# Semiparametric Regression

Semiparametric regression is concerned with the flexible incorporation of nonlinear functional relationships in regression analyses. Any application area that uses regression analysis can benefit from semiparametric regression. Assuming only a basic familiarity with ordinary parametric regression, this user-friendly book explains the techniques and benefits of semiparametric regression in a concise and modular fashion. The authors make liberal use of graphics and examples plus case studies taken from environmental, financial, and other applications. They include practical advice on implementation and pointers to relevant software.

This book is suitable as a textbook for students with little background in regression as well as a reference book for statistically oriented scientists – such as biostatisticians, econometricians, quantitative social scientists, and epidemiologists – with a good working knowledge of regression and the desire to begin using more flexible semiparametric models. Even experts on semiparametric regression should find something new here.

David Ruppert is the Andrew Schultz, Jr., Professor of Engineering (School of Operations Research and Industrial Engineering) and Professor of Statistical Science at Cornell University. He has served as editor for a number of prestigious series and journals and has published some 80 articles of his own as well as co-authoring two popular books, *Transformation and Weighting in Regression* and *Measurement Error in Nonlinear Models.* He is also winner of the Wilcoxon Prize for best practical applications paper in technometrics and an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics.

M. P. Wand is Professor of Statistics at the University of New South Wales in Sydney, Australia. He has held faculty appointments at Harvard University, Rice University, and Texas A&M University. Dr. Wand is a Fellow of the American Statistical Association and has served as an associate editor for the *Journal of the American Statistical Association* and *Biometrika*. He is winner of the P. A. P. Moran Medal for statistical research.

R. J. Carroll is Distinguished Professor of Statistics, Nutrition and Toxicology at Texas A&M University. Among his many honors are the COPSS Presidents' Award, the Fisher Lecture, the Snedecor Award, and the Wilcoxon Prize. He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics as well as an elected member of the International Statistical Institute.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS



This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

*Already published*

1. *Bootstrap Methods and Their Application,* by A. C. Davison and D. V. Hinkley
2. *Markov Chains,* by J. Norris
3. *Asymptotic Statistics,* by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis,* by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods,* by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation,* by Sara van de Geer
7. *Numerical Methods of Statistics,* by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability,* by David Pollard
9. *The Estimation and Tracking of Frequency,* by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R,* by John Maindonald and John Braun
11. *Statistical Models,* by A. C. Davison

# Semiparametric Regression

DAVID RUPPERT
*Cornell University*

M. P. WAND
*Harvard University*

R. J. CARROLL
*Texas A&M University*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by
Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521780506



First published 2003
Reprinted 2005, 2006, 2008, 2009

*A catalogue record for this publication is available from the British Library*

*Library of Congress Cataloguing in Publication data*

Ruppert, David, 1948–
Semiparametric regression / David Ruppert, M. P. Wand, R. J. Carroll.
p.cm.
Includes bibliographical references and index.
ISBN 0-521-78050-0 – ISBN 0-521-78516-2 (pb.)
1. Regression analysis. 2. Nonparametric statistics. I. Wand, M. P. (Matthew P.).
II. Carroll, Raymond J. III. Title.
QA278.2.R87 2003
519.5´36 – dc21 2002041460

ISBN 978-0-521-78050-6 Hardback
ISBN 978-0-521-78516-7 Hardback

*To Anne, with love*
— David

*To my wife's parents, Ayhan and Recep*
— Matt

*To Brett and Jeb*
— Raymond

# Contents

# Preface

The primary aim of this book is to guide researchers needing to flexibly incorporate nonlinear relationships into their regression analyses. Flexible nonlinear regression is traditionally known as *nonparametric regression*; it differs from parametric regression in that the shape of the functional relationships are not predetermined but can adjust to capture unusual or unexpected features of the data.

Almost all existing regression texts treat either parametric or nonparametric regression exclusively. The level of exposition between books of either type differs quite alarmingly. In this book we argue that nonparametric regression can be viewed as a relatively simple extension of parametric regression and treat the two together. We refer to this combination as *semiparametric regression*. Our approach to semiparametric regression is based on penalized regression splines and mixed models. Indeed, every model in this book is a special case of the linear mixed model or its generalized counterpart. This makes the methodology modular and is in keeping with our general philosophy of *minimalist statistics* (see Section 19.2), where the amount of methodology, terminology, and so on is kept to a minimum. This is the first smoothing book that makes use of the mixed model representation of smoothers.

Unlike many other texts on nonparametric regression, this book is very much problem-driven. Examples from our collaborative research (and elsewhere) have driven the selection of material and emphases and are used throughout the book.

The book is suitable for several audiences. One audience consists of students or working scientists with only a moderate background in regression, though familiarity with matrix and linear algebra is assumed. Marginal notes and the appendices are intended for beginners, especially those from interface disciplines. We make liberal use of graphics because visualization is a particularly effective tool for acquiring intuition in a new subject.

Another audience that we are aiming at consists of statistically oriented scientists (e.g., biostatisticians, econometricians, quantitative social scientists, and epidemiologists) who have a good working knowledge of linear models and the desire to begin using more flexible semiparametric models. There are many connections between linear and nonparametric regression. Our goal is to exploit them and the reader's knowledge of linear models to provide a foundation for understanding nonparametric modeling.

There is enough new material to be of interest even to experts on smoothing, and they are a third possible audience.

There are several competing approaches to nonparametric modeling: smoothing splines (e.g., Eubank 1988, 1999; Wahba 1990; Green and Silverman 1994); series-based smoothers, including wavelets (Tarter and Lock 1993; Ogden 1996); kernel methods, including local regression (Wand and Jones 1995; Fan and Gijbels 1996); and regression splines (Friedman 1991; Stone et al. 1997; Hansen and Kooperberg 2002). All four approaches can be used effectively and have their devotees. We believe that the nature of the data should play a role in the choice among them. For example, wavelets are more suited to highly oscillatory functions. Apart from this, the choice of a nonparametric regression method is a matter somewhat of individual taste and background. Based on our motivating applications and personal tastes, the approach to nonparametric regression used throughout this book is what we call *penalized splines,* although they are also labeled as *P-splines, pseudosplines,* and *low-rank spline smoothers* in the literature. Penalized splines are quite similar to smoothing splines; in fact, they are a generalization of smoothing splines that allow more flexible choices of the spline model, the basis functions for that model, and the penalty.

Penalized splines have close ties with ridge regression, mixed models, and Bayesian statistics, ties that were discovered by researchers working on smoothing splines. These ties allow techniques from mixed models – for example, (restricted) maximum likelihood estimation and likelihood ratio tests – to be added to penalized spline methodology. Similarly, Bayesian techniques based on Markov chain Monte Carlo provide what we believe to be the most satisfactory approach to fitting complex semiparametric models as well as the direction that semiparametric regression is most likely to take in the future. This book includes introductions to mixed models and to Bayesian modeling.

*Acknowledgments*

We are especially grateful to Ciprian Crainiceanu and Bhaswati Ganguli for their assistance in the preparation of this book. Ciprian wrote the WinBugs program in Appendix B and wrote the programs used for simulations-based $p$-values for likelihood ratio tests. Several other of our colleagues and collaborators have contributed to the book in various ways. We would like to thank Marc Aerts, Babette Brumback, Tianxi Cai, Gerda Claeskens, Brent Coull, Maria Durban, Garrett Fitzmaurice, Jonathan French, Robert Gentleman, Bob Gray, Nick Horton, Joe Ibrahim, Erin Kammann, Göran Kauermann, Robert Kohn, Nan Laird, Nick Lange, Mary Lindstrom, Long Ngo, Doug Nychka, Michael O'Connell, Helen Parise, José Pinheiro, Louise Ryan, Misha Salganik, Joel Schwartz, John Staudenmayer, Sally Thurston, Carrie Wager, Naisyin Wang, Jim Ware, Antonella Zanobetti, and Yihua Zhao for their collaboration, interest, and comments.

We thank Lauren Cowles for being a very supportive and patient editor. We would like to express our gratitude to Tom Ryan and Misha Salganik for sending us errata.

The second author lovingly acknowledges the support of his wife, Handan, and children, Declan and Jaida, throughout this project. Support of the Department of Biostatistics, Harvard University, is also gratefully acknowledged.

# Guide to Notation

This chapter gives a brief overview of notational conventions used in the book. Please see the Notation Index for more specialized notation.

The symbol "$\equiv$" means "equal by definition".

We use both lower- and uppercase letters (e.g., $x$, $X$, and $\lambda$) to denote scalar quantities, either fixed or random. Lowercase bold letters (e.g., $\mathbf{x}$ and $\boldsymbol{\lambda}$) will be used for vectors. Uppercase bold fonts (e.g., $\mathbf{X}$ and $\boldsymbol{\Lambda}$) will denote matrices. The entries of a vector or matrix use the same letter and case as the vector or matrix itself but are not bold. Thus,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

If a matrix is partitioned then the submatrices are in bold; for example,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

We will indicate the row index of a matrix to the right and the column index below, as in:

$$\mathbf{C} = \left[ c_{ik} \right]_{1 \le k \le K}{}_{1 \le i \le n}.$$

The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}^{\mathsf{T}}$. If $\mathbf{A}$ is an invertible square matrix, then $\mathbf{A}^{-1}$ denotes its inverse. Any vector is assumed to be a column, so its transpose is a row.

The *norm* of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|$; that is,

$$\|\mathbf{x}\| \equiv \sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{x}}.$$

The real line will be denoted by $\mathbb{R}$, and $d$-dimensional space will be denoted by $\mathbb{R}^d$.

For a function $f(x)$ of a scalar $x$,

$$f^{(r)}(x) \equiv (d^r/dx^r) f(x),$$

the $r$th derivative of $f(x)$.

If $f(\mathbf{x})$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$ then the *derivative vector* is a $1 \times d$ row vector with $j$th entry equal to $(\partial/\partial x_j) f(\mathbf{x})$, the partial derivative of $f(\mathbf{x})$ with respect to $x_j$, and is denoted by

$$\mathsf{D} f(\mathbf{x}).$$

The *Hessian matrix* is a $d \times d$ matrix whose $(i, j)$ entry is equal to

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x});$$

it is denoted by

$$\mathsf{H} f(\mathbf{x}).$$

If $x$ and $y$ are random variables, then $\mathsf{E}(x)$, $\mathrm{var}(x)$, and $\mathrm{st.dev.}(x)$ are the mean, variance, and standard deviation of $x$, and $\mathrm{cov}(x, y)$ is the covariance between $x$ and $y$. $\mathrm{Cov}(\mathbf{x})$ is the covariance matrix of a random vector $\mathbf{x}$; see Appendix A for its definition.