UNIVERSITI PUTRA MALAYSIA

A MODEL FOR BINARY RESPONSE VARIABLE WITH TIME-CENSORED OBSERVATIONS

HANAN HASSAN ALI ADLAN

FSAS 2000 2

A MODEL FOR BINARY RESPONSE VARIABLE WITH TIME-CENSORED OBSERVATIONS

By

HANAN HASSAN ALI ADLAN

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

May 2000

To my Husband, ……

Husam & Aala

To

my father, mother, brothers, and sisters

♥ with love ♥

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

A MODEL FOR BINARY RESPONSE VARIABLE WITH TIME-CENSORED OBSERVATIONS

By

HANAN H. A. ADLAN

May 2000

Chairman : Assoc. Prof. Dr. Isa Daud

Faculty : Science and Environmental Studies

The work in this thesis is concerned with the investigation and development of a new model for analyzing censored binary observations. Binary data is always expressed in the form 0 and 1. In other words, the response variable can only take on the value 0 or 1, which represent the non-response or response of the variable under consideration. In the case of grouped binary data, the response is expressed in the form of proportions. In certain follow-up studies, especially during the study period, some variables may result in censored observations. This becomes a problem especially in the case of ungrouped binary observations, where 0 is assigned to the non-responding individuals, or those who have got a chance to respond but not within the study period. This gives rise to the probit-exponential model.

As a basis, the probit model is used to analyze grouped binary data. It is also used in comparison with the suggested probit-exponential model in the case of ungrouped data. Comparison with the logistic fit is performed for the model with grouped data. The model diagnostic is done through the Pearson residuals, and standardized residuals.

Newton-Raphson algorithm is used to find the parameter estimates. A half normal plot is employed to investigate the usefulness of the two models, and to compare the suitability of the models when the observations are censored. Validity of the probit-exponential assumptions is also investigated through simulation.

The work above is carried out within the generalized linear model framework, which accommodate the new model as a member of this family.

Abstrak yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Master Sains.

MODEL UNTUK PEMBOLEHUBAH RESPON DEDUA DENGAN CERAPAN TAPISAN-MASA

Oleh

HANAN H. A. ADLAN

May 2000

Pengerusi : Profesor Madya Dr. Isa Daud

Fakulti : Sains dan Pengajian Alam Sektiar

Skop dalam thesis ini tertumpu kepada siasasatan dan perkembangan suatu model baru untuk menganalisis data tertapis bagi cerapan dedua. Data dedua lazim dinyatakan dalam bentuk 0 dan 1. Dengan kata lain, pembolehubah respon hanya boleh mengambil nilai 0 atau 1, yang mewakili tiada sambutan atau ada sambutan dari pembolehubah yang dipertimbangkan. Dalam kes data dedua terkumpul, respon dinyatakan dalam bentuk kadaran. Dalam beberapa kajian susulan tertentu, terutamanya dalam tempoh jangkamasa kajian, sebahagian daripada pembolehubah mungkin menghasilkan cerapan tertapis. Ini menimbulkan masalah terutamanya

dalan kes cerapan dedua tak terkumpul, dengan 0 diambil kira sebagai tiada sambutan.

Sebagai asas, model probit digunakan untuk menganalisis data dedua terkumpul. Ia juga digunakan sebagai perbandingan dengan model probit-ekspoen tersaran bagi kes data yang tidak terkumpul. Perbandingan dengan cucukan logit dilakukan untuk model dari data terkumpul. Dignostik model dijalankan melalui reja Pearson dan reja dipiawaikan.

Untuk mencari penganggar parameter kami tumpukan kepada pengunaan algorithma Newton-Raphson. Plot normal separuh digunakan untuk menyiasat kebaikan diantara kedua model dan untuk membandingkan kesesuaian model apabila cerapan adalah tertapis.

Kesahihan andaian probit-eksponen juga disiasat melalui simulasi. Kesemua kerja diatas dijalankan dalam kerangka model linear teritlak yang mana model baru ini tersirat sebagai ahli famili model.

ACKNOWLEDGEMENTS

Firstly and foremost, praise be to God, for giving me the strength and patience to complete this work.

I am deeply indebted to my supervisor Assoc. Prof. Dr. Isa Daud for his invaluable guidance, assistance, suggestions, helpful discussions and continuous encouragement during this work.

My thanks also to the members of my supervisory committee, Assoc. Prof. Dr. Mat Yusoff Abdullah, and Dr. Noor Akma Ibrahim.

Thanks are also due to Assoc. Prof. Dr Harun Budin, for allocating me a very comfortable place, when he was the head of the department.

Finally I would like to thank University of Khartoum for granting me study leave, and the government of Sudan for the award of scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GLM	generalized linear models
GLIM4	generalized linear interactive modelling
pdf	probability density function
iid	independent and identically distributed
df	degrees of freedom
LF	likelihood function
ML	maximum likelihood
CHD	coronary heart disease
r.v.	random variable
BAN	best asymptotically normal

CHAPTER I

INTRODUCTION

General Overview

Modelling in Statistics in the recent years becomes the key for building systems in all aspects of sciences. The models are built and checked using modern methods for computing parameter estimates, and estimating errors. Fitting a model to data set enables the understanding of the essential features of the data. Of course, the model fitted will not completely describe reality but it can be thought of as an approximation to the true situation. The so called asymptotic theory is employed in looking for optimal approximate solutions. This makes the study of the analysis become the most important area of research for many years. Also the emergence of the analysis of incomplete data sets refreshes this area, especially the analysis of censored data which has become a popular field in recent researches. Old methods for building models are improved and re-evaluated so that it can deal with censoring.

In this thesis, our concern is to develop and build a probit-exponential regression model for censored data, although the work on the probit model has an older application and a lot of work has been done on it, but its application to censored data taking into account the time factor is of new interest. The problem is to usc the transformed probits in building a survival model to censored observation, in

order to implement a method for analysing binary variables within the generalised linear models framework.

Chapter Overview

In this chapter we highlight some background for the scope of the research. Some important key words are defined. General description of the formulation of the problem, and the purpose of the research are outlined.

Key Words and Definitions

Statistical Modelling

The term statistical modelling referred to modelling observations of characteristics of individuals or groups, which are assumed to have random elements.

The set of observations will be a sample representing some population. The variable characteristics are assumed to respond to mechanisms involving the various circumstances of the different individuals. This is referred to as the response variables. The aim of modelling is to explain how the variation in the observed values of the response variable is explained by differences in circumstances between individual cases or groups of cases. These differences may result from nature (gender, age), nurture (social environment, education), exposure (diet, pollution, media) or treatment (drug, fertilisers, educational techniques), and are termed explanatory variables.

GLM as a single class. GLM has been introduced by Nelder and Wedderburn (1972) as a unifying family of models for non-normal cross-sectional regression analysis with non-normal responses. Their further development had a major influence on statistical modelling in a wider sense (Fahrmeir and Tutz ,1994; Francis , Green, and Payne,1994).

Binary Data

In many areas of application of statistical principles and procedures, one encounters observations made on individual experimental units that take one of two possible forms. For example, a patient in a clinical trial to compare alternative forms of treatment may or may not experience relief from symptoms; an insect in an insecticidal trial may survive or die when exposed to a particular dose of an insecticide; a seed may germinate or fail to germinate under certain experimental conditions; an integrated circuit may be defective or non-defective. Such data are said to be binary. The two possible forms are always referred to as success and failure.

The term grouped binary data refer to the number of successes out of the total number of units exposed to a particular set of experimental conditions. Binary data are assumed to have Bernoulli distribution, while data in the form of proportions can be modelled using the binomial distribution. Special form of regression model is required for analyzing such data. For example, an experiment in which it is noted that there is an increasing proportions with increasing dose leads to an S-shaped

curve relating the proportions to the dose. It would be possible to transform the response variable and fit the resulting variable using one of the statistical techniques.

Modelling of proportions involves the modern method of Probit analysis (Bliss, 1935) remains popular in this field of applications, also the logistic regression (Cox 1972) is of wider application (D.Collett, 1991; Francis, Green, and Payne, 1994) .

Consider the case of ungrouped binary responses, coded by zero and one. Given the covariate x, a binary variable Y is completely determined by its response probability

$$E(\boldsymbol{Y} / \boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{Y} = 1 / \boldsymbol{x}) = \pi \tag{1.1}$$

implying $$\text{var}(\boldsymbol{Y} / \boldsymbol{x}) = \pi(1 - \pi), \tag{1.2}$$

for the case of grouped data. Let $\bar{y}$ denotes the relative frequency of observed 1's for the say m independent binary observations – with the same covariate vector $\boldsymbol{x}$. The absolute frequencies $m\bar{y}$ are binomially distributed with

$$E(m\bar{\boldsymbol{Y}} / \boldsymbol{x}) = m\pi \tag{1.3}$$

$$\text{var}(m\bar{\boldsymbol{Y}} / \boldsymbol{x}) = m\pi(1 - \pi) . \tag{1.4}$$

The relative frequencies are scaled binomial, i.e. they take the corresponding values 0,1/m, 2/m, …,1 with the same binomial probabilities as $m\bar{y}$ and

$$E(\bar{\boldsymbol{Y}} / \boldsymbol{x}) = \pi . \tag{1.5}$$

$$\text{var}(\boldsymbol{Y} / \boldsymbol{x}) = \frac{\pi(1 - \pi)}{m} . \tag{1.6}$$

Thus for grouped binary data, response functions are the same as for individual binary responses, with the variance function has to be divided by m i.e. $w = m$. When

the individual response (given $\boldsymbol{x}$) is binomially distributed with $Y \sim B(m,\pi)$, the scaled binomial or relative frequency y/m will be considered as response, since y/m can act as an average of individual independent binary observations, and can be treated as occurring from grouped observations. Then the variance

$$\text{var}(Y / \boldsymbol{x}) = \frac{\pi(1-\pi)}{m} \tag{1.7}$$

is the same as for grouped binary observations.

Models for binary and binomial responses are determined by relating the response probability π to a linear predictor $\eta = \beta X$ via some response function $\pi = h(\eta)$, link function $g(\pi) = \eta$. The most common models in this field are linear probability model, the probit, and the logit model.

Our concern in this thesis based on implementation of the probit model, which can be defined by $\pi = \Phi(\eta) = \Phi(\beta X)$ where Φ is the standard normal distribution function, no restrictions on η.

The Logit model corresponds to the natural link function

$$g(\pi) = \log(\frac{\pi}{1-\pi}) = \eta \tag{1.8}$$

with the logistic distribution function $\pi = h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, (1.9)

as the resulting response function. The logistic distribution function also has support on the entire real axis and is symmetric, but it has some heavier tails than the standard normal. Apart from π values near zero or one which correspond to the tails, fits by probit or logit models are generally quite similar (Fahrmeir and Tutz, 1994).

Survival Analysis:

Survival analysis is concerned with studying the time between entry to a study and a subsequent event. Originally the analysis was concerned with time from treatment until death, hence the name, but survival analysis is applicable to many areas as well as mortality. Recent examples include time to discontinuation of a contraceptive, maximum dose of bronchoconstrictor required to reduce a patient's lung function to 80% of baseline, time taken to exercise to maximum tolerance, time that a transdermal patch can be left in place, time for a leg fracture to heal.

When the outcome of a study is the time between one event and another, a number of problems can occur.

1.The times are most unlikely to be Normally distributed.

2.We cannot afford to wait until events have happened to all the subjects, for example until all are dead. Some patients might have left the study early - they are lost to follow up . Thus the only information we have about some patients is that they were still alive at the last follow up. These are termed censored observations .

Let the survival time, T for individuals in a population have density function $f(t)$. The corresponding distribution function $F(t) = \int_{-\infty}^{t} f(s)ds$ is the function of the population dying by time t. The complementary function 1- $F(t)$ (survivor function) is the fraction still surviving at time t. The hazard function $\lambda(t)$ measures the instantaneous risk, and it means that $\lambda(t)\delta_t$ is the probability of dying in the next small interval δ_t given survived to time t. These functions can be expressed as

$$\Pr(\text{survival to } t + \delta_t) = \Pr(\text{survival to } t)\ \Pr(\text{survival for } \delta_t/\text{survival to } t)$$

$$1 - F(t + \delta_t) = (1 - F(t))(1-\lambda(t)\delta_t) \qquad (1.10)$$

hence $\delta_t F'(t) = (1-F(t))\lambda(t)\delta_t$. and we have $\lambda(t) = f(t)/(1-F(t))$ (1.11)

(Cox, and Oakes, 1984; Miller,1981; McCullagh, and Nelder, 1989).

Censoring

Censoring represent a mechanism which is considered a special feature of survival data. It occurs when some subjects in the study have not experienced the event of interest. At the end of the study, the exact survival times of these subjects are unknown. And they are referred to as censored observations or censored times. In other words in an experiment in which subjects are followed over time until the event of interest (such as death or other type of failure) occurs, it is not always possible to follow every subject until the event is observed. Subjects may drop out of the study and be lost to follow-up, or be deliberately withdrawn, or the end of the data collection period may arrive before the event is observed to happen. For such a subject, all that is known is that the time to the event is at least as long as the time to when the subject was last observed. The observed time to the event under such circumstances is censored. As mentioned before censoring may occur from the right that is the observation stops before the event is observed, as in censorship for survival analysis, or from the left when the observations does not begin until after the event has occurred.

The types of right censoring

1- type I censoring

Subjects are entered the study for a fixed period in the calendar. At the end of the study, subjects that don't experience the event of interest or lost to follow-up are censored.