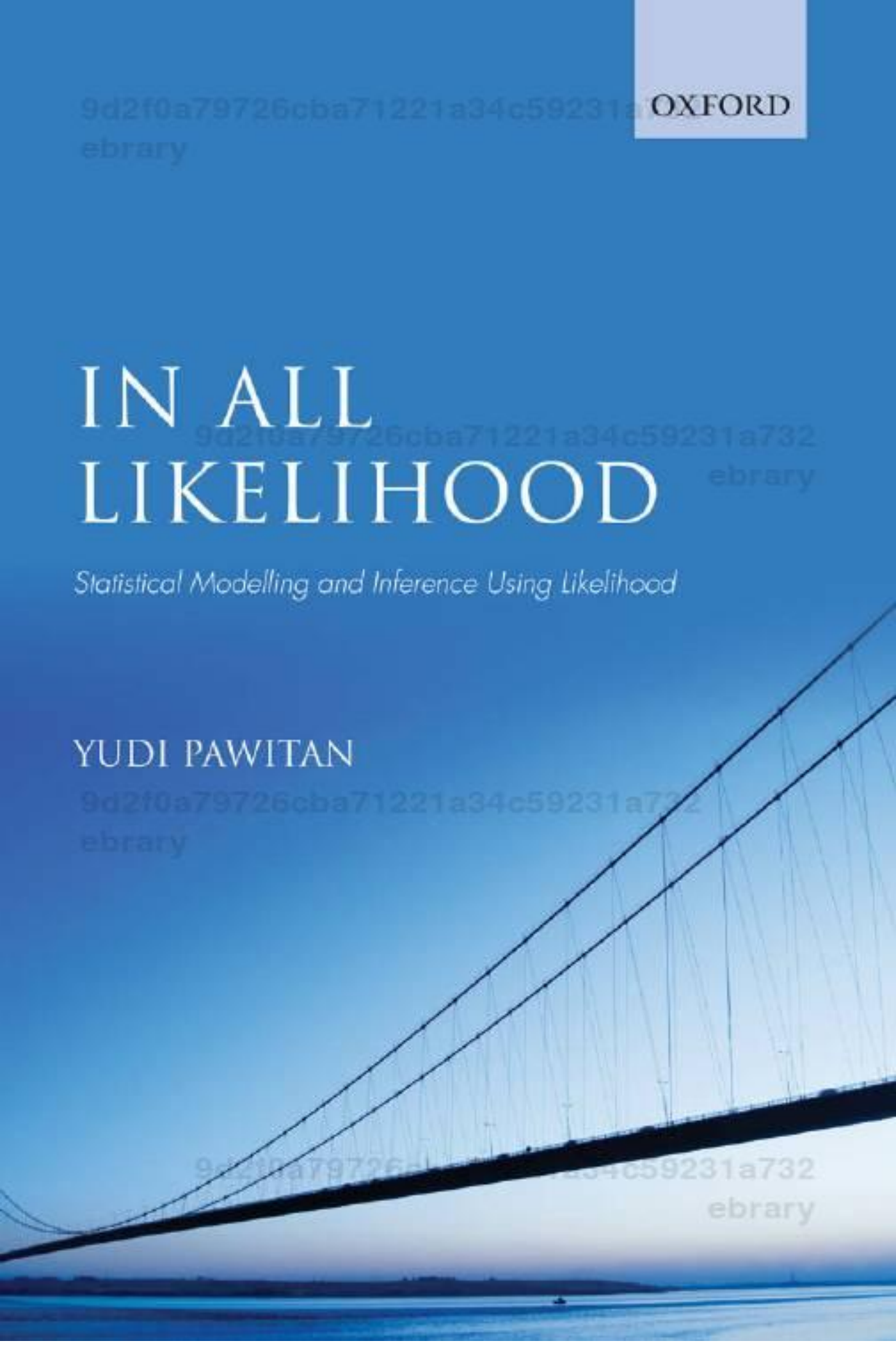
OXFORD

IN ALL LIKELIHOOD

Statistical Modelling and Inference Using Likelihood

YUDI PAWITAN



In All Likelihood

Statistical Modelling and Inference Using Likelihood

Yudi Pawitan
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
yudi.pawitan@ki.se

CLARENDON PRESS · OXFORD
2001

This page intentionally left blank

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First published in paperback 2013

Impression: 2 4 6 8 10 9 7 5 3 1

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

ISBN 978–0–19–850765–9 (Hbk.)
ISBN 978–0–19–967122–9 (Pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

This page intentionally left blank

Preface

Likelihood is the central concept in statistical modelling and inference. *In All Likelihood* covers the essential aspects of likelihood-based modelling as well as likelihood's fundamental role in inference. The title is a gentle reminder of the original meaning of 'likelihood' as a measure of uncertainty, a Fisherian view that tends to be forgotten under the weight of likelihood's more technical role.

Fisher coined the term 'likelihood' in 1921 to distinguish the method of maximum likelihood from the Bayesian or inverse probability argument. In the early days its application was fairly limited; few statistical techniques from the 1920s to 1950s could be called 'likelihood-based'. To see why, let us consider what we mean by 'statistical activities':

- *planning:* making decisions about the study design or sampling protocol, what measurements to take, stratification, sample size, etc.
- *describing:* summarizing the bulk of data in few quantities, finding or revealing meaningful patterns or trends, etc.
- *modelling:* developing mathematical models with few parameters to represent the patterns, or to explain the variability in terms of relationship between variables.
- *inference:* assessing whether we are seeing a real or spurious pattern or relationship, which typically involves an evaluation of the uncertainty in the parameter estimates.
- *model checking:* assessing whether the model is sensible for the data. The most common form of model checking is residual analysis.

A lot of early statistical works was focused on the first two activities, for which likelihood thinking does not make much contribution. Often the activity moved directly from description to inference with little modelling in between. Also, the early modelling scene was dominated by the normal-based linear models, so statisticians could survive with least-squares, and t tests or F tests (or rank tests if the data misbehaved).

The emergence of likelihood-based modelling had to wait for both the advent of computing power and the arrival of more challenging data analysis problems. These problems typically involve nonnormal outcome data, with possible complexities in their collection such as censoring, repeated measures, etc. In these applications, modelling is important to impose

structure or achieve simplification. This is where the likelihood becomes indispensable.

Plan of the book

The chapters in this book can be categorized loosely according to

- modelling: Chapters 4, 6, 11, 14, 17, 18;
- inference: Chapters 2, 3, 5, 7, 10, 13, 15, 16.

The inference chapters describe the anatomy of likelihood, while the modelling chapters show its physiology or functioning. The other chapters are historical (Chapter 1) or technical support (Chapters 8, 9, 12).

There is no need to proceed sequentially. Traditionally, likelihood inference requires the large sample theory covered in Chapter 9, so some instructors might feel more comfortable to see the theory developed first. *Some sections are starred* to indicate that they can be skipped on first reading, or they are optional as teaching material, or they involve ideas from future sections. In the last case, the section is there more for organizational reasons, so some 'nonlinear' reading might be required.

There is much more material here than can be covered in two semesters. In about 50 lectures to beginning graduate students I covered a selection from Chapters 2 to 6, 8 to 11, 13 and 14. Chapter 1 is mostly for reading; I use the first lecture to discuss the nature of statistical problems and the different schools of statistics. Chapter 7 is also left as reading material. Chapter 12 is usually covered in a separate statistical computing course. Ideally Chapter 15 is covered together with Chapters 13 and 14, while the last three chapters also form a unit on mixed models. So, for a more leisurely pace, Chapters 13 to 14 can be removed from the list above, and covered separately in a more advanced modelling course that covers Chapters 13 to 18.

Prerequisites

This book is intended for senior students of statistics, which include advanced undergraduate or beginning graduate students. Students taking this course should already have

- two semesters of introductory applied statistics. They should be familiar with common statistical procedures such as z, t, and χ^2 tests, P-value, simple linear regression, least-squares principle and analysis of variance.
- two semesters of introduction to probability and theory of statistics. They should be familiar with standard probability models such as the binomial, negative binomial, Poisson, normal, exponential, gamma, etc.; with the concepts of conditional expectation, Bayes theorem, transformation of random variables; with rudimentary concepts of estimation, such as bias and the method of moments; and with the central limit theorem.

- two semesters of calculus, including partial derivatives, and some matrix algebra.
- some familiarity with a flexible statistical software package such as `Splus` or `R`. Ideally this is learned in conjunction with the applied statistics course above.

The mathematical content of the book is kept relatively low (relative to what is possible). I have tried to present the whole spectrum of likelihood ideas from both applied and theoretical perspectives, both showing the depth of the ideas. To make these accessible I am relying (most of the time) on a nontechnical approach, using heuristic arguments and encouraging intuitive understanding. What is intuitive for me, however, may not be so for the reader, so sometimes the reader needs to balance the personal words with the impersonal mathematics.

Computations and examples

Likelihood-based methods are inherently computational, so computing is an essential part of the course. Inability to compute impedes our thought processes, which in turn will hamper our understanding and willingness to experiment. For this purpose it is worth learning a statistical software package. However, not all packages are created equal; different packages have different strengths and weaknesses. In choosing a software package for this course, bear in mind that here we are not trying to perform routine statistical analyses, but to learn and understand what is behind them, so graphics and programming flexibility are paramount.

All the examples in this book can be programmed and displayed quite naturally using `R` or `Splus`. `R` is *free* statistical programming software developed by a dedicated group of statisticians; it can be downloaded from `http://cran.r-project.org`.

Most educators tell us that understanding is best achieved through direct experience, in effect letting the knowledge pass through the fingers rather than the ears and the eyes only. Students can get such an experience from verifying or recreating the examples, solving the exercises, asking questions that require further computations, and, best still, trying out the methodology with their own data. To help, I have put all the `R` programs I used for the examples in `http://www.meb.ki.se/~yudpaw`.

Acknowledgements

This book is an outgrowth of the lecture notes for a mathematical statistics course in University College Dublin and University College Cork. I am grateful to the students and staff who attended the course, in particular Phil Boland, John Connolly, Gabrielle Kelly and Dave Williams, and to University College Dublin for funding the sabbatical leave that allowed me to transform the notes into tidy paragraphs. Jim Lindsey was generous with comments and encouragement. Kathleen Loughran and Áine Allen were of enormous help in hunting down various errors; I would kindly ask the reader

to please let me know if they spot any remaining errors. It is impossible to express enough gratitude to my parents. During the sabbatical leave my mother took a special interest in the book, checking every morning whether it was finished. In her own way she made it possible for me to concentrate fully on writing. And finally, *buíochas le mo bhean chéile* Marie Reilly.

Y.P.
Cork
April, 2001

Notes on the 2003 corrected edition

Since the book was published in 2001 I received many kind words as well as corrections, which I took as indications that the book is being read and that the effort of writing it had been worthwhile. Special thanks go to John Nelder and Hiroshi Okamura for such a careful reading of the book. I am also grateful to Pat Altham, Harry Southworth, Aji Hamim Wigena, Jin Peng and various other people for comments and for informing me of errors. The current website for R programs, datasets and list of errors is moved to `http://www.mep.ki.se/~yudpaw`.

Y.P.
Stockholm
June, 2003

Notes on the 2008 corrected edition

I have corrected around 15 errors found since 2003 (these are listed in my website); the good news is that – assuming people are still reading this book – the rate of errors found has reduced dramatically. The current website for R programs, datasets and list of errors is moved to `http://www.meb.ki.se/~yudpaw`.

Y.P.
Stockholm
September, 2008

Notes on the paperback edition

I am very happy that OUP decided to publish the book in paperback and eBook forms, as these will make it more affordable. Recently found typos are in `http://www.meb.ki.se/~yudpaw/likelihood/errors.htm`.

Y.P.
Stockholm
June, 2012