بسم الله الرحمن الرحيم

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الالكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأفلام قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأفلام بعيدا عن الغبار

في درجة حرارة من ١٥-٢٥ مئوية ورطوبة نسبية من ٢٠-٤٠%

To be Kept away from Dust in Dry Cool place of 15-25- c and relative humidity 20-40%

بعض الوثائق
الاصلية تالفة

بالرسالة صفحات

لم ترد بالاصل

Cairo University
Faculty of Economics and Political Science
Department of Statistics

Analysis of Longitudinal data with intermittent Missing Values Using the Stochastic EM Algorithm

By

Abeer Saleh Ahmed Abd-Alla

Supervised by

Professor Dr. Maged Osman
Professor of Statistics
Faculty of Economics and Political Science
Cairo University

Dr. Laila El-Zeini
Associate Professor of Statistics
Faculty of Economics and Political Science
Cairo University

A Thesis Submitted to the Faculty of Economics and Political Science (Department of Statistics) in partial Fulfillment of the Requirements for the Ph. D. in Statistics

2005

بسم الله الرحمن الرحيم

وَمَآ أُوتِيتُم مِّنَ العِلمِ إِلا قَلِيلاً

صدق الله العظيم

سورة الإسراء الآية ٨٥

DEDICATION

Abeer, Yasmeen, Dr. Ahmed Gad and My Family

Name: Abeer saleh ahmed Abd-Alla

Nationality: Egyptian

Date and Place of birth: 29/6/ Cairo

Degree: Ph.D.

Specialization: Statistics

Supervisors:

Professor Dr. Maged Osman
Professor of Statistics

Dr. Laila El-Zeini
Associate Professor of Statistics

Title of the Thesis: "Analysis of Longitudinal Data with Intermittent Missing Values using The Stochastic EM Algorithm"

Summary:

In longitudinal studies each unit is measured on several occasions. In practice, some intended measurements may not available for some subjects resulting in a missing data pattern. Dropout pattern occurs when some subject are withdrawn from the study prematurely. The missing data pattern is defined as intermittent if a missing value followed by an observed value. When the probability of missingness depends on the missing value, and may be on the observed values, the missing data mechanism termed as nonrandom. Ignoring the missing values mechanism in this case leads to biased inference. In this study, the stochastic EM algorithm is proposed and developed to find parameters estimates to normal response distribution in the presence of intermittent missing values. Also, in this setting, the Monte Carlo method is developed to find the standard errors of parameters estimates. Assumptions about the distribution of the response can have a large effect on inference about parameters that govern the dropout process. So, it is important to consider the sensitivity of conclusions to distributional assumptions. We apply the proposed method to non-normal response distribution. This enables us to examine the effect of the distributional assumptions of the response on the parameters that characterize the dropout process. Finally, the proposed techniques are applied to a real data from the International Breast Cancer Study Group.

Supervisors:

Professor Dr. Maged Osman
Professor of Statistics

Dr. Laila El-Zeini
Associate Professor of Statistics

Abstract

In longitudinal studies each unit is measured on several occasions. In practice, some intended measurements may not be available for some subjects resulting in a missing data pattern. Dropout pattern occurs when some subject are withdrawn from the study prematurely. The missing data pattern is defined as intermittent if a missing value is followed by an observed value. When the probability of missingness depends on the missing value, and may be on the observed values, the missing data mechanism termed as nonrandom. Ignoring the missing values mechanism in this case leads to biased inference. In this study, the stochastic EM algorithm is proposed and developed to find parameters estimates to normal response distribution in the presence of intermittent missing values. Also, in this setting, the Monte Carlo method is developed to find the standard errors of parameters estimates. Assumptions about the distribution of the response can have a large effect on inference about parameters that govern the dropout process. So, it is important to consider the sensitivity of conclusions to distributional assumptions. We apply the proposed method to non-normal response distributions. This enables us to examine the effect of the distributional assumptions of the response on the parameters that characterize the dropout process. Finally, the proposed techniques are applied to a real data from the International Breast Cancer Study Group.

Key words: Longitudinal data; Nonrandom missing values; Intermittent missing data; The stochastic EM algorithm; MCMC methods; Standard errors.

Supervisors:

Professor Dr. Maged Osman
Professor of Statistics

Dr. Laila El-Zeini
Associate Professor of Statistics

Acknowledgements

First of all, ultimate thanks are due to ALLAH. Many thanks to my mother who without her aid this work could not be done.

I wish to express my sincere appreciation and gratefulness to Dr. Ahmed M. Gad, Department of Statistics, Faculty of Economics and Political Science, Cairo University for suggesting this point for research, supervising, valuable help and constructive criticism. His effort in revising the manuscript is much appreciated.

Deep thanks and gratitude are also due to Prof. Dr. Maged Osman, professor of statistics, Department of Statistics, Faculty of Economics and Political Science, Cairo University for supervising.

Thanks are also extended to Dr. Laila El-Zeini, associate professor of statistics, Department of Statistics, Faculty of Economics and Political Science, Cairo University for sharing the supervision.

I would like to thank Quality of Life Committee of the International Breast Cancer Study Group for permission to use the quality-of-life data.

My sincerest thanks go to my family. My family has been a constant source of support, emotional and moral, during my preparation for this thesis.

Contents

List of Tables v

List of Abbreviations vi

Chapter 1: Longitudinal Data Analysis: An Overview 1

1.1 Introduction 1

1.2 Longitudinal Data with Missing Values 3

1.2.1 Notation 3

1.2.2 Missing Data Patterns 4

1.2.3 Missing Data Mechanisms 5

1.2.4 Methods for Handling Missingness 5

1.3 Parametric Models for Covariance Structure 7

1.4 Modelling the Longitudinal Data with Missing Values 10

1.5 A General Class of Selection Models 14

1.6 Likelihood Estimation for Incomplete Longitudinal Data (Ignorable Case) 15

1.6.1 The Newton-Raphson Algorithm 15

1.6.2 The Scoring Method 16

1.6.3 The Expectation-Maximization (EM) Algorithm 16

1.6.4 Jennrich and Schluchter Algorithm 19

1.7 Likelihood Estimation for Incomplete Longitudinal Data (Nonignorable Case) 21

1.8 Sensitivity Analysis 24

1.9 Structure of this Dissertation 25

Chapter 2: The Stochastic EM Algorithm 27

2.1 Introduction 27

2.2 Markov Chain Monte Carlo (MCMC) Methods 28

2.3 The Metropolis-Hastings Algorithm 30

2.4 The Gibbs sampling 31

2.5 Stochastic Variants of the EM Algorithm........ 32
2.5.1 The Stochastic EM (SEM) Algorithm 33
2.5.2 The Monte Carlo EM (MCEM) Algorithm........ 34
2.5.3 The Stochastic Approximation EM (SAEM) Algorithm........ 35
2.6 The Stochastic EM Estimates........ 36
2.6.1 The Mean of the Stationary Distribution........ 36
2.6.2 The Largest Log-Likelihood........ 37
2.7 Monitoring Convergence........ 37
2.7.1 The Gelman and Rubin Method........ 38
2.7.2 The Brooks and Gelman Method........ 41
2.7.3 The Yu and Mykland Method........ 41
2.7.4 The Brooks Method........ 42
2.8 Standard Error........ 43
2.8.1 Direct Computation........ 44
2.8.2 Missing Information Principle........ 44
2.8.3 The Louis Method........ 47
2.8.4 The Monte Carlo Method........ 47
2.8.5 The Bootstrap Method........ 48

Chapter 3: Normal Model for Longitudinal Data........ 49
3.1 Introduction 49
3.2 The Stochastic EM Algorithm for Normal Longitudinal Data with Nonrandom Intermittent Missing Values........ 51
3.3 Standard Errors for Longitudinal Data with Nonrandom Intermittent Missing Values........ 54
3.3.1 The Log-Likelihood Function........ 54
3.3.2 The Log-Likelihood Derivatives........ 55
3.3.3 The Monte Carlo Method........ 56
3.4 Application........ 57

Chapter 4: Non-normal Models for Longitudinal Data........ 65
4.1 Introduction 65
4.2 The Methods of Sensitivity Analysis........ 66

4.2.1 Change Distributional Assumptions.................................... 67

4.2.2 Fitting Several Models for the Missing Data Mechanism.............. 69

4.2.3 Pattern-Mixture Modeling Approach................................ 70

4.2.4 Local Influence Method.. 71

4.3 The EM Algorithm for Non-normal Models 73

4.3.1 Introduction.. 73

4.3.2 The Univariate Case.. 73

4.3.3 The Mutivariate Case... 76

4.4 The Stochastic EM Algorithm for Non-normal Longitudinal data with Nonrandom Intermittent Missing Values.. 78

4.4.1 The General Case... 78

4.4.2 The t Distribution.. 82

4.5 Application.. 84

Chapter 5: Discussion and Conclusions.. 87

References 91

Summary (in Arabic) 1

List of Tables

Table 3.1: The SEM estimates for PACIS, AR(1) 61

Table 3.2: The Shrink factor of SEM estimates, AR(1) 61

Table 3.3: The SEM estimates for PACIS, unstructured model 62

Table 3.4: The Shrink factor of SEM estimates for PACIS, unstructured model 63

Table 3.5: The SE for PACIS, AR(1) 63

Table 4.1: The SEM estimates for PACIS, AR(1) 85

Table 4.2: The SE for PACIS 86