Preface

Multivariable Model-Building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables is principally written for scientists (including statisticians, researchers and graduate students) working with regression models in all branches of application. Our general objective is to provide a readable text giving the rationale of, and practical advice on, a unified approach to multivariable modelling which aims to make such models simpler and more effective. Specifically, we focus on the selection of important variables and the determination of functional form for continuous predictors. Since our own background is in biostatistics and clinical research, inevitably there is a focus on applications in medicine and the health sciences, but the methodology is much more widely useful. The topic of multivariable model-building is very broad; we doubt if it is possible to cover all the relevant topics in a single book. Therefore, we have concentrated on what we see as a few key issues. No multivariable model-building strategy has rigorous theoretical underpinnings. Even those approaches most used in practice have not had their properties studied adequately by simulation. In particular, handling continuous variables in a multivariable context has largely been ignored. Since there is no consensus among researchers on the 'best' strategy, a pragmatic approach is required. Our book reflects our views derived from wide experience. The text assumes a basic understanding of multiple regression modelling, but it can be read without detailed mathematical knowledge.

Multivariable regression models are widely used in all areas of science in which empirical data are analysed. We concentrate on normal-errors models for continuous outcomes, logistic regression for binary outcomes and Cox regression for censored time-to-event data. Our methodology is easily transferred to more general regression models. As expressed in a very readable paper by Chatfield (2002), we aim to 'encourage and guide practitioners, and also to counterbalance a literature that can be overly concerned with theoretical matters far removed from the day-to-day concerns of many working statisticians'. The main focus is the modelling of continuous covariates by using fractional polynomials. The methods are illustrated by the analysis of many datasets, mainly from clinical epidemiology, ranging from prognostic factors in breast cancer and treatment in kidney cancer to risk factors in heart disease.

WHAT IS IN OUR BOOK

Our main concern is how to build a multivariable regression model from several candidate predictors, some of which are continuous.We are more interested in explanatory models (that is, in assessing the effects of individual variables in a multivariable context) than in deriving

a ‘good’ predictor without regard to its components. The basic techniques that are dealt with in many textbooks on regression analysis are not repeated.

Chapters 2 and 3 deal mainly with the selection of variables and coping with different types of variables in a modelling context. Relationships with continuous covariates are assumed linear. The importance of the coding chosen for categorical covariates is discussed. Chapters and 5 provide a reasonably comprehensive account of univariate fractional polynomial (FP) models, our preferred method of working with continuous predictors.We introduce the function selection procedure (FSP). Chapter 6, the heart of the book, introduces multivariable FP (MFP) modelling, combining backward elimination with the FSP. In Chapter 7, FP modelling is extended to include interactions between predictors, both categorical-by-continuous and continuous-by-continuous. Chapter 8 looks at techniques for assessing the stability of multivariable models. Bootstrap resampling is the key tool here. Chapter 9 briefly outlines spline models. We introduce two multivariable modelling procedures in which the FSP is adapted for splines, and we compare the results with FP models in several examples. Chapter 10 is a fairly self-contained guide to working with MFPs, taking a problem-oriented approach using an artificial but realistic dataset. A practitioner with some experience in regression modelling should be able to take in the principles and practice of MFP modelling from this chapter. As throughout the book, frequent use is made of model criticism, particularly of plots of fitted functions and of smoothed residuals, and of techniques for assessing the effects of influential observations on the selected model. Chapter 11 is a brief tour of further applications of FP methodology. Chapter 12 gives our recommendations for practice, briefly discusses some topics not dealt with in our book, and points to further research.We lay stress on deriving parsimonious models that make sense from a subject-matter viewpoint. We are more concerned with getting the ‘big picture’ right than in refining the minor details of a model fit.

HOW TO READ OUR BOOK

The chapters have been organized such that the ideas unfold in a logical sequence, with Chapter 1 providing motivation and a flavour of what is to follow. However, to grasp the core ideas of our book more rapidly, we suggest as a bare minimum reading the following segments:

- Section 1.7 defines our approach to modelling in general terms.
- Section 2.6 discusses stepwise and other procedures for selecting variables.
- Sections 4.2–4.10 introduce FP functions and show how they are used in modelling a single continuous predictor. Section 4.14 contains a worked example. For an experienced modeller, the example may be a sufficient guide to the main principles.
- Sections 6.1, 6.2, 6.3 and 6.5 describe the key parts of the MFP method of multivariable

model-building.

• Chapter 10 is particularly recommended to the practitioner who wants an appreciation of how to use MFP. We include material on some of the pitfalls that may be avoided using simple diagnostic techniques. Sections 10.5.6 and 10.8.3 on interactions may be omitted at a first reading.

• Chapter 12 summarizes some recommendations for practice.

SOFTWARE AND DATA

For practical use it is important that the necessary software is generally available. Software for the basic MFP method has been implemented in Stata, SAS and R. Special-purpose programs for Stata are available on our book's website http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book for all the extensions we describe. In some of the examples, we showthat use of the software is simple if basic principles of the methodology are understood. To assist the reader in developing their own experience in multivariable model-building, some of the datasets used in the book are available on the website.

EDUCATIONAL RESOURCES

Supplementary materials, including datasets, software, exercises and relevant Web links, are available on the website. Many of the issues in multivariable model-building with continuous covariates discussed in our book are explored in Chapter 10, where an artificial dataset (the 'ART study') is described and analysed. The dataset and details of the design and Stata programs used to create it are available on the website, allowing the data to be modified for different purposes. Exercises based on the ART study are suggested. We would encourage others to extend and develop the exercises as part of the material used to teach MFP methodology. Slide presentations are available as a starting point for preparing talks and teaching the material.

ACKNOWLEDGEMENTS

We are indebted to our colleagues at the MRC Clinical Trials Unit, London, and at the Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg for discussion and encouragement. Our research and the book have benefited from constructive comments from many people, including particularly the following: Doug Altman, Harald Binder, Gareth Ambler, Carol Coupland, Christel Faes, David Hosmer, Tony Johnson, Paul

Lambert, Rumana Omar, Michael Schemper, Martin Schumacher, Simon Thompson, and Hans van Houwelingen.We thank the following for kind permission to use their valuable datasets in our book: Lyn Chitty (fetal growth), Tim Cole (triceps), John Foekens and Maxime Look (Rotterdam breast cancer), Amy Luke (research body fat), John Matthews and Maeve O'Sullivan (nerve conduction), Alastair Ritchie and Mahesh Parmar (kidney cancer), Philip Rosenberg (oral cancer), Martin Shipley (Whitehall I). We are grateful to Lena Barth, Karina Gitina, Georg Koch and Edith Motschall for technical assistance. Finally, we owe very special thanks to the director and staff of the Mathematisches Forschungsinstitut Oberwolfach, Germany. The excellent atmosphere and working conditions during visits there over several years were conducive to the development of many research ideas and papers which led up to our book.

London and Freiburg
November 2007