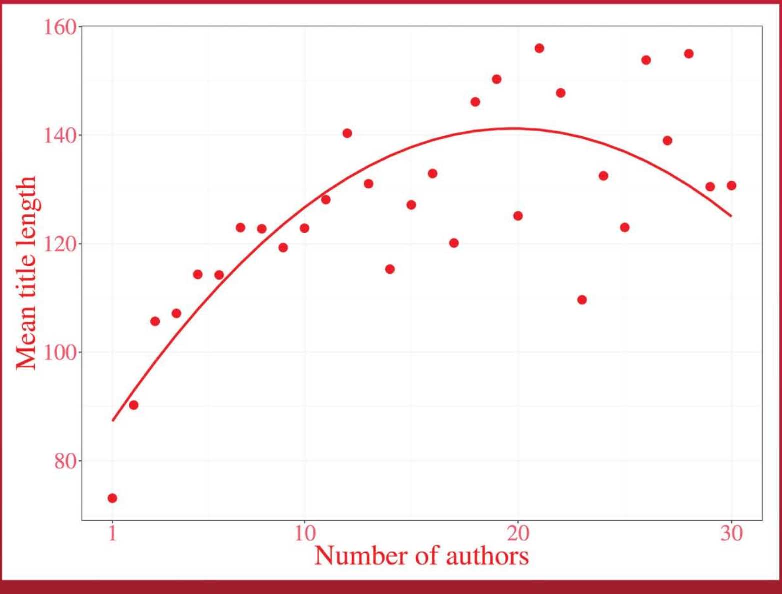
Texts in Statistical Science

An Introduction to Generalized Linear Models

Fourth Edition



Annette J. Dobson
Adrian G. Barnett

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

An Introduction to Generalized Linear Models

Fourth Edition

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Nonlinear Time Series: Theory, Methods, and Applications with R Examples
R. Douc, E. Moulines, and D.S. Stoffer

Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists
F.J. Samaniego

Introduction to Multivariate Analysis: Linear and Nonlinear Modeling
S. Konishi

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Bayesian Networks: With Examples in R
M. Scutari and J.-B. Denis

Linear Models with R, Second Edition
J.J. Faraway

Introduction to Probability
J. K. Blitzstein and J. Hwang

Analysis of Categorical Data with R
C. R. Bilder and T.M. Loughin

Statistical Inference: An Integrated Approach, Second Edition
H. S. Migon, D. Gamerman, and F. Louzada

Modelling Survival Data in Medical Research, Third Edition
D. Collett

Design and Analysis of Experiments with R
J. Lawson

Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition
P. J. Bickel and K. A. Doksum

Statistics for Finance
E. Lindström, H. Madsen, and J. N. Nielsen

Spatio-Temporal Methods in Environmental Epidemiology
G. Shaddick and J.V. Zidek

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II
P. J. Bickel and K. A. Doksum

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II
P. J. Bickel and K. A. Doksum

Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data
M. Friendly and D. Meyer

Statistical Rethinking: A Bayesian Course with Examples in R and Stan
R. McElreath

Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data, Second Edition
R. Christensen

Essentials of Probability Theory for Statisticians
M.A. Proschan and P.A. Shaw

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition
J.J. Faraway

Modeling and Analysis of Stochastic Systems, Third Edition
V.G. Kulkarni

Pragmatics of Uncertainty
J.B. Kadane

Stochastic Processes: From Applications to Theory
P.D Moral and S. Penev

Modern Data Science with R
B.S. Baumer, D.T Kaplan, and N.J. Horton

Logistic Regression Models
J.M. Hilbe

Generalized Additive Models: An Introduction with R, Second Edition
S. Wood

Design of Experiments: An Introduction Based on Linear Models
Max Morris

Introduction to Statistical Methods for Financial Models
T. A. Severini

Statistical Regression and Classification: From Linear Models to Machine Learning
N. Matloff

Introduction to Functional Data Analysis
P. Kokoszka and M. Reimherr

Stochastic Processes: An Introduction, Third Edition
P.W. Jones and P. Smith

An Introduction to Generalized Linear Models

Fourth Edition

By

Annette J. Dobson

and

Adrian G. Barnett

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20180306

International Standard Book Number-13: 978-1-138-74168-3 (Hardback)
International Standard Book Number-13: 978-1-138-74151-5 (Paperback)

Library of Congress Cataloging-in-Publication Data

Names: Dobson, Annette J., 1945- author. | Barnett, Adrian G., author.
Title: An introduction to generalized linear models / by Annette J. Dobson, Adrian G. Barnett.
Other titles: Generalized linear models
Description: Fourth edition. | Boca Raton : CRC Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018002845| ISBN 9781138741683 (hardback : alk. paper) | ISBN 9781138741515 (pbk. : alk. paper) | ISBN 9781315182780 (e-book : alk. paper)
Subjects: LCSH: Linear models (Statistics)
Classification: LCC QA276 .D589 2018 | DDC 519.5--dc23
LC record available at https://lccn.loc.gov/2018002845

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Beth.

Contents

Preface

The original purpose of the book was to present a unified theoretical and conceptual framework for statistical modelling in a way that was accessible to undergraduate students and researchers in other fields.

The second edition was expanded to include nominal and ordinal logistic regression, survival analysis and analysis of longitudinal and clustered data. It relied more on numerical methods, visualizing numerical optimization and graphical methods for exploratory data analysis and checking model fit.

The third edition added three chapters on Bayesian analysis for generalized linear models. To help with the practical application of generalized linear models, Stata, R and WinBUGS code were added.

This fourth edition includes new sections on the common problems of model selection and non-linear associations. Non-linear associations have a long history in statistics as the first application of the least squares method was when Gauss correctly predicted the non-linear orbit of an asteroid in 1801.

Statistical methods are essential for many fields of research, but a widespread lack of knowledge of their correct application is creating inaccurate results. Untrustworthy results undermine the scientific process of using data to make inferences and inform decisions. There are established practices for creating reproducible results which are covered in a new Postface to this edition.

The data sets and outline solutions of the exercises are available on the publisher's website: http://www.crcpress.com/9781138741515. We also thank Thomas Haslwanter for providing a set of solutions using Python: https://github.com/thomas-haslwanter/dobson.

We are grateful to colleagues and students at the Universities of Queensland and Newcastle, Australia, and those taking postgraduate courses through the Biostatistics Collaboration of Australia for their helpful suggestions and comments about the material.

Annette J. Dobson and Adrian G. Barnett
Brisbane, Australia

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 nonrandom 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:

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. Nowa-

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

days 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. This chapter also includes a section on **model selection** that is also applicable for other types of generalized linear models

Chapter 7 is about methods for analyzing binary response data. The most common one is **logistic regression** which is used to model associations 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 number of motor vehicles registered or the distances travelled by the drivers. Modelling methods are based on assuming that the distribution of counts can be described by the Poisson distribution, at least approximately. These methods include **Poisson regression** and **log-linear models**.

Survival analysis is the usual term for methods of analyzing failure time data. The parametric methods described in Chapter 10 fit into the framework of generalized linear models although the probability distribution assumed for the failure times may not belong to the exponential family.

Generalized linear models have been extended to situations where the responses are correlated rather than independent random variables. This may occur, for instance, if they are **repeated measurements** on the same subject or measurements on a group of related subjects obtained, for example, from **clustered sampling**. The method of **generalized estimating equations** (GEEs) has been developed for analyzing such data using techniques analogous to those for generalized linear models. This method is outlined in Chapter 11 together with a different approach to correlated data, namely **multilevel modelling** in which some parameters are treated as random variables rather than fixed but unknown constants. Multilevel modelling involves both fixed and random effects (mixed models) and relates more closely to the Bayesian approach to statistical analysis.

The main concepts and methods of Bayesian analysis are introduced in Chapter 12. In this chapter the relationships between classical or frequentist methods and Bayesian methods are outlined. In addition the software WinBUGS which is used to fit Bayesian models is introduced.

Bayesian models are usually fitted using computer-intensive methods based on Markov chains simulated using techniques based on random numbers. These methods are described in Chapter 13. This chapter uses some examples from earlier chapters to illustrate the mechanics of Markov chain Monte Carlo (MCMC) calculations and to demonstrate how the results allow much richer statistical inferences than are possible using classical methods.

Chapter 14 comprises several examples, introduced in earlier chapters, which are reworked using Bayesian analysis. These examples are used to illustrate both conceptual issues and practical approaches to estimation, model fitting and model comparisons using WinBUGS.

Finally there is a Postscript that summarizes the principles of good statistical practice that should always be used in order to address the "**reproducibility crisis**" that plagues science with daily reports of "breakthroughs" that turn out to be useless or untrue.

Further examples of generalized linear models are discussed in the books by McCullagh and Nelder (1989), Aitkin et al. (2005) and Myers et al. (2010). Also there are many books about specific generalized linear models such as Agresti (2007, 2013), Collett (2003, 2014), Diggle et al. (2002), Goldstein (2011), Hilbe (2015) and Hosmer et al. (2013).

1.3 Notation

Generally we follow the convention of denoting random variables by uppercase italic letters and observed values by the corresponding lowercase letters. For example, the observations $y_1, y_2, ..., y_n$ are regarded as realizations of the random variables $Y_1, Y_2, \ldots, Y_n$. Greek letters are used to denote parameters and the corresponding lowercase Roman letters are used to denote estimators and estimates; occasionally the symbol $\widehat{}$ is used for estimators or estimates. For example, the parameter β is estimated by $\widehat{\beta}$ or b. Sometimes these conventions are not strictly adhered to, either to avoid excessive notation in cases where the meaning should be apparent from the context, or when there is a strong tradition of alternative notation (e.g., e or ε for random error terms).

Vectors and matrices, whether random or not, are denoted by boldface lower- and uppercase letters, respectively. Thus, $\mathbf{y}$ represents a vector of observations

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

or a vector of random variables

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$\boldsymbol{\beta}$ denotes a vector of parameters and $\mathbf{X}$ is a matrix. The superscript ^{T} is used for a matrix transpose or when a column vector is written as a row, e.g., $y = [Y_1, \ldots, Y_n]^T$.

The probability density function of a continuous random variable Y (or the probability mass function if Y is discrete) is referred to simply as a **probability distribution** and denoted by

$$f(y; \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ represents the parameters of the distribution.

We use dot ($\cdot$) subscripts for summation and bars ($^-$) for means; thus,

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\, y_{\cdot}\,.$$

The expected value and variance of a random variable Y are denoted by $\mathrm{E}(Y)$ and $\mathrm{var}(Y)$, respectively. Suppose random variables $Y_1, \ldots, Y_N$ are independent with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \sigma_i^2$ for $i = 1, \ldots, n$. Let the random variable W be a **linear combination** of the Y_i's

$$W = a_1 Y_1 + a_2 Y_2 + \ldots + a_n Y_n, \tag{1.1}$$

where the a_i's are constants. Then the expected value of W is

$$\mathrm{E}(W) = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n \tag{1.2}$$

and its variance is

$$\mathrm{var}(W) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2. \tag{1.3}$$

1.4 Distributions related to the Normal distribution

The sampling distributions of many of the estimators and test statistics used in this book depend on the Normal distribution. They do so either directly because they are derived from Normally distributed random variables or asymptotically, via the Central Limit Theorem for large samples. In this section we give definitions and notation for these distributions and summarize the relationships between them. The exercises at the end of the chapter provide practice in using these results which are employed extensively in subsequent chapters.

1.4.1 Normal distributions

1. If the random variable Y has the Normal distribution with mean μ and variance σ^2, its probability density function is

   $$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right].$$

   We denote this by $Y \sim \mathrm{N}(\mu,\sigma^2)$.
2. The Normal distribution with $\mu = 0$ and $\sigma^2 = 1$, $Y \sim \mathrm{N}(0,1)$, is called the **standard Normal distribution**.
3. Let $Y_1,\ldots,Y_n$ denote Normally distributed random variables with $Y_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$ for $i = 1,\ldots,n$ and let the covariance of Y_i and Y_j be denoted by

   $$\mathrm{cov}(Y_i,Y_j) = \rho_{ij}\sigma_i\sigma_j,$$

   where ρ_{ij} is the correlation coefficient for Y_i and Y_j. Then the joint distribution of the Y_i's is the **multivariate Normal distribution** with mean vector $\boldsymbol{\mu} = [\mu_1,\ldots,\mu_n]^T$ and variance-covariance matrix $\mathbf{V}$ with diagonal elements σ_i^2 and non-diagonal elements $\rho_{ij}\sigma_i\sigma_j$ for $i \neq j$. We write this as $\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\mu},\mathbf{V})$, where $\mathbf{y} = [Y_1,\ldots,Y_n]^T$.
4. Suppose the random variables $Y_1,\ldots,Y_n$ are independent and Normally distributed with the distributions $Y_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$ for $i = 1,\ldots,n$. If

   $$W = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n,$$

   where the a_i's are constants, then W is also Normally distributed, so that

   $$W = \sum_{i=1}^{n} a_iY_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i\mu_i,\ \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

   by Equations (1.2) and (1.3).

1.4.2 Chi-squared distribution

1. The **central chi-squared distribution** with n degrees of freedom is defined as the sum of squares of n independent random variables $Z_1,\ldots,Z_n$ each with the standard Normal distribution. It is denoted by

$$X^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

In matrix notation, if $\mathbf{z} = [Z_1,\ldots,Z_n]^T$, then $\mathbf{z}^T\mathbf{z} = \sum_{i=1}^{n} Z_i^2$ so that $X^2 = \mathbf{z}^T\mathbf{z} \sim \chi^2(n)$.

2. If X^2 has the distribution $\chi^2(n)$, then its expected value is $\mathrm{E}(X^2) = n$ and its variance is $\mathrm{var}(X^2) = 2n$.

3. If $Y_1,\ldots,Y_n$ are independent, Normally distributed random variables, each with the distribution $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, then

$$X^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n) \tag{1.4}$$

because each of the variables $Z_i = (Y_i - \mu_i)/\sigma_i$ has the standard Normal distribution $\mathrm{N}(0,1)$.

4. Let $Z_1,\ldots,Z_n$ be independent random variables each with the distribution $\mathrm{N}(0,1)$ and let $Y_i = Z_i + \mu_i$, where at least one of the μ_i's is non-zero. Then the distribution of

$$\sum Y_i^2 = \sum (Z_i + \mu_i)^2 = \sum Z_i^2 + 2\sum Z_i\mu_i + \sum \mu_i^2$$

has larger mean $n + \lambda$ and larger variance $2n + 4\lambda$ than $\chi^2(n)$ where $\lambda = \sum \mu_i^2$. This is called the **non-central chi-squared distribution** with n degrees of freedom and **non-centrality parameter** λ. It is denoted by $\chi^2(n,\lambda)$.

5. Suppose that the Y_i's are not necessarily independent and the vector $\mathbf{y} = [Y_1,\ldots,Y_n]^T$ has the multivariate Normal distribution $\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\mu},\mathbf{V})$ where the variance–covariance matrix $\mathbf{V}$ is non-singular and its inverse is $\mathbf{V}^{-1}$. Then

$$X^2 = (\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi^2(n). \tag{1.5}$$

6. More generally if $\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\mu},\mathbf{V})$, then the random variable $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$ has the non-central chi-squared distribution $\chi^2(n,\lambda)$ where $\lambda = \boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu}$.

7. If $X_1^2,\ldots,X_m^2$ are m independent random variables with the chi-squared distributions $X_i^2 \sim \chi^2(n_i,\lambda_i)$, which may or may not be central, then their sum

also has a chi-squared distribution with $\sum n_i$ degrees of freedom and non-centrality parameter $\sum \lambda_i$, that is,

$$\sum_{i=1}^{m} X_i^2 \sim \chi^2\left(\sum_{i=1}^{m} n_i, \sum_{i=1}^{m} \lambda_i\right).$$

This is called the **reproductive property** of the chi-squared distribution.

8. Let $\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{V})$, where $\mathbf{y}$ has n elements but the Y_i's are not independent so that the number k of linearly independent rows (or columns) of $\mathbf{V}$ (that is, the rank of $\mathbf{V}$) is less than n and so $\mathbf{V}$ is singular and its inverse is not uniquely defined. Let $\mathbf{V}^-$ denote a generalized inverse of $\mathbf{V}$ (that is a matrix with the property that $\mathbf{V}\mathbf{V}^-\mathbf{V} = \mathbf{V}$). Then the random variable $\mathbf{y}^T\mathbf{V}^-\mathbf{y}$ has the non-central chi-squared distribution with k degrees of freedom and non-centrality parameter $\lambda = \boldsymbol{\mu}^T\mathbf{V}^-\boldsymbol{\mu}$.
   For further details about properties of the chi-squared distribution see Forbes et al. (2010).
9. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be n independent random vectors each of length p and $\mathbf{y}_n \sim \mathrm{MVN}(\mathbf{0}, \mathbf{V})$. Then $\mathbf{S} = \sum_{i=i}^{n} \mathbf{y}_i \mathbf{y}_i^T$ is a $p \times p$ random matrix which has the Wishart distribution $\mathrm{W}(\mathbf{V}, n)$. This distribution can be used to make inferences about the covariance matrix $\mathbf{V}$ because $\mathbf{S}$ is proportional to $\mathbf{V}$. In the case $p = 1$ the Y_i's are independent random variables with $Y_i \sim \mathrm{N}(0, \sigma^2)$, so $Z_i = Y_i/\sigma \sim \mathrm{N}(0,1)$. Hence, $\mathbf{S} = \sum_{i=1}^{n} Y_i^2 = \sigma^2 \sum_{i=1}^{n} Z_i^2$ and therefore $\mathbf{S}/\sigma^2 \sim \chi^2(n)$. Thus, the Wishart distribution can be regarded as a generalisation of the chi-squared distribution.

1.4.3 t-distribution

The **t-distribution** with n degrees of freedom is defined as the ratio of two independent random variables. The numerator has the standard Normal distribution and the denominator is the square root of a central chi-squared random variable divided by its degrees of freedom; that is,

$$T = \frac{Z}{(X^2/n)^{1/2}} \tag{1.6}$$

where $Z \sim \mathrm{N}(0,1)$, $X^2 \sim \chi^2(n)$ and Z and X^2 are independent. This is denoted by $T \sim \mathrm{t}(n)$.

1.4.4 F-distribution

1. The **central F-distribution** with n and m degrees of freedom is defined as the ratio of two independent central chi-squared random variables, each

divided by its degrees of freedom,

$$F = \frac{X_1^2}{n} \Big/ \frac{X_2^2}{m}\,, \tag{1.7}$$

where $X_1^2 \sim \chi^2(n), X_2^2 \sim \chi^2(m)$ and X_1^2 and X_2^2 are independent. This is denoted by $F \sim \mathrm{F}(n,m)$.

2. The relationship between the t-distribution and the F-distribution can be derived by squaring the terms in Equation (1.6) and using definition (1.7) to obtain

$$T^2 = \frac{Z^2}{1} \Big/ \frac{X^2}{n} \sim \mathrm{F}(1,n)\,, \tag{1.8}$$

that is, the square of a random variable with the t-distribution, $t(n)$, has the F-distribution, $F(1,n)$.

3. The **non-central F-distribution** is defined as the ratio of two independent random variables, each divided by its degrees of freedom, where the numerator has a non-central chi-squared distribution and the denominator has a central chi-squared distribution, that is,

$$F = \frac{X_1^2}{n} \Big/ \frac{X_2^2}{m}\,,$$

where $X_1^2 \sim \chi^2(n,\lambda)$ with $\lambda = \boldsymbol{\mu}^T\mathbf{V}^{-1}\boldsymbol{\mu}$, $X_2^2 \sim \chi^2(m)$, and X_1^2 and X_2^2 are independent. The mean of a non-central F-distribution is larger than the mean of central F-distribution with the same degrees of freedom.

1.4.5 Some relationships between distributions

We summarize the above relationships in Figure 1.1. In later chapters we add to this diagram and a more extensive diagram involving most of the distributions used in this book is given in the Appendix. Asymptotic relationships are shown using dotted lines and transformations using solid lines. For more details see Leemis (1986) from which this diagram was developed.

1.5 Quadratic forms

1. A **quadratic form** is a polynomial expression in which each term has degree 2. Thus, $y_1^2 + y_2^2$ and $2y_1^2 + y_2^2 + 3y_1y_2$ are quadratic forms in y_1 and y_2, but $y_1^2 + y_2^2 + 2y_1$ or $y_1^2 + 3y_2^2 + 2$ are not.

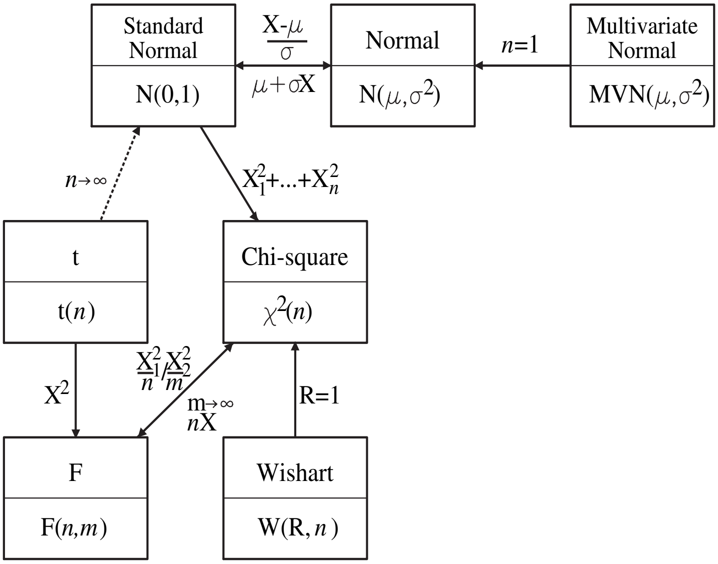


Figure 1.1 *Some relationships between common distributions related to the Normal distribution, adapted from Leemis (1986). Dotted line indicates an asymptotic relationship and solid lines a transformation.*

2. Let **A** be a symmetric matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},$$

where $a_{ij} = a_{ji}$; then the expression $\mathbf{y}^T\mathbf{A}\mathbf{y} = \sum_i \sum_j a_{ij} y_i y_j$ is a quadratic form in the y_i's. The expression $(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is a quadratic form in the terms $(y_i - \mu_i)$ but not in the y_i's.

3. The quadratic form $\mathbf{y}^T\mathbf{A}\mathbf{y}$ and the matrix **A** are said to be **positive definite** if $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$ whenever the elements of **y** are not all zero. A necessary and sufficient condition for positive definiteness is that all the determinants $|A_1| = a_{11}, |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots,$ and $|A_n| = \det \mathbf{A}$ are positive. If a matrix is positive definite, then it can be inverted and also it has a square root matrix $\mathbf{A}^*$ such that $\mathbf{A}^*\mathbf{A} = \mathbf{A}$. These

properties are useful for the derivation of several theoretical results related to estimation and the probability distributions of estimators.

4. The rank of the matrix $\mathbf{A}$ is also called the degrees of freedom of the quadratic form $Q = \mathbf{y}^T \mathbf{A} \mathbf{y}$.
5. Suppose $Y_1, \ldots, Y_n$ are independent random variables each with the Normal distribution $\mathrm{N}(0, \sigma^2)$. Let $Q = \sum_{i=1}^n Y_i^2$ and let $Q_1, \ldots, Q_k$ be quadratic forms in the Y_i's such that

$$Q = Q_1 + \ldots + Q_k,$$

where Q_i has m_i degrees of freedom ($i = 1, \ldots, k$). Then $Q_1, \ldots, Q_k$ are independent random variables and $Q_1/\sigma^2 \sim \chi^2(m_1)$, $Q_2/\sigma^2 \sim \chi^2(m_2), \ldots$, and $Q_k/\sigma^2 \sim \chi^2(m_k)$, if and only if

$$m_1 + m_2 + \ldots + m_k = n.$$

This is Cochran's theorem. A similar result also holds for non-central distributions. For more details see Forbes et al. (2010).

6. A consequence of Cochran's theorem is that the difference of two independent random variables, $X_1^2 \sim \chi^2(m)$ and $X_2^2 \sim \chi^2(k)$, also has a chi-squared distribution

$$X^2 = X_1^2 - X_2^2 \sim \chi^2(m-k)$$

provided that $X^2 \geq 0$ and $m > k$.

1.6 Estimation

1.6.1 Maximum likelihood estimation

Let $\mathbf{y} = [Y_1, \ldots, Y_n]^T$ denote a random vector and let the joint probability density function of the Y_i's be

$$f(\mathbf{y}; \boldsymbol{\theta})$$

which depends on the vector of parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$.

The **likelihood function** $L(\boldsymbol{\theta}; \mathbf{y})$ is algebraically the same as the joint probability density function $f(\mathbf{y}; \boldsymbol{\theta})$ but the change in notation reflects a shift of emphasis from the random variables $\mathbf{y}$, with $\boldsymbol{\theta}$ fixed, to the parameters $\boldsymbol{\theta}$, with $\mathbf{y}$ fixed. Since L is defined in terms of the random vector $\mathbf{y}$, it is itself a random variable. Let Ω denote the set of all possible values of the parameter vector $\boldsymbol{\theta}$; Ω is called the **parameter space**. The **maximum likelihood estimator** of θ is the value $\widehat{\boldsymbol{\theta}}$ which maximizes the likelihood function, that is,

$$L(\widehat{\boldsymbol{\theta}}; \mathbf{y}) \geq L(\boldsymbol{\theta}; \mathbf{y}) \qquad \text{for all } \boldsymbol{\theta} \text{ in } \Omega.$$

Equivalently, $\widehat{\boldsymbol{\theta}}$ is the value which maximizes the **log-likelihood function** $l(\boldsymbol{\theta};\mathbf{y}) = \log L(\boldsymbol{\theta};\mathbf{y})$ since the logarithmic function is monotonic. Thus,

$$l(\widehat{\boldsymbol{\theta}};\mathbf{y}) \geq l(\boldsymbol{\theta};\mathbf{y}) \qquad \text{for all } \boldsymbol{\theta} \text{ in } \Omega.$$

Often it is easier to work with the log-likelihood function than with the likelihood function itself.

Usually the estimator $\widehat{\boldsymbol{\theta}}$ is obtained by differentiating the log-likelihood function with respect to each element θ_j of $\boldsymbol{\theta}$ and solving the simultaneous equations

$$\frac{\partial l(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_j} = 0 \qquad \text{for } j = 1,\ldots,p. \tag{1.9}$$

It is necessary to check that the solutions do correspond to maxima of $l(\boldsymbol{\theta};\mathbf{y})$ by verifying that the matrix of second derivatives

$$\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_j \partial \theta_k}$$

evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ is negative definite. For example, if $\boldsymbol{\theta}$ has only one element θ, this means it is necessary to check that

$$\left[\frac{\partial^2 l(\theta,y)}{\partial \theta^2}\right]_{\theta=\widehat{\theta}} < 0.$$

It is also necessary to check if there are any values of $\boldsymbol{\theta}$ at the edges of the parameter space Ω that give local maxima of $l(\boldsymbol{\theta};\mathbf{y})$. When all local maxima have been identified, the value of $\widehat{\boldsymbol{\theta}}$ corresponding to the largest one is the maximum likelihood estimator. (For most of the models considered in this book there is only one maximum and it corresponds to the solution of the equations $\partial l/\partial \theta_j = 0,\ j = 1,\ldots,p$.)

An important property of maximum likelihood estimators is that if $g(\boldsymbol{\theta})$ is any function of the parameters $\boldsymbol{\theta}$, then the maximum likelihood estimator of $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$. This follows from the definition of $\widehat{\boldsymbol{\theta}}$. It is sometimes called the **invariance property** of maximum likelihood estimators. A consequence is that we can work with a function of the parameters that is convenient for maximum likelihood estimation and then use the invariance property to obtain maximum likelihood estimates for the required parameters.

In principle, it is not necessary to be able to find the derivatives of the likelihood or log-likelihood functions or to solve Equation (1.9) if $\widehat{\boldsymbol{\theta}}$ can be found numerically. In practice, numerical approximations are very important for generalized linear models.

Other properties of maximum likelihood estimators include consistency, sufficiency, asymptotic efficiency and asymptotic normality. These are discussed in books such as Cox and Hinkley (1974) or Forbes et al. (2010).

1.6.2 Example: Poisson distribution

Let $Y_1, \ldots, Y_n$ be independent random variables each with the Poisson distribution

$$f(y_i;\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}, \qquad y_i = 0,1,2,\ldots$$

with the same parameter θ. Their joint distribution is

$$\begin{aligned} f(y_1,\ldots,y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta) &= \frac{\theta^{y_1} e^{-\theta}}{y_1!} \times \frac{\theta^{y_2} e^{-\theta}}{y_2!} \times \cdots \times \frac{\theta^{y_n} e^{-\theta}}{y_n!} \\ &= \frac{\theta^{\Sigma y_i}\, e^{-n\theta}}{y_1! y_2! \ldots y_n!}. \end{aligned}$$

This is also the likelihood function $L(\theta;y_1,\ldots,y_n)$. It is easier to use the log-likelihood function

$$l(\theta;y_1,\ldots,y_n) = \log L(\theta;y_1,\ldots,y_n) = (\sum y_i)\log\theta - n\theta - \sum(\log y_i!).$$

To find the maximum likelihood estimate $\widehat{\theta}$, use

$$\frac{dl}{d\theta} = \frac{1}{\theta}\sum y_i - n.$$

Equate this to zero to obtain the solution

$$\widehat{\theta} = \sum y_i/n = \overline{y}.$$

Since $d^2l/d\theta^2 = -\sum y_i/\theta^2 < 0$, l has its maximum value when $\theta = \widehat{\theta}$, confirming that $\overline{y}$ is the maximum likelihood estimate.

1.6.3 Least squares estimation

Let $Y_1, \ldots, Y_n$ be independent random variables with expected values $\mu_1, \ldots, \mu_n$, respectively. Suppose that the μ_i's are functions of the parameter vector that we want to estimate, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$; $p < n$. Thus

$$E(Y_i) = \mu_i(\boldsymbol{\beta}).$$

The simplest form of the **method of least squares** consists of finding the estimator $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of squares of the differences between Y_i's and their expected values

$$S = \sum [Y_i - \mu_i(\boldsymbol{\beta})]^2 .$$

Usually $\widehat{\boldsymbol{\beta}}$ is obtained by differentiating S with respect to each element β_j of $\boldsymbol{\beta}$ and solving the simultaneous equations

$$\frac{\partial S}{\partial \beta_j} = 0, \qquad j = 1, \ldots, p.$$

Of course it is necessary to check that the solutions correspond to minima (i.e., the matrix of second derivatives is positive definite) and to identify the global minimum from among these solutions and any local minima at the boundary of the parameter space.

Now suppose that the Y_i's have variances σ_i^2 that are not all equal. Then it may be desirable to minimize the weighted sum of squared differences

$$S = \sum w_i [Y_i - \mu_i(\boldsymbol{\beta})]^2 ,$$

where the weights are $w_i = (\sigma_i^2)^{-1}$. In this way, the observations which are less reliable (i.e., the Y_i's with the larger variances) will have less influence on the estimates.

More generally, let $\mathbf{y} = [Y_1, \ldots, Y_n]^T$ denote a random vector with mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$ and variance–covariance matrix $\mathbf{V}$. Then the **weighted least squares estimator** is obtained by minimizing

$$S = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}).$$

1.6.4 Comments on estimation

1. An important distinction between the methods of maximum likelihood and least squares is that the method of least squares can be used without making assumptions about the distributions of the response variables Y_i beyond specifying their expected values and possibly their variance–covariance structure. In contrast, to obtain maximum likelihood estimators we need to specify the joint probability distribution of the Y_i's.
2. For many situations maximum likelihood and least squares estimators are identical.
3. Often numerical methods rather than calculus may be needed to obtain parameter estimates that maximize the likelihood or log-likelihood function or minimize the sum of squares. The following example illustrates this approach.

1.6.5 Example: Tropical cyclones

Table 1.2 shows the number of tropical cyclones in northeastern Australia for the seasons 1956–7 (season 1) through 1968–9 (season 13), a period of fairly consistent conditions for the definition and tracking of cyclones (Dobson and Stewart 1974).

Table 1.2 *Numbers of tropical cyclones in 13 successive seasons.*

Season	1	2	3	4	5	6	7	8	9	10	11	12	13
No. of cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

Let Y_i denote the number of cyclones in season i, where $i = 1, \ldots, 13$. Suppose the Y_i's are independent random variables with the Poisson distribution with parameter θ. From Example 1.6.2, $\widehat{\theta} = \overline{y} = 72/13 = 5.538$. An alternative approach would be to find numerically the value of θ that maximizes the log-likelihood function. The component of the log-likelihood function due to y_i is

$$l_i = y_i \log \theta - \theta - \log y_i!.$$

The log-likelihood function is the sum of these terms

$$l = \sum_{i=1}^{13} l_i = \sum_{i=1}^{13} (y_i \log \theta - \theta - \log y_i!)\,.$$

Only the first two terms in the brackets involve θ and so are relevant to the optimization calculation because the term $\sum_1^{13} \log y_i!$ is a constant. To plot the log-likelihood function (without the constant term) against θ, for various values of θ, calculate $(y_i \log \theta - \theta)$ for each y_i and add the results to obtain $l^* = \sum (y_i \log \theta - \theta)$. Figure 1.2 shows l^* plotted against θ.

Clearly the maximum value is between $\theta = 5$ and $\theta = 6$. This can provide a starting point for an iterative procedure for obtaining $\widehat{\theta}$. The results of a simple bisection calculation are shown in Table 1.3. The function l^* is first calculated for approximations $\theta^{(1)} = 5$ and $\theta^{(2)} = 6$. Then subsequent approximations $\theta^{(k)}$ for $k = 3, 4, \ldots$ are the average values of the two previous estimates of θ with the largest values of l^*(for example, $\theta^{(6)} = \frac{1}{2}(\theta^{(5)} + \theta^{(3)})$). After 7 steps, this process gives $\widehat{\theta} \simeq 5.54$ which is correct to 2 decimal places.

1.7 Exercises

1.1 Let Y_1 and Y_2 be independent random variables with $Y_1 \sim \mathrm{N}(1,3)$ and $Y_2 \sim \mathrm{N}(2,5)$. If $W_1 = Y_1 + 2Y_2$ and $W_2 = 4Y_1 - Y_2$, what is the joint distribution of W_1 and W_2?

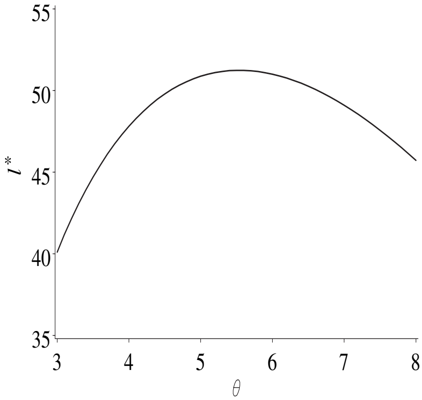


Figure 1.2 *Graph showing the location of the maximum likelihood estimate for the data in Table 1.2 on tropical cyclones.*

Table 1.3 *Successive approximations to the maximum likelihood estimate of the mean number of cyclones per season.*

k	$\theta^{(k)}$	l^*
1	5	50.878
2	6	51.007
3	5.5	51.242
4	5.75	51.192
5	5.625	51.235
6	5.5625	51.243
7	5.5313	51.24354
8	5.5469	51.24352
9	5.5391	51.24360
10	5.5352	51.24359

1.2 Let Y_1 and Y_2 be independent random variables with $Y_1 \sim \mathrm{N}(0,1)$ and $Y_2 \sim \mathrm{N}(3,4)$.

a. What is the distribution of Y_1^2?

b. If $\mathbf{y} = \begin{bmatrix} Y_1 \\ (Y_2-3)/2 \end{bmatrix}$, obtain an expression for $\mathbf{y}^T\mathbf{y}$. What is its distribution?

c. If $\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ and its distribution is $\mathbf{y} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{V})$, obtain an expression for $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$. What is its distribution?

1.3 Let the joint distribution of Y_1 and Y_2 be MVN$(\boldsymbol{\mu}, \mathbf{V})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix}.$$

a. Obtain an expression for $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu})$. What is its distribution?

b. Obtain an expression for $\mathbf{y}^T \mathbf{V}^{-1} \mathbf{y}$. What is its distribution?

1.4 Let $Y_1, \ldots, Y_n$ be independent random variables each with the distribution $\mathrm{N}(\mu, \sigma^2)$. Let

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

a. What is the distribution of $\overline{Y}$?

b. Show that $S^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} (Y_i - \mu)^2 - n(\overline{Y} - \mu)^2 \right]$.

c. From (b) it follows that $\sum (Y_i - \mu)^2/\sigma^2 = (n-1)S^2/\sigma^2 + \left[(\overline{Y} - \mu)^2 n/\sigma^2 \right]$. How does this allow you to deduce that $\overline{Y}$ and S^2 are independent?

d. What is the distribution of $(n-1)S^2/\sigma^2$?

e. What is the distribution of $\frac{\overline{Y} - \mu}{S/\sqrt{n}}$?

1.5 This exercise is a continuation of the example in Section 1.6.2 in which $Y_1, \ldots, Y_n$ are independent Poisson random variables with the parameter θ.

a. Show that $\mathrm{E}(Y_i) = \theta$ for $i = 1, \ldots, n$.

b. Suppose $\theta = e^{\beta}$. Find the maximum likelihood estimator of β.

c. Minimize $S = \sum \left(Y_i - e^{\beta} \right)^2$ to obtain a least squares estimator of β.

1.6 The data in Table 1.4 are the numbers of females and males in the progeny of 16 female light brown apple moths in Muswellbrook, New South Wales, Australia (from Lewis, 1987).

a. Calculate the proportion of females in each of the 16 groups of progeny.

b. Let Y_i denote the number of females and n_i the number of progeny in each group $(i = 1, \ldots, 16)$. Suppose the Y_i's are independent random variables each with the Binomial distribution

$$f(y_i; \theta) = \binom{n_i}{y_i} \theta^{y_i} (1 - \theta)^{n_i - y_i}.$$

Find the maximum likelihood estimator of θ using calculus and evaluate it for these data.

c. Use a numerical method to estimate $\widehat{\theta}$ and compare the answer with the one from (b).

Table 1.4 *Progeny of light brown apple moths.*

Progeny group	Females	Males
1	18	11
2	31	22
3	34	27
4	33	29
5	27	24
6	33	29
7	28	25
8	23	26
9	33	38
10	12	14
11	19	23
12	25	31
13	14	20
14	4	6
15	22	34
16	7	12

Chapter 2

Model Fitting

2.1 Introduction

The model fitting process described in this book involves four steps:

1. Model specification—a model is specified in two parts: an equation linking the response and explanatory variables and the probability distribution of the response variable.
2. Estimation of the parameters of the model.
3. Checking the adequacy of the model—how well it fits or summarizes the data.
4. Inference—for classical or frequentist inference this involves calculating confidence intervals, testing hypotheses about the parameters in the model and interpreting the results.

In this chapter these steps are first illustrated using two small examples. Then some general principles are discussed. Finally there are sections about notation and coding of explanatory variables which are needed in subsequent chapters.

2.2 Examples

2.2.1 Chronic medical conditions

Data from the Australian Longitudinal Study on Women's Health (Lee et al. 2005) show that women who live in country areas tend to have fewer consultations with general practitioners (family physicians) than women who live near a wider range of health services. It is not clear whether this is because they are healthier or because structural factors, such as shortage of doctors, higher costs of visits and longer distances to travel, act as barriers to the use of general practitioner (GP) services. Table 2.1 shows the numbers of chronic medical conditions (for example, high blood pressure or arthritis) reported

Table 2.1 *Number of chronic medical conditions of 26 town women and 23 country women with similar use of general practitioner services.*

Town
0 1 1 0 2 3 0 1 1 1 1 2 0 1 3 0 1 2 1 3 3 4 1 3 2 0
n = 26, mean = 1.423, standard deviation = 1.172, variance = 1.374
Country
2 0 3 0 0 1 1 1 1 0 0 2 2 0 1 2 0 0 1 1 1 0 2
n = 23, mean = 0.913, standard deviation = 0.900, variance = 0.810

by samples of women living in large country towns (town group) or in more rural areas (country group) in New South Wales, Australia. All the women were aged 70–75 years, had the same socio-economic status and had three or fewer GP visits during 1996. The question of interest is: Do women who have similar levels of use of GP services in the two groups have the same need as indicated by their number of chronic medical conditions?

The Poisson distribution provides a plausible way of modelling these data as they are count data and within each group the sample mean and variance are similar. Let Y_{jk} be a random variable representing the number of conditions for the kth woman in the jth group, where $j=1$ for the town group and $j=2$ for the country group and $k=1,\ldots,K_j$ with $K_1=26$ and $K_2=23$. Suppose the Y_{jk}'s are all independent and have the Poisson distribution with parameter θ_j representing the expected number of conditions.

The question of interest can be formulated as a test of the null hypothesis $\mathrm{H}_0 : \theta_1=\theta_2=\theta$ against the alternative hypothesis $\mathrm{H}_1 : \theta_1 \neq \theta_2$. The model fitting approach to testing H_0 is to fit two models, one assuming H_0 is true, that is

$$\mathrm{E}(Y_{jk})=\theta; \quad Y_{jk} \sim \mathrm{Po}(\theta), \tag{2.1}$$

and the other assuming it is not, so that

$$\mathrm{E}(Y_{jk})=\theta_j; \quad Y_{jk} \sim \mathrm{Po}(\theta_j), \tag{2.2}$$

where $j=1$ or 2. Testing H_0 against H_1 involves comparing how well Models (2.1) and (2.2) fit the data. If they are about equally good, then there is little reason for rejecting H_0. However, if Model (2.2) is clearly better, then H_0 would be rejected in favor of H_1.

If H_0 is true, then the log-likelihood function of the Y_{jk}'s is

$$l_0 = l(\theta;\mathbf{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K_j}(y_{jk}\log\theta - \theta - \log y_{jk}!), \tag{2.3}$$

where $J=2$ in this case. The maximum likelihood estimate, which can be obtained as shown in the example in Section 1.6.2, is

$$\widehat{\theta}=\sum\sum y_{jk}/N,$$

where $N=\sum_j K_j$. For these data the estimate is $\widehat{\theta}=1.184$ and the maximum value of the log-likelihood function, obtained by substituting this value of $\widehat{\theta}$ and the data values y_{jk} into (2.3), is $\widehat{l_0}=-68.3868$.

If H_1 is true, then the log-likelihood function is

$$\begin{aligned} l_1=l(\theta_1,\theta_2;\mathbf{y}) &= \sum_{k=1}^{K_1}(y_{1k}\log\theta_1-\theta_1-\log y_{1k}!) \\ &+\sum_{k=1}^{K_2}(y_{2k}\log\theta_2-\theta_2-\log y_{2k}!). \end{aligned} \tag{2.4}$$

(The subscripts on l_0 and l_1 in (2.3) and (2.4) are used to emphasize the connections with the hypotheses H_0 and H_1, respectively). From (2.4) the maximum likelihood estimates are $\widehat{\theta}_j=\sum_k y_{jk}/K_j$ for $j=1$ or 2. In this case $\widehat{\theta}_1=1.423$, $\widehat{\theta}_2=0.913$ and the maximum value of the log-likelihood function, obtained by substituting these values and the data into (2.4), is $\widehat{l_1}=-67.0230$.

The maximum value of the log-likelihood function l_1 will always be greater than or equal to that of l_0 because one more parameter has been fitted. To decide whether the difference is statistically significant, we need to know the sampling distribution of the log-likelihood function. This is discussed in Chapter 4.

If $Y\sim\mathrm{Po}(\theta)$ then $\mathrm{E}(Y)=\mathrm{var}(Y)=\theta$. The estimate $\widehat{\theta}$ of $\mathrm{E}(Y)$ is called the **fitted value** of Y. The difference $Y-\widehat{\theta}$ is called a **residual** (other definitions of residuals are also possible, see Section 2.3.4). Residuals form the basis of many methods for examining the adequacy of a model. A residual is usually standardized by dividing by its standard error. For the Poisson distribution an approximate standardized residual is

$$r=\frac{Y-\widehat{\theta}}{\sqrt{\widehat{\theta}}}.$$

The standardized residuals for Models (2.1) and (2.2) are shown in Table 2.2 and Figure 2.1. Examination of individual residuals is useful for assessing certain features of a model such as the appropriateness of the probability distribution used for the responses or the inclusion of specific explanatory variables. For example, the residuals in Table 2.2 and Figure 2.1 exhibit some skewness, as might be expected for the Poisson distribution.