UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Fitting of survival functions for grouped data on insurance policies

by

Elizabeth Magrietha Louw

Submitted in fulfilment of part of the requirements
for the degree of Doctor of Philosophiae
in the Faculty of Natural and Agricultural Sciences
University of Pretoria

Supervisor: Professor N.A.S. Crowther

January 2002

© University of Pretoria

Preface

I am extremely grateful to my supervisor, Professor N.A.S. Crowther, for his careful supervision, positive guidance and constructive comments throughout.

I appreciated the enthusiastic encouragement of my family and friends. The patience and understanding and continuous support of my husband, Johann, and daughters, Hannelie and Annemarie, kept me going.

Contents

1 INTRODUCTION 1

1.1 **Background** 1

1.2 **The Research Problem** 2

1.3 **Generalization of the Statistical Procedures** 4

1.4 **Future Research** 4

1.5 **Outlay of chapters** 4

2 SURVIVAL FUNCTIONS AND DISTRIBUTIONS 6

2.1 **Introduction** 6

2.2 **Notation** 6

2.2.1 Right-censored continuous survival data 6

2.2.2 Discrete data 8

2.2.3 Interval-censored data 8

2.3 **Basic Survival Functions** 10

2.3.1 Introduction 10

2.3.2 Probability distribution and distribution function 10

2.3.3 Survivor function 11

2.3.4 Hazard rate function 11

2.3.5 Cumulative hazard function 13

2.3.6 Median lifetime, mean lifetime and mean residual lifetime 13

2.4 **Survival Models** 14

2.4.1 Introduction 14

2.4.2 Weibull distribution 14

2.4.3 The log-logistic distribution 15

2.4.4 The lognormal distribution 15

2.4.5 Location-scale parameter survival models 16

2.5 **Relationships between Survival Distributions** 18

2.5.1 Introduction 18

2.5.2 Relationship between the Weibull and extreme value distributions . 18

2.5.3 Relationship between the log-logistic and logistic distributions . . . 19

2.5.4 Relationship between the lognormal and normal distributions 20

2.5.5 A linear model representation in log-time 20

2.6 **Construction of Likelihood Functions** 22

2.6.1 Introduction 22

2.6.2 Likelihood function for random right-censored continuous data . . . 22

2.6.3 Likelihood function for interval-censored data 24

2.6.4 Likelihood function for right-censored grouped data 25

3 PARAMETRIC MODEL FOR A SINGLE SAMPLE FROM A HOMOGENEOUS POPULATION 27

3.1 Introduction 27

3.2 Standard Method of Maximum Likelihood Estimation 28

3.2.1 Introduction 28

3.2.2 Likelihood function for the linear model in log-time 28

3.2.3 Maximum likelihood estimators of the log-linear parameters 29

3.3 Maximum Likelihood Estimation subject to Constraints - A Fixed Censoring Time 30

3.3.1 Introduction 30

3.3.2 Notation for a fixed censoring time 31

3.3.3 Fitting of a Weibull distribution to grouped survival data 35

3.3.4 Fitting of a log-logistic distribution to grouped survival data 37

3.3.5 Fitting of a lognormal distribution to grouped survival data 39

3.3.6 A measure to compare the fit of survival distributions 41

3.4 Maximum Likelihood Estimation subject to Constraints - Staggered Entry of Policies 43

3.4.1 Introduction 43

3.4.2 Notation for staggered entry 43

3.4.3 Definition of Constraints 49

3.4.4 Method of maximum likelihood estimation subject to constraints: staggered entry . . . 52

3.4.5 Fitting of one survival distribution to the four histograms . . . 53

3.4.6 Fitting of four survival distributions to the four histograms . . . 60

3.4.7 Fitting of a joint histogram to the four histograms . . . 69

3.4.8 Estimated survivor and hazard functions and percentiles . . . 71

3.5 **Simulation Studies** . . . 73

4 PARAMETRIC REGRESSION MODELS FOR SURVIVAL DATA WITH COVARIATES **87**

4.1 **Notation** . . . 87

4.2 **Three Approaches to Regression Modelling** . . . 87

4.3 **Comparison between the AFM and the PHM** . . . 93

4.4 **Comparison between the AFM and the POM** . . . 94

4.5 **Log-linear Presentation of Regression Models for Survival Data** . . . 95

4.5.1 A linear regression model for the log of the hazard ratio . . . 95

4.5.2 A linear regression model for the log of the odds ratio . . . 95

4.5.3 A linear regression model for log-time . . . 96

4.6 **Maximum Likelihood Estimation** . . . 97

4.6.1 Introduction . . . 97

4.6.2 Likelihood function for random right-censored continuous data . . . 97

4.6.3 Likelihood function for right-censored grouped data . . . 98

4.6.4 Likelihood function for the linear model in log-time 99

4.6.5 Maximum likelihood estimators 101

4.7 Maximum Likelihood Estimation subject to Constraints 103

4.7.1 Introduction 103

4.7.2 Notation for a regression model with one risk factor at staggered entry 103

4.7.3 The log-logistic regression model: staggered entry, shape parameter remains constant 106

4.7.4 The log-logistic regression model: staggered entry, shape parameter alters 110

4.7.5 Deriving of indices and risk scores from the log-logistic regression model 113

4.7.6 The Weibull regression model: staggered entry, shape parameter remains constant 115

4.7.7 The Weibull regression model: staggered entry, shape parameter alters 118

4.7.8 Deriving of indices and risk scores from the Weibull regression model 121

4.7.9 The fitting of a regression model with a continuous risk factor . . . 124

4.7.10 Notation for a regression model with two risk factors at staggered entry 129

4.7.11 The log-logistic regression model with two risk factors at staggered entry 131

4.7.12 The Weibull regression model with two risk factors at staggered entry 135

5 APPLICATION TO DATA FROM THE INSURANCE INDUSTRY 140

5.1 Description of the Data Set 140

5.1.1 Introduction 140

5.1.2 The raw data set of policies 140

5.1.3 The grouped data set of policies 143

5.2 **Fitting of a Lifetime Distribution** 145

5.2.1 A fixed censoring time 145

5.2.2 Staggered entry 148

5.3 **Fitting of Parametric Regression Models** 160

5.3.1 Introduction 160

5.3.2 A survival model for each level of a risk factor 160

5.3.3 Deriving of indices and risk scores from log-logistic regression model 169

5.3.4 Deriving of indices and risk scores from Weibull regression model . . 181

5.3.5 The fitting of a regression model with a continuous predictor 193

5.3.6 A survival model for each combination of levels of two risk factors . 195

5.3.7 Relationship between the indices of the regression and logit model . 211

5.3.8 Median lifetimes of the nine survival distributions 215

6 RESUMÉ 216

A COMPUTER PROGRAMS 219

A.1 **Chapter 3: Maximum Likelihood Estimation** 219

A.1.1 A Fixed Censoring Time - Standard Program using PROC LIFEREG 219

A.1.2 Staggered Entry of Policies - Standard Program using PROC LIFEREG220

A.2 **Chapter 3: Maximum Likelihood Estimation subject to Constraints** . 221

A.2.1 A Fixed Censoring Time - IML Programs 221

A.2.2 Staggered Entry of Policies - IML Programs 225

A.3 **Chapter 3: Simulation Studies** 235

A.3.1 Program to simulate continuous right-censored lifetime data 235

A.3.2 Program to simulate grouped right-censored lifetime data 239

A.4 **Chapter 4: Maximum Likelihood Estimation** 244

A.4.1 Staggered Entry of Policies - Standard Programs using PROC LIFEREG 244

A.5 **Chapter 4: Maximum Likelihood Estimation subject to Constraints** . 249

A.5.1 Staggered Entry of Policies - IML Programs 249

List of Tables

3.1 Relative frequency distribution of survival data - fixed censoring time 32

3.2 Relative frequency distributions of survival data - staggered entry . . 48

3.3 Weibull Simulation - Continuous Data 75

3.4 Weibull Simulation - Grouped Data 77

3.5 Log-logistic Simulation - Continuous Data 79

3.6 Log-logistic Simulation - Grouped Data 81

3.7 Lognormal Simulation - Continuous Data 83

3.8 Lognormal Simulation - Grouped Data 85

4.1 Number of entries per cell in cross table of entry period and risk factor A 105

4.2 Number of entries per cell in cross table of entry period and risk factors A and B 130

5.1 Frequency distributions of the four samples 143

5.2 Maximum likelihood estimation subject to constraints: a fixed censoring time 146

5.3 Fitting of one survival model to the four histograms 150

5.4 Fitting of one survival model to the four histograms 151

5.5 Fitted joint relative frequency distribution to the four samples 152

5.6 Percentiles estimated from Weibull and log-logistic regression models 159

5.7 Multi-dimensional frequency table of grouped data set with one risk factor 161

5.8 Fitting a regression model (constant shape) to grouped data with one risk factor 162

5.9 Parameters of a survival model (constant shape) for each level of risk factor AGE 162

5.10 Fitting a regression model (shape alters) to grouped data with one risk factor 163

5.11 Parameters of a survival model (shape alters) for each level of risk factor AGE 164

5.12 Fitted joint frequency distributions for the three AGE levels 165

5.13 Comparison of indices: log-logistic regression model and logit model 177

5.14 Predicted indices from log-logistic regression model (constant shape) 178

5.15 Predicted indices from log-logistic regression model (shape alters) . 178

5.16 Predicted risk scores from log-logistic regression model (constant shape) 179

5.17 Predicted risk scores from log-logistic regression model (shape alters) 179

5.18 Lifetime percentiles estimated from log-logistic regression model . . . 180

5.19 Comparison of indices: Weibull regression model and logit model . . 189

5.20 Predicted indices from Weibull regression model (constant shape) . 190

5.21 Predicted indices from Weibull regression model (shape alters) . . . 190

5.22 Predicted risk scores from Weibull regression model (constant shape) 191

5.23 Predicted risk scores from Weibull regression model (shape alters) . 191

5.24 Lifetime percentiles estimated from Weibull regression model 192

5.25 Fitting a regression model (constant shape) to grouped data with one continuous predictor 193

5.26 Parameters of a survival model (constant shape) for each age group 194

5.27 Cross table of Age and Score 195

5.28 Multi-dimensional frequency table of grouped data set with two risk factors 196

5.29 Fitting a regression model with two risk factors 198

5.30 Parameters of a survival model for each combination of AGE and SCORE levels 199

5.31 Fitted joint frequency distributions for the nine combinations of AGE and SCORE levels 200

5.32 Logit model indices for three age levels and three score levels obtained from the logit model 212

5.33 Comparison of odds ratios (indices): log-logistic and Weibull regression models and logit model 214

5.34 Median lifetimes of the nine survival distributions 215

List of Figures

3.1 Experimental design for staggered entry of policies 44

3.2 Constraints imposed by the experimental design 51

5.1 Experimental design for illustrative data set 142

5.2 Histograms of the four relative frequency distributions 144

5.3 June 1998 policies: histogram and fitted survival distributions 147

5.4 Constraints imposed by the experimental design 149

5.5 Joint histogram over the four histograms 153

5.6 Joint histogram and fitted survival distributions 154

5.7 Survival curves of fitted Weibull and log-logistic models 157

5.8 Graphs of hazard rates of fitted Weibull and log-logistic models . . . 158

5.9 Joint histogram and fitted survival distributions for age group [18;35) 166

5.10 Joint histogram and fitted survival distributions for age group [35;45) 167

5.11 Joint histogram and fitted survival distributions for age group [45+) 168

5.12 Survival curves of fitted log-logistic age group models 175

5.13 Graphs of hazard rates of fitted log-logistic age group models 176

5.14 Survival curves of fitted Weibull age group models 187

5.15 Graphs of hazard rates of fitted Weibull age group models 188

5.16 Joint histogram and fitted survival distributions for age group [18;35) and low score 202

5.17 Joint histogram and fitted survival distributions for age group [18;35) and medium score 203

5.18 Joint histogram and fitted survival distributions for age group [18;35) and high score 204

5.19 Joint histogram and fitted survival distributions for age group [35;45) and low score 205

5.20 Joint histogram and fitted survival distributions for age group [35;45) and medium score 206

5.21 Joint histogram and fitted survival distributions for age group [35;45) and high score 207

5.22 Joint histogram and fitted survival distributions for age group [45+) and low score 208

5.23 Joint histogram and fitted survival distributions for age group [45+) and medium score 209

5.24 Joint histogram and fitted survival distributions for age group [45+) and high score 210