

Predictive Statistics

Analysis and Inference beyond Models

All scientific disciplines prize predictive success. Conventional statistical analyses, however, treat prediction as secondary, instead focusing on modeling and hence on estimation, testing, and detailed physical interpretation, tackling these tasks before the predictive adequacy of a model is established. This book outlines a fully predictive approach to statistical problems based on studying predictors; the approach does not require that predictors correspond to a model although this important special case is included in the general approach. Throughout, the point is to examine predictive performance before considering conventional inference. These ideas are traced through five traditional subfields of statistics, helping readers to refocus and adopt a directly predictive outlook. The book also considers prediction via contemporary ‘blackbox’ techniques and emerging data types and methodologies, where conventional modeling is so difficult that good prediction is the main criterion available for evaluating the performance of a statistical method. Well-documented open-source R code in a Github repository allows readers to replicate examples and apply techniques to other investigations.

BERTRAND S. CLARKE is Chair of the Department of Statistics at the University of Nebraska, Lincoln. His research focuses on predictive statistics and statistical methodology in genomic data. He is a fellow of the American Statistical Association, serves as editor or associate editor for three journals, and has published numerous papers in several statistical fields as well as a book on data mining and machine learning.

JENNIFER L. CLARKE is Professor of Food Science and Technology, Professor of Statistics, and Director of the Quantitative Life Sciences Initiative at the University of Nebraska, Lincoln. Her current interests include statistical methodology for metagenomics and also prediction, statistical computation, and multitype data analysis. She serves on the steering committee of the Midwest Big Data Hub and is Co-Principal Investigator on an award from the NSF focused on data challenges in digital agriculture.

CAMBRIDGE SERIES IN STATISTICAL AND
 PROBABILISTIC MATHEMATICS

Editorial Board

- Z. Ghahramani (Department of Engineering, University of Cambridge)
 R. Gill (Mathematical Institute, Leiden University)
 F. P. Kelly (Department of Pure Mathematics and Mathematical Statistics,
 University of Cambridge)
 B. D. Ripley (Department of Statistics, University of Oxford)
 S. Ross (Department of Industrial and Systems Engineering, University of Southern California)
 M. Stein (Department of Statistics, University of Chicago)

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.

A complete list of books in the series can be found at www.cambridge.org/statistics.
 Recent titles include the following:

20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort et al.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition)*, by Rick Durrett
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability (Second Edition)*, by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
39. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, by Jianfeng Yao, Shurong Zheng and Zhidong Bai
40. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, by Evarist Giné and Richard Nickl
41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks (Volume 1)*, by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vlasos Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke

Predictive Statistics

Analysis and Inference beyond Models

Bertrand S. Clarke

University of Nebraska, Lincoln

Jennifer L. Clarke

University of Nebraska, Lincoln



CAMBRIDGE
UNIVERSITY PRESS



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107028289

DOI: 10.1017/9781139236003

© Bertrand S. Clarke and Jennifer L. Clarke 2018

This publication is in copyright. Subject to statutory exception and to the provisions
of relevant collective licensing agreements, no reproduction of any part may take
place without the written permission of Cambridge University Press & Assessment.

First published 2018

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

ISBN 978-1-107-02828-9 Hardback

Additional resources for this publication at www.cambridge.org/predictivestatistics

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

Contents

<i>Expanded Contents</i>	<i>page</i>	vi
<i>Preface</i>		xi
Part I The Predictive View		1
1 Why Prediction?		3
2 Defining a Predictive Paradigm		34
3 What about Modeling?		67
4 Models and Predictors: A Bickering Couple		86
Part II Established Settings for Prediction		123
5 Time Series		125
6 Longitudinal Data		161
7 Survival Analysis		206
8 Nonparametric Methods		249
9 Model Selection		307
Part III Contemporary Prediction		359
10 Blackbox Techniques		361
11 Ensemble Methods		449
12 The Future of Prediction		524
<i>References</i>		605
<i>Index</i>		635

Expanded Contents

<i>Preface</i>	xi
Part I The Predictive View	1
1 Why Prediction?	3
1.1 Motivating the Predictive Stance	4
1.2 Some Examples	11
1.2.1 Prediction with Ensembles rather than Models	12
1.2.2 Hypothesis Testing as Prediction	21
1.2.3 Predicting Classes	26
1.3 General Issues	32
2 Defining a Predictive Paradigm	34
2.1 The Sunrise Problem	34
2.2 Parametric Families	41
2.2.1 Frequentist Parametric Case	41
2.2.2 Bayesian Parametric Case	43
2.2.3 Interpretation	46
2.3 The Abstract Version	47
2.3.1 Frequentism	48
2.3.2 Bayes Approach	51
2.3.3 Survey Sampling	56
2.3.4 Predictivist Approach	58
2.4 A Unified Framework for Predictive Analysis	63
3 What about Modeling?	67
3.1 Problem Classes for Models and Predictors	68
3.2 Interpreting Modeling	73
3.3 The Dangers of Modeling	75
3.4 Modeling, Inference, Prediction, and Data	78
3.5 Prequentialism	80
4 Models and Predictors: A Bickering Couple	86
4.1 Simple Nonparametric Cases	87
4.2 Fixed Effects Linear Regression	94
4.3 Quantile Regression	101
4.4 Comparisons: Regression	104

Expanded Contents

vii

4.5	Logistic Regression	108
4.6	Bayes Classifiers and LDA	111
4.7	Nearest Neighbors	115
4.8	Comparisons: Classification	116
4.9	A Look Ahead to Part II	119
Part II Established Settings for Prediction		123
5	Time Series	125
5.1	Classical Decomposition Model	125
5.2	Box–Jenkins: Frequentist SARIMA	128
5.2.1	Predictor Class Identification	129
5.2.2	Estimating Parameters in an ARMA(p, q) Process	132
5.2.3	Validation in an ARMA(p, q) Process	133
5.2.4	Forecasting	135
5.3	Bayes SARIMA	139
5.4	Computed Examples	142
5.5	Stochastic Modeling	150
5.6	Endnotes: Variations and Extensions	156
5.6.1	Regression with an ARMA(p, q) Error Term	157
5.6.2	Dynamic Linear Models	159
6	Longitudinal Data	161
6.1	Predictors Derived from Repeated-Measures ANOVA	167
6.2	Linear Models for Longitudinal Data	172
6.3	Predictors Derived from Generalized Linear Models	180
6.4	Predictors Using Random Effects	184
6.4.1	Linear Mixed Models	184
6.4.2	Generalized Linear Mixed Models	193
6.4.3	Nonlinear Mixed Models	194
6.5	Computational Comparisons	194
6.6	Endnotes: More on Growth Curves	201
6.6.1	A Fixed Effect Growth Curve Model	203
6.6.2	Another Fixed Effect Technique	204
7	Survival Analysis	206
7.1	Nonparametric Predictors of Survival	208
7.1.1	The Kaplan–Meier predictor	208
7.1.2	Median as a Predictor	216
7.1.3	Bayes Version of the Kaplan–Meier Predictor	219
7.1.4	Discrimination and Calibration	221
7.1.5	Predicting with Medians	222
7.2	Proportional Hazards Predictors	226
7.2.1	Frequentist Estimates of h_0 and β in PH Models	228
7.2.2	Frequentist PH Models as Predictors	231
7.2.3	Bayes PH Models	233
7.2.4	Continuing the Example	236
7.3	Parametric Models	239
7.4	Endnotes: Other Models	245

7.4.1	Accelerated Failure Time (AFT) Models	245
7.4.2	Competing Risks	246
8	Nonparametric Methods	249
8.1	Predictors Using Orthonormal Basis Expansions	252
8.2	Predictors Based on Kernels	260
8.2.1	Kernel Density Estimation	260
8.2.2	Kernel Regression: Deterministic Designs	266
8.2.3	Kernel Regression: Random Design	270
8.3	Predictors Based on Nearest Neighbors	275
8.3.1	Nearest Neighbor Density Estimation	275
8.3.2	Nearest Neighbor Regression	281
8.3.3	Beyond the Independence Case	285
8.4	Predictors from Nonparametric Bayes	286
8.4.1	Polya Tree Process Priors for Distribution Estimation	288
8.4.2	Gaussian Process Priors for Regression	291
8.5	Comparing Nonparametric Predictors	294
8.5.1	Description of the Data, Methods, and Results	295
8.5.2	\mathcal{M} -Complete or \mathcal{M} -Open?	300
8.6	Endnotes	302
8.6.1	Smoothing Splines	303
8.6.2	Nearest Neighbor Classification	304
8.6.3	Test-Based Prediction	304
9	Model Selection	307
9.1	Linear Models	312
9.2	Information Criteria	320
9.3	Bayes Model Selection	327
9.4	Cross-Validation	334
9.5	Simulated Annealing	339
9.6	Markov Chain Monte Carlo and the Metropolis–Hastings Algorithm	344
9.7	Computed Examples: SA and MCMC–MH	348
9.8	Endnotes	353
9.8.1	DIC	354
9.8.2	Posterior Predictive Loss	354
9.8.3	Information-Theoretic Model Selection Procedures	355
9.8.4	Scoring Rules and BFs Redux	356
Part III	Contemporary Prediction	359
10	Blackbox Techniques	361
10.1	Classical Nonlinear Regression	364
10.2	Trees	368
10.2.1	Finding a Good Tree	371
10.2.2	Pruning and Selection	379
10.2.3	Bayes Trees	383
10.3	Neural Nets	386
10.3.1	‘Fitting’ a Good NN	388
10.3.2	Choosing an Architecture for an NN	393

Expanded Contents

ix

10.3.3	Bayes NNs	394
10.3.4	NN Heuristics	397
10.3.5	Deep Learning, Convolutional NNs, and All That	399
10.4	Kernel Methods	405
10.4.1	Bayes Kernel Predictors	409
10.4.2	Frequentist Kernel Predictors	416
10.5	Penalized Methods	422
10.6	Computed Examples	429
10.6.1	Doppler Function Example	429
10.6.2	Predicting a Vegetation Greenness Index	433
10.7	Endnotes	443
10.7.1	Projection Pursuit	443
10.7.2	Logic Trees	445
10.7.3	Hidden Markov Models	446
10.7.4	Errors-in-Variables Models	447
11	Ensemble Methods	449
11.1	Bayes Model Averaging	454
11.2	Bagging	462
11.3	Stacking	471
11.4	Boosting	480
11.4.1	Boosting Classifiers	481
11.4.2	Boosting and Regression	486
11.5	Median and Related Methods	489
11.5.1	Different Sorts of ‘Median’	489
11.5.2	Median and Other Components	494
11.5.3	Heuristics	495
11.6	Model Average Prediction in Practice	497
11.6.1	Simulation Study	497
11.6.2	Reanalyzing the Vegout Data	507
11.6.3	Mixing It Up	518
11.7	Endnotes	519
11.7.1	Prediction along a String	520
11.7.2	No Free Lunch	522
12	The Future of Prediction	524
12.1	Recommender Systems	526
12.1.1	Collaborative Filtering Recommender Systems	526
12.1.2	Content-Based (CB) Recommender Systems	530
12.1.3	Other Methods	533
12.1.4	Evaluation	536
12.2	Streaming Data	537
12.2.1	Key Examples of Procedures for Streaming Data	538
12.2.2	Sensor Data	547
12.2.3	Streaming Decisions	551
12.3	Spatio-Temporal Data	556
12.3.1	Spatio-Temporal Point Data	559
12.3.2	Remote Sensing Data	562
12.3.3	Spatio-Temporal Point Process Data	565

12.3.4	Areal Data	568
12.4	Network Models	570
12.4.1	Static Networks	572
12.4.2	Dynamic Networks	581
12.5	Multitype Data	585
12.5.1	'Omics Data	586
12.5.2	Combining Data Types	592
12.6	Topics that Might Have Been Here . . . But Are Not	599
12.7	Predictor Properties that Remain to be Studied	600
12.8	Whither Prediction?	602
	<i>References</i>	605
	<i>Index</i>	635

Preface

This book grew out of a nagging dissatisfaction with the various schools of thought in statistics and their increasing disjunction. Each one – frequentist, Bayes, survey sampling, information-theoretic, etc. – has its strengths and weaknesses, and comparisons amongst their different approaches to inference has energized statistical thinking. This dynamic has only grown stronger over the last decade as more challenging data types have become commonplace. Moreover, in contrasting the techniques advocated by the different schools of thought on harder problems, such as working with big data, high-dimensional data, or complex data, the nagging doubts have only become more insistent. Otherwise stated, the less data (or other information) relative to the believed complexity of the data generator that is available, the more the modeling contributes to an analysis and therefore the more the differences in schools of thought, which largely rest on modeling, become apparent.

Concisely, the era of big data, whether high-dimensional, streaming, multitype or otherwise ‘big’, is forcing us all to rethink statistics and its philosophy. Questions about how to measure the distance between points in high dimensions have to be addressed since that is one version of the curse of dimensionality, likewise, questions of sparsity – when it holds, when it fails, and how to deal with it in either case – and questions of data sets that have information which is not extractable within the traditional formulation of a ‘random variable on a measure space’ or a valid sample from a well-defined population. In these contexts, this book is a small first step to reorganize some of what we know in order to focus on predictive structure, which is one of the few properties that cuts across all the new, exciting, developments challenging us and our field.

In our field, we have relied too much on our models by not assessing them as extensively as we should. We have not looked enough at their stability. We have not, as a rule, considered a sufficient number of alternative models to be sure that the model we used was reasonable. With few exceptions, we have not done sequential searches over modeling strategies to find a reasonable model, given a certain amount of data, and then modified it in view of getting more data. Also, we have not assessed the robustness of our inferences to our modeling strategies sufficiently. The present authors are as guilty of this as anyone else. In short, we have contented ourselves with the bromide that even if the model is wrong it may be useful, in the hope that if there is a true model (and here we argue that often there isn’t) we have found at least a part of it. However, that ain’t necessarily so.

This book is an attempt to focus more heavily on the data than the formalism and to focus more heavily on the performance of predictors rather than the fit or physical interpretation of a model or other construct. As a consequence, testing and estimation are given short

shrift. In fact, the general enterprise of inference by modeling, testing, and estimation seems premature until a lot more is known about a data generator than that it is described by a simple model that may be useful even when it's not true. In reality, the situation is usually worse than that for conventional analysis because the inferences are generated from one data set and a pile of assumptions, often dubious. Granted, in the hands of capable statisticians with enough persistence, most schools of thought will yield useful inferences. However, such success reflects the insight and doggedness of the statistician more than the efficacy of the methods. Consequently, it is hoped that one effect of focusing on the data, here via prediction, will be to energize the debate about what the central goals of statistics should be and how to go about achieving them.

An idea that recurs throughout this book is the concept of a problem class based on the relationship between the data generator and a class of predictors. The emphasis is on predictors that do not correspond to a perfect model for the data generator. In particular, cases in which the model used is only an approximation, or in which there is no true model, are frequently considered.

This book is in three parts. Part I outlines a general approach to statistics based on prediction. No claim is made that it is complete, merely that it is an alternative to various established schools of thought and deserves more attention than it has received. It is based on the prequential (predictive sequential) ideas that emerged in the early 1980s from A. P. Dawid and M. West, amongst others. There are four chapters, outlining the importance of prediction, defining a predictive paradigm, explaining why modeling, while sometimes useful, is not as good an approach as prediction, and finally looking at some familiar predictors. The view here is also more general: other schools of thought are incorporated into the predictive approach by regarding them as techniques for generating predictors that may be tested and studied. Thus, other schools of thought are not 'wrong' so much as incomplete: one school's techniques may not yield good predictors as readily as those of another.

Part II is a review of five major fields within statistics (time series, longitudinal data, survival analysis, nonparametrics, and model selection) from the predictive standpoint. The material is not new; the perspective on it is. The point of Part II is to demonstrate the feasibility of the predictive view: that it is a valid way to think about traditional branches of statistics and is computationally feasible. The five specific subfields were chosen because they are quite different from each other, suggesting the wide applicability of a general predictive view. They are also fields where the problems are so complicated that prediction is obviously important.

Part III is brings prediction up to the present. Starting with prediction in more contemporary model classes such as trees, neural nets, kernel methods, and penalized methods, it moves on to a chapter on ensemble methods, including Bayes model averaging, bagging, stacking, boosting and median methods. Even more than in previous chapters, computing is stressed to verify that the perspective advocated here is feasible. The final chapter is intended to bring predictive concepts to branches of statistics that have either recently emerged or recently changed character through, e.g., big data, changes in data collection, or new applications that have made prediction more important. Having dealt with terrestrial matters, the last chapter also indulges in some moon-gazing, speculating on which problems become more interesting when a predictive view is taken.

On the one hand this book does not require much mathematical background; a strong, determined MS student in statistics, mathematics, engineering, computer science, or other highly quantitative field should be able to follow the formal derivations. On the other hand, the book is primarily conceptual and so makes demands on the prospective reader that likely require more sophistication than a typical MS student, even a strong one, would have. Thus, our primary target audience is mid-career PhD students, practicing statisticians, and researchers in statistical fields. The authors sincerely hope that, whether or not these audiences agree with the perspective expressed in this book, they will find this perspective worth their time to understand.

For those interested in examining the R code or data used for the many examples in this text, please visit the catalog page on the Cambridge University Press website:

www.cambridge.org/predictivestatistics

This page includes a link to the github repository containing all relevant R code and data. The repository is structured so that each chapter has a branch. All code is provided under GNU Public License 3.0.

As with every book, there are people who should be thanked. First, all the people who supplied the data sets we used for examples. Second, all the people who, over the past four years put up with us obsessing over this book; we apologize for endlessly bending your ear. Third, Diana Gillooly of Cambridge University Press, with whom we had many conversations about the content, organization, and orientation of this book. Fourth, those colleagues who encouraged us in our folly. (You know who you are!) We forbear from mentioning names for fear they will regret encouraging us.

Finally, we have consistently tried to be engaging and sometimes provocative. Of course, some people will disagree with us and some errors may remain despite our best efforts. We are reminded of the (possibly mythical) story of a French physicist who, when asked about a colleague's work, pondered a few moments and finally responded: 'It's not even wrong.' In the spirit of that witticism, we apologize in advance for any errors that remain, whether technical or philosophical, hoping that they will at least be interesting.

Bertrand S. Clarke
Jennifer L. Clarke