或多或少地存在各种缺失值, 还有不合逻辑或不一致等问题. 这需要进行预处理. 这些工作很可能非常费时而且极其琐碎, 但必须去做, 否则后续的分析是不可能展开的. 填补缺失值有很多方法, 最简单的就是删除, 或者用同一变量其他值的均值或中位数填补, 或者在各个变量之间建立模型 (比如线性模型、最近邻方法等) 来填补.①

寻找适合的模型 有了数据, 我们需要的是模型, 其目的或者是预测, 或者是理解产生数据的机制. 为了寻找模型, 首先要对数据进行探索性分析, 利用图形、各种统计量、或者比较复杂的探索方法来查看数据的关联性、线性性、异方差性、多重共线性、聚类特征、平衡特征、分布形状等. 有了对数据的粗略认识之后, 就要寻找适合的模型, 无论是传统意义上的模型还是以算法为基础的模型. 首先寻找现成的模型, 比较各种模型的计算结果, 如果现有模型不能满足需要, 新的数据分析方法就应该产生了. 模型的选择贯穿整个数据分析过程.

比较模型的标准 在传统统计中, 通常要对分布和模型形式作出假定, 在这些假定下确定损失函数, 并依此得到各种判别准则, 这些准则包括各种检验、

① R 网站有一个名叫 missForest 的很好的填补缺失值的软件包, 可以使用随机森林的方法, 同时自动填补定量变量和分类变量.

一些统计量的临界值等. 但绝对不要忘记, 所有这些都是在对数据分布及描述变量之间关系的模型所做的假定之下得到的.[①] 如果这些假定不满足, 这些准则也没有什么意义. 在使用算法模型时, 由于没有传统模型的那些假定, 判断模型的好坏通常都用交叉验证 (cross validation) 的方法, 也就是说, 拿一部分数据作为训练集 (training set), 得到模型, 再用另一部分数据 (称为测试集, testing set) 来看误差是多少. 有时需要进行 k 折交叉验证 (k-fold cross validation), 即把数据分成 k 份, 每次拿 $k-1$ 份作为训练集, 用剩下的一份作为测试集, 重复 k 次, 得到 k 个误差作出平均, 以避免仅用一个测试集可能出现的偏差. 显然, 交叉验证的方法也适用于传统模型之间以及传统模型和算法模型之间的比较.

对结果的解释 最终目的不是选择模型, 而是解释模型所产生的结果, 结果则必须是应用领域的结果, 必须有实际意义. 仅仅用统计术语说某个模型较好、某个变量显著之类的话是不够的.

§1.3 数据的形式以及可能用到的模型

数据形式多种多样, 但大部分可以放到二维数据文件中, 比如, n 维列联表在 R 中是一个 n 维数组, 但总可以转换成一个二维数据阵. 其行数为 n 个变量各个水平的全部组合数, 如果第 i 个变量有 L_i 个水平, 则行数为 $\prod_1^n L_i$. 其列数为 $n+1$: n 个变量加上一个频数变量. 当维数较大时, n 维列联表数据很可能有许多水平组合为空值 (稀疏). 一些数据, 比如某些空间数据, 则可能需要多个数据文件来描述. 本书所用的所有数据都可以从网上下载, 对于单一文件的方阵型数据, 通常每一行代表一个观测值, 每一列代表一个变量.

根据研究目的, 数据中的一些变量可以被人为地作为因变量或自变量, 这样的情况往往出现在人们想要用一部分变量描述或预测另一部分变量的时候, 例如统计中的回归或分类 (判别) 问题. 在一些情况下, 人们不太关心预测, 但想知道数据变量之间的其他关系, 比如在多元分析课程中的聚类及因子分析等内容. 还有一些情况, 人们想用较长的历史值预测未来, 诸如时间序列数据的分析. 一些数据被称为横截面数据, 那里每个对象仅观测一次. 而有些时候一

① 这些假定都是无法用确定性方法验证的, 最多只能尝试用显著性检验来拒绝, 如果没有充分理由拒绝, 也不能 "证明" 这些假定是对的, 它们仍然属于假定.

个对象观测若干次, 但又不同于时间序列的长期观测, 称为纵向数据. 纵向数据的一些特例在某些领域中被称为面板数据. 在诸如医学、质量控制、心理学、社会学等领域, 人们关心某事件 (比如死亡, 失效等) 是否发生及什么时候发生的问题, 相应的数据则被称为生存数据. 调查问卷的结果往往做成列联表的形式, 也被称为列联表数据. 这些数据类型绝对不是排他的, 比如, 有些数据既是生存数据, 又是纵向数据.

对每一种数据类型, 人们根据目的以及数据的一些特殊性, 找到各种数学模型来处理. 下面就本书后面章节可能涉及的各种数据分析的内容加以简单介绍. 由于实际数据是复杂的, 可能必须经过多次尝试及对比才能确定什么模型不适用, 什么模型适用. 所有的传统方法都有一定的假定, 必须注意这些假定的合理性.

1.3.1 横截面数据: 因变量为实轴上的数量变量

当因变量为数量变量时, 人们首先想到的是回归, 教科书中最先介绍的是假定模型中误差项独立同正态分布的线性回归. 由于误差项很可能不满足这个假定, 还有可能有多重共线性等问题, 这样就产生了诸如加权回归、稳健回归、偏最小二乘回归、lasso 回归、岭回归、主成分回归、Box-Cox 变换 (或其他变换)、多项式回归、分位数回归等模型, 还产生了相应的各种检验及判断方法, 诸如最优子集、逐步回归、回归诊断等. 在线性或其他假定不满足时, 又出现了非线性回归、非参数回归、广义线性模型、随机效应混合模型、可加模型、广义可加模型等. 生存分析也包含了回归的内容. 近年来, 神经网络、决策树的回归树、boosting、bagging、最近邻方法、随机森林、支持向量机等算法模型也被广泛应用于回归. 第 2 章将会介绍上面所述的大部分内容.

1.3.2 横截面数据: 因变量为分类 (定性) 变量或者频数

当因变量为分类变量时, 一般的回归方法就不能使用了. 在传统的统计领域, 如果因变量是二分变量, 则可以尝试用广义线性模型中的 logistic 回归和 probit 回归, 如果自变量是数量型的变量, 则可以尝试线性判别分析, 这里有可能要求自变量满足正态分布. 但最新的一些基于算法的模型则没有关于分布的假定, 比如决策树的分类树、神经网络、boosting、bagging、随机森林、最近邻方法、支持向量机等能够很好地处理分类问题.

在许多情况下, 数据的形式是频数, 比如列联表数据, 这不是常规意义上

的连续型数值变量. 如果把这些频数作为关注的因变量, 而把那些形成列联表的分类变量以及可能得到的数量变量作为自变量, 则可以应用多项分布对数线性模型, 多项 logit 模型及 Poisson 对数线性模型 (作为广义线性模型的特例), 在使用 Poisson 模型时, 往往可能出现 Poisson 模型假定不允许的过散布 (overdispersion, 即方差大于均值) 或欠散布 (underdispersion, 即方差小于均值) 情况. 这时就要进行矫正或者利用一些双参数模型. 当然, "非正统" 的机器学习的方法在这里也可以用.

上述内容将在第 3 章介绍.

1.3.3 纵向数据, 多水平数据, 面板数据, 重复观测数据

当一个对象有多次重复观测时, 得到的就不是横截面数据, 而是纵向数据, 经济金融领域的这一类数据有时被称为面板数据, 属于纵向数据的特例. 应对这种数据有许多不同的方法, 这些方法也成为相应处理纵向数据的课程名称, 比如多层模型、随机效应混合模型等. 此外, 很多纵向数据还与生存数据有关, 很多广义线性模型可处理的数据也是纵向数据, 这就产生了更多的处理方法. 一些机器学习方法也被用到纵向数据的分析之中. 第 4 章将介绍线性随机效应混合模型、广义线性随机效应混合模型、决策树的应用及纵向生存分析等内容.

1.3.4 多元数据各变量之间的关系: 多元分析

在没有预测任务, 对数据进行分析的目的仅在于确定各个变量之间关系时, 就不需要确定自变量及因变量了. 这时, 如果变量全部都是服从正态分布的数量变量, 这个问题属于传统多元分析的范畴, 主要方法包括主成分分析、因子分析、聚类分析、典型相关分析等. 有点另类的对应分析是一种图描述方法, 并不被认为是经典方法, 但也出现于一些教科书中. 对于复杂的分层数据, 或者包含分类 (定性) 变量的数据, 传统的多元分析方法就不能使用了, 新的基于算法的模型可以应对非常复杂的多元分析课题. 第 5 章首先回顾经典的多元分析方法, 然后介绍可视化很强的新发展的多元分析方法, 最后介绍没有丝毫经典统计味道的纯粹机器学习方法: 关联规则分析.

1.3.5 路径模型/结构方程模型

还有一种调查是在已经确定的模型基础上设计的, 例如顾客满意度调查是