行政院國家科學委員會專題研究計畫成果報告

計畫名稱(中文): 重複事件資料之邊際迴歸分析

(英文): Marginal Linear Regression Analysis for Recurrent Event Data

計畫類別: 個別型
計畫編號: NSC89-2118-M002-003
執行期限: 88年8月1日至89年7月31日

主持人: 張淑惠副教授
台大公共衛生學院公共衛生學系
Email: shuhui@ccms.ntu.edu.tw

中文摘要

重複事件資料是在長期觀察研究經常被收集到的一種資料型態。在不同的領域包括生物醫學, 工業工程, 人口學, 經濟學, 及其他領域等經常出現此類資料。當分析事件重複發生的資料時，則兩連續事件的間隔時間的長短與其他因素的關係是此類資料的研究目的之一。因此本研究有興趣的結果變數是兩個連續事件的間隔時間且所考慮的模式是一半母數之邊際線性迴歸模式。利用重複事件依序發生的特性探討在適當的假設下對所考慮的線性模式發展適當的迴歸參數的估計方法。並利用模擬研究與實例來說明此迴歸參數的估計方法。

關鍵詞：加速失敗時間模式，對數等級統計量，半母數模式

MARGINAL LINEAR REGRESSION ANALYSIS FOR RECURRENT EVENT DATA

Shu-Hui Chang
Department of Public Health College of Public Health National Taiwan University
1 Jen-Ai Road, Section 1, Taipei 10018, Taiwan
(shuhui@ccms.ntu.edu.tw)

Key Words: estimating equation, log-rank statistic, recurrent event times.

Abstract:

Recurrent event data are commonly encountered in longitudinal studies when failure events can occur repeatedly over time for each study subject. The aim of this article is to examine the covariate effects on the time interval between two successive events. Consider semiparametric linear regression models on the time interval between two successive events provided that the logarithm of the time interval of interest is linearly related to its covariates without specifying the joint distribution of the observations within a subject. From the ordinal nature of recurrent events, estimating procedures for the covariate effects based on a population-averaged model approach can be developed. Examples are conducted to illustrate the estimating methods studied in this article.

1. Introduction

Recurrent event data are common in biomedical research. For example, carcinogenesis experiments may result in the appearance of multiple tumors in each animal from the day of injection. In a longitudinal study, epilepsy patients were followed since their onset of seizure and they may suffer the recurrences of the illness during the study period. Many other examples include asthma attacks, bladder cancer and infections in AIDS patients. In the examples, the scientific interests may center on the effects of covariates on the risk of occurrence of the events. In the literature, various statistical regression methods have been developed for recurrent event data. Conditional regression analysis using semiparametric hazards models for times to events and times between successive events is developed by Prentice, Williams, and Peterson (1981). Wei, Lin and Weissfeld (1989) and Pepe and Cai (1993) developed marginal and semi-conditional hazards regression models for the times to events. Intensity models and recurrent rate models are considered by Anderson and Gill (1982), Lawless (1987), Lawless and Nadeau (1995).

Alternatively to the above proportional risk models, semiparametric linear regression models are considered in this paper. Lin and Wei (1992) considered the linear regression models on the logarithms of the times to multiple events. However, the times of interest in the paper are the times between two successive events. Specifically, the logarithm of the times between two successive events is linearly related to its covariates and the corresponding covariate effects in the model are assumed to be the same for each episode of events. The aim of this paper is to consider the estimation approaches for the marginal covariate effects without specifying the joint distribution of the multiple times of interest. First, assume that for each subject the random errors in the model are exchangeable and have the same unspecified marginal distribution function. Then, a weighted log-rank type estimating function based on the observed errors is considered and this estimation method is an extension of the one-sample estimation approach proposed by Wang and Chang (1999). However, the distributions of the random errors may not be the same for different episodes of events. Under this situation, we consider a stratified log-rank estimating function for the marginal covariate effects, where the stratification variable is the episodes of the events. The stratified estimating approach is the same as the log-rank estimating function for the times to multiple events proposed by Lin and Wei (1992). The model assumptions and estimation procedures for the covariate effects will be presented in section 2. In this section, the asymptotic properties for the corresponding estimators of the marginal covariate effects are also discussed. In section 3, a simulation study and analysis of a real data are conducted to illustrate the performance of the estimators considered in the previous section.

This work is supported by NSC 89-22118-M-002-003.

2. Model and Estimation

There are n independent subjects in the study, set $T_{i0} = 0$, which is the time of the initial event for subject i in a longitudinal study or the starting time in an animal experiment or a clinical trial, $i = 1, 2, \ldots, n$. Let the random variables T_{ij} represent the times between the $(j-1)$st and jth recurrences for $j = 1, 2, \ldots$ and C_i the time to independent censoring, i.e. the end of the recurrence process. Let k_i be the number of recurrences for subject i under observation, that is,

$$\sum_{j=1}^{k_i} T_{ij} \leq C_i \text{ and } \sum_{j=1}^{k_i+1} T_{ij} > C_i.$$

The bounded covariate vector for subject i is denoted by Z_i for $i = 1, 2, \ldots, n$. The censoring time C_i is assumed to be conditionally independent on $(T_{i1}, T_{i2}, \ldots)$ given Z_i. Under censoring, the observed data consists of $\{z_i, x_{ij}, \delta_{ij}, j = 1, 2, \ldots, k_i + 1\}$ for $i = 1, 2, \ldots, n$, where $x_{ij} = min(0, t_{ij}, c_i - \sum_{\ell=1}^{j} t_{i\ell})$ and $\delta_{ij} = I(\sum_{\ell=1}^{j} t_{i\ell} \leq c_i)$, where $I(\cdot)$ is the indicator function. Note that if T_{ij} is censored then $T_{i\ell}$ for $\ell \geq j+1$ will also be censored because the events of interest are ordered, that is, if $\delta_{ij} = 0$ then $\delta_{i\ell} = 0$ for $\ell \geq j+1$.

Consider the times between two successive events, the T_{ij}'s, follow a linear regression model so that

$$\log T_{ij} = \beta' Z_i + \epsilon_{ij}, j = 1, 2, \ldots; i = 1, 2, \ldots, n$$

where β is a $p \times 1$ vector of covariate effects. Assume that for each i, the random errors $\epsilon_{i1}, \epsilon_{i2}, \ldots$ are exchangeable and have a common marginal distribution function F, which is an unknown function. Then, the vectors $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots)'$, $i = 1, 2, \ldots, n$, are independently identically distributed. Note that the survival function of ϵ_{ij}, $1 - F(\cdot)$, can be estimated by the weighted Kaplan-Meier estimator proposed by Wang and Chang (1999). To obtain an estimating function for β based on the observed errors , $e_{ij} = \log x_{ij} - \beta' z_i$ and δ_{ij} for $j = 1, 2, \ldots, k_i + 1$ and $i = 1, 2, \ldots, n$, consider the weighted log-rank statistic

$$U_{1n}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{k_i^*} \sum_{j=1}^{k_i^*} \delta_{ij} \left(z_i - \frac{S_1(\beta, e_{ij})}{S_0(\beta, e_{ij})} \right), \quad (1)$$

where $k_i^* = 1$ if $k_i = 0$ and $k_i^* = k_i$ if $k_i \geq 1$;

$$S_0(\beta, e_{ij}) = \sum_{\ell=1}^{n} \frac{1}{k_\ell^*} \sum_{m=1}^{k_\ell^*} I(e_{\ell m} \geq e_{ij})/n \text{ and}$$

$$S_1(\beta, e_{ij}) = \sum_{\ell=1}^{n} \frac{1}{k_\ell^*} \sum_{m=1}^{k_\ell^*} z_\ell I(e_{\ell m} \geq e_{ij})/n.$$

The estimate of β, denoted by $\hat{\beta}_1$, can be derived by solving $U_{1n}(\beta) = 0$. The asymptotic normality of $U_{1n}(\beta)$ and $\hat{\beta}_1$ can be established as discussed in the appendix of Wang and Chang (1999).

Suppose that the vectors $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots,)'$, $i = 1, 2, \ldots, n$, are be independently identically distributed, but the random errors $\epsilon_{i1}, \epsilon_{i2}, \ldots$ do not have a common marginal distribution function. Then, the above estimating functions, $U_{1n}(\beta)$, are not applicable to this model assumption. Under this alternative model assumption, consider the transformations of the observed data, $e_{ij}^* = \min\{log(\sum_{\ell=1}^{j} T_{i\ell}) - \beta' z_i, \log C_i - \beta' z_i\}$ and $\delta_{ij}^* = I(log(\sum_{\ell=1}^{j} T_{i\ell}) - \beta' z_i, \log C_i - \beta' z_i)$ and then the corresponding estimating function of β is

$$U_{2n}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{k_i} \delta_{ij}^* \left(z_i - \frac{\sum_{\ell=1}^{n} z_\ell I(e_{\ell j}^* \geq e_{ij}^*)}{\sum_{\ell=1}^{n} I(e_{\ell j}^* \geq e_{ij}^*)} \right). \quad (2)$$

Let $\hat{\beta}_2$ be the estimate of β derived by $U_{2n}(\beta) = 0$. Note that the estimating function (2) is same as the proposed estimating function of Lin & Wei (1992) considering the linear model on the logarithm of the times to the multiple events. Lin and Wei (1992) have shown that $U_{2n}(\beta)$ weakly converges to a multivariate normal distribution and $\sqrt{n}(\hat{\beta}_2 - \beta)$ is asymptotically normally distributed (Ying, 1993).

In addition, the asymptotic covariance matrices for $\hat{\beta}_1$ and $\hat{\beta}_2$ are difficult to estimate directly. The corresponding confidence intervals can be obtained by test-based statistics considered by Wei & Gail (1983) and Wei, Ying & Lin (1990)) and the resampling method developed by Parzen, Wei & Ying (1994).

3. Simulation

A simulation study is conducted to illustrate the estimating methods considered in section 2. In the simulation, consider n subjects randomly assigned into two groups. Let $a_1, a_2, \ldots, a_n$ be the frailty values from the Gamma distribution with a unit mean and variance α, denoted by $Gamma(1/\alpha, \alpha)$. Given a_i, the random errors $(\epsilon_{i1}, \epsilon_{i2}, \ldots)$, for $i = 1$ to n, are generated from the Weibull distribution with the survival function $exp\{-a_i t^2\}$. Suppose that the true covariate effect $\beta = 2$ and the fixed censoring times are equal to 3 in both groups. Then, subjects in the group with longer expected time between two successive events are heavily censored. For comparison, we consider two naive logrank-type estimating functions of β: one is based on the (e_{i1}, δ_{i1})'s only and another one is to use all the e_{ij}'s and δ_{ij}'s with the same weights as the usual independence case.

Table 1 gave the simulation results based on 1000 replicates of samples generated from the above simulation procedures. The displays in table 1 includes the mean estimates of β and the corresponding standard deviations for these estimating methods. Both estimators, $\hat{\beta}_1$ and $\hat{\beta}_2$ are more precious than the naive estimator based on the (e_{i1}, δ_{i1})'s only. The magnitude of the bias of the naive estimators by combining all the (e_{ij}, δ_{ij})'s equally increases when the correlation among recurrence times is stronger. Note that in the simulation study the average number of events for the group with longer expected times is less than 0.2 and the results show that $\hat{\beta}_1$ is more efficient than $\hat{\beta}_2$.

4. Data Anaylsis for Tumor Data

The example is an animal experimental data set given in Table1 1 of Gail, Snatner and Brown (1980). In the experiment, 76 rats were injected with a carcinogen for mammary cancer at day zero, and then all animals were given retinyl acetate to prevent cancer for sixty days. After sixty days, the 48 animals which remained tumor-free were randomly assigned to continued retinoid prophylaxis (treatment group) or control (control group). Rats were palpated for tumors twice weekly and observation ended 182 days after the initial carcinogen injection. 20 of 23 rats in the treatment group and all 25 rats in the control group have at least one tumor occurring in the experiment.

The purpose of conducting the animal experiment is to study the common treatment effect of the log-linear model on the times between two successive tumors. Table 2 displays the four estimates of the treatment effect as considered in the above simulation study. In table 2, the test-based confidence intervals for the treatment effect are obtained by using the asymptotic properties of the corresponding estimating statistics developed by Wei & Gail (1983) and Wei, Ying & Lin (1990). Based on the results in table 2, one may conclude that on average the time between two successive tumors for the treatment group is statistically significantly (0.3 to 0.8 times) longer than that for the control group. However, the naive estimate using all the (e_{ij}, δ_{ij})'s may overemphasize the treatment effect on prolonging the length between two successive tumors. Note that for th tumor data the assumption of the identically distributed random errors for each rat may be suspected since the 95 % confidence interval based on the estimating function (1) is much wider than those from the other unbiased estimating methods.

Reference

1. Anderson, P. K., and Gill, R. D. 'Cox's regression model for counting processes: A large sample study'. *The Annals of Statistics*, **10**, 1100-1120 (1982).
2. Gail, M.H., Snatner, T.J. and Brown, C.C. 'An analysis of comparative carcinogenesis experiements based on multiple times to tumor'. *Biometrics*, **36**, 255-266 (1980).
3. Lawless, J. F. 'Regression methods for Poisson process data'. *Journal of the American Statistical Association*, **82**, 808-815 (1987).
4. Lawless, J. F. and Nadeau, C. 'Some simple robust methods for the analysis of recurrent events'. *Technometrics*, **37**, 158-168 (1995).
5. Lin J. S. and Wei L. J. 'Linear regression analysis for multivariate failure time observations'. *Journal of the American Statistical Association*, **87**, 1091-1097 (1992).
6. Parzen, M. I., Wei, L. J. and Ying, Z. 'A resampling method based on pivotal estimating functions'. *Biometrika*, **81**, 341-350 (1994).
7. Pepe, M.S. and Cai, J. 'Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates'. *Journal of American Statistical Association*, **88**, 811-820 (1993).
8. Prentice, R. L., Williams, B. J., and Peterson, A. V. 'On the regression analysis of multivariate failure time data', *Biometrika*, **68**, 373-379 (1981).
9. Wei, L. J., Lin, D. Y., and Weissfeld, L. 'Regression Analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065-1073 (1989).
10. Wang, M-C and Chang, S-H. 'Nonparametric estimation of a recurrent survival function'. *Journal of the American Statistical Association*, **94**, 146-153 (1999).
11. Wei, L.J. and Gail, M.H. 'Nonparametric estimation for a scale-change with censored observations'. *J. Am. Statist. Assoc.* **78**, 382-388 (1983).
12. Wei, L. J., Ying, Z. and Lin, D. Y. 'Linear regression analysis of censored survival data based on rank tests'. *Biometrika*, **77**, 845-851 (1990).

Table 1: Simulation

Methods	Frailty distributions	
	$Gamma(1/2, 2)$	$Gamma(1/4, 4)$
	estimate (s.d.)	estimate (s.d.)
Naive (1st time)	2.013 (0.247)	2.020 (0.285)
Method (1)	2.011 (0.202)	2.013 (0.237)
Method (2)	2.015 (0.238)	2.021 (0.257)
Naive (all times)	2.230 (0.362)	2.482 (0.535)

Table 2: Analysis of Tumor Data

Methods	estimate	95% confidence interval
Naive (1st time)	0.262	(0.050, 0.511)
Method (1)	0.579	(0.214, 1.176)
Method (2)	0.373	(0.306, 0.486)
Naive (all times)	0.847	(0.510, 1.292)

13. Ying, Z. 'A large sample study of rank estimation for censored regression data'. *Ann. Statist.* **21**, 76-99 (1993) .