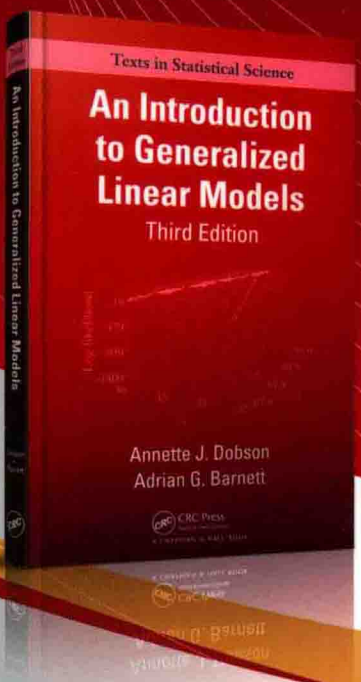
国外实用统计丛书

广义线性模型导论

（英文导读版·原书第3版）

An Introduction to Generalized Linear Models

[澳] 安妮特 J. 杜布森（Annette J. Dobson）
艾德里安 G. 巴奈特（Adrian G.Barnett） 著



CRC Press
Taylor & Francis Group

机械工业出版社
CHINA MACHINE PRESS

国外实用统计丛书

广义线性模型导论

（英文导读版·原书第3版）

[澳] 安妮特 J. 杜布森（Annette J. Dobson）
昆士兰大学（University of Queensland）
艾德里安 G. 巴奈特（Adrian G. Barnett）
昆士兰科技大学（Queensland University of Technology）
著

王星 注释
中国人民大学

机械工业出版社

本书首先介绍了广义线性模型的理论背景，其次着重分析特定类型的数据，其中包含：正态分布、泊松分布和二项分布；线性回归模型；经典的估计和模型拟合方法；以及统计推断的方法。在此基础上，作者又探究了线性回归、方差分析（ANOVA）、逻辑斯谛回归、对数线性模型、生存分析、多水平建模、贝叶斯分析和马尔可夫链蒙特卡罗方法（MCMC）。书中为统计建模提供了一个紧密的框架，更强调数值和图像方法，并增加了Stata、R和WinBUGS软件的代码以及三个有关贝叶斯分析的章节。

本书适合作为大学本科统计专业教材，或相关科研人员的参考书

An introduction to Generalized Linear Models, Third Edition/Annette J. Dobson /ISBN: 978-1-58488-950-2

图书在版编目（CIP）数据

广义线性模型导论=An introduction to Generalized Linear Models：英文导读版：第3版：英文/（澳）杜布森（Dobson, A. J.），（澳）巴奈特（Barnett, A. G.）著；王星注释.—北京：机械工业出版社，2015.8

（国外实用统计丛书）

ISBN 978-7-111-50318-7

Ⅰ.①广… Ⅱ.①杜…②巴…③王… Ⅲ.①线性模型—研究—英文 Ⅳ.①O212

中国版本图书馆CIP数据核字（2015）第107378号

机械工业出版社（北京市百万庄大街22号 邮政编码100037）

策划编辑：汤 嘉 责任编辑：汤 嘉

封面设计：张 静 责任印制：乔 宇

北京铭成印刷有限公司印刷

2015年6月第1版第1次印刷

184mm x 260mm · 21.75印张 · 1插页 · 507千字

标准书号：ISBN 978-7-111-50318-7

定价：49.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务		网络服务	
服务咨询热线：	（010）88361066	机 工 官 网：	www.cmpbook.com
读者购书热线：	（010）68326294	机 工 官 博：	weibo.com/cmp1952
	（010）88379203	金 书 网：	www.golden-book.com
封面无防伪标均为盗版		教育服务网：	www.cmpedu.com

前　言

编写本书的初衷是以本科生和其他领域的研究人员能够理解的方式，展现统计建模的统一理论和概念框架。

本书的第 2 版扩充了名义型变量、序数型变量的逻辑斯谛回归，生存分析，以及纵向数据、聚类数据分析等内容，同时更多地依赖数值方法、可视化数值优化和图形方法来进行探索性的数据分析和模型拟合检验。这些内容在第 3 版中会有更加深入的介绍。

第 3 版包含了关于贝叶斯分析的三个新章节。基础的贝叶斯理论基础早在传统统计理论发展之前就有所记载，然而实用的贝叶斯分析却是最近才出现。它的出现主要归功于我们将在第 13 章介绍的马尔可夫链蒙特卡罗方法。贝叶斯方法越来越强的可操作性意味着更多懂经典统计理论的人在尝试使用贝叶斯方法来求解广义线性模型。贝叶斯分析具备比传统方法更大的优势，因为它正式地引入了先验信息，所以具有更大的灵活性，可以解决更复杂的问题。

本版还更新了 Stata 和 R 软件代码，会对广义线性模型的实际应用有所帮助。贝叶斯分析的章节还包含了 R 和 WinBUGS 代码。

本书中的数据集和练习题的简要解答可以在出版社网站上获得：http://www.crcpress.com/eproducts/downloads/。

来自澳大利亚昆士兰大学和纽卡斯尔大学的同仁和同学们以及在澳大利亚生物统计合作协会上过研究生课程的诸位同学都给本书提出了许多中肯的建议并对本书中的材料给出了意见，在此我们表示感谢。

第 1 章

介绍

本章作为全书第 1 章，主要介绍本书各章节的主要内容并为本书后续内容介绍了必要的背景知识，对书中的主要符号进行了解释。本章简要讲解了两类变量和三种数据类型，并介绍了与正态分布相关的三个重要分布、二次型、对极大似然估计和最小二乘估计以及二者之间的联系进行比较。

目 录

Contents

CHAPTER 1

Introduction

1.1 Background

This book is designed to introduce the reader to generalized linear models; these provide a unifying framework for many commonly used statistical techniques. They also illustrate the ideas of statistical modelling.

The reader is assumed to have some familiarity with classical statistical principles and methods. In particular, understanding the concepts of estimation, sampling distributions and hypothesis testing is necessary. Experience in the use of t-tests, analysis of variance, simple linear regression and chi-squared tests of independence for two-dimensional contingency tables is assumed. In addition, some knowledge of matrix algebra and calculus is required.

The reader will find it necessary to have access to statistical computing facilities. Many statistical programs, languages or packages can now perform the analyses discussed in this book. Often, however, they do so with a different program or procedure for each type of analysis so that the unifying structure is not apparent.

Some programs or languages which have procedures consistent with the approach used in this book are **Stata**, **R**, **S-PLUS**, **SAS** and **Genstat**. For Chapters 13 to 14 programs to conduct Markov chain Monte Carlo methods are needed and WinBUGS has been used here. This list is not comprehensive as appropriate modules are continually being added to other programs.

In addition, anyone working through this book may find it helpful to be able to use mathematical software that can perform matrix algebra, differentiation and iterative calculations.

1.2 Scope

The statistical methods considered in this book all involve the analysis of relationships between measurements made on groups of subjects or objects. For example, the measurements might be the heights or weights and the ages of boys and girls, or the yield of plants under various growing conditions. We use the terms **response**, **outcome** or **dependent variable** for measurements that are free to vary in response to other variables called **explanatory variables** or **predictor variables** or **independent variables**—although this last term can sometimes be misleading. Responses are regarded as random variables. Explanatory variables are usually treated as though they are non-

random measurements or observations; for example, they may be fixed by the experimental design.

Responses and explanatory variables are measured on one of the following scales.

1. **Nominal** classifications: e.g., red, green, blue; yes, no, do not know, not applicable. In particular, for **binary**, **dichotomous** or **binomial** variables there are only two categories: male, female; dead, alive; smooth leaves, serrated leaves. If there are more than two categories the variable is called **polychotomous, polytomous** or **multinomial**.

2. **Ordinal** classifications in which there is some natural order or ranking between the categories: e.g., young, middle aged, old; diastolic blood pressures grouped as ≤ 70, 71–90, 91–110, 111–130, ≥ 131 mmHg.

3. **Continuous** measurements where observations may, at least in theory, fall anywhere on a continuum: e.g., weight, length or time. This scale includes both **interval scale** and **ratio scale** measurements—the latter have a well-defined zero. A particular example of a continuous measurement is the time until a specific event occurs, such as the failure of an electronic component; the length of time from a known starting point is called the **failure time**.

Nominal and ordinal data are sometimes called **categorical** or **discrete variables** and the numbers of observations, **counts** or **frequencies** in each category are usually recorded. For continuous data the individual measurements are recorded. The term **quantitative** is often used for a variable measured on a continuous scale and the term **qualitative** for nominal and sometimes for ordinal measurements. A qualitative, explanatory variable is called a **factor** and its categories are called the **levels** for the factor. A quantitative explanatory variable is sometimes called a **covariate**.

Methods of statistical analysis depend on the measurement scales of the response and explanatory variables.

This book is mainly concerned with those statistical methods which are relevant when there is just *one response variable* although there will usually be several explanatory variables. The responses measured on different subjects are usually assumed to be statistically independent random variables although this requirement is dropped in Chapter 11, which is about correlated data, and in subsequent chapters. Table 1.1 shows the main methods of statistical analysis for various combinations of response and explanatory variables and the chapters in which these are described.The last three chapters are devoted to Bayesian methods which substantially extend these analyses.

The present chapter summarizes some of the statistical theory used throughout the book. Chapters 2 through 5 cover the theoretical framework that is common to the subsequent chapters. Later chapters focus on methods for analyzing particular kinds of data.

Chapter 2 develops the main ideas of classical or frequentist statistical modelling. The modelling process involves four steps:

Table 1.1 *Major methods of statistical analysis for response and explanatory variables measured on various scales and chapter references for this book. Extensions of these methods from a Bayesian perspective are illustrated in Chapters 12–14.*

Response (chapter)	Explanatory variables	Methods
Continuous (Chapter 6)	Binary	t-test
	Nominal, >2 categories	Analysis of variance
	Ordinal	Analysis of variance
	Continuous	Multiple regression
	Nominal & some continuous	Analysis of covariance
	Categorical & continuous	Multiple regression
Binary (Chapter 7)	Categorical	Contingency tables Logistic regression
	Continuous	Logistic, probit & other dose-response models
	Categorical & continuous	Logistic regression
Nominal with >2 categories (Chapters 8 & 9)	Nominal	Contingency tables
	Categorical & continuous	Nominal logistic regression
Ordinal (Chapter 8)	Categorical & continuous	Ordinal logistic regression
Counts (Chapter 9)	Categorical	Log-linear models
	Categorical & continuous	Poisson regression
Failure times (Chapter 10)	Categorical & continuous	Survival analysis (parametric)
Correlated responses (Chapter 11)	Categorical & continuous	Generalized estimating equations Multilevel models

1. Specifying models in two parts: equations linking the response and explanatory variables, and the probability distribution of the response variable.
2. Estimating fixed but unknown parameters used in the models.
3. Checking how well the models fit the actual data.
4. Making inferences; for example, calculating confidence intervals and testing hypotheses about the parameters.

The next three chapters provide the theoretical background. Chapter 3 is about the **exponential family of distributions**, which includes the Normal, Poisson and Binomial distributions. It also covers **generalized linear models** (as defined by Nelder and Wedderburn 1972). Linear regression and many other models are special cases of generalized linear models. In Chapter 4 methods of classical estimation and model fitting are described.

Chapter 5 outlines frequentist methods of statistical inference for generalized linear models. Most of these methods are based on how well a model describes the set of data. For example, **hypothesis testing** is carried out by first specifying alternative models (one corresponding to the null hypothesis and the other to a more general hypothesis). Then test statistics are calculated which measure the "goodness of fit" of each model and these are compared. Typically the model corresponding to the null hypothesis is simpler, so if it fits the data about as well as a more complex model it is usually preferred on the grounds of parsimony (i.e., we retain the null hypothesis).

Chapter 6 is about **multiple linear regression** and **analysis of variance** (ANOVA). Regression is the standard method for relating a continuous response variable to several continuous explanatory (or predictor) variables. ANOVA is used for a continuous response variable and categorical or qualitative explanatory variables (factors). **Analysis of covariance** (ANCOVA) is used when at least one of the explanatory variables is continuous. Nowadays it is common to use the same computational tools for all such situations. The terms **multiple regression** or **general linear model** are used to cover the range of methods for analyzing one continuous response variable and multiple explanatory variables.

Chapter 7 is about methods for analyzing binary response data. The most common one is **logistic regression** which is used to model relationships between the response variable and several explanatory variables which may be categorical or continuous. Methods for relating the response to a single continuous variable, the dose, are also considered; these include **probit analysis** which was originally developed for analyzing dose-response data from bioassays. Logistic regression has been generalized to include responses with more than two nominal categories (**nominal**, **multinomial**, **polytomous** or **polychotomous logistic regression**) or ordinal categories (**ordinal logistic regression**). These methods are discussed in Chapter 8.

Chapter 9 concerns **count** data. The counts may be frequencies displayed in a **contingency table** or numbers of events, such as traffic accidents, which need to be analyzed in relation to some "exposure" variable such as the num-