EM APPROACH ON INFLUENCE MEASURES IN COMPETING RISKS VIA PROPORTIONAL HAZARD REGRESSION MODEL

FAIZ. A. M. ELFAKI

FSAS 2000 5

EM APPROACH ON INFLUENCE MEASURES IN COMPETING RISKS VIA PROPORTIONAL HAZARD REGRESSION MODEL

FAIZ. A. M. ELFAKI

MASTER OF SCIENCE
UNIVERSITI PUTRA MALAYSIA
2000

EM APPROACH ON INFLUENCE MEASURES IN COMPETING RISKS VIA PROPORTIONAL HAZARD REGRESSION MODEL

By

FAIZ. A. M. ELFAKI

Thesis Submitted in Fulfilment of the Requirements for the
Degree of Master of Science in the Faculty of
Science and Environmental Studies
Universiti Putra Malaysia

June 2000

Dedicated to my late mother, Etayah Mohamed Abdullah,
May Allah rest her soul in heaven

Abstract of thesis submitted to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science

EM APPROACH ON INFLUENCE MEASURES IN COMPETING RISKS VIA PROPORTIONAL HAZARD MODEL

By

FAIZ AHMED MOHAMED ELFAKI

June 2000

Chairman: **Noor Akma Ibrahim, Ph.D.**

Faculty: **Science and Environmental Studies**

In a conventional competing risks model, the time to failure of a particular experimental unit might be censored and the cause of failure can be known or unknown. In this thesis the analysis of this particular model was based on the cause-specific hazard of Cox model. The Expectation Maximization (EM) was considered to obtain the estimate of the parameters. These estimates were then compared to the Newton-Raphson iteration method. A generated data where the failure times were taken as exponentially distributed was used to further compare these two methods of estimation. From the simulation study for this particular case, we can conclude that the EM algorithm proved to be more superior in terms of mean value of parameter estimates, bias and root mean square error.

To detect irregularities and peculiarities in the data set, the residuals, Cook distance and the likelihood distance were computed. Unlike the residuals, the perturbation method of Cook's distance and the likelihood distance were effective in the detection of observations that have influenced on the parameter estimates. We considered both the EM approach and the ordinary maximum likelihood estimation (MLE) approach in the computation of the influence measurements. For the ultimate results of influence measurements we utilized the approach of the one-step. The EM one-step and the maximum likelihood (ML) one-step gave conclusions that are analogous to the full iteration distance measurements. In comparison, it was found that EM one-step gave better results than the ML one-step with respect to the value of Cook's distance and likelihood distance. It was also found that Cook's distance is better than the likelihood distance with respect to the number of observations detected.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Master Sains

PENDEKATAN PEMAKSIMUMAN JANGKAAN TERHADAP UKURAN PENGARUH DALAM RISIKO BERSAING MENERUSI MODEL KADARAN BAHAYA

Oleh

FAIZ AHMED MOHAMED ELFAKI

Jun 2000

Pengerusi: **Noor Akma Ibrahim, Ph.D.**

Fakulti: **Sains dan Pengajian Alam Sekitar**

Dalam model risiko bersaing konvensional, masa kegagalan dari unit ujikaji tertentu boleh jadi tertapis dengan punca kegagalan mungkin diketahui atau tidak diketahui. Dalam tesis ini analisis model risiko bersaing adalah berlandaskan model bahaya punca-tertentu Cox. Pemaksimuman Jangkaan (PJ) dipertimbangkan untuk memperolehi anggaran bagi parameter. Anggaran ini dibandingkan dengan anggaran yang diperolehi dari kaedah lelaran Newton-Raphson. Data yang dijana dengan masa kegagalannya tertabur secara eksponen digunakan selanjutnya untuk membandingkan kedua-dua kaedah anggaran ini. Dari kajian simulasi khususnya bagi masalah ini, didapati algoritma PJ mempunyai kelebihan terhadap anggaran parameter berdasarkan nilai min, kepincangan dan punca kuasa dua min ralat (PKMR).

Untuk melihat ketidaktentuan dan keganjilan data dalam model, reja, jarak Cook dan jarak kebolehjadian dihitung. Tidak seperti reja, kaedah jarak Cook dan jarak kebolehjadian adalah berkesan dalam menentukan cerapan yang mempengaruhi anggaran parameter. Kedua-dua pendekatan iaitu PJ dan anggaran maksimum kebolehjadian dilaksanakan dalam perhitungan ukuran pengaruh. Sebagai keputusan muktamad ukuran pengaruh, satu-langkah digunakan. PJ satu-langkah dan kebolehjadian maksimum (KM) satu-langkah memberikan kesimpulan yang sama dengan ukuran jarak lelaran penuh. Secara perbandingan, didadapati bahawa PJ satu-langkah memberikan keputusan yang lebih baik daripada KM satu-langkah berdasarkan nilai jarak Cook dan jarak kebolehjadian yang diperolehi. Juga didapati bahawa jarak Cook adalah lebih baik daripada jarak kebolehjadian dari segi bilangan cerapan yang dikesan sebagai berpengaruh.

ACKNOWLEDGEMENTS

Praise be to ALLAH for giving me the strength and patience to complete this work. I would like to single out the particular and tremendous contribution of Dr. Noor Akma Ibrahim, the chairman of supervisory committee, for her persistent inspiration, constant guidance, wise counsel, encouragement, kindness, financial help and various logistic supports during all the stages of my study. I highly appreciate her effort to give first hand knowledge about a very interesting area of statistic. Her command on the subject matter, together with her research experiences, have been highly valuable to my study. In spite of her busy schedule, she managed to find enough time for my discussions and provided necessary direction in order to develop my study. Her enthusiasm and patience have left a feeling of indebtedness which can not be fully expressed.

My deepest appreciation and sincere gratitude also to Assoc. Prof. Dr. Isa Daud, Head of the Mathematics Department and member of my supervisory committee, for his kind co-operation and thoughtful suggestions. I owe a great deal of gratitude and appreciation to Mrs Fauziah Maarof, member of the supervisory committee, for her supervision and helpful comments.

I also would like to expand my thanks to Assoc. Prof. Dr. Harun bin Budin for his help and continuous encouragement. And all the members of the Mathematics Department for their kind assistance during my studies, particularly Dr. Jamal I. Daoud., Dr. Habshah Midi., Dr. Farris Assim, Mrs Fadzilah Ali and Miss Zainuridah Yusof. For my friends Iing Lukman, Idi Fulayi, Ahlam Abdel Hadi, Lawan Ahmed, Hanan Hassan whatever I stated in the acknowledgement will remain under the true credit of their genuine contribution to the success of this study. I am also thankful to Salah Madni and his wife Ghada, Hassan Doka, Yassin Mohd, Natrah Mohd, Verna Taylor and Ahmed Elyas, for their strong support and fast response whenever I needed their help.

Last but not the least, my heartfelt thanks should go to my father, Ahmed, my mother, Eltayah, my brothers Abdul Gadir, Nadir, Mohamed, Mostafa and sisters, Saza, Sara, Eltayah, Fridah, for their sacrifices, devotion and understanding, which have always been a source of inspiration and strength throughout my life up to this moment. Also my thanks go to all my friends in Yemen and Sudan.

Studying abroad for a prolonged period is unpleasant, if not impossible, without friends. For all those whose names were not mentioned here, the moral support and friendship they offered will be remembered.

I certify that an Examination Committee met on 7 June, 2000 to conduct the final examination of Faiz Ahmed Mohamed Elfaki on his Master of Science thesis entitled "EM Approach On Influence Measures In Competing Risks Via Proportional Hazard Model" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulation 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Mat Yusoff Abdullah, Ph.D.
Associate Professor
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Chairman)

Noor Akma Ibrahim, Ph.D.
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Member)

Isa Bin Daud, Ph.D.
Associate Professor/Head
Department Of Mathematics
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Member)

Fauziah Maarof, M.Sc.
Faculty of Science and Environmental Studies
Universiti Putra Malaysia
(Member)

MOHD. GHAZALI MOHAYIDIN, Ph.D.
Professor/Deputy Dean of Graduate School

Date: 12 JUN 2000

This thesis was submitted to the Senate of Universiti Putra Malaysia and was accepted as fulfilment of the requirements for the degree of Master of Science.

Kamis Awang

KAMIS AWANG, Ph.D.
Associate Professor,
Dean of Graduate School,
Universiti Putra Malaysia

Date: 13 JUL 2000

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions

(FAIZ AHMED MOHAMED ELFAKI)

Date: 12-6-2000

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

LIST OF FIGURE

CHAPTER I

INTRODUCTION

GENERAL OVERVIEW

In the early concept of regression expansion, many researchers concentrated on the residuals to detect weaknesses in models. Residuals were also used to indicate odd data points. Plots like residual plots versus projection values, and residual plots versus projection variables were recommended. Tests on residuals were practiced in most statistical analyses with the help of computer programmes. However, problems still arise whereby residual failed to fulfil normal assumptions. These problems initiated the use of other techniques on regression problems. Some of these techniques were able to improve the results of the estimation.

In the later years, efforts were directed towards the identification of isolated points and extreme cases. This procedure was known as regression diagnostic, and it helped to detect potential cases that could influence estimates of the regression. The procedure was also designed to assist researchers in making the decision whether the assumptions made on the model are suitable and valid. Literatures by

Cook (1977, 1979), Andrews and Pregibon (1978), Cook and Weisberg (1980), Belsley et al. (1980), and Cook and Weisberg (1982), introduced several diagnostic measurements in order to detect and identify influential individual or group cases with respect to the parameter estimates.

Cook proposed that the influence of data point be tested using distance measurement,

$$D_i = [(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})]/(s\sigma^2) \qquad (1.01)$$

$$i = 1, \ldots, n$$

where $\hat{\beta}$ indicates an estimate for β with full data. Full data in this context refers to the failure time[1] for all observations that can be obtained until the study is completed, while $\hat{\beta}_{(i)}$ indicates estimate for β by deleting data point i, $X'X$ is a positive (semi-) definite matrix, s is the parameter number, and σ^2 is the variance. Equation (1.01) becomes the basis for most distance measurements in detecting the influence of an observation or a case.

Influence diagnostics which have been popular in terms of their implementations are Cook's D, DFBETAS and DFFITS (see Belsley et al 1980; Cook and Weisberg, 1982). These distance measurements are formed through standardized residuals and diagonal matrix for observation from Hessian matrix ($H = X(X'X)^{-1}X'$). The diagnostic of influence that is built based on the least

[1] The time observed on individual or object from one original point to the time an anticipated event occurs.

square method needs to be adjusted in order to accommodate non-linear model. Pregibon (1981) and Cook and Weisberg (1982) contributed a lot towards the analysis of influence for models involving non-linear models. Cook (1986) also introduced the method of global measures to assess small distractions in models and applied it to linear regression analysis. The application of global measure analysis to specific problems has been described in several recent publications. Reid and Crepeau (1985) treated the influence function for proportional hazard regression model (PHRM), Bin Daud (1987) and Barlow (1997) used PHRM to analyse global measures, Bechman, Nashtshsheim, and Cook (1987) described applications to mixed model analysis of variance. Escobar and Meeker (1988) described several methods using SAS macros for local influence analyses with censored data and parametric regression models. Thomas and Cook (1989, 1990) applied local influence methods to generalized linear model, while Pettitt and Bin Daud (1989) did the same for the PHRM. Weissfed and Schneider (1990) compared numerical results of local influence analysis methods and case deletion methods for Weibull regression analysis with censored data. Wellman and Gunst (1991) proposed one-step approximation to Cook's distance to identify influential points within the context of linear measurement error models, and Escobar and Meeker (1992) described new interpretations for some local influence statistics and showed how these statistics can be extended and complemented to the traditional case deletion influence statistics for linear least squares.

Studies on diagnostic and influence in regression originally involved full data. In survival analysis[2], where most observations have to be censored, the study of the compatibility of the models and influence diagnostic becomes necessary.

Survival models, like other statistical models, can also be considered as situational estimates to a more complex process, and may, therefore, give a less definite result. This can give rise to doubts about the models. A variation study on the results of the analysis with small modifications on the data is then necessary. Therefore, one important factor in statistical analysis is to conduct a study on result suitability. Residual value and Hessian matrix are useful components in detecting extreme points, but, they cannot be used to assess the effect on model suitability in general, and parameter estimate, in particular. In this research, we extend the techniques of studying result suitability of a survival model focusing on competing risks model.

Several researchers have used competing risks in their studies. Kimball (1969) compared two models for the estimation of competing risks from grouped data. Gail (1975) compared actuarial model with other models of competing risk in analysis for failure time data. Prentice et al. (1978) discussed the analysis of failure times in the presence of competing risks based on Cox model. Holt (1978) compared two models of competing risks with special reference to matched pair experiments. Larson (1984) used log-linear model. Larson and Dinse (1985) and Kuk (1992) fitted more complex models incorporating different failure types. Lubin

[2] Analysis for failure time data

(1985) and Kay (1986) analysed competing risks via PHRM for prostate cancer data. Farewell (1986) considered a mixture of logistic regression and Weibull regression. Dinse (1986) developed a likelihood-based approach, which leads to nonidentifiability and breaks down if the hazard functions of the competing risks are proportional. Gray (1988) used competing risks analysis in reliability study for comparing the probability of failures of a certain type being observed among different groups. Robins and Greenland (1989), and Bagai et al. (1989) used non-parametric approach on two independent risks. Heckman and Honorore (1989) discussed threats to competing risk model. Benichou and Gail (1990) looked into estimated absolute cause-specific risk in cohort studies. Goetghebeur and Ryan (1990) derived a modified logrank test to compare survival in two groups while Dewanji (1992) suggested a modification of that approach. Narendranathan and Stewart (1991) described simple methods for testing various hypotheses of proportionality between the cause-specific hazards in competing risks model. Taylor (1995) studied a logistic regression with a Kaplan-Meier estimator. Goetghebeur and Ryan (1995) used PHRM to analyse competing risks survival data when failure types are missing for some individuals. Lunn and McNeil (1995) and Flehinger et al. (1998) analysed competing risks by using PHRM and the hazard function, respectively. Flehinger et al. (1996) discussed masking failure situation, whereby failure times are assumed to be irrelevant. Lam (1998) suggested distribution-free tests for the equality of k cause-specific hazard rates in a competing risks model and Chao (1998) used mixture models for fitting long-term survival data with competing risks.