KYBERNETIKA — VOLUME *51* (2015), NUMBER 2, PAGES 268–275

ADDITIVE HAZARDS REGRESSION WITH CASE-COHORT SAMPLED CURRENT STATUS DATA

Wei Chen, Fengling Ren and Guosheng Tang

In a case-cohort design, covariate histories are measured only on cases and a subcohort that is randomly selected from the entire cohort. This design has been widely used in large epidemiologic studies, especially when the exposures of interest are expensive to assemble for all the subjects. In this paper, we propose statistical procedures for analyzing case-cohort sampled current status data under the additive hazards model. Asymptotical properties of the proposed estimator are described and we suggest a resampling method to estimate the variances. Simulation studies show that the proposed method works well for finite sample sizes, and one data set is analyzed for illustrative purposes.

Keywords: additive hazards model, case-cohort, current status data, estimating equations, simple random sampling

Classification: 62N02, 62N01

1. INTRODUCTION

The case-cohort design, originally proposed by Prentice [18], is a cost-effective means in large epidemiological studies and disease prevention trials in which the outcome of interest is time to event and covariates are measured only on the cases and a subsample selected from the entire cohort. The reduction of cost offered by this design is specially prominent if the assembly of covariate histories may be prohibitively expensive on the entire cohort members or the disease of interest occurs infrequently.

The statistical inference method for analyzing the case-cohort data has been well studied by many authors. Prentice [18] proposed a "pseudolikelihood" procedure for the relative risk model, and later some asymptotic results for this approach were established by Self and Prentice [19]. Under the additive hazards (AH) model, Kulich and Lin [8] constructed an estimating equation and showed that their estimator was consistent and asymptotically normal. Recently, Lu and Tsiatis [14] and Kong et al. [6] studied the transformation model, Nan et al. [16] and Kong and Cai [7] studied the accelerated failure time model. Note that all of these methods mentioned above are developed to tackle right censored case-cohort data. To the best knowledge of ours, there are few studies focusing on interval censored case-cohort data. Gilbert et al. [3] directly approximated the interval censored case-cohort data by the right censored version and then the

DOI: 10.14736/kyb-2015-2-0268

existing approaches for analyzing right censored case-cohort data can be employed. Ma [15] investigated the AH model with case-cohort sampled current status data. Under the Cox model, Li et al. [9] considered the interval censoring mechanism in case-cohort studies, but presumed that the inspection time intervals are fixed. Recently, Li and Nan [10] addressed the relative risk regression for current status data in the case-cohort design.

In this paper, we consider the additive hazards model with case-cohort sampled current status data. Current status data, also called the "case 1" interval censored data, arise in studies in which the time of occurrence of some event is of interest, but we only know whether the event has occurred or not at the time the sample is collected. For a detailed discussion, see Groeneboom and Wellner [4], and Sun [20]. As said before, the additive hazards model for the current status data in the case-cohort design has been studied by Ma [15], where partly motivated by the work of Chen and Lo [2], he proposed a class of estimating equations for estimating the regression parameters. It should be specially pointed out that his estimating functions cannot be calculated in general, because they involve some unobservable covariates, as presented in the next section. Therefore, it is necessary to develop a feasible approach to make inference for parameters in the AH model.

Motivated by the construction of estimating equations in Kulich and Lin [8] for analyzing the right censored case-cohort data, we propose in Section 2 an estimating function for estimating regression parameters in the additive hazards model with current status data under the simple random sampling without replacement. In Section 3, we conduct simulations to assess the finite sample performance of the proposed estimator. A data from the ED_{01} experiment is analyzed to illustrate our proposed method in Section 4. A brief discussion concludes this paper.

2. ESTIMATION

Let T and C denote the failure time and censoring time, and $Z(\cdot)$ be the d-vector of possibly time-dependent covariates. Suppose that the failure time T follows the additive hazards model (Lin and Ying [8], Lin et al. [11]) which assumes that the hazard function of T at time t associated with the history of $Z(\cdot)$ up to t is given by

$$\lambda(t \mid Z(\cdot)) = \lambda_0(t) + \beta_0' Z(t), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β_0 is an unknown d-vector of regression parameter. The additive hazards model is a useful alternative to the Cox model, and convenient when the absolute effects of covariates on the hazard function are of interest. We assume that C is independent of T and $Z(\cdot)$, which is also adopted in Ma [15].

For current status data, the exact value of T is never observed, rather one only can observe $O = (C, \delta, Z(\cdot))$, where $\delta = I(C \leq T)$ is the censoring indicator. Let $(T_i, C_i, Z_i(\cdot))$ $(i = 1, \ldots, n)$ be independent copies of $(T, C, Z(\cdot))$. Then the observed data consist of $(C_i, \delta_i, Z_i(\cdot))$, which are i.i.d. copies of O. Under the case-cohort design, covariate histories are available only on cases, i. e., those with $\delta_i = 0$ (Ma [15, p. 598]), and a random subset of the entire cohort, i. e., the subcohort. Let ξ_i be the subcohort

indicator, taking value 1 if the ith subject is in the subcohort and 0 otherwise. Suppose that we select the subcohort of size $\tilde{n}$ by simply random sampling without replacement from the entire cohort. Denote $p_n = \tilde{n}/n$ as the subcohort proportion, which converges to $p \in (0,1)$ as $n \to \infty$.

2.1. Ma [15] method

In a full cohort study, Lin et al. [11] proposed to estimate β_0 in model (1) with randomly sampled current status data by solving $U(\beta) = 0$, denote the solution by $\hat{\beta}$, where

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i^*(t) - \frac{S^1(\beta,t)}{S^0(\beta,t)} \right\} \mathrm{d}N_i(t), \tag{2}$$

where $Z_i^*(t) = \int_0^t Z_i(s)\,\mathrm{d}s$, $S^k(\beta,t) = 1/n \sum_{j=1}^n Y_j(t)\exp(-\beta' Z_j^*(t)) Z_j^*(t)^{\otimes k} (k = 0,1)$, $a^{\otimes 0} = 1, a^{\otimes 1} = a$, and $Y_i(t) = I(C_i \geq t)$, $N_i(t) = \delta_i I(C_i \leq t)$.

Before introducing Ma [15] method, we first give some notations needed herein. Define $\tilde{R}_t^1$ and R_t^0 as the index sets of subjects still at risk at time t in the subcohort with $\delta_i = 1$ and in the entire cohort with $\delta_i = 0$, respectively. Due to the fact that the covariate histories $Z_i(t)$'s are not available for each subject, calculation of $S^k(\beta,t)$ in (2) becomes impossible. Note that $S^1(\beta,t)/S^0(\beta,t)$ is a consistent estimator of

$$\mathrm{E}[Z^*(t) \mid C = t, \delta = 1] =$$
$$\frac{q\mathrm{E}[Y(t)\exp(-\beta' Z^*(t)) Z^*(t) \mid \delta = 1] + (1-q)\mathrm{E}[Y(t)\exp(-\beta' Z^*(t)) Z^*(t) \mid \delta = 0]}{q\mathrm{E}[Y(t)\exp(-\beta' Z^*(t)) \mid \delta = 1] + (1-q)\mathrm{E}[Y(t)\exp(-\beta' Z^*(t)) \mid \delta = 0]},$$

where $q = P(\delta = 1)$ and $1 - q$ is the prevalence ratio in the cohort. Based on several estimators $\hat{q}$ of q, and following the approach of Chen and Lo [2], Ma [15] suggested a class of estimating functions as follows,

$$U^M(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i^*(t) - \frac{\tilde{S}^1(\beta,t)}{\tilde{S}^0(\beta,t)} \right\} \mathrm{d}N_i(t), \tag{3}$$

where $\tilde{S}^k(\beta,t) = [\hat{q}/n_1^s \sum_{j \in \tilde{R}_t^1} + (1-\hat{q})/n_1^s \sum_{j \in R_t^0}] \exp(-\beta' Z_j^*(t)) Z_j^*(t)^{\otimes k}$ $(k = 0,1)$, n_1 and n_1^s are the numbers of subjects with $\delta_i = 1$ in the entire cohort and the subcohort, respectively. Here $\hat{q}$ can be $n_1^s/\tilde{n}$ or n_1/n and so on (Ma [15]). In fact, the expression in (3) can be rewritten as

$$U^M(\beta) = \sum_{i=1}^{n} \delta_i \{ Z_i^*(C_i) - \frac{\tilde{S}^1(\beta, C_i)}{\tilde{S}^0(\beta, C_i)} \}. \tag{4}$$

It should be specially pointed out that computation of (4) needs all the covariate information of subjects with $\delta_i = 1$. However, under the case-cohort design, it is usually impractical to do so, since it is possible that only part of subjects with $\delta_i = 1$ are selected into the subcohort. Therefore, other estimation procedures need to be developed. In our opinion, (4) can be modified as

$$U^M(\beta) = \sum_{i=1}^{n} \delta_i \xi_i \left\{ Z_i^*(C_i) - \frac{\tilde{S}^1(\beta, C_i)}{\tilde{S}^0(\beta, C_i)} \right\}.$$

However, we will not discuss it explicitly here.

2.2. The proposed method

When dealing with case-cohort data, one common technique is to choose a desired case-cohort weight to modify the initial estimating function based on the entire cohort data. Here we denote $h_i = (1-\delta_i) + \delta_i\xi_i/p_n$, which is also used in analyzing right censored case-cohort data under the additive hazards model (Kulich and Lin [9]) and accelerated failure time model (Kong and Cai [7]). Explicitly, the proposed estimating function has the form

$$U^N(\beta) = \sum_{i=1}^{n} \int_0^\infty h_i \left\{ Z_i^*(t) - \frac{S_h^1(\beta,t)}{S_h^0(\beta,t)} \right\} \mathrm{d}N_i(t), \tag{5}$$

where $S_h^k(\beta,t) = 1/n \sum_{j=1}^n h_j Y_j(t) \exp(-\beta' Z_j^*(t)) Z_j^*(t)^{\otimes k} (k = 0, 1)$. The resulting estimator of β_0 is defined to be solution to $U^N(\beta) = 0$, denoted by $\hat{\beta}^N$. After some manipulation, (5) can be rewritten as

$$U^N(\beta) = \sum_{i=1}^{n} \frac{\delta_i\xi_i}{p_n} \{ Z_i^*(C_i) - \frac{S_h^1(\beta,C_i)}{S_h^0(\beta,C_i)} \}. \tag{6}$$

Unlike Equation (4), the expressions (5) and (6) can be straightforwardly computed without involving possibly unobserved quantities. If the exact subcohort proportion p is known, we can substitute p_n in h_i of (5) by p, and obtain a analogous estimating equation to (5). Denote its solution by $\tilde{\beta}^N$, which is used to compare with $\hat{\beta}^N$ in the following simulation studies.

To establish the the asymptotic properties of $\hat{\beta}^N$, following arguments in Kulich and Lin [8, p. 76] and those in Self and Prentice [19], we can decompose $n^{1/2}U^N(\beta)$ in (5) into three parts: the first part is just the equation in (2) up to the multiple constant $\sqrt{n}$; the third part is $o_P(1)$; and the second part is $n^{1/2} \sum_{i=1}^n (1-\delta_i)(\frac{\xi_i}{p_n} - 1)\{S^1(\beta,C_i)/S^0(\beta,C_i)\}$. For the first and second parts, following similar arguments in appendix of Ma [15] and those in Self and Prentice [19], it can be shown that they are asymptotically uncorrelated and normally distributed; furthermore, $\hat{\beta}^N$ is $\sqrt{n}$-consistent and asymptotically normal, where the limiting covariance matrix, denoted by Σ, consists of sum of two terms as occurred in other statistical inference problems for case-cohort data.

Due to the complex structure of Σ, we suggest one resampling method as follows. Let η_i, $i = 1, \ldots, n$, be i.i.d. positive random variables with $\mathrm{E}[\eta_i] = \mathrm{Var}(\eta_i) = 1$. Define

$$\tilde{U}^N(\beta) = \sum_{i=1}^{n} \frac{\delta_i\xi_i\eta_i}{p_n} \left\{ Z_i^*(C_i) - \frac{S_h^1(\beta,C_i)}{S_h^0(\beta,C_i)} \right\}. \tag{7}$$

One similar technique has been adopted by Jin et al. [5]. For a realization of η_i's, the solution of (7) provides one draw of $\hat{\beta}^N$ from its limiting distribution. Then through repeating this process a large number M_0 times, we can estimate the variance matrix Σ of $\hat{\beta}^N$ directly by the sampling variance matrix of the bootstrap sample of $\hat{\beta}^N$. In applications, we set $M_0 = 100$, which behaves well from our experience.

3. SIMULATION STUDIES

To investigate the performance of the proposed estimator for sample size commonly encountered in practice, simulation studies are conducted. Failure time T was generated from model (1) with

$$\lambda(t \mid Z) = 0.5 + \beta_0 Z,$$

where $Z \sim$ Uniform $(0, 12^{1/2})$, $\beta_0 = 0.5$ or 1, respectively. The censoring times are generated independently of T and Z, and have exponential distribution with varying rates to make the overall proportion of cases, i. e., the expected prevalence rate $P(\delta = 1)$ close to 0.1. The subcohort was selected by independent Bernoulli sampling with the subcohort proportion was 0.4 and 0.2, respectively. Simulation results based on cohort sizes 3000 and 5000 replications are obtained by R software and shown in Tables 1 and 2.

As shown in Tables 1 and 2, the proposed estimator of the regression parameter are approximately unbiased for all the scenarios, in terms of small biases and sample standard deviations. Both Bias and SD decrease as the size of cohort increases. What's more, the performance of proposed estimator $\hat{\beta}^N$ is very close to that of the full cohort estimator $\hat{\beta}$. Although there exists a difference, the difference between them is small and acceptable, rather the cost of collecting covariate histories is greatly reduced compared with the full cohort situation. The relative efficiency of the case-cohort estimator is defined as

$$RE = \frac{MSE(\hat{\beta})}{MSE(\hat{\beta}^N)} \doteq \frac{\sum(\hat{\beta} - \beta_0)^2}{\sum(\hat{\beta}^N - \beta_0)^2},$$

where the summation is over the 1000 replications. For the $\beta_0 = 0.5$ case, RE of $\hat{\beta}^N$ (and $\tilde{\beta}^N$) under the 40% and 20% of the cohort size $n = 3000$ varies from 0.6767(0.6750) to 0.4686(0.4671). Under the situation $n = 5000$, RE varies from 0.7008(0.7008) to 0.4836(0.4836). Similar trend also can be found for the $\beta_0 = 1$ case. It seems that the estimator based on the estimator of proportion of the subcohort is compared with that based on the exact value of that proportion. Moreover, the relative efficiency increases as the size of cohort increases, which may be caused by the increased number of observations of covariates and in agree with our expectation.

4. DATA ANALYSIS

In this section, we illustrate the proposed method using the data summarized by month in Tables 1 and 2 of Lindsey and Ryan [13], which is a subset of data from the ED_{01} study. The ED_{01} experiment was conducted at the National Center for Toxicological Research and involved 24,000 female mice which were randomly assigned to enter a control group or one of seven dose levels of the known carcinogen 2-acetylaminofluorene(24). There were eight interim sacrifice times, and a terminal sacrifice at 33 months. The data analyzed here involves 671 mice with bladder and lung tumors from one room considering control and high-dose groups only, where the control group and high-dose groups contained 387 and 284 mice, respectively, and we just studied the onset of bladder tumors. As to the onset of bladder tumors, a total of 124 mice were left-censored ($\delta = 0$), among which 13 mice were of the control group and 111 were of the high-dose group.

n		$\beta_0 = 0.5$			$\beta_0 = 1$		
3000		Bias	SD	MSE	Bias	SD	MSE
	$\hat{\beta}$	0.0142	0.2320	0.0538	0.0191	0.3519	0.1237
	$\hat{\beta}^N$	0.0153	0.2817	0.0795	0.0222	0.4389	0.1929
	$\tilde{\beta}^N$	0.0150	0.2820	0.0797	0.0213	0.4396	0.1935
5000		Bias	SD	MSE	Bias	SD	MSE
	$\hat{\beta}$	0.0100	0.1796	0.0328	0.0153	0.2747	0.0756
	$\hat{\beta}^N$	0.0147	0.2160	0.0468	0.0251	0.3342	0.1122
	$\tilde{\beta}^N$	0.0149	0.2160	0.0468	0.0255	0.3344	0.1123

Tab. 1. Simulation study. Subcohort sizes are equal to 40 % of the cohort size n. $\hat{\beta}$ is the full cohort estimator. $\hat{\beta}^N$ and $\tilde{\beta}^N$ are the case-cohort estimators using p_n and p in subsection 2.2, respectively. SD: standard deviation. MSE: mean squared error.

n		$\beta_0 = 0.5$			$\beta_0 = 1$		
3000		Bias	SD	MSE	Bias	SD	MSE
	$\hat{\beta}$	-0.0105	0.2443	0.0597	-0.0106	0.3745	0.1402
	$\hat{\beta}^N$	0.0270	0.3571	0.1274	0.0326	0.5695	0.3245
	$\tilde{\beta}^N$	0.0275	0.3577	0.1278	0.0340	0.5707	0.3259
5000		Bias	SD	MSE	Bias	SD	MSE
	$\hat{\beta}$	0.0005	0.1840	0.0338	0.0008	0.2706	0.0731
	$\hat{\beta}^N$	0.0061	0.2631	0.0699	0.0219	0.4172	0.1749
	$\tilde{\beta}^N$	0.0062	0.2631	0.0699	0.0221	0.4175	0.1753

Tab. 2. Simulation study. Subcohort sizes are equal to 20% of the cohort size n. $\hat{\beta}$ is the full cohort estimator. $\hat{\beta}^N$ and $\tilde{\beta}^N$ are the case-cohort estimators using p_n and p in subsection 2.2, respectively. SD: standard deviation. MSE: mean squared error.

Like that in Chen et al. [1], we consider only the univariate Z_i, taking value 1 for the high-dose group and 0 for the control group. For the full cohort data, we fit the the model (1) and obtain $\hat{\beta} = 0.0225$ with an estimated standard error 0.0081, which implies that the hazard of the mice in the high-dose group is significantly higher than that of the control group. The similar finding has also been observed by Lindsey and Ryan [13] and Chen et al. [1] in other modeling settings. To make a comparison between it and our method for the case-cohort data, we selected the subcohort by independent Bernoulli sampling and the resultant subcohort proportion is about 0.27. For such set of case-cohort data, the estimate $\hat{\beta}^N = 0.0409$ with a estimated standard error 0.0177, indicating that the mice in the high-dose group also have a statistically significantly

higher hazard than those in the control group. This is in accordance with the findings mentioned above.

5. CONCLUSION AND DISCUSSION

In this paper, we developed a weighted estimating equation to fit the current status data from case-cohort studies with an additive hazards model. We also sketched that the proposed estimator was consistent and asymptotically normally distributed. Compared with the full cohort estimator, the efficiency loss of the case-cohort estimator in our simulation studies remained acceptable compared to the sample size reduction.

In our study, we focus on the case-cohort design with simple random sampling without replacement, which is one example of the general two-phase sampling schemes introduced by Neyman [17]. From the construction of our estimating equation, it is straightforward to extend our method to other schemes under which one subcohort is selected. One possible difficulty may be the derivation of theoretical properties of the resultant estimators. Moreover, other procedures that will improve the relative efficiency are still desirable and will be pursued in the future. As argued by Lin et al. [11], the assumption that the censoring time is independent of covariates is stringent in some applications. Therefore, to develop procedures for case-cohort sampled current status data allowing dependence of censoring time on covariates under the additive hazards model is of interest and we are studying this issue.

(Received September 11, 2014)

REFERENCES

[1] C. M. Chen, T. F. C. Lu, M. H. Chen, and C. M. Hsu: Semiparametric transformation models for current status data with informative censoring. Biometrical J. *54* (2012), 641–656. DOI:10.1002/bimj.201100131

[2] K. Chen and S. H. Lo: Case-cohort and case-control analysis with Cox's model. Biometrika *86* (1999), 755–764. DOI:10.1093/biomet/86.4.755

[3] P. B. Gilbert, M. L. Peterson, D. Follmann, M. G. Hudgens, D. P. Francis, M. Gurwith, W. L. Heyward, D. V. Jobes, V. Popovic, S. G. Self, F. Sinangil, D. Burke, and P. W. Berman: Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. J. Infectious Diseases *191* (2005), 666–677. DOI:10.1086/428405

[4] P. Groeneboom and J A. Wellner: Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhauser, Basel 1992. DOI:10.1007/978-3-0348-8621-5

[5] Z. Jin, D. Y. Lin, L. J. Wei, and Z. Ying: Rank-based inference for the accelerated failure time model. Biometrika *90* (2003), 341–353. DOI:10.1093/biomet/90.2.341

[6] L. Kong, J. Cai, and P. K. Sen: Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. Statistica Sinica *16* (2006), 135–151.

[7] L. Kong and J. Cai: Case-cohort analysis with accelerated failure time model. Biometrics *65* (2009), 135–142. DOI:10.1111/j.1541-0420.2008.01055.x

[8] M. Kulich and D. Y. Lin: Additive hazards regression for case-cohort studies. Biometrika *87* (2000), 73–87. DOI:10.1093/biomet/87.1.73

[9] Z. Li, P. Gilbert, and B. Nan: Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. Biometrics *64* (2008), 1247–1255. DOI:10.1111/j.1541-0420.2008.00998.x

[10] Z. Li and B. Nan: Relative risk regression for current status data in case-cohort studies. Canadian J. Statist. *39* (2011), 557–577. DOI:10.1002/cjs.10111

[11] D. Y.Lin, D. Oakes, and Z. Ying: Additive hazards regression with current status data. Biometrika *85* (1998), 289–298. DOI:10.1093/biomet/85.2.289

[12] D. Y. Lin and Z. Ying: Semiparametric analysis of the additive risk model. Biometrika *81* (1994), 61–71. DOI:10.1093/biomet/81.1.61

[13] J. C. Lindsey and L. M. Ryan: A comparison of continuous- and discrete-time three-state models for rodent tumorigenicity experiments. Environmental Health Perspectives Supplements *102* (1994), 9–17. DOI:10.1289/ehp.94102s19

[14] W. Lu and A. A. Tsiatis: Semiparametric transformation models for the case-cohort study. Biometrika *93* (2006), 207–214. DOI:10.1093/biomet/93.1.207

[15] S. Ma: Additive risk model with case-cohort sampled current status data. Statistical Papers *48* (2007), 595–608. DOI:10.1007/s00362-007-0359-y

[16] B. Nan, M. Yu, and J. D. Kalbfleisch: Censored linear regression for case-cohort studies. Biometrika *93* (2006), 747–762. DOI:10.1093/biomet/93.4.747

[17] J. Neyman: Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc. *33* (1938), 101–116. DOI:10.1080/01621459.1938.10503378

[18] R. L. Prentice: A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika *73* (1986), 1–11. DOI:10.1093/biomet/73.1.1

[19] S. G. Self and R. L. Prentice: Asymptotic distribution theory and efficiency results for case-cohort studies. Ann. Statist. *16* (1988), 64–81. DOI:10.1214/aos/1176350691

[20] J. Sun: The Statistical Analysis of Interval-Censored Failure Time Data (Vol. 2). Springer, New York 2006. DOI:10.1007/0-387-37119-2

Wei Chen, School of Zhangjiagang, Jiangsu University of Science and Technology, Zhangjiagang 215600. P. R. China.
e-mail: novicejlu@gmail.com

Fengling Ren, School of Computer Science and Engineering, Xinjiang University of Finance and Economics, Urumqi 830012. P. R. China.
e-mail: 909205543@qq.com

Guosheng Tang, School of Zhangjiagang, Jiangsu University of Science and Technology, Zhangjiagang 215600. P. R. China.
e-mail: 504951453@qq.com