

## Maximum Likelihood for Social Science

This volume provides a practical introduction to the method of maximum likelihood as used in social science research. Michael D. Ward and John S. Ahlquist focus on applied computation in  $\mathcal{R}$  and use real social science data from actual, published research. Unique among books at this level, it develops simulation-based tools for model evaluation and selection alongside statistical inference. The book covers standard models for categorical data, as well as counts, duration data, and strategies for dealing with data missingness. By working through examples, math, and code the authors build an understanding about the contexts in which maximum likelihood methods are useful and develop skills in translating mathematical statements into executable computer code. Readers will not only be taught to use likelihood-based tools and generate meaningful interpretations, but they will also acquire a solid foundation for continued study of more advanced statistical techniques.

MICHAEL D. WARD is Professor Emeritus at Duke University. He has taught at Northwestern University, the University of Colorado, and the University of Washington. He worked as a principal research scientist at the WZB Berlin Social Science Center and held a municipal chair at the University of Pierre Mendès France (Grenoble II). His work began with a study of the links between global and national inequalities, continued with seminal articles on the conflict processes in the Cold War, and more recently turned to analyses of networks of conflict and cooperation in the contemporary era. At Duke, he established an innovative research lab of graduate and undergraduate students focusing on conflict prediction. One of the first political scientists to focus on the role of prediction in scholarly and policy work, he continues these efforts in his company, Predictive Heuristics, a data analytics firm that provides risk analysis for commercial and institutional clients.

JOHN S. AHLQUIST is Associate Professor of Political Economy at UC San Diego's School of Global Policy and Strategy and a 2017–18 Fellow at Stanford's Center for Advanced Study in the Behavioral Sciences. He previously held faculty positions at the University of Wisconsin, Madison, and Florida State University. His work has focused on the political structure and actions of labor unions, as well as the politics of redistribution and social insurance in a globalized economy. His methodological interests have spanned statistical models for network data, machine learning and cluster analysis, and the analysis of survey list experiments. He is author of more than twenty journal articles appearing in a variety of outlets, including the *American Journal of Political Science*, *American Political Science Review*, *Journal of Politics*, and *Political Analysis*. His most recent book (with Margaret Levi; 2013) is *In the Interest of Others*. He is a past winner of a variety of prizes, including the Mancur Olson Award, the Michael Wallerstein Award, and the APSA Labor Project Best Book Award. Ahlquist holds a PhD from the University of Washington and B.A. from UC Berkeley.

## Analytical Methods for Social Research

Analytical Methods for Social Research presents texts on empirical and formal methods for the social sciences. Volumes in the series address both the theoretical underpinnings of analytical techniques as well as their application in social research. Some series volumes are broad in scope, cutting across a number of disciplines. Others focus mainly on methodological applications within specific fields such as political science, sociology, demography, and public health. The series serves a mix of students and researchers in the social sciences and statistics.

### Series Editors

R. Michael Alvarez, *California Institute of Technology*  
Nathaniel L. Beck, *New York University*  
Lawrence L. Wu, *New York University*

### Other Titles in the Series

*Time Series Analysis for the Social Sciences*, by Janet M. Box-Steffensmeier, John R. Freeman, Jon C.W. Pevehouse and Matthew Perry Hitt  
*Event History Modeling: A Guide for Social Scientists*, by Janet M. Box-Steffensmeier and Bradford S. Jones  
*Ecological Inference: New Methodological Strategies*, edited by Gary King, Ori Rosen, and Martin A. Tanner  
*Spatial Models of Parliamentary Voting*, by Keith T. Poole  
*Essential Mathematics for Political and Social Research*, by Jeff Gill  
*Political Game Theory: An Introduction*, by Nolan McCarty and Adam Meirowitz  
*Data Analysis Using Regression and Multilevel/Hierarchical Models*, by Andrew Gelman and Jennifer Hill  
*Counterfactuals and Causal Inference*, by Stephen L. Morgan and Christopher Winship

# Maximum Likelihood for Social Science

## *Strategies for Analysis*

MICHAEL D. WARD

*Duke University*

JOHN S. AHLQUIST

*University of California, San Diego*



CAMBRIDGE  
UNIVERSITY PRESS

Cambridge University Press  
978-1-107-18582-1 — Maximum Likelihood for Social Science  
Michael D. Ward, John S. Ahlquist  
Frontmatter  
[More Information](#)

## CAMBRIDGE UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, 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  
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781107185821](http://www.cambridge.org/9781107185821)  
DOI: 10.1017/9781316888544

© Michael D. Ward and John S. Ahlquist 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.

First published 2018

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

### *Library of Congress Cataloging-in-Publication Data*

Names: Ward, Michael Don, 1948– author. | Ahlquist, John S., author.

Title: Maximum likelihood for social science : strategies for analysis/  
Michael D. Ward, John S. Ahlquist.

Description: 1 Edition. | New York : Cambridge University Press, 2018. |

Series: Analytical methods for social research

Identifiers: LCCN 2018010101 | ISBN 9781107185821 (hardback) |

ISBN 9781316636824 (paperback)

Subjects: LCSH: Social sciences—Research. | BISAC: POLITICAL SCIENCE / General.

Classification: LCC H62 .W277 2018 | DDC 300.72—dc23

LC record available at <https://lccn.loc.gov/2018010101>

ISBN 978-1-107-18582-1 Hardback

ISBN 978-1-316-63682-4 Paperback

Cambridge University Press 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>List of Figures</i>	<i>page xi</i>
<i>List of Tables</i>	xv
<i>Preface</i>	xvii
<i>Acknowledgments</i>	xxiii
<i>Notes on Notation</i>	xxv

### PART I CONCEPTS, THEORY, AND IMPLEMENTATION

1	Introduction to Maximum Likelihood	3
1.1	<i>Introduction to Maximum Likelihood</i>	3
1.2	<i>Coin Flips and Maximum Likelihood</i>	3
1.3	<i>Samples and Sampling Distributions</i>	6
1.4	<i>Maximum Likelihood: An Overview</i>	9
1.4.1	<i>Maximum Likelihood: Specific</i>	9
1.4.2	<i>The Least Squares Approach</i>	16
1.5	<i>Maximum Likelihood: General</i>	17
1.6	<i>Conclusion</i>	19
1.7	<i>Further Reading</i>	20
2	Theory and Properties of Maximum Likelihood Estimators	21
2.1	<i>The Likelihood Function: A Deeper Dive</i>	21
2.1.1	<i>Likelihood and Bayesian Statistics</i>	22
2.1.2	<i>Regularity</i>	23
2.1.3	<i>Score</i>	24
2.1.4	<i>Fisher Information</i>	25
2.2	<i>Properties of Maximum Likelihood Estimators</i>	28
2.2.1	<i>Invariance</i>	28
2.2.2	<i>Large-Sample Properties</i>	28

2.3	<i>Diagnostics for Maximum Likelihood</i>	32
2.3.1	<i>Likelihood Ratios</i>	32
2.3.2	<i>Score and Wald Tests</i>	35
2.3.3	<i>Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)</i>	37
2.4	<i>What If the Likelihood Is Wrong?</i>	38
2.4.1	<i>Robust Variance and Standard Errors</i>	39
2.5	<i>Conclusion</i>	40
2.6	<i>Further Reading</i>	41
3	<b>Maximum Likelihood for Binary Outcomes</b>	43
3.1	<i>Binary Responses</i>	43
3.2	<i>Binary Data</i>	44
3.2.1	<i>Odds, Odds Ratios, and Relative Risk</i>	45
3.3	<i>The Linear Probability Model</i>	47
3.4	<i>The Logit Model, a.k.a. Logistic Regression</i>	50
3.4.1	<i>The Logit Model</i>	51
3.4.2	<i>Estimation</i>	52
3.4.3	<i>Output and Interpretation</i>	54
3.4.4	<i>Estimation Uncertainty around the MLE</i>	58
3.4.5	<i>Graphical Presentation of Effects in Logistic Regression</i>	60
3.5	<i>An Example: US Senate Voting</i>	62
3.5.1	<i>Model Evaluation</i>	64
3.6	<i>Some Extensions</i>	71
3.6.1	<i>Latent Variable Formulation for Binary Data</i>	71
3.6.2	<i>Heteroskedastic Probit</i>	72
3.6.3	<i>Complimentary Logistic Model, a.k.a. cloglog</i>	73
3.6.4	<i>An Example: Civil Conflict</i>	74
3.7	<i>Summary of General Approach to Binary Variables</i>	76
3.8	<i>Further Reading</i>	77
4	<b>Implementing MLE</b>	79
4.1	<i>The Likelihood Surface</i>	79
4.1.1	<i>Examples</i>	80
4.1.2	<i>Profile Likelihood</i>	82
4.2	<i>Numerical Optimization</i>	82
4.2.1	<i>Newton-Raphson</i>	83
4.2.2	<i>Other Algorithms</i>	84
4.2.3	<i>Expectation Maximization (EM)</i>	85
4.3	<i>Estimation Challenges</i>	86
4.3.1	<i>Optimizer Fails to Converge</i>	86
4.3.2	<i>Starting Values</i>	86
4.3.3	<i>Algorithm Choice</i>	86
4.3.4	<i>Scaling of Variables</i>	87
4.3.5	<i>Step Sizes</i>	87

<i>Contents</i>	vii
4.3.6 <i>Flat Likelihoods</i>	87
4.3.7 <i>Absence of Single Maximum</i>	88
4.4 <i>Different Software Yields “Different” Results</i>	88
4.5 <i>Modeling Challenges Appear as Estimation Problems</i>	89
4.5.1 <i>(Near-)Perfect Separation</i>	89
4.5.2 <i>Rare Events (and Small Samples)</i>	91
4.6 <i>Conclusion</i>	91
4.7 <i>Further Reading</i>	92
PART II MODEL EVALUATION AND INTERPRETATION	
5 <i>Model Evaluation and Selection</i>	97
5.1 <i>Current Practice</i>	97
5.1.1 <i>BUTON</i>	98
5.1.2 <i>Graphical Displays</i>	99
5.1.3 <i>What’s Left Out?</i>	100
5.1.4 <i>Sign, Significance, and the Perils of p-Values</i>	102
5.2 <i>The Logic of Prediction, Out-of-Sample</i>	104
5.2.1 <i>The Process of Evaluating Out-of-Sample Predictions</i>	105
5.2.2 <i>Variations on a Theme</i>	110
5.2.3 <i>Out-of-Sample Prediction: Trade and the WTO, Act II</i>	112
5.2.4 <i>Benefits of Cross-Validation</i>	115
5.3 <i>Conclusion</i>	115
5.4 <i>Further Reading</i>	116
6 <i>Inference and Interpretation</i>	119
6.1 <i>The Mechanics of Inference</i>	119
6.2 <i>Interpreting Models and Producing Scenarios of Interest</i>	122
6.2.1 <i>Quantities of Interest</i>	122
6.2.2 <i>Scenarios</i>	123
6.2.3 <i>Interpolation, Extrapolation, and Convex Hulls</i>	124
6.2.4 <i>Simulating Estimation Uncertainty</i>	126
6.3 <i>Conclusion</i>	130
6.4 <i>Further Reading</i>	130
PART III THE GENERALIZED LINEAR MODEL	
7 <i>The Generalized Linear Model</i>	135
7.1 <i>The Generalized Linear Model</i>	135
7.2 <i>The Exponential Family</i>	136
7.3 <i>Past Examples as GLMs</i>	137
7.3.1 <i>GLMs in <math>\mathcal{R}</math></i>	138
7.4 <i>“Quasi-” and “Pseudo-”likelihood</i>	138
7.5 <i>Conclusion</i>	140
7.6 <i>Further Reading</i>	140

8	Ordered Categorical Variable Models	141
8.1	<i>Motivation</i>	141
8.2	<i>The Ordered Logit Model</i>	142
8.2.1	<i>What about the Ordered Probit Model?</i>	145
8.2.2	<i>Results and Interpretation</i>	145
8.3	<i>Example: Categorical Assessments of Political Ideology</i>	146
8.4	<i>Parallel Regressions</i>	149
8.4.1	<i>Example: Genocide Severity</i>	151
8.4.2	<i>Parallel Regressions and Genocide Severity</i>	152
8.4.3	<i>Extensions</i>	154
8.5	<i>Ordered Categorical Variables as Regressors</i>	154
8.5.1	<i>An Example: Child Mortality and Democracy</i>	157
8.6	<i>Further Reading</i>	160
9	Models for Nominal Data	161
9.1	<i>Introduction</i>	161
9.2	<i>Multinomial Logit</i>	162
9.2.1	<i>A Latent Variable Formulation</i>	164
9.2.2	<i>IIA</i>	166
9.3	<i>An Example: Australian Voters in 2013</i>	168
9.3.1	<i>Evaluation</i>	168
9.3.2	<i>Interpretation</i>	173
9.4	<i>Conditional Multinomial Logit</i>	179
9.4.1	<i>A Note on Data Structure</i>	180
9.5	<i>Extensions</i>	184
9.5.1	<i>Heteroskedastic Multinomial Logit</i>	184
9.5.2	<i>Nested Logit</i>	185
9.5.3	<i>Multinomial Probit</i>	187
9.5.4	<i>Random Coefficients and Mixed Logit</i>	187
9.6	<i>Conclusion</i>	188
9.7	<i>Further Reading</i>	189
10	Strategies for Analyzing Count Data	190
10.1	<i>Introduction</i>	190
10.2	<i>The Poisson Distribution</i>	191
10.2.1	<i>An Example: Means and Mediation</i>	192
10.2.2	<i>Interpretation</i>	195
10.3	<i>Dispersion</i>	197
10.3.1	<i>Diagnosing Over-Dispersion</i>	198
10.4	<i>Modeling Over-Dispersion</i>	201
10.4.1	<i>Quasipoisson</i>	202
10.4.2	<i>The Negative Binomial Model</i>	202
10.4.3	<i>Mediations</i>	205
10.4.4	<i>Under-Dispersion</i>	207

<i>Contents</i>	ix
10.5 <i>Hurdling 0s and 1s</i>	207
10.5.1 <i>How Many 0s Are Too Many?</i>	209
10.5.2 <i>Hurdle Models</i>	210
10.5.3 <i>Zero Inflation</i>	211
10.5.4 <i>Example: Droughts and Atrocities</i>	212
10.6 <i>Summary</i>	213
10.7 <i>Further Reading</i>	215
PART IV ADVANCED TOPICS	
11 <i>Strategies for Temporal Dependence: Duration Models</i>	219
11.1 <i>Introduction</i>	219
11.2 <i>Data Structures and Thinking about Time</i>	220
11.2.1 <i>Discrete and Continuous Time</i>	221
11.2.2 <i>Organizing Data as a Spell-Level</i>	221
11.2.3 <i>Continuous Time with Time-Varying Covariates:                 Counting Process</i>	222
11.3 <i>Discrete Time: The BTSCS Approach</i>	224
11.4 <i>Duration Models</i>	226
11.4.1 <i>Survivor and Hazard Functions</i>	226
11.4.2 <i>A Likelihood</i>	227
11.4.3 <i>Writing Down a Duration Model</i>	228
11.5 <i>The Cox Model</i>	230
11.5.1 <i>The Partial Likelihood</i>	231
11.5.2 <i>An Example: Cabinet Durations</i>	233
11.5.3 <i>Pros and Cons of the Cox Model</i>	235
11.6 <i>A Split-Population Model</i>	237
11.6.1 <i>Interpretation of Split-Duration Estimates</i>	241
11.6.2 <i>An Example</i>	243
11.7 <i>Conclusion</i>	245
11.8 <i>Further Reading</i>	247
12 <i>Strategies for Missing Data</i>	249
12.1 <i>Introduction</i>	249
12.1.1 <i>Concepts and Notation</i>	250
12.2 <i>Finding Missingness</i>	251
12.3 <i>A Taxonomy of Missing Data</i>	256
12.4 <i>Common Strategies</i>	257
12.4.1 <i>Complete Case Analysis</i>	257
12.4.2 <i>Available Case Analysis</i>	258
12.4.3 <i>Mean Imputation</i>	258
12.4.4 <i>Conditional Mean Imputation</i>	258
12.5 <i>Multiple Imputation</i>	259
12.5.1 <i>MLE-EM and Bayesian Multiple Imputation</i>	259
12.5.2 <i>MICE</i>	260

12.5.3	<i>Copula Methods</i>	261
12.5.4	<i>Combining Results Across Imputations</i>	263
12.6	<i>Building an Imputation Model</i>	264
12.7	<i>Evaluating Imputation Models</i>	265
12.7.1	<i>Empirical Replications: Sorens and Ruger</i>	265
12.8	<i>Conclusion</i>	268
12.9	<i>Further Reading</i>	269
PART V A LOOK AHEAD		
13	<i>Epilogue</i>	273
13.1	<i>The Broader Landscape</i>	274
	<i>Bibliography</i>	277
	<i>Index</i>	293

## Figures

1.1	The likelihood/probability of getting two heads in three coin tosses, over various values of $\theta$ .	<i>page 7</i>
1.2	Illustrating the Central Limit Theorem with a histogram of the means of 1,000 random samples of size 10 drawn from a population with mean of 10 and variance of 1.	8
1.3	2012 GDP per capita and CO <sub>2</sub> emissions. The prediction equation is shown as a straight line, with intercept and slope as reported in Table 1.2.	10
1.4	Likelihood as a function of possible values for $\beta_1$ .	14
2.1	Geometrical interpretation of the likelihood ratio (LR), Lagrange multiplier/score (LM), and the Wald (W) test statistics.	37
3.1	Diagnostic plots of the linear probability model described in Table 3.4.	48
3.2	The logistic function is nonlinear with range bounded by 0 and 1.	51
3.3	Plot displaying the 95% confidence bands for the predicted probability of LFP across different wage rates for women with and without young children.	61
3.4	A coefficient plot of the logit regression of US Senate votes on the Coburn Amendment to eliminate NSF funding for political science.	65
3.5	Receiver Operator Characteristics (ROC) curve for the logit models of the vote on the Coburn Amendment.	67
3.6	Separation plots for the partisan-only and full models of US Senate voting on the Coburn Amendment.	69
3.7	The predicted probability of voting “nay” on the Coburn Amendment as reelection approaches for Democrat and non-Democrat US senators.	70

3.8	The (inverse) logit, probit, and complementary log-log functions, mapping the linear term into $\theta$ .	74
3.9	ROC curves for the five models reported in Table 3.11.	76
4.1	The likelihood surface for $\mathcal{L}(\theta   \mathbf{x})$ where the probability model is $X \sim \text{Unif}(-\theta, \theta)$ .	80
4.2	The likelihood surface for $\mathcal{L}(\theta   \mathbf{x})$ where the probability model is $X \sim \text{Unif}(\theta - 2, \theta + 2)$ .	81
5.1	ROC plot of twofold cross-validation of the Fearon and Laitin replication from Table 5.1.	109
6.1	Visualizing the convex hull in 2 dimensions.	125
6.2	The relationship between world trade and the probability of a democratic transition in labor-abundant and labor-scarce autocracies.	129
8.1	Visualizing how cutpoints divide the density of a continuous latent variable into discrete ordered categories.	144
8.2	Party identification and perceptions of Obama's conservatism, calculated from the ordered logistic regression reported in Table 8.2.	148
8.3	First differences: Comparing weak democrats against independents in their predicted probabilities of identifying Obama as "moderate" (solid) and "liberal" (broken), calculated from the ordered logistic regression reported in Table 8.2.	149
8.4	Plot of the conditional means of the regressors at different levels of the response variable, magnitude.	155
8.5	Plot of the estimated regression coefficients from $M - 1$ regressions on $\tilde{Y}_m$ , the indicators that $Y \leq m$ .	156
8.6	Child mortality and constraints on the executive.	158
9.1	"One-v.-all" ROC curve diagnostics for the multinomial logit in Table 9.1.	173
9.2	The three-dimensional unit simplex.	174
9.3	Using ternary plots to interpret multinomial logit models when $M = 3$ .	175
9.4	Predicted vote choice in the 2013 Australian federal elections across age cohorts.	176
9.5	Predicted vote choice in the 2013 Australian federal elections for different income levels.	177
9.6	An example of sequential choices confronting voters, from the 2003 California gubernatorial recall election.	186
10.1	Diagnostic plots for the standard OLS regression and the OLS on $\log(y_i + 10)$ reported in Table 10.1.	194
10.2	The expected and predicted number of international mediations as a function of per capita GDP and UNSC membership.	196
10.3	A Poissonness plot for international mediations data from Bercovitch and Schneider (2000).	199

<i>List of Figures</i>	xiii
10.4 A hanging rootogram plotting expected versus actual counts.	200
10.5 The observed frequency of rebel atrocities against civilians (on the square root scale) between 1995 and 2008 in 26,566 0.5° × 0.5° grid cells.	209
11.1 An illustration of duration data using conflicts, onsets, durations, spells.	224
11.2 Schoenfeld residuals against adjusted time.	234
11.3 Expected government survival times depending on whether the cabinet requires a vote of investiture, with all over covariates held at central tendencies.	236
11.4 Country A is at risk; Country B is not at risk.	239
11.5 Conditional hazard rates for the split-population Weibull and log-logistic model of coups with all covariates held at sample means.	245
11.6 Plots of the hazard rate for the log-logistic model of coups.	245
11.7 Out-of-sample separation plots.	246
12.1 A “missingness map” displaying missing data by variable for an example data set.	252
12.2 The left panel displays the proportion of observations missing for a selection of variables.	253
12.3 Visualizing missingness in TSCS data for a selection of countries and variables.	254
12.4 Polity IV uses numeric values to encode different kinds of missing data.	255
12.5 Using <code>spbcop</code> to impute missing polity scores.	268

## Tables

1.1	Choosing a restaurant with three flips of a fair coin?	<i>page 5</i>
1.2	Standard output for the OLS regression of log CO <sub>2</sub> emissions on log per capita GDP.	17
3.1	Alternative data structures for binary data.	44
3.2	Select variables from Long's (1997) reexamination of Mroz's (1987) female labor force participation study.	45
3.3	Labor force participation and the number of young children.	46
3.4	Linear probability model of labor force participation (LFP) as a function of five independent variables plus an intercept.	48
3.5	Logit estimation of female labor force participation (LFP).	55
3.6	The central tendencies (medians and modes) for the variables included in the analysis of labor force participation in Table 3.5.	56
3.7	The matrix $\text{cov}(\hat{\beta})$ from the Mroz estimation in Table 3.5.	59
3.8	Replication of Table 1 in Uscinski and Klobstad (2010).	64
3.9	Predicted versus observed votes for the Coburn Amendment, with $t = 0.5$ for mapping probabilities into event space.	66
3.10	Odds ratios for the replication of table 1 in Uscinski and Klobstad (2010).	70
3.11	Replication and extension of table II in Cook and Savun (2016).	75
4.1	Estimation of the Rauchhaus (2009) logit model.	90
4.2	Contingency table showing the incidence of war between dyad-years of different levels of nuclear capacity from Rauchhaus (2009).	90
5.1	Logistic regression of civil war onset, replicating Fearon and Laitin (2003).	109
5.2	Linear regression of bilateral trade flows 1948–99, replicating Tomz et al. (2007).	114

5.3	Out-of-sample predictive performance for models of international trade.	115
6.1	Probit regression of democratic transitions 1875–2001, a reestimation of Ahlquist and Wibbels (2012).	121
6.2	Model dependence and the Ahlquist and Wibbels interpretation scenarios.	126
8.1	Selected variables and descriptors from 2016 ANES pilot study.	146
8.2	Ordered logit analysis of 2016 ANES data	147
8.3	The distribution of observations over Krain’s 11-point scale of genocide magnitude.	152
8.4	The correlates of genocide severity, a reanalysis of Krain (2005).	153
8.5	Coding Krain’s response variable for the cumulative logit.	154
8.6	OLS regression of (log) infant mortality on different encodings of executive constraints (XCONST).	159
9.1	Multinomial logistic regression on vote choice in the 2013 Australian elections.	169
9.2	Confusion matrix for the classifications from the multinomial logistic regression in Table 9.1.	171
9.3	The category-by-category “one-versus-all” confusion matrices.	171
9.4	Per-class error.	172
9.5	Conditional and mixed conditional-multinomial logistic regression on vote choice for the 2013 Australian elections.	181
9.6	A typical rectangular data structure for a multinomial model.	182
9.7	A grouped-response data structure data structure enabling conditional and mixed logit estimation.	183
10.1	OLS and Poisson regression of the number of times a country served as a mediator in an international conflict between 1950 and 1990.	193
10.2	Quasipoisson and negative binomial regression of the number of times a country served as a mediator in an international conflict between 1950 and 1990.	205
10.3	NB2, zero-inflated, and hurdle negative binomial regressions on counts of anti-civilian atrocities in the developing world, 1995–2008.	214
11.1	Spell organization of data.	222
11.2	Counting process data, from Ahlquist (2010b).	223
11.3	Cox model of government cabinet duration.	233
11.4	Formal tests of the correlation of residuals with time.	235
11.5	Weibull and split-population Weibull and log-logistic regression models of coups, 1960–2000.	244
12.1	Missingness in Sorens and Ruger data.	266
12.2	Root mean square error on imputations of Sorens and Ruger (2012) data.	267

## Preface

This project began many years ago at the University of Washington's Center for Statistics and the Social Sciences (CSSS). There two ambitious graduate students, John S. Ahlquist and Christian Breunig (now at the University of Konstanz), asked Michael D. Ward if he would supervise their training in maximum likelihood methods so that they could be better prepared for taking more advanced CSSS courses as well as those in the statistics and biostatistics departments. Ward gave them a stack of materials and asked them to start by preparing a lecture on ordinal regression models. Ward subsequently developed a class on maximum likelihood methods, which he has taught at the University of Washington (where it is still taught by Christopher Adolph) and, more recently, at Duke University. Ahlquist has gone on to teach a similar course at Florida State, the University of Wisconsin, and UC San Diego.

The point of the course was singular, and this book has a simple goal: to introduce social scientists to the maximum likelihood principle in a practical way. This praxis includes (a) being able to recognize where maximum likelihood methods are useful, (b) being able to interpret results from such analyses, and (c) being able to implement these methods both in terms of creating the likelihood and in terms of specifying it in a computational language that permits empirical analysis to be undertaken using the developed model.

The text is aimed at advanced PhD students in the social sciences, especially political science and sociology. We assume familiarity with basic probability concepts, the application of multivariate calculus to optimization problems, and the basics of matrix algebra.

### OUR APPROACH

We take a resolutely applied perspective here, emphasizing core concepts, computation, and model evaluation and interpretation. While we include a

chapter that introduces some of the important theoretical results and their derivations, we spend relatively little space discussing formal statistical properties. We made this decision for three reasons. First, there are several ways to motivate the likelihood framework. We find that a focus on a method's "desirable properties" in a frequentist setting to be a less persuasive reason to study maximum likelihood estimators (MLE). Instead we prefer to emphasize the powerful conceptual jump that likelihood-based reasoning represents in the study of statistics, one that enables us to move to a Bayesian setting relatively easily. Second, the statistical theory underlying the likelihood framework is well understood; it has been for decades. The requisite theorems and proofs are already collected in other excellent volumes, so we allocate only a single chapter to recapitulating them here. Rather, we seek to provide something that is missing: an applied text emphasizing modern applications of maximum likelihood in the social sciences. Third, and perhaps most important, we find that students learn more and have a more rewarding experience when the acquisition of new technical tools is directly bound to the substantive applications motivating their study.

Many books and even whole graduate training programs start with so-called Ordinary Least Squares (OLS). There is a certain logic to that. OLS is easy to teach, implement, and utilize while introducing a variety of important statistical concepts. OLS was particularly attractive in a world before powerful modern computers fit in our pockets. But OLS can be viewed as a special case of a more general class of models. Practically speaking, a limited range of social science data fit into this special case. Data in the social sciences tend to be lumpier, often categorical. Nominal, truncated, and bounded variables emerge not just from observational datasets but in researcher-controlled experiments as well (e.g., treatment selection and survival times). Indeed, the vast majority of social science data comes in forms that are profitably analyzed without resort to the special case of OLS. While OLS is a pedagogical benchmark, you will have to look hard for recent, state-of-the-art empirical articles that analyze observational data based on this approach. After reading this book and working through the examples, student should be able to fit, choose, and interpret many of the statistical models that appear in published research. These models are designed for binary, categorical, ordered, and count data that are neither continuous nor distributed normally.

#### WHAT FOLLOWS

We have pruned this book down from versions that appeared earlier online. We wanted the main text to focus entirely on the method and application of maximum likelihood principles. The text is divided into four parts.

Part I (Chapters 1–4) introduces the concept of likelihood and how it fits with both classical and Bayesian statistics. We discuss OLS only in passing,

highlighting how, under certain assumptions, it can be thought of as a maximum likelihood estimator. We identify and derive the major theoretical results and then show how to apply them in the context of binary response variables. Chapter 4 provides a discussion of how MLE is implemented computationally, with a particular emphasis on the  $\mathcal{R}$  computational environment.

Part II (Chapters 5 and 6) is the core of this volume. Its two chapters cover model selection and interpretation. Unique among texts at this level, we emphasize that model selection must occur prior to any inference about estimated parameters or other quantities. We argue explicitly for a wide application of an out-of-sample predictive heuristic in this area, something that is seeing increased attention with the machine learning revolution. In Chapter 6, we discuss how we might use the models we fit, and we focus on the power of modern computation to present nuanced and detailed interpretation of our statistical findings. We de-emphasize mechanical hypothesis testing against arbitrary null values, instead focusing on estimating meaningful quantities of interest and discussing our uncertainty around these estimates. In both chapters we include reflections on the mechanics and aesthetics of constructing tables and displays for effective communication, as well as thoughts on improving research transparency and credibility. While the material covered in this section is in no way unique to the study of maximum likelihood, we view this section as critical to continued progress in studying both maximum likelihood and more advanced statistical and computational topics.

Part III (Chapters 7–10) covers the Generalized Linear Model (GLM). Chapter 7 is short, introducing the basic structure of the GLM and some terminology and concepts. Chapters 8–10 present models for categorical variables, both ordered and nominal, as well as integer counts. Unlike some other texts for categorical data, these chapters are designed to be approached in a particular order and all rely on concepts and computational tools developed in Parts I and II.

In Part IV (Chapters 11 and 12) of the book we introduce more advanced topics. In Chapter 11 we begin the process of relaxing the standard assumption of conditional independence by presenting an introduction to duration models. This chapter is somewhat idiosyncratic, glossing over many of the details and complications one might expect in a full-fledged treatment of survival analysis, not to mention time series. Instead, we focus on how we can develop models for data that are inherently connected in time using likelihood tools and principles. Chapter 12 takes on the ubiquitous problem of missing data. We view this subject as woefully understudied in graduate training, while also presenting the pedagogical opportunity to discuss model construction and computation from a different perspective. We have also found that many of the most successful student replication projects came from critical interrogation of the earlier scholars' treatments of missing data.

Covering all the material in this book in a 15-week semester with beginning graduate students is certainly a challenge; doing so in a 10-week academic

quarter is even more demanding. In a quarter-length course or with first-year students we have found that Chapters 2, 4, and 7 are better left as reference, instead emphasizing intuition, computation, and examples. When the temporal budget constraint binds, we typically allow student interest to determine whether we focus on duration models or missing data.

#### SPECIAL FEATURES

This volume contains several special features and sections that deserve further elaboration.

#### Real Examples from Published Research

Each chapter contains at least one example drawn from actual published social science research. These examples use real data drawn from scholars' data repositories to illustrate the models, highlight the modeling assumptions involved, and present detailed interpretations. All these datasets are archived in the online repository accompanying this volume.

#### $\mathcal{R}$ Code

This is an applied, computational text. We are particularly interested in helping students transform mathematical statements into executable computer code.  $\mathcal{R}$  has become the dominant language in statistical computing because it is object-oriented, based on vectors; still has the best statistical graphics; and is open-source, meaning it is free to students and has a large network of contributors submitting new libraries almost daily. The newest statistical tools generally appear in  $\mathcal{R}$  first.

We include code directly in the text in offset and clearly marked boxes. We include our own comments in the code chunks so students can see annotation clarifying computational steps. We also include  $\mathcal{R}$  output and warnings in various places to aid in interpreting actual  $\mathcal{R}$  output as well as trouble-shooting. All analysis and graphics are generated in  $\mathcal{R}$ . The online repository contains the  $\mathcal{R}$  code needed to reproduce all tables and graphics.

#### “In case you were wondering ...”

Throughout the text there are special boxes labeled “In case you were wondering ...” The purpose of the boxes is to provide basic information about important mathematical tools and statistical distributions. These are things not easily defined in the main text and likely already familiar to some readers while appearing *de novo* to others. Our goal is to provide this information at the point of need while setting it off from the main text and marking it as “supplemental.”

We do not refer to the boxes directly in the main text, unlike equations, tables, figures, and code chunks. The title of the boxes reflects their function and status; they present supplemental information for the curious.

### “Further Reading”

Each chapter ends with a “further reading” section. These sections all follow a similar format, with subheadings for “applications,” “past work,” “advanced study,” and “software notes,” depending on the context these have for different topics.

The “applications” section highlights two to four studies using the tools discussed in that chapter and published in major social science journals in the last four years. These studies are meant to be examples of the types of papers students might consider when choosing replication projects.

The “past work” section is designed to provide pointers to the major contributors to the development and popularization of these tools in the social sciences. The “advanced study” section collects references to more advanced texts and articles where interested students can look for more detail on the math or computational algorithms. We consulted many of these texts in writing this book.

In the “software notes” sections we collect references to the major  $\mathcal{R}$  libraries that we found useful in preparing the book or in conducting analysis ourselves. Since  $\mathcal{R}$  is open-source, these references will surely become stale. We nevertheless thought it beneficial to collect references to  $\mathcal{R}$  packages in a single place in each chapter.

### NOTATION GLOSSARY

In our experience students often find mathematical notation a particularly frustrating barrier. To mitigate that problem we have included a notation “glossary” at the beginning of the book

### ONLINE RESOURCES

The online repository, [maxlikebook.com](http://maxlikebook.com), accompanying this volume contains

- all datasets used in this volume,
- $\mathcal{R}$  code for producing all tables and graphics,
- suggested problem sets and partial solutions, and
- some of our teaching slides.

We expect that repository content will evolve as we continue to teach this material and receive feedback from other instructors and students.

## Acknowledgments

We would like to thank colleagues and former students worldwide, including Kathryn Alexander, Andrew Ballard, Sanuel Bagg, Jeanette Birnbaum, Anders Bjorn, Sarah Bouchat, Xun Cao, Chris Carrington, Mariana Carvalho Barbosa, Hannah Chapman, Cindy Cheng, Hsiao Chi, Andrew Cockrell, Amelia Cronan, Chris de Sante, Yao-yao Dai, Jared Daugherty, Matt Dickenson, Nhat-Dang Do, Zhang Dong, Cassy Dorff, Michael Duda, Mark Dudley, Josh Eastin, Idil Edes, Brian Engelsma, Brad Epperly, Amy Finnegan, Laura Frankel, Marcela García-Castañón, Asaph Glosser, Brian Greenhill, Ana Guzman, Andrew Heiss, John Holbein, Molly Hogan, Hsiao-Chi Hsu, Ashley Jochim, Arzu Kibris, Jin-Young Kim, Dimitrii Kofanov, Tobias Konitzer, Sabino Kornrich, Sophie Lee, Tae-dong Lee, Rosanna Shuk-Yin Lee, Brian Ho-Yin Lee, Ning Leng, Josh Lerner, Yuting Li, Nimah Mazaheri, Shahryar Minhas, Zoe Nemerever, Alexandra Oprea, Mathias Orłowski, Victoria Paniagua, Francisco Pedraza, Ryan Powers, Barry Pump, Ben Radford, Carlisle Rainey, Eric Schmidt, Kristan Seibel, Jeffry Smith, David Sparks, Won Steinbach, Tanja Srebotnjak, Rebecca Szper, Aaron Tsang, Ashley Thirkill, Jason Thomas, Linda Tran, Peter Vining, Samantha Vortherms, Austin Wang, Zach Warner, Simon Weschle, Undes Wen, Michelle Wolfe, Xinyan Xuan, McKenzie Young, Steve Zech, and countless others for  $\beta$ -testing this material in an unusual format. They were old-school: reading text, asking questions, finding mistakes, asking questions, doing homework, asking questions, etc.

Several colleagues gave us valuable feedback on the manuscript and the content of the courses on which it was based. We would like to thank Christian Breunig, Seth Hill, Susan Holmes, Will Moore, Molly Roberts, Alex Tahk, and Yiqing Xu. We are grateful to Michael Alvarez and Neal Beck for their advice, diligence, and high standards. Ethan J. Davis provided superb feedback as well.

In very different ways both authors received institutional support from the Center for Statistics and the Social Sciences (CSSS) at the University of

Washington, under the leadership of Adrian Raftery and subsequently by Thomas Richardson. This book would not have happened with the intellectual environment CSSS provided. Ahlquist benefited from the institutional support of the University of Wisconsin–Madison and from the School of Global Policy and Strategy as well as the department of political science at UC San Diego. Final editing of the book occurred while Ahlquist was a fellow at the Center for Advanced Studies in the Behavioral Sciences at Stanford University. Joshua Penney was extremely helpful in the production process. We are grateful for his patience and skill in navigating the final process.

## Notes on Notation

We generally follow notational standards common in applied statistics. But to a student, notation can often prove a barrier. This notation “glossary” is meant to ease the transition to reading notation-heavy material and provide a place to look up unfamiliar symbols. The underlying assumption is that students have already been introduced to basic probability, calculus, and linear algebra concepts.

Random variables and sets are denoted using script capitals. Thus, for example,  $X = \{\dots, -2, 0, 2, \dots\}$  denotes the set of even integers.  $Y \sim f_N(0, 1)$  states that  $Y$  is random variable that is distributed according to a Gaussian normal distribution with mean of 0 and variance of 1. We will denote the set of admissible values for  $X$  (its support) as  $\mathcal{X}$ .

Both upper- and lowercase letters can represent functions. When both upper- and lowercase versions of the same letter are used, the uppercase function typically represents the integral of the lowercase function, e.g.,  $G(x) = \int_{-\infty}^x g(u)du$ .

To conserve notation we will use  $f_s(\cdot; \theta)$  to represent the probability distribution and mass functions commonly used in building Generalized Linear Models.  $\theta$  denotes generic parameters, possibly vector-valued. The subscript will denote the specific distribution:

- $f_B$  is the Bernoulli distribution
- $f_b$  is the binomial distribution
- $f_\beta$  is the Beta distribution
- $f_c$  is the categorical distribution
- $f_e$  is the exponential distribution
- $f_{EV_1}$  is the type-I extreme value distribution
- $f_\Gamma$  is the Gamma distribution
- $f_{GEV}$  is the generalized extreme value distribution
- $f_L$  is the logistic distribution

- $f_{IL}$  is the log-logistic distribution
- $f_m$  is the multinomial distribution
- $f_{\mathcal{N}}$  is the Gaussian (Normal) distribution
- $f_{Nb}$  is the negative binomial
- $f_P$  is the Poisson distribution
- $F_W$  is the Weibull distribution

To conform with conventional terminology and notation in  $\mathcal{R}$ , we refer to one-dimensional vectors as *scalars*. Scalars and observed realizations of random variables are denoted using lowercase math script.  $\Pr(Y_i \leq y_i)$  denotes the probability that some random variable,  $Y_i$ , takes a value no greater than some realized level,  $y_i$ .

Matrices are denoted using bolded capital letters;  $\mathbf{X}_{n \times k}$  is the matrix with  $n$  rows and  $k$  columns. The symbol  $\top$  denotes matrix or vector transposition, as in  $\mathbf{X}^\top$ . Vectors are represented with bolded lowercase letters, e.g.,  $\mathbf{x}_i$ . In our notation we implicitly treat all vectors as *column* vectors unless otherwise stated. For example,  $\mathbf{x}_i$  is a column vector even though it may represent a row in the  $\mathbf{X}_{n \times k}$  matrix. “Barred” items denote the sample mean e.g.,  $\bar{y}$ .

Lowercase Greek letters are typically reserved for parameters of models and statistical distributions. These parameters could be either scalars or vectors. Vectors will be expressed in bold font. Where more specificity is needed we will subscript.

“Hatted” objects denote fitted or estimated quantities; when used in the context of an MLE then hatted objects are the MLE. For example,  $\beta$  might be a regression parameter and  $\hat{\beta}$  is the estimated value of that parameter.

Common functions, operators, and objects:

- $\propto$  means “is proportional to”
- $\dot{\sim}$  means “approximately distributed as”
- $\xrightarrow{d}$  means “convergence in distribution.”
- $\xrightarrow{p}$  means “convergence in probability,” what some texts denote *plim*.
- $\mathbb{1}(\cdot)$  is the indicator function that returns a 1 if true and a 0 otherwise.
- $\text{cov}(\cdot, \cdot)$  is the covariance function
- $\det(\cdot)$  is the determinant of a square matrix
- $E[\cdot]$  is the expectation operator
- $\exp(\cdot)$  is the exponential function
- $\Gamma(\cdot)$  is the Gamma function
- $\mathbf{I}_n$  is the  $n \times n$  identity matrix
- $\mathcal{I}(\cdot)$  is the expected Fisher information
- $I(\cdot)$  is the observed Fisher information
- *iid* means “independently and identically distributed.”
- $\Lambda(\cdot)$  is the logistic cumulative distribution function
- $\log$  is the logarithm. If no base is given, then it denotes the natural logarithm (base  $e$ )

*Notes on Notation*

xxvii

- $\nabla$  is the gradient vector of some function.
- $\Phi(\cdot)$  is the standard Normal cumulative distribution function
- $\phi(\cdot)$  is the standard Normal density function
- $\Pr(\cdot)$  denotes probability
- $\text{var}(\cdot)$  is the variance function