Springer Actuarial Lecture Notes

Michel Denuit
Donatien Hainaut
Julien Trufin

Effective Statistical Learning Methods for Actuaries I

GLMs and Extensions

Springer

Springer Actuarial

Springer Actuarial Lecture Notes

Editors-in-Chief

Hansjoerg Albrecher, University of Lausanne, Lausanne, Switzerland
Michael Sherris, UNSW, Sydney, NSW, Australia

Series Editors

Daniel Bauer, University of Wisconsin-Madison, Madison, WI, USA
Stéphane Loisel, ISFA, Université Lyon 1, Lyon, France
Alexander J. McNeil, University of York, York, UK
Antoon Pelsser, Maastricht University, Maastricht, The Netherlands
Ermanno Pitacco, Università di Trieste, Trieste, Italy
Gordon Willmot, University of Waterloo, Waterloo, ON, Canada
Hailiang Yang, The University of Hong Kong, Hong Kong, Hong Kong

This subseries of Springer Actuarial includes books with the character of lecture notes. Typically these are research monographs on new, cutting-edge developments in actuarial science; sometimes they may be a glimpse of a new field of research activity, or presentations of a new angle in a more classical field.

In the established tradition of Lecture Notes, the timeliness of a manuscript can be more important than its form, which may be informal, preliminary or tentative.

More information about this subseries at http://www.springer.com/series/15682

Michel Denuit · Donatien Hainaut ·
Julien Trufin

Effective Statistical Learning Methods for Actuaries I

GLMs and Extensions

Springer

Michel Denuit
Institut de Statistique, Biostatistique et Sciences Actuarielles
Université Catholique de Louvain
Louvain-la-Neuve, Belgium

Donatien Hainaut
Institut de Statistique, Biostatistique et Sciences Actuarielles
Université Catholique de Louvain
Louvain-la-Neuve, Belgium

Julien Trufin
Département de Mathématiques
Université Libre de Bruxelles
Brussels, Belgium

ISSN 2523-3262 ISSN 2523-3270 (electronic)
Springer Actuarial
ISSN 2523-3289 ISSN 2523-3297 (electronic)
Springer Actuarial Lecture Notes
ISBN 978-3-030-25819-1 ISBN 978-3-030-25820-7 (eBook)
https://doi.org/10.1007/978-3-030-25820-7

Mathematics Subject Classification (2010): C1, C46, C23, C35, C8, 62P05, 62-XX, 68-XX

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The present material is written for students enrolled in actuarial master programs and practicing actuaries, who would like to gain a better understanding of insurance data analytics. It is built in three volumes, starting from the celebrated Generalized Linear Models, or GLMs and continuing with tree-based methods and neural networks.

After an introductory chapter, this first volume starts with a recap' of the basic statistical aspects of insurance data analytics and summarizes the state of the art using GLMs and their various extensions: GAMs, mixed models and credibility, and some nonlinear versions, or GNMs. Analytical tools from Extreme Value Theory are also presented to deal with tail events that arise in liability insurance or survival analysis. This book also goes beyond mean modeling, considering volatility modeling (double GLMs) and the general modeling of location, scale and shape parameters (GAMLSS).

Throughout this book, we alternate between methodological aspects and numerical illustrations or case studies to demonstrate practical applications of the proposed techniques. The numerous examples cover all areas of insurance, not only property and casualty but also life and health, being based on real data sets from the industry or collected by regulators.

The R statistical software has been found convenient to perform the analyses throughout this book. It is a free language and environment for statistical computing and graphics. In addition to our own R code, we have benefited from many R packages contributed by the members of the very active community of R-users. We provide the readers with information about the resources available in R throughout the text as well as in the closing section to each chapter. The open-source statistical software R is freely available from https://www.r-project.org/.

The technical requirements to understand the material are kept at a reasonable level so that this text is meant for a broad readership. We refrain from proving all results but rather favor an intuitive approach with supportive numerical illustrations, providing the reader with relevant references where all justifications can be found, as well as more advanced material. These references are gathered in a dedicated section at the end of each chapter.

The three authors are professors of actuarial mathematics at the universities of Brussels and Louvain-la-Neuve, Belgium. Together, they accumulate decades of teaching experience related to the topics treated in the three books, in Belgium and throughout Europe and Canada. They are also scientific directors at Detralytics, a consulting office based in Brussels.

Within Detralytics as well as on behalf of actuarial associations, the authors have had the opportunity to teach the material contained in the three volumes of “Effective Statistical Learning Methods for Actuaries” to various audiences of practitioners. The feedback received from the participants to these short courses greatly helped to improve the exposition of the topic. Throughout their contacts with the industry, the authors also implemented these techniques in a variety of consulting and R&D projects. This makes the three volumes of “Effective Statistical Learning Methods for Actuaries” the ideal support for teaching students and CPD events for professionals.

Louvain-la-Neuve, Belgium Michel Denuit
Louvain-la-Neuve, Belgium Donatien Hainaut
Brussels, Belgium Julien Trufin
June 2019

Notation

Here are a few words on the notation and terminology used throughout the book. For the most part, the notation conforms to what is usual in mathematical statistics as well as insurance mathematics. Unless stated otherwise,

$n =$ Number of observations (or data points).

$y_i =$ The ith observed value of the response (or outcome, dependent variable), observed response for individual i, realization of the random variable $Y_i, i = 1, \ldots, n$.

$\widehat{y}_i =$ Fitted value, predicted response for the ith individual.

$\bar{y} =$ Average, or sample mean of the n observed responses $y_1, \ldots, y_n$.

$\boldsymbol{y} =$ Column vector of all n response values. All vectors are column ones, by convention.

$p =$ Number of features (sometimes also called covariates, independent variables, regressors, as well as explanatory variables or predictors when they influence the response).

$x_{ij} =$ Value of the jth feature for the ith data point, $i = 1, \ldots, n$, $j = 1, \ldots, p$, realization of the random variable X_{ij}.

$\boldsymbol{x}_i =$ Column vector of the p features for the ith data point.

$\boldsymbol{X} =$ Matrix of the p features for all data points, known as the design matrix with n rows and p columns ($\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}$)
or random vector with components X_j, $j = 1, \ldots, p$ (clear from the context).

$\boldsymbol{X}^\top =$ The transpose of the design matrix $\boldsymbol{X}$, with p rows and n columns.

$\mathrm{I}[\cdot] =$ Indicator function of an event (equal to 1 if the event appearing within the brackets is realized and to 0 otherwise).

$\mathrm{P}[\cdot] =$ Probability of an event.

$\mathrm{E}[\cdot] =$ Expectation of a random variable.

$\mathrm{Var}[\cdot] =$ Variance of a random variable.

$\mathrm{Cov}[\cdot, \cdot] =$ Covariance of a pair of random variables.

$\widehat{\theta} =$ Estimator or estimate (clear from the context) of the unknown parameter θ (parameters are denoted by Greek letters).

The real line is denoted as $(-\infty, \infty)$. We use the symbol $\sim$ to mean "is distributed as", $\approx$ to mean "is approximately equal to" or "is approximately distributed as" (clear from the context), and $\propto$ to mean "is proportional to".

Contents

Part I Loss Models

1 Insurance Risk Classification 3
- 1.1 Introduction 3
- 1.2 Technical Versus Commercial Price List 4
  - 1.2.1 Selling Price and Operating Cost 4
  - 1.2.2 Risk Factor Versus Rating Factor 5
  - 1.2.3 Top-Down Pricing 6
  - 1.2.4 Prevention 7
- 1.3 The Risk Classification Problem 8
  - 1.3.1 Insurance Risk Diversification 8
  - 1.3.2 Why Classifying Risks? 9
  - 1.3.3 The Need for Regression Models 10
  - 1.3.4 Observable Versus Hidden Risk Factors 11
  - 1.3.5 Risk Sharing in Segmented Tariff: Random Versus Subsidizing Solidarity 12
  - 1.3.6 Insurance Ratemaking Versus Loss Prediction 13
  - 1.3.7 A Priori Versus a Posteriori Variables 14
- 1.4 Insurance Data 15
  - 1.4.1 Claim Data 15
  - 1.4.2 Propensities 16
  - 1.4.3 Frequency-Severity Decomposition 17
  - 1.4.4 Observational Data 19
  - 1.4.5 Format of the Data 21
  - 1.4.6 Data Quality Issues 23
- 1.5 Bibliographic Notes and Further Reading 25
- References 25

2 Exponential Dispersion (ED) Distributions 27
2.1 Introduction 27
2.2 From Normal to ED Distributions 28
2.2.1 Normal Distribution 28
2.2.2 ED Distributions 31
2.2.3 Some Distributions Related to the Normal 33
2.3 Some ED Distributions 36
2.3.1 Gamma Distribution 36
2.3.2 Inverse Gaussian Distribution 40
2.3.3 Binomial Distribution 44
2.3.4 Poisson Distribution 51
2.4 Some Useful Properties 57
2.4.1 Averages 58
2.4.2 Mean 60
2.4.3 Variance 61
2.4.4 Variance Function 63
2.4.5 Weights 64
2.5 Tweedie Distributions 65
2.5.1 Power Variance Function 65
2.5.2 Scale Transformations 66
2.6 Bibliographic Notes and Further Reading 68
References 68

3 Maximum Likelihood Estimation 69
3.1 Introduction 69
3.2 Statistical Inference 70
3.2.1 Estimation 70
3.2.2 Properties of Estimators 71
3.3 Likelihood-Based Statistical Inference 73
3.3.1 Maximum-Likelihood Estimator 73
3.3.2 Derivation of the Maximum-Likelihood Estimate 74
3.3.3 Properties of the Maximum-Likelihood Estimators 75
3.4 Fisher's Score and Information 77
3.4.1 Limited Taylor Expansion of the Log-Likelihood 77
3.4.2 Observed Information 77
3.4.3 Fisher's Score and Expected Information 78
3.4.4 Alternative Expression 79
3.5 Sampling Distribution 80
3.6 Numerical Illustrations 80
3.6.1 Death Counts 80
3.6.2 Claim Frequencies 84

3.7 Bootstrap 89
3.7.1 Estimation Accuracy 89
3.7.2 Plug-In Principle 90
3.7.3 Bootstrap Estimate 91
3.7.4 Nonparametric Versus Parametric Bootstrap 91
3.7.5 Applications 92
3.8 Bibliographic Notes and Further Reading 92
References 94

Part II Linear Models

4 Generalized Linear Models (GLMs) 97
4.1 Introduction 97
4.2 Structure of GLMs 98
4.2.1 Response Distribution 99
4.2.2 Linear Score 100
4.2.3 Design Matrix 101
4.2.4 Link Function 103
4.2.5 Interactions 110
4.2.6 Recap' of the Assumptions for Regression with GLMs 112
4.3 Likelihood Equations in GLMs 112
4.3.1 General Form 112
4.3.2 Fisher Information 114
4.3.3 Impact of the Variance Function 114
4.3.4 Likelihood Equations and Fisher's Information with Canonical Link 115
4.3.5 Individual or Grouped Data 117
4.3.6 Marginal Totals Under Canonical Link 120
4.3.7 Offsets 121
4.4 Solving the Likelihood Equations in GLMs 124
4.4.1 Back to the Normal Linear Regression Model 124
4.4.2 IRLS Algorithm 125
4.4.3 Choice of the Base Levels for Categorical Features 129
4.4.4 Correlated Features, Collinearity 130
4.4.5 Cramer's V 132
4.4.6 Omitted Variable Bias 133
4.4.7 Numerical Illustrations 135
4.5 Deviance 151
4.5.1 Null Model 151
4.5.2 Full, or Saturated Model 151
4.5.3 Deviance 152

4.5.4 Bernoulli Case 154
4.5.5 Covariance Penalties and AIC 155
4.6 Estimation of the Dispersion Parameter 160
4.7 Sampling Distribution of Estimated Regression Coefficients 161
4.7.1 Sampling Distribution 161
4.7.2 Back to the Normal Linear Regression Model 161
4.7.3 Sampling Distribution in GLMs 162
4.7.4 Wald Confidence Intervals 163
4.7.5 Variance Inflation Factor 164
4.7.6 Testing Hypothesis on the Parameters 165
4.7.7 Numerical Illustrations 168
4.8 Influence Measures 172
4.8.1 General Definition 172
4.8.2 Diagnostic Measures Based on the Hat Matrix 172
4.8.3 Leave-One-Out Estimators and Cook's Distance 173
4.9 Raw, Standardized and Studentized Residuals 174
4.9.1 Back to the Normal Linear Regression Model 174
4.9.2 Residuals in GLMs 176
4.9.3 Individual or Grouped Residuals 178
4.9.4 Numerical Illustrations 178
4.10 Risk Classification in Motor Insurance 181
4.10.1 Description of the Data Set 181
4.10.2 Marginal Impact of Features on Claim Counts 184
4.10.3 Poisson GLM Analysis for Claim Counts 184
4.10.4 Alternative Modeling 187
4.11 Quasi-Likelihood, M-Estimation and Pseudo-Likelihood 191
4.11.1 Quasi-Likelihood Estimation 191
4.11.2 M-Estimation 192
4.11.3 Pseudo-Maximum Likelihood 193
4.12 Conclusion About GLMs 194
4.13 Bibliographic Notes and Further Reading 194
References 195

5 Over-Dispersion, Credibility Adjustments, Mixed Models, and Regularization 197
5.1 Introduction 197
5.2 Mixed Distributions 199
5.2.1 Heterogeneity and Overdispersion 199
5.2.2 From Homogeneous ED to ED Mixtures 201
5.2.3 Mixed Poisson Distribution 202
5.3 Structure of GLMMs 209

5.4 Inference in GLMMs for Cross-Sectional Data . . . 211
5.4.1 Maximum-Likelihood Estimation . . . 211
5.4.2 Numerical Illustrations . . . 212
5.5 Panel Data . . . 216
5.5.1 Panel Data . . . 216
5.5.2 Modeling Panel Data with Static Random Effects . . . 217
5.5.3 Modeling Panel Data with Dynamic Random Effects . . . 218
5.5.4 Generalized Estimating Equations (GEEs) . . . 219
5.6 From Mixed Models to Experience Rating . . . 223
5.6.1 Predictive Distribution . . . 223
5.6.2 ED Family with Conjugate Priors and Credibility Premiums . . . 226
5.6.3 Poisson-Gamma Credibility Model . . . 228
5.7 Usage-Based Motor Insurance Pricing . . . 231
5.7.1 Context . . . 231
5.7.2 Mixed Poisson Model for Annual Claim Frequencies . . . 231
5.7.3 Single Behavioral Variable, or Signal . . . 233
5.7.4 Multiple Signals . . . 234
5.7.5 Credibility Updating Formulas . . . 235
5.7.6 Presentation of the Data Set . . . 236
5.7.7 Fitted Models . . . 237
5.7.8 A Posteriori Corrections . . . 240
5.8 Mixed Models, Penalized Log-Likelihood, and Regularization . . . 247
5.9 Bibliographic Notes and Further Reading . . . 249
References . . . 250

Part III Additive Models

6 Generalized Additive Models (GAMs) . . . 253
6.1 Introduction . . . 253
6.2 Structure of GAMs . . . 254
6.2.1 Response Distribution . . . 254
6.2.2 Additive Scores . . . 255
6.2.3 Link Function . . . 258
6.3 Inference in GAMs . . . 260
6.3.1 Local Polynomial GLMs . . . 260
6.3.2 Splines . . . 275
6.3.3 Extension to Multiple Continuous Features . . . 280
6.4 Risk Classification in Motor Insurance . . . 282
6.4.1 Description of the Data Set . . . 282
6.4.2 Modeling Claim Counts . . . 284

6.5 Poisson Regression for Risk Classification in Life and Health Insurance . . . 298
6.5.1 Life Table Modeling . . . 298
6.5.2 Mortality Statistics . . . 298
6.5.3 Actual Likelihood . . . 299
6.5.4 Poisson Likelihood . . . 300
6.5.5 Dual Modeling . . . 301
6.5.6 Poisson GLM with Aggregated Data . . . 301
6.5.7 Poisson Regression with Individual Data . . . 303
6.6 Numerical Illustration . . . 306
6.7 Graduation of Morbidity Rates . . . 309
6.7.1 Multi-state Modeling . . . 309
6.7.2 Transition Probabilities and Intensities . . . 310
6.7.3 Likelihood Function . . . 311
6.7.4 Application to Medical Expenses Cover . . . 312
6.8 Boosting . . . 318
6.8.1 General Principle . . . 318
6.8.2 Gradient Boosting . . . 319
6.8.3 Likelihood-Based Boosting . . . 319
6.8.4 Feature Selection . . . 320
6.9 Projection Pursuit Regression . . . 321
6.10 Regression Trees and Multivariate Adaptive Regression Splines . . . 322
6.10.1 General Approach . . . 322
6.10.2 Calibration . . . 323
6.11 Bibliographic Notes and Further Reading . . . 325
References . . . 326

7 Beyond Mean Modeling: Double GLMs and GAMs for Location, Scale and Shape (GAMLSS) . . . 329
7.1 Introduction . . . 329
7.2 Tweedie Compound Poisson Distributions . . . 330
7.2.1 Compound Distributions . . . 330
7.2.2 Compound Poisson Distributions . . . 331
7.2.3 Tweedie and Compound Poisson with Gamma Summands . . . 332
7.2.4 Exponent Parameter . . . 334
7.2.5 Limitation of Compound Poisson Tweedie GLM . . . 335
7.3 Dispersion Modeling: Double GLMs . . . 336
7.4 Dispersion Modeling in Mortality Graduation with Duplicates . . . 338
7.4.1 Duplicates . . . 338
7.4.2 Numbers of Contracts, Lives, Deaths and Claims . . . 338
7.4.3 Mean-Variance Relationship . . . 339

7.4.4 Overdispersed Binomial Distribution 340
7.4.5 Dispersion Modeling 341
7.5 Beyond Dispersion 342
7.5.1 The GAMLSS Framework 342
7.5.2 GAMLSS for Count Data 343
7.5.3 Finite Mixture Distributions 346
7.5.4 Zero-Augmented Distributions 347
7.6 Beyond the Tweedie Loss Reserving Model 348
7.6.1 Motivation 348
7.6.2 Notation and Data 348
7.6.3 Collective Model for Losses 351
7.7 Bibliographic Notes and Further Reading 356
References 358

Part IV Special Topics

8 Some Generalized Non-linear Models (GNMs) 363
8.1 Introduction 363
8.2 Mortality Projection 365
8.2.1 Observed Mortality Trends 365
8.2.2 Log-Bilinear Decomposition 369
8.2.3 Identifiability Constraints 371
8.2.4 Model Interpretation 372
8.2.5 Calibration 373
8.2.6 Poisson Maximum Likelihood 373
8.2.7 Cairns-Blake-Dowd Model 383
8.3 Projection Models for Health Expenses 386
8.3.1 Context 386
8.3.2 Data Description 388
8.3.3 Modeling Approach 389
8.3.4 Model Specifications 390
8.3.5 Model Selection 392
8.4 Bibliographic Notes and Further Reading 395
References 398

9 Extreme Value Models 401
9.1 Introduction 401
9.2 Basics of EVT 402
9.2.1 Sample Averages and Maxima 402
9.2.2 Extreme Value Distribution 403
9.2.3 Fisher-Tippett Theorem 404
9.2.4 Extreme Lifetimes 406

9.3 Excess Over Threshold and Mean Excess Function 408
- 9.3.1 Excess Distribution 408
- 9.3.2 Mean Excess Function 408
- 9.3.3 Illustration in Nonlife Insurance 409
- 9.3.4 Illustration in Life Insurance 412

9.4 Generalized Pareto Distribution 418
- 9.4.1 Definition 418
- 9.4.2 Properties 419

9.5 Peak Over Threshold Approach 420
- 9.5.1 Principle 420
- 9.5.2 Tail Formulas 421
- 9.5.3 Tail Estimators 422
- 9.5.4 Applications to Remaining Lifetimes 423
- 9.5.5 Selection of the Generalized Pareto Threshold 425

9.6 Poisson-Generalized Pareto Model for Excesses Over a High Threshold 433
- 9.6.1 Distribution of the Maximum of a Generalized Pareto Sample of Poisson Size 433
- 9.6.2 Probable Maximal Loss 435
- 9.6.3 Highest Age at Death 436
- 9.6.4 Stability Property with Increasing Thresholds 437

9.7 Bibliographic Notes and Further Reading 437

References 440