DETECTING OUTLIERS AND INFLUENTIAL OBSERVATIONS IN SURVIVAL MODEL

NOR AKMAL BT. MD NOH

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2010

DETECTING OUTLIERS AND INFLUENTIAL OBSERVATIONS IN SURVIVAL MODEL

NOR AKMAL BT. MD NOH

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2010

ABSTRACT

This study proposes outlier and influential observation detection procedures for Cox proportional hazard model. In the estimation process, the parameters for Cox proportional hazard model are estimated using partial likelihood method, while the baseline hazard estimates are obtained using Nelson-Aalen method.

The procedure of outlier detection is based on three types of residuals; deviance, log-odd and normal deviate residuals. We study their properties and compare their performance in detecting outliers via simulation. On the other hand, we propose a procedure of identifying influential observation using forward search method. The method has been shown to be effective in detecting influential observations in linear regression and, more importantly, generalized linear models. The later motivates us to extend the method to Cox proportional hazard model due to their close resemblance. We compare the results with that of case-deletion method.

As for illustration, the proposed procedure is applied on the prostate cancer data and two cohorts of local breast cancer patients diagnosed at the Breast Cancer Center, University of Malaya Medical Center from year 1993 to year 2002. In both cases, the proposed procedures have successfully detected outliers and influential observations based on the best fitted Cox proportional hazard model of the full data set.

TABLE OF CONTENT

		PAGE
ABSTRACT		**i**
TABLE OF CONTENT		**ii**
ACKNOWLEDGEMENT		**v**
LIST OF TABLES		**vi**
LIST OF FIGURES		**viii**
LIST OF SYMBOL AND ABBREVIATION		**x**
1. INTRODUCTION		1
1.1	Survival Analysis	3
1.2	Survival Model	5
1.3	Outlier and Influential Observation	6
1.4	Breast Cancer	9
1.5	Problem Statement	12
1.6	Research objectives	12
1.7	Thesis Outline	13
2. LITERATURE REVIEW		15
2.1	Modeling Survival Data	15
2.2	Proportional Hazard Assumption	17
2.3	Outliers and Influential Observations	19
2.4	Forward Search Method	21
2.5	Breast Cancer	22
3. PROPORTIONAL HAZARDS MODEL		24
3.1	Estimating the PHM parameters	26
3.2	Estimating the Baseline Hazard Function of PHM	27
3.2.1	No Tied Death Time	28
3.2.2	Tied Death Time	29
3.2.3	Alternative Method	30
3.3	Variable Selection Procedure	32
4. UMMC BREAST CANCER DATA: DESCRIPTION AND SURVIVAL ANALYSIS		34
4.1	Description of Data	34
4.1.1	Data Summary	40
4.2	Survival Analysis	42
4.2.1	Survival Probability of Breast Cancer Patients	44

4.3 Modeling the Local Breast Cancer Data 51
4.3.1 Proportional Hazard Assumption 51
4.3.2 Modeling 52
4.3.3 Cox Proportional Hazard Model Analysis 55
4.4 Summary 57

5. OUTLIERS AND INFLUENTIAL OBSERVATIONS IN SURVIVAL DATA 59
5.1 Outliers Detection Procedure 60
5.1.1 Deviance Residual 64
5.1.2 Normal Deviate Residual 65
5.1.3 Log-odds Residual 66
5.1.4 Cut Point for Detecting Outliers 66
5.2 Simulation Study 67
5.2.1 Sampling Scheme 67
5.2.2 Effect of Percentage of Censoring on the Residuals Distribution 69
5.2.3 Sampling Behavior of the Maximum / Minimum Values of Residuals 75
5.3 Delete-case Method for Detecting Influential Observation 75
5.4 Analysis on the Prostate Cancer Data 81
5.4.1 Outliers Detection 82
5.4.2 Influential Observations Detection 85
5.5 Summary 87

6. FORWARD SEARCH METHOD 89
6.1 Cox PHM FS Method 90
6.1.1 Different types of Cox PHM FS Method 93
6.1.2 Measure of influential observations 94
6.2 Real Data Analysis - Prostate Cancer 95
6.2.1 Comparison of Technique for Selecting Initial Subset 96
6.2.2 Influential Observation 98
6.2.3 Discussion 102
6.3 Summary 103

7. UMMC BREAST CANCER DATA: DETECTION OF OUTLIERS AND INFLUENTIAL OBSERVATIONS 106
7.1 The First Cohort of Local Breast Cancer Data 106
7.1.1 Outliers Detection 106

7.1.2 Influential Observations Detection 109
Delete-case Method 110
Cox PHM FS Method 112
7.2 The Second Cohort of Local Breast Cancer Data 120
7.2.1 Outlier Detection 120
7.2.2 Influential Observations 124
Delete-case Method 124
Cox PHM FS Method 126
7.3 Summary 133

8. CONCLUSION 134
8.1 Summary of the Study 134
8.2 Significant of the Study 135
8.3 Further Research 135

REFERENCES 136
APPENDIX A 141
APPENDIX B 153
APPENDIX C 169

ACKNOWLEDGEMENT

Syukur Alhamdulillah this research project is made out with help of those who are very kind, enthusiastic and full of desire to see the study completed in a manner that will give better understanding on the problem of breast cancer in Malaysia.

Special thanks to Assoc. Prof. Dr. Ibrahim bin Mohamed and Assoc. Prof. Dr. Nur Aishah binti Mohd Taib for providing the opportunity to conduct this research project and giving me invaluable guidance and advice throughout my master work.

I would like to thank my family and friends, especially to my parents for their greatest support and being patient throughout my life. Their encouragement gives me the strength to complete this project.

Last but not least, to all other persons who were involved in the preparation of this report, thanks to all of you and wish you all the best of luck.

List of Figures

		Page
Figure 1.1	Breast Structure with Cancer	11
Figure 4.1	Three spheres representing the size of tumor growth	35
Figure 4.2	Normal Cells and Cancer Cells Structure	35
Figure 4.3	Estrogen receptor in the cells	36
Figure 4.4	The lymph nodes near breast area	37
Figure 4.5	Kaplan-Meier plot of overall probability for each cohort	45
Figure 4.6	Kaplan-Meier plot of variables for first cohort	46
Figure 4.7	Kaplan-Meier plot of variables for second cohort	47
Figure 5.1	Martingale residuals vs. risk score plot on generate sample 200 with no censoring	61
Figure 5.2	Deviance residuals vs. risk score plot on generate sample 200 with no censoring	62
Figure 5.3	Log-odds residuals vs. risk score plot on generate sample 200 with no censoring	63
Figure 5.4	Normal deviate residuals vs. risk score plot on generate sample 200 with no censoring	63
Figure 5.5	Empirical distribution of three suggested residuals from the last observation in each sample size in 10,000 trials without censoring	71
Figure 5.6	Empirical distribution of deviance residual from the last observation in difference ρ and n in 10,000 trials	72
Figure 5.7	Empirical distribution of normal deviate residual from the last observation in difference ρ and n in 10,000 trials	73
Figure 5.8	Empirical distribution of log-odds residual from the last observation in difference ρ and n in 10,000 trials	74
Figure 5.9	The *dfbetas* plot on *pf* factors in prostate cancer data	81
Figure 5.10	r_{Di} vs. prognostic index plot for prostate cancer data	84
Figure 5.11	r_{L_i} vs. prognostic index plot for prostate cancer data	84
Figure 5.12	r_{N_i} vs. prognostic index plot for prostate cancer data	85
Figure 5.13	Delta-betas for each factor of prostate cancer data	86
Figure 6.1	Progression plot on *wt* factor of prostate cancer patients in FS method	93
Figure 6.2	*IM* plot on *wt* factor of prostate cancer data	95
Figure 6.3	Progression plot on prostate cancer data using FS1 method	99

Figure 6.4	IM plot on prostate cancer data using FS1 method	100
Figure 7.1	r_{Di} vs. prognostic index plot for first cohort	107
Figure 7.2	r_{N_i} vs. prognostic index plot for the first cohort	108
Figure 7.3	r_{L_i} vs. prognostic index plot for the first cohort	109
Figure 7.4	*dfbetas* residuals plot for first cohort	111
Figure 7.5	The progression plots for FS1 procedures on the first cohort of local breast cancer data	115
Figure 7.6	The *IM* plots for FS1 procedures on the first local breast cancer data	116
Figure 7.7	r_{Di} vs. prognostic index plot for second cohort	121
Figure 7.8	r_{Ni} vs. prognostic index plot for second cohort	123
Figure 7.9	r_{Li} vs. prognostic index plot for second cohort	124
Figure 7.10	*dfbetas* residuals plot for second cohort	125
Figure 7.11	The progression plots for FS1 procedures on the second local breast cancer data	129
Figure 7.12	The *IM* plots for FS1 procedures on the second local breast cancer data	130

List of Tables

		Page
Table 4.1	Description of data	41
Table 4.2	Number of patients with breast cancer in two cohorts	42
Table 4.3	Five-year probability of overall survival	44
Table 4.4	Five-year probability of survival for both cohorts	49
Table 4.5	Cox's time dependant covariate test for ER, size and age factors in the first cohort	52
Table 4.6	Variables selection on first cohort	53
Table 4.7	Variables selection in second cohort	54
Table 4.8	Cox PHM result for both cohort respect to significance variables	56
Table 5.1	Percentile of min and max of residuals in 10,000 trials without censoring	76
Table 5.2	Percentile of min and max of residuals in 10,000 trials with 20% censoring	77
Table 5.3	Percentile of min and max of residuals in 10,000 trials with 40% censoring	78
Table 5.4	Percentile of min and max of residuals in 10,000 trials with 60% censoring	79
Table 5.5	Descriptive analysis of prostate cancer data	82
Table 5.6	Profile of individuals defined as outlier identified by deviance residual	83
Table 5.7	Profile of outliers detected by normal deviate and log-odd residuals	85
Table 5.8	Observations considered as influential observations using delta-betas method	87
Table 5.9	Cox PHM result	88
Table 6.1	Techniques in forming the initial subset in the Cox PHM FS method	91
Table 6.2	Combination of procedures in Cox PHM FS method	94
Table 6.3	Proportion of patients selected by seven different initial subset techniques, a) similar proportion b) non-similar proportion	97
Table 6.4	Patients selected as Influential observation by seven different FS procedures	101
Table 6.5	Patients identified as influential observation by different γ values on FS method	102