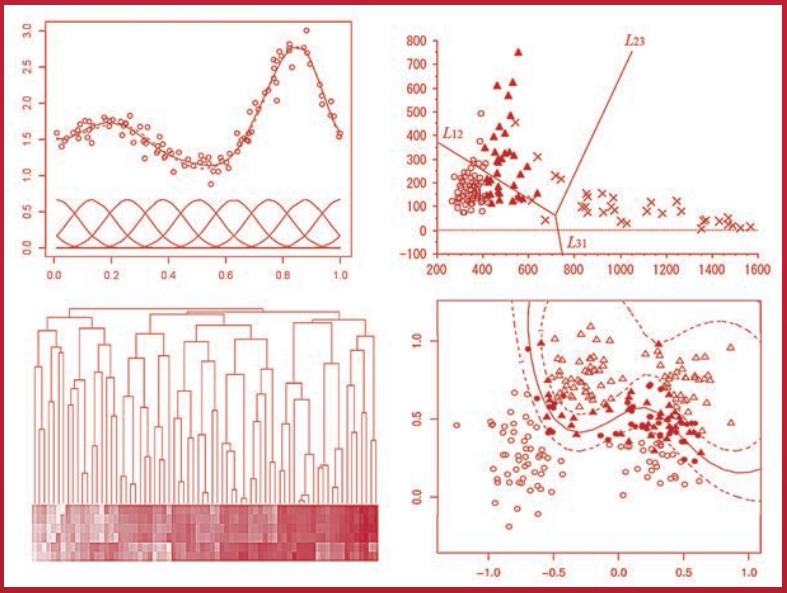
Texts in Statistical Science

Introduction to Multivariate Analysis

Linear and Nonlinear Modeling



Sadanori Konishi

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Introduction to Multivariate Analysis

Linear and Nonlinear Modeling

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

Interpreting Data: A First Course in Statistics
A.J.B. Anderson

Introduction to Probability with R
K. Baclawski

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

Statistical Methods for SPC and TQM
D. Bissell

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Second Edition
R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition
C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition
C. Chatfield

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition
D. Collett

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

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples
D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks
M. Crowder

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

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

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

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

A Course in Large Sample Theory
T.S. Ferguson

Multivariate Statistics: A Practical Approach
B. Flury and H. Riedwyl

Readings in Decision Analysis
S. French

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Bayesian Data Analysis, Third Edition
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin

Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Practical Data Analysis for Designed Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Logistic Regression Models
J.M. Hilbe

Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects
J.S. Hodges

Statistics for Epidemiology
N.P. Jewell

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

The Theory of Linear Models
B. Jørgensen

Principles of Uncertainty
J.B. Kadane

Graphics for Statistics and Data Analysis with R
K.J. Keen

Mathematical Statistics
K. Knight

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

Nonparametric Methods in Statistics with SAS Applications
O. Korosteleva

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

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Design and Analysis of Experiments with SAS
J. Lawson

A Course in Categorical Data Analysis
T. Leonard

Statistics for Accountants
S. Letchford

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Statistical Theory, Fourth Edition
B.W. Lindgren

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

Time Series Analysis
H. Madsen

Pólya Urn Models
H. Mahmoud

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistics in Engineering: A Practical Approach
A.V. Metcalfe

Beyond ANOVA: Basics of Applied Statistics
R.G. Miller, Jr.

A Primer on Linear Models
J.F. Monahan

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Elements of Simulation
B.J.T. Morgan

Probability: Methods and Measurement
A. O'Hagan

Introduction to Statistical Limit Theory
A.M. Polansky

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Statistics in Research and Development, Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Introduction to Statistical Process Control
P. Qiu

Sampling Methodologies with Applications
P.S.R.S. Rao

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Essential Statistics, Fourth Edition
D.A.G. Rees

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

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Decision Analysis: A Bayesian Approach
J.Q. Smith

Analysis of Failure and Survival Data
P. J. Smith

Applied Statistics: Handbook of GENSTAT Analyses
E.J. Snell and H. Simpson

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Data Driven Statistical Methods
P. Sprent

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Survival Analysis Using S: Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

Statistical Process Control: Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Generalized Additive Models: An Introduction with R
S. Wood

Epidemiology: Study Design and Data Analysis, Third Edition
M. Woodward

Experiments
B.S. Yandell

Texts in Statistical Science

Introduction to Multivariate Analysis

Linear and Nonlinear Modeling

Sadanori Konishi
Chuo University
Tokyo, Japan

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

Originally published in Japanese by Iwanami Shoten, Publishers, Tokyo, 2010. This English language edition published in 2014 by Chapman & Hall/CRC, Boca Raton, FL, U.S.A., by arrangement with the author c/o Iwanami Shoten, Publishers, Tokyo.

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

Version Date: 20140508

International Standard Book Number-13: 978-1-4665-6729-0 (eBook - PDF)

Contents

This page intentionally left blank

List of Figures

This page intentionally left blank

List of Tables

This page intentionally left blank

Preface

The aim of statistical science is to develop the methodology and the theory for extracting useful information from data and for reasonable inference to elucidate phenomena with uncertainty in various fields of the natural and social sciences. The data contain information about the random phenomenon under consideration and the objective of statistical analysis is to express this information in an understandable form using statistical procedures. We also make inferences about the unknown aspects of random phenomena and seek an understanding of causal relationships.

Multivariate analysis refers to techniques used to analyze data that arise from multiple variables between which there are some relationships. Multivariate analysis has been widely used for extracting useful information and patterns from multivariate data and for understanding the structure of random phenomena. Techniques would include regression, discriminant analysis, principal component analysis, clustering, etc., and are mainly based on the linearity of observed variables.

In recent years, the wide availability of fast and inexpensive computers enables us to accumulate a huge amount of data with complex structure and/or high-dimensional data. Such data accumulation is also accelerated by the development and proliferation of electronic measurement and instrumentation technologies. Such data sets arise in various fields of science and industry, including bioinformatics, medicine, pharmaceuticals, systems engineering, pattern recognition, earth and environmental sciences, economics, and marketing. Therefore, the effective use of these data sets requires both linear and nonlinear modeling strategies based on the complex structure and/or high-dimensionality of the data in order to perform extraction of useful information, knowledge discovery, prediction, and control of nonlinear phenomena and complex systems.

The aim of this book is to present the basic concepts of various procedures in traditional multivariate analysis and also nonlinear techniques for elucidation of phenomena behind observed multivariate data, focusing primarily on regression modeling, classification and discrimination, dimension reduction, and clustering. Each chapter includes many figures

and illustrative examples to promote a deeper understanding of various techniques in multivariate analysis.

In practice, the need always arises to search through and evaluate a large number of models and from among them select an appropriate model that will work effectively for elucidation of the target phenomena. This book provides comprehensive explanations of the concepts and derivations of the AIC, BIC, and related criteria, together with a wide range of practical examples of model selection and evaluation criteria. In estimating and evaluating models having a large number of predictor variables, the usual methods of separating model estimation and evaluation are inefficient for the selection of factors affecting the outcome of the phenomena. The book also reflects these aspects, providing various regularization methods, including the L_1 norm regularization that gives simultaneous model estimation and variable selection.

The book is written in the hope that, through its fusion of knowledge gained in leading-edge research in statistical multivariate analysis, machine learning, and computer science, it may contribute to the understanding and resolution of problems and challenges in this field of research, and to its further advancement.

This book might be useful as a text for advanced undergraduate and graduate students in statistical sciences, providing a systematic description of both traditional and newer techniques in multivariate analysis and machine learning. In addition, it introduces linear and nonlinear statistical modeling for researchers and practitioners in various scientific disciplines such as industrial and systems engineering, information science, and life science. The basic prerequisites for reading this textbook are knowledge of multivariate calculus and linear algebra, though they are not essential as it includes a self-contained introduction to theoretical results.

This book is basically a translation of a book published in Japanese by Iwanami Publishing Company in 2010. I would like to thank Uichi Yoshida and Nozomi Tsujimura of the Iwanami Publishing Company for giving me the opportunity to translate and publish in English.

I would like to acknowledge with my sincere thanks Yasunori Fujikoshi, Genshiro Kitagawa, and Nariaki Sugiura, from whom I have learned so much about the seminal ideas of statistical modeling. I have been greatly influenced through discussions with Tomohiro Ando, Yuko Araki, Toru Fujii, Seiya Imoto, Mitsunori Kayano, Yoshihiko Maesono, Hiroki Masuda, Nagatomo Nakamura, Yoshiyuki Ninomiya, Ryuei Nishii, Heewon Park, Fumitake Sakaori, Shohei Tateishi, Takahiro Tsuchiya, Masayuki Uchida, Takashi Yanagawa, and Nakahiro Yoshida.

I would also like to express my sincere thanks to Kei Hirose, Shuichi Kawano, Hidetoshi Matsui, and Toshihiro Misumi for reading the manuscript and offering helpful suggestions. David Grubbs patiently encouraged and supported me throughout the final preparation of this book. I express my sincere gratitude to all of these people.

Sadanori Konishi

Tokyo, January 2014

This page intentionally left blank

Chapter 1

Introduction

The highly advanced computer systems and progress in electronic measurements and instrumentation technologies have together facilitated the acquisition and accumulation of data with complex structure and/or high-dimensional data in various fields of science and industry. Data sets arise in such areas as genome databases in life science, remote-sensing data from earth-observing satellites, real-time recorded data of motion process in system engineering, high-dimensional data in character recognition, speech recognition, image analysis, etc. Hence, it is desirable to research and develop new statistical data analysis techniques to efficiently extract useful information as well as elucidate patterns behind the data in order to analyze various phenomena and to yield knowledge discovery. Under the circumstances linear and nonlinear multivariate techniques are rapidly developing by fusing the knowledge in statistical science, machine learning, information science, and mathematical science.

The objective of this book is to present the basic concepts of various procedures in the traditional multivariate analysis and also nonlinear techniques for elucidation of phenomena behind the observed multivariate data, using many illustrative examples and figures. In each chapter, starting from an understanding of the traditional multivariate analysis based on the linearity of multivariate observed data, we describe nonlinear techniques, focusing primarily on regression modeling, classification and discrimination, dimension reduction, and clustering.

1.1 Regression Modeling

Regression analysis is used to model the relationship between a response variable and several predictor (explanatory) variables. Once a model has been identified, various forms of inferences such as prediction, control, information extraction, knowledge discovery, and risk evaluation can be done within the framework of deductive argument. Thus, the key to solving various real-world problems lies in the development and construction of suitable linear and nonlinear regression modeling.

1.1.1 Regression Models

Housing prices vary with land area and floor space, but also with proximity to stations, schools, and supermarkets. The quantity of chemical products is sensitive to temperature, pressure, catalysts, and other factors. In Chapter 2, using *linear regression* models, which provide a method for relating multiple factors to the outcomes of such phenomena, we describe the basic concept of *regression modeling*, including model specification based on data reflecting the phenomena, model estimation of the specified model by least squares or maximum likelihood methods, and model evaluation of the estimated model. Throughout this modeling process, we select a suitable one among competing models.

The volume of extremely high-dimensional data that are observed and entered into databases in biological, genomic, and many other fields of science has grown rapidly in recent years. For such data, the usual methods of separating model estimation and evaluation are ineffectual for the selection of factors affecting the outcome of the phenomena, and thus effective techniques are required to construct models with high reliability and prediction. This created a need for work on modeling and has led, in particular, to the proposal of various regularization methods with an L_1 penalty term (the sum of absolute values of regression coefficients), in addition to the sum of squared errors and log-likelihood functions. A distinctive feature of the proposed methods is their capability for simultaneous model estimation and variable selection. Chapter 2 also describes various regularization methods, including *ridge* regression (Hoerl and Kennard, 1970) and the least absolute shrinkage and selection operator (*lasso*) proposed by Tibshirani (1996), within the framework of linear regression models.

Figure 1.1 shows the results of an experiment performed to investigate the relation between falling time (x sec) and falling distance (y m) of a body. The figure suggests that it should be possible to model the relation using a polynomial. There are many phenomena that can be modeled in this way, using polynomial equations, exponential functions, or other specific nonlinear functions to relate the outcome of the phenomenon and the factors influencing that outcome.

Figure 1.2, however, poses new difficulties. It shows the measured impact y (in acceleration, g) on the head of a dummy in repeated experimental crashes of a motorcycle into a wall, with a time lapse of x (msec) as measured from the instant of collision (Härdle, 1990). For phenomena with this type of apparently complex nonlinear structure, it is quite difficult to effectively capture the structure by modeling with specific

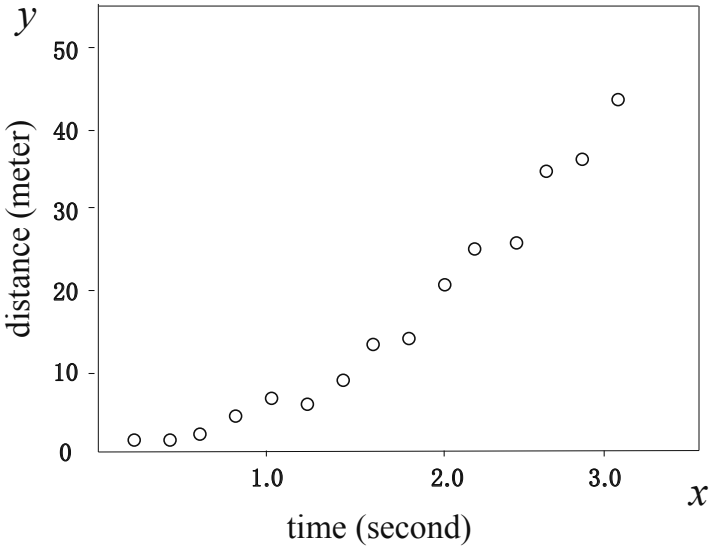


Figure 1.1 *The relation between falling time (x sec) and falling distance (y m) of a body.*

nonlinear functions such as polynomial equations and exponential functions.

Chapter 3 discusses *nonlinear regression* modeling for extracting useful information from data containing complex nonlinear structures. It introduces models based on more flexible splines, *B*-splines, and radial basis functions for modeling complex nonlinear structures. These models often serve to ascertain complex nonlinear structures, but their flexibility often prevents their effective function in the estimation of models with the traditional least squares and maximum likelihood methods. In such cases, these estimation methods are replaced by regularized least squares and regularized maximum likelihood methods.

The latter two techniques, which are generally referred to as *regularization* methods, are effectively used to reduce over-fitting of models to data and thus prevent excessive model complexity, and are known to contribute for reducing the variability of the estimated models. This chapter also describes regularization methods within the framework of nonlinear regression modeling.

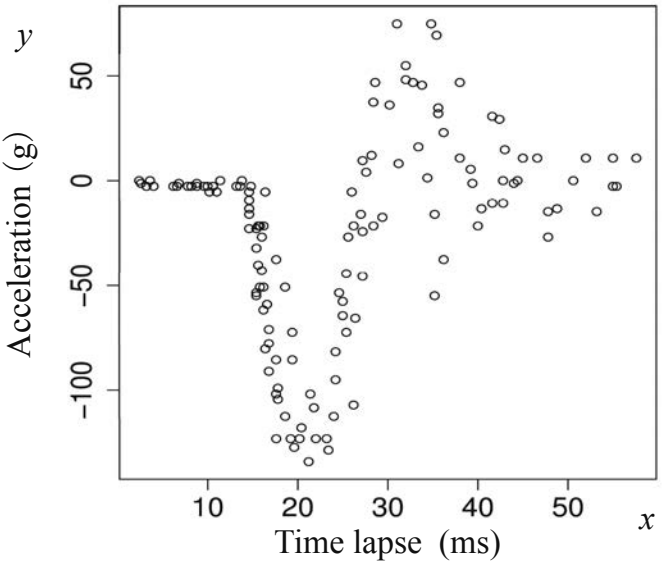


Figure 1.2 *The measured impact y (in acceleration, g) on the head of a dummy in repeated experimental crashes of a motorcycle with a time lapse of x (msec).*

1.1.2 Risk Models

In today's society marked by complexity and uncertainty, we live in a world exposed by various types of risks. The risk may be associated with occurrences such as traffic accidents, natural disasters such as earthquakes, tsunamis, or typhoons, or development of a lifestyle disease, with transactions such as credit card issuance, or with many other occurrences too numerous to enumerate. It is possible to gauge the magnitude of risk in terms of probability based on past experience and information gained in life in society, but often with only a limited accuracy.

All of this poses the question of how to probabilistically assess unknown risks for a phenomenon using information obtained from data. For example, in searching for the factors that induce a certain disease, the problem is in how to construct a model for assessing the probability of its occurrence based on observed data. The effective probabilistic model for assessing the risk may lead to its future prevention. Through such risk modeling, moreover, it may also be possible to identify important disease-related factors.

Chapter 4 presents an answer to this question, in the form of model-

ing for the risk evaluation, and in particular describes the basic concept of *logistic regression modeling*, together with its extension from linear to nonlinear modeling. This includes models to assess risks based on binary data {0, 1} expressing the presence or absence of response in an individual or object on exposure to various levels of stimulus, as shown in Figure 1.3.

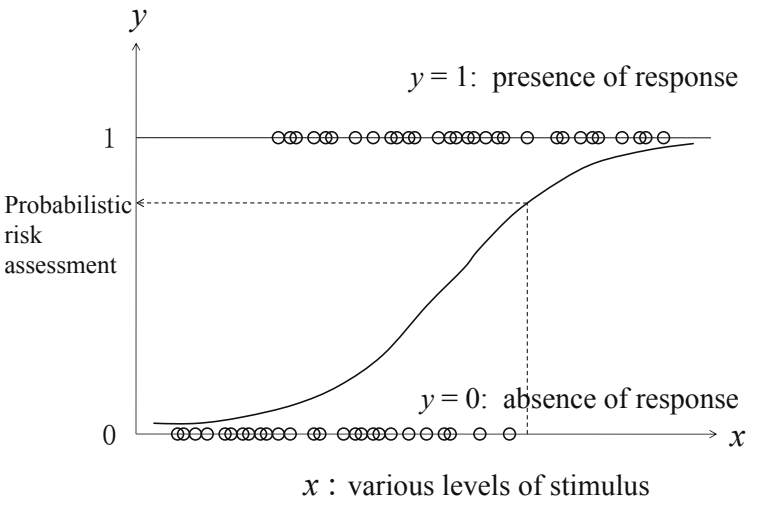


Figure 1.3 *Binary data {0, 1} expressing the presence or absence of response in an individual on exposure to various levels of stimulus.*

1.1.3 Model Evaluation and Selection

Figure 1.4 shows a process consisting essentially of the conceptualization of *regression modeling*; the specification of models that approximates the structure of a phenomenon, the estimation of their parameters, and the evaluation and selection of estimated models.

In relation to the data shown in Figure 1.1 for a body dropped from a high position, for example, it is quite natural to consider a polynomial model for the relation between the falling time and falling distance and to carry out polynomial model fitting. This represents the processes of model specification and parameter estimation. For elucidation of this

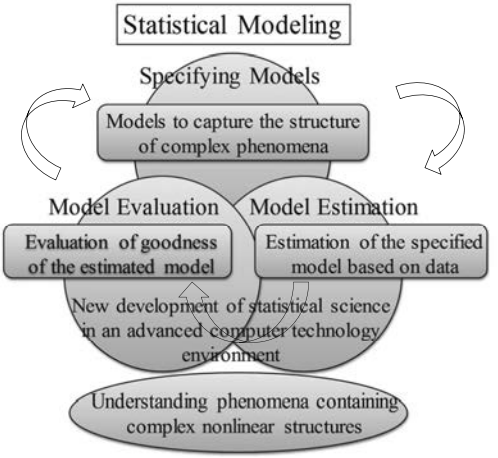


Figure 1.4 *Regression modeling; the specification of models that approximates the structure of a phenomenon, the estimation of their parameters, and the evaluation and selection of estimated models.*

physical phenomenon, however, a question may remain as to the optimum degree of the polynomial model. In the prediction of housing prices with linear regression models, moreover, a key question is what factors to include in the model. Furthermore, in considering nonlinear regression models, one is confronted by the availability of infinite candidate models for complex nonlinear phenomena controlled by smoothing parameters, and the need for selection of models that will appropriately approximate the structures of the phenomena, which is essential for their elucidation.

In this way, the need always arises to search through and evaluate a large number of models and from among them select one that will work effectively for elucidation of the target phenomena, based on the information provided by the data. This is commonly referred to as the *model evaluation and selection* problem.

Chapter 5 focuses on the model evaluation and selection problems, and presents various model selection criteria that are widely used as indicators in the assessment of the *goodness* of a model. It begins with a description of evaluation criteria proposed as estimators of prediction error, and then discusses the AIC (Akaike information criterion) based

on Kullback-Leibler information and the BIC (Bayesian information criterion) derived from a Bayesian view point, together with fundamental concepts that serve as the bases for derivation of these criteria.

The AIC, proposed in 1973 by Hirotugu Akaike, is widely used in various fields of natural and social sciences and has contributed greatly to elucidation, prediction, and control of phenomena. The BIC was proposed in 1978 by Gideon E. Schwarz and is derived based on a Bayesian approach rather than on information theory as with the AIC, but like the AIC it is utilized throughout the world of science and has played a central role in the advancement of modeling. Chapters 2 to 4 of this book show the various forms of expression of the AIC for linear, nonlinear, logistic, and other models, and give examples for model evaluation and selection problems based on the AIC.

Model selection from among candidate models constructed on the basis of data is essentially the selection of a single model that best approximates the data-generated probability structure. In Chapter 5, the discussion is further extended to include the concept of *multimodel inference* (Burnham and Anderson, 2002) in which the inferences are based on model aggregation and utilization of the relative importance of constructed models in terms of their weighted values.

1.2 Classification and Discrimination

Classification and discrimination techniques are some of the most widely used statistical tools in various fields of natural and social sciences. The primary aim in discriminant analysis is to assign an individual to one of two or more classes (groups) on the basis of measurements on feature variables. It is designed to construct linear and nonlinear decision boundaries based on a set of training data.

1.2.1 Discriminant Analysis

When a preliminary diagnosis concerning the presence or absence of a disease is made on the basis of data from blood chemistry analysis, information contained in the blood relating to the disease is measured, assessed, and acquired in the form of qualitative data. The diagnosis of normality or abnormality is based on multivariate data from several test results. In other words, it is an assessment of whether the person examined is included in a group consisting of normal individuals or a group consisting of individuals who exhibit a disease-related abnormality.

This kind of assessment can be made only if information from test re-