Vol. 38 (2018)
No. 2

数 学 杂 志
J. of Math. (PRC)

A NOTE ON POWER CALCULATION FOR GENERALIZED CASE-COHORT SAMPLING WITH ACCELERATED FAILURE TIME MODEL

SHI Yue-yong[1,3], CAO Yong-xiu[2], JIAO Yu-ling[2], YU Ji-chang[2]

(*1.School of Economics and Management, China University of Geosciences, Wuhan 430074, China*)

(*2.School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China*)

(*3.Center for Resources and Environmental Economic Research, China University of Geosciences, Wuhan 430074, China*)

Abstract: In this paper, we study the power calculation for the generalized case-cohort sampling. By using the smoothed weighted Gehan estimating equation method, we obtain the unbiased estimators of the unknown regression parameters in the accelerated failure time model. The simulation studies and the real data analysis show the good performances of the proposed method.

Keywords: accelerated failure time model; generalized case-cohort sampling; induced smooth; power calculation

2010 MR Subject Classification: 62D05; 62J99; 62N01

Document code: A **Article ID:** 0255-7797(2018)02-0200-09

1 Introduction

In many epidemiological studies, the meaningful results can be obtained through observing thousands of subjects for a long time. Due to the financial limitation or technical difficulties, it needs to develop the cost-effective design for selecting subjects in the underlying cohort to observe their expensive covariates. The case-cohort sampling (Prentice, 1986) is a well-known cost-effective design with the response subject to censoring, where the expensive covariates are measured only for a subcohort randomly selected from the cohort and additional failures outside the subcohort. The statistical methods for case-cohort sampling were well studied in the literature (e.g., Self and Prentice, 1988; Chen and Lo, 1999; Kulich and Lin, 2000; Kong, Cai and Sen, 2004; Kong and Cai, 2009).

Aforementioned works show the case-cohort sampling is especially useful when the failure rate is low. However, the failure rate may be high in practice. Therefore, it's unpractical

* **Received date:** 2017-04-08 **Accepted date:** 2017-07-10

Foundation item: Supported by National Natural Science Foundation of China (11501578; 11701571; 11501579; 41572315); Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGW150809).

Biography: Shi Yueyong (1984–), male, born at Luzhou, Sichuan, lecturer, major in biostatistics.

Correspongding author: Yu Jichang.

to assemble covariates of all failures due to the fixed budget. Under such situations, the generalized case-cohort (GCC) sampling is proposed, which selects a subset of failures instead of all the failures in the case-cohort design. For example, Chen (2001) proposed the GCC design and studied its statistical properties. Kang and Cai (2009) studied the GCC design with the multivariate failure time. Cao, Yu and Liu (2015) studied the optimal GCC design through the power function of a significant test. The aforementioned works are all under the framework of Cox's proportional hazards model (Cox, 1972). Yu et al. (2014), Cao and Yu (2017) studied the GCC design under the additive hazards model (Lin and Ying, 1994).

Both the Cox proportional hazards model and additive hazards model are based on modeling the hazards function. However, it is important to directly model the failure time in some applications. Recently, the accelerated failure time (AFT) model which linearly relates the logarithm of the failure time to the covariates gained more and more attention. Kong and Cai (2009) studied the case-cohort sampling under the AFT model. Chiou, Kang and Yan (2014) proposed a fast algorithm for the AFT model under the case-cohort sampling. Cao et al. (2017) studied the GCC sampling under the AFT model and discussed the optimal subsample allocation by the asymptotic relative efficiency between the proposed estimators and the estimators from the simple random sampling scheme.

In order to design a GCC study in practice, there is an important question for the principal investigators that how to calculate the power function under a fixed budget. To the best of our knowledge, no such consideration is given under the generalized case-cohort design. Therefore, we will fill this gap under the accelerated failure time model in this paper.

The article is organized as follows. In Section 2, we propose the generalized case-cohort sampling, use the smoothed weighted Gehan estimating equation approach to estimate the unknown regression parameters in the accelerated failure time model, and give the corresponding asymptotic properties. In Section 3, we study the power calculation under a fixed budget. In Section 4, we conduct the simulation studies to evaluate the performances of the proposed methods. A real data analysis is analyzed through the proposed method in Section 5. Some concluding remarks are presented in Section 6.

2 Generalized Case–Cohort Sampling and Inference Procedures

2.1 Model

Let $\tilde{T}$ and C denote the failure time and the corresponding censoring time, respectively. Due to the right censoring, we only observe $T = \min(\tilde{T}, C)$ and $\delta = I(\tilde{T} \leq C)$, where $I(\cdot)$ is an indicator function. Let Z_e be a d_1-dimensional vector of covariates which are expensive to measure and Z_c be a d_2-dimensional vector of covariates which are cheap or easily to measure. It is assumed that given the covariates (Z_e, Z_c), $\tilde{T}$ and C are independent. We consider the following accelerated failure time model

$$\log(\tilde{T}) = \beta_0^{'} Z_e + \gamma_0^{'} Z_c + \epsilon, \tag{2.1}$$

where β_0 and γ_0 are unknown regression parameters and ϵ is the random error with an unknown distribution function.

2.2 Generalized Case-Cohort Sampling

Suppose the underlying population has n subjects and $\{T_i, \delta_i, Z_{e,i}, Z_{c,i}, i = 1, \cdots, n\}$ are the independent copies of (T, δ, Z_e, Z_c). In the generalized case-cohort sampling, binary random variable ξ_i denotes whether or not the i-th subject is selected into the subcohort and the corresponding successful probability is p. Let η_i be the selection indicator for whether or not the j-th subject is selected into supplemental failure samples and the conditional probability $P(\eta_j = 1|\xi_j = 0, \delta_j = 1) = q$. In the GCC sampling, the covariates Z_e are only observed on the selected subjects. Hence, the observed data structure is given as follows:

$$\{T_i, \delta_i, Z_{e,i}\left[\xi_i + (1-\xi_i)\delta_i\eta_i\right], Z_{c,i}, i = 1, \cdots, n\}.$$

2.3 Inference Procedures

Define $\theta = (\beta', \gamma')'$, $\theta_0 = (\beta_0', \gamma_0')'$, $X_i = (Z_{e,i}', Z_{c,i}')'$, and $e_i(\theta) = \log(T_i) - \beta' Z_{e,i} - \gamma' Z_{c,i}, i = 1, \cdots, n$. Let $N_i(t,\theta) = I(e_i(\theta) \le t, \delta_i = 1)$ and $Y_i(t,\theta) = I(e_i(\theta) \ge t)$ denote the counting process and at risk process, respectively. If the data $\{T_i, \delta_i, Z_{e,i}, Z_{c,i}, i = 1, \cdots, n\}$ are completely observed, the unknown regression parameters in model (2.1) can be estimated by solving the following estimating equations

$$U_{n,\psi}(\theta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \psi(t,\theta)[X_i - \bar{X}(t,\theta)]dN_i(t,\theta) = 0, \tag{2.2}$$

where $\psi(\cdot)$ is a possible data-dependent weight function and $\bar{X}(t,\theta) = S^{(1)}(t,\theta)/S^{(0)}(t,\theta)$ with $S^{(d)}(t,\theta) = n^{-1}\sum\limits_{j=1}^{n} Y_j(t,\theta)X_j^d$ for $d = 0, 1$. The weight function $\psi(t,\theta) = 1$ and $S^{(0)}(t,\theta)$ are corresponding to the log-rank and Gehan statistics, respectively.

Unfortunately, in the GCC sampling, the covariates Z_e are only observed for selected subject and the distribution of selected supplemental failures is different from the distribution of the underlying population. Therefore, the inverse probability weight method (Horvitz and Thompson, 1951) is needed to adjust for the biased sampling mechanism of the GCC sampling

$$W_i = \delta_i\xi_i + (1-\delta_i)\xi_i/p + (1-\xi_i)\delta_i\eta_i/q,\ i = 1, \cdots, n. \tag{2.3}$$

Then, the true regression parameters θ_0 in model (2.1) can be estimated by solving the following weighted estimating equations

$$\tilde{U}_{n,\tilde{\psi}}(\theta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \tilde{\psi}(t,\theta)W_i[X_i - \tilde{X}(t,\theta)]dN_i(t,\theta) = 0, \tag{2.4}$$

where $\tilde{\psi}(\cdot)$ is also a possible data-dependent weight function and $\tilde{X}(t,\theta) = \tilde{S}^{(1)}(t,\theta)/\tilde{S}^{(0)}(t,\theta)$ with

$$\tilde{S}^{(d)}(t,\theta) = n^{-1}\sum_{j=1}^{n} W_j Y_j(t,\theta)X_j^d$$

for $d = 0, 1$. In this paper, we consider Gehan statistics, $\tilde{\psi}(t,\theta) = \tilde{S}^{(0)}(t,\theta)$. Hence, the weighted Gehan estimating equations can be re-written as

$$\tilde{U}_{n,G}(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i W_i W_j (X_i - X_j) I(e_i(\theta) \le e_j(\theta)) = 0, \tag{2.5}$$

which are monotone in each component of θ and let $\tilde{\theta}_n$ denote the estimators obtained by solving (2.5).

Due to the fact that the weighted Gehan estimating equations are not continuous, induced smoothing procedure is adopted to smooth the weighted Gehan estimating equations (Brown and Wang, 2007; Cao, Yang and Yu, 2017). The smoothed weighted Gehan estimating equations can be re-written as

$$\bar{U}_{n,G}(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i W_i W_j (X_i - X_j)\Phi\left(\frac{e_j(\theta) - e_i(\theta)}{r_{ij}}\right) = 0, \tag{2.6}$$

where $r_{ij}^2 = n^{-1}(X_j - X_i)^{'}(X_j - X_i)$ and $\Phi(\cdot)$ is a distribution function of the standard normal distribution. As n goes to infinity, the distribution function $\Phi(\cdot)$ will convergent to the indicator function. Let $\widehat{\theta}_n$ denote the estimators by solving estimating equation (2.6).

2.4 Asymptotic Properties

In this subsection, we will show the consistency and asymptotic distribution of the $\widehat{\theta}_n$. Furthermore, the asymptotic distribution of $\widehat{\theta}_n$ is also the same as that of $\tilde{\theta}_n$. Define $M_i(t,\theta) = N_i(t,\theta) - \Lambda_i(t,\theta)$, $\Lambda_i(t,\theta) = \int_{-\infty}^{t} Y_i(s,\theta)\lambda(u)du$, and $\lambda(\cdot)$ is the common hazard function of the error term and $a^{\otimes 2} = aa^{'}$ for a vector a.

Theorem 2.1 Under some regular conditions,

$$n^{-1/2}\tilde{U}_{n,G}(\theta_0) = n^{-1/2}\bar{U}_{n,G}(\theta_0) + o_p(1),$$

$\widehat{\theta}_n$ is strong consistency, and $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution to zero-mean normal with covariance

$$\Sigma_A^{-1}(\theta_0)(\Sigma_F(\theta_0) + \frac{1-p}{p}\Sigma_C(\theta_0) + \frac{(1-p)(1-q)}{q}\Sigma_G(\theta_0))(\Sigma_A^{-1}(\theta_0))^{'},$$

where the matrix $\Sigma_A(\theta_0)$ is the limit of

$$\begin{aligned} &n^{-1}\partial\bar{U}_{n,G}(\theta_0)/\partial\theta, \Sigma_F(\theta_0) = E[H_1(\theta_0)^{\otimes 2}], \\ &\Sigma_C(\theta_0) = E[(1-\delta_1)H_1(\theta_0)^{\otimes 2}], \Sigma_G(\theta_0) = E[\delta_1 H_1(\theta_0)^{\otimes 2}] \end{aligned}$$

with

$$H_1(\theta_0) = \int_{-\infty}^{\infty}\tilde{\psi}(t,\theta_0)[X_1 - e_X(t,\theta_0)]dM_1(t,\theta_0).$$

The regularity conditions and the proof of Theorem 2.1 can be founded in [15].

3 Power Calculation

In this section, we consider the power calculation for GCC sampling with a fixed budget. In order to simplify the notations, let $\mathbf{B}$ denote the fixed budget, C_c denote the unit price to measure the observed failure time, censoring indicator and cheap covariates $\{T_i, \delta_i, Z_{c,i}\}$, and C_e denote the unit price to measure expensive covariates $Z_{e,i}$. Hence,

$$\mathbf{B} = n\{C_c + [p + (1-p)\pi q]C_e\}, \tag{3.1}$$

where $\pi = P(\delta = 1)$. In practice, n, $\mathbf{B}$, C_c and C_e are known, π can be estimated by $n^{-1}\sum_{i=1}^{n}\delta_i$, which is equal to $p + (1-p)\pi q$ fixed. Let $\rho_v = p + (1-p)\pi q$, which is the proportion of the validation data set in the missing data literature, where all the data is completely observed.

We consider the following significant test

$$H_0 : \beta_0 = 0 \quad \text{VS} \quad H_1 : \beta_0 = k, \tag{3.2}$$

where k is a non-zero d_1-dimensional constant. Let $\widehat{\beta}_n$ denote the proposed estimator of β_0 and α denote the type I error, respectively. From Theorem 2.1, the reject region of the test (3.2) at the significant level α is

$$W = \{|\widehat{\beta}_n - \beta_0| \geq \Psi^{-1}(1-\frac{\alpha}{2})\Sigma(\widehat{\beta}_n)^{1/2}\},$$

where

$$\begin{aligned}\Sigma(\widehat{\beta}_n) &= \frac{1}{n}[\Sigma_A^{-1}(\theta_0)\Sigma_F(\theta_0)\Sigma_A^{-1}(\theta_0)]_{d_1\times d_1} + \frac{1-p}{np}[\Sigma_A^{-1}(\theta_0)\Sigma_C(\theta_0)\Sigma_A^{-1}(\theta_0)]_{d_1\times d_1}\\ &\quad + \frac{(1-p)(1-q)}{nq}[\Sigma_A^{-1}(\theta_0)\Sigma_G(\theta_0)\Sigma_A^{-1}(\theta_0)]_{d_1\times d_1},\end{aligned}$$

$\Psi^{-1}(1-\frac{\alpha}{2})$ is a d_1-dimensional vector with same element $(\Phi^{-1}(1-\frac{\alpha}{2}), \cdots, \Phi^{-1}(1-\frac{\alpha}{2}))'$ with $\Phi(\cdot)$ being the distribution function of the standard norm distribution, and $[A]_{d_1\times d_1}$ is the upper-left $d_1 \times d_1$ submatrix of matrix A. Obviously, the power function for the significant test is a function of (p, q) given as follows

$$\begin{aligned}Power(p,q) &= P(\widehat{\beta}_n \in W | H_1 : \beta_0 = k)\\ &= 1 - \Psi(\Psi^{-1}(1-\frac{\alpha}{2}) - k\Sigma(\widehat{\beta}_n)^{-1/2}) + \Psi(-\Psi^{-1}(1-\frac{\alpha}{2}) - k\Sigma(\widehat{\beta}_n)^{-1/2}).\end{aligned}$$

When we calculate the power function, due to constrain (3.1), we need to consider the following optimization problem through the Lagrange multiplier argument

$$\begin{aligned}L_n(p,q,\lambda) &= \|1 - \Psi(\Psi^{-1}(1-\frac{\alpha}{2}) - k\Sigma(\widehat{\beta}_n)^{-1/2}) + \Psi(-\Psi^{-1}(1-\frac{\alpha}{2}) - k\Sigma(\widehat{\beta}_n)^{-1/2})\|_1\\ &\quad -\lambda\{\mathbf{B} - nC_c - n[p + (1-p)\pi q]C_e\},\end{aligned}$$

where $\|\cdot\|_1$ denote the L_1 norm. Because the power function is positive, the optimal solution (p^*, q^*) can be easy to obtain and the corresponding power is Power(p^*, q^*).

4 Simulation Study

In this section, the simulation studies are conducted to evaluate the finite sample performances of the proposed method. We generate the failure time from the accelerated failure time model

$$\log T = \beta_0 Z_e + \gamma_0 Z_c + \epsilon, \tag{4.1}$$

where Z_e follows a standard normal distribution, Z_c follows a Bernoulli distribution with a successful probability of 0.5, the regression parameters $\beta_0 = 0$ and $\gamma_0 = 0.5$, and the error term ϵ follows a standard normal distribution or a standard extreme value distribution, which will result a log-norm distribution or a Weibull distribution for the failure time, respectively. The censoring time is generated from the uniform distribution over the interval $[0, c]$, where c is chosen to yield around 80% censoring rate, respectively.

We consider the following test at the significant level α being 0.05:

$$H_0 : \beta_0 = 0 \quad \text{VS} \quad H_1 : \beta_0 = 0.3.$$

The size of the underlying population is $n = 600$. We investigate different scenarios for sampling probabilities (p, q) under constraint (3.1), which is equal to ρ_v being fixed. For each configuration, we generate 1000 simulated data sets. The results of the simulation studies are summarized in Figure 1. From Figure 1, we can obtain following results

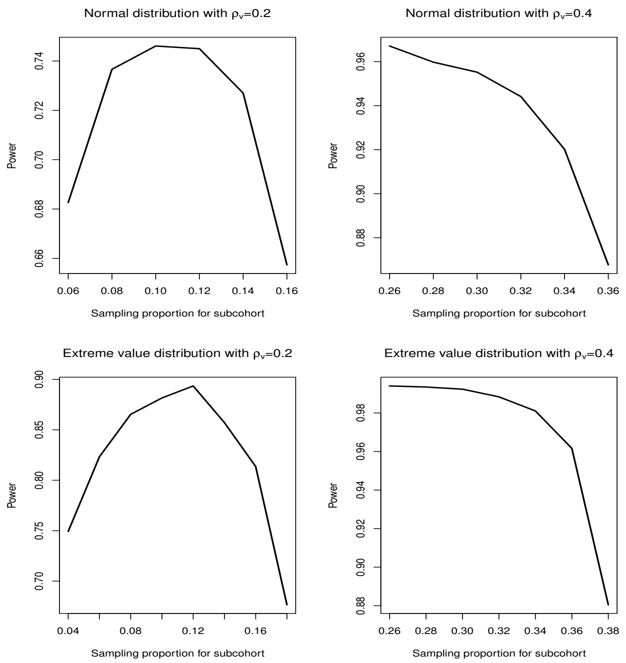

Figure 1: Power function with fixed ρ_v and different sampling probability p

(I) When the error term follows the standard normal distribution, the powers are 0.746 and 0.967 with ρ_v being 0.200 and 0.400, respectively, and the corresponding sampling probability (p, q) are $(0.100, 0.556)$ and $(0.260, 0.946)$, respectively.

(II) When the error term follows the extreme value distribution, the powers are 0.893 and 0.994 with ρ_v being 0.200 and 0.400, respectively, and the corresponding sampling probability (p, q) are $(0.120, 0.455)$ and $(0.260, 0.946)$, respectively.

5 National Wilm's Tumor Study Group

The national Wilm's tumor study group (NWTSG) is a cancer research which was conducted to improve the survival of children with Wilms' tumor by evaluating the relationship between the time to tumor relapse and the tumor histology (Green et al., 1998). However, the tumor histology is difficult and expensive to measure. According to the cell type, the tumor histology can be classified into two categories, named as favorable and unfavorable. Let the variable *histol* denote the category of the tumor histology. We also consider other covariates including the patient age, the disease stages and the study group.

We consider the accelerated failure time model

$$\log T = \alpha_1 histol + \alpha_2 age + \alpha_3 stage2 + \alpha_4 stage3 + \alpha_5 stage4 + \alpha_6 study + \epsilon,$$

where the covariates $stage2$, $stage3$, $stage4$ indicate the disease stages and the variable $study$ indicates the study group. There are 4028 subjects in the full cohort and 571 subjects subject to tumor relapse. We randomly selected a subcohort by $p = 0.166$ and select a subset of the failures outside subcohort through $q = 0.400$. We compare the proposed estimator $\widehat{\alpha}_G$ with $\widehat{\alpha}_S$, which is based on the simple random sampling design with the same sample size as GCC design. The results are summarized in Table 1.

Table 1: Analysis results for NWTSG

Method	Covariate	Estimate	SD	p_{value}
$\widehat{\alpha}_S$	*histol*	-3.244	0.332	0.000
	age	-0.100	0.061	0.051
	*stage*2	-0.976	0.488	0.023
	*stage*3	-1.735	0.423	0.000
	*stage*4	-2.125	0.557	0.000
	study	-0.485	0.396	0.110
$\widehat{\alpha}_G$	*histol*	-2.783	0.208	0.000
	age	-0.113	0.046	0.007
	*stage*2	-1.284	0.366	0.000
	*stage*3	-1.130	0.375	0.001
	*stage*4	-2.175	0.356	0.000
	study	-0.027	0.244	0.456

From Table 1, both the two methods confirm that tumor histology is significant to the cancer relapse. The proposed method shows the age is significant to cancer relapse which is different from the result from $\widehat{\alpha}_S$.

6 Concluding Remarks

In this paper, we study the power calculation for the generalized case-cohort (GCC) design under the accelerated failure time model. Due to the biased sampling mechanism of GCC, the weighted Gehan estimating equations are adopted to estimate the regression coefficients. The induced smoothing procedure is introduced to overcome the discontinuous of the smoothed weighted Gehan estimating equation, which could lead to continuously differentiable estimating equations and can be solved by the standard numerical methods. The simulation studies are conducted to evaluate the finite sample performances of the proposed method and we also analyze a real data set from national Wilm's tumor study group.

In this paper, we consider the covariates which are time-invariant. Next, we will consider power calculation in the accelerated failure time model under the GCC design with time-dependent covariates. Finally, it will be interesting to evaluate the performance of stratified sampling in the subcohort to enhance the efficiency. Study along this directions is currently under way.

References

[1] Prentice R L. A case-cohort design for epidemiologic cohort studies and disease prevention trials[J]. Biometrika, 1986, 73(1): 1–11.

[2] Self S G, Prentice R L. Asymptotic distribution theory and efficiency results for case-cohort studies[J]. Ann. Stat., 1988, 16(1): 64–81.

[3] Chen K, Lo S H. Case-cohort and case-control analysis with Cox's model[J]. Biometrika, 1999, 86(4): 755–764.

[4] Kulich M, Lin D Y. Additive hazards regression with covariate measurement error[J]. J. Amer. Stat. Assoc., 2000, 95(449): 238–248.

[5] Kong L, Cai J, Sen P K. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design[J]. Biometrika, 2004, 91(2): 305–319.

[8] Kong L, Cai J. Case-cohort analysis with accelerated failure time model[J]. Biometrics, 2009, 65(1): 135–142.

[7] Chen K. Generalized case-cohort sampling[J]. J. R. Stat. Soc. Ser. B Stat. Meth., 2001, 63(4): 791–809.

[8] Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes[J]. Biometrika, 2009, 96(4): 887–901.

[9] Cao Y X, Yu J C, Liu Y Y. Optimal generalized case-cohort analysis with Cox's proportional hazards model[J]. Acta Math. Appl. Sin. Engl. Ser., 2015, 31(3): 841–854.

[10] Cox D R. Regression models and life-tables (with discussion)[J]. J. R. Stat. Soc. Ser. B Stat. Meth., 1972, 34(2): 187-220.

[11] Yu J C, Shi Y Y, Yang Q L, Liu Y Y. Additive hazards regression under generalized case-cohort sampling[J]. Acta Math. Sin. Engl. Ser., 2014, 30(2): 251–260.

[12] Cao Y, Yu J. Optimal generalized case-cohort sampling design under the additive hazard model[J]. Comm. Statist. The. Meth., 2017, 46(9): 4484–4493.

[13] Lin D, Ying Z. Semiparametric analysis of additive hazards model[J]. Biometrika, 1994, 81(1): 61–71.

[14] Chiou S H, Kang S, Yan J. Fast accelerated failure time modeling for case-cohort data[J]. Stat. Comput., 2014, 24(4): 559–568.

[15] Cao Y, Yang Q, Yu J. Optimal generalized case-cohort analysis with accelerated failure time model[J]. J. Korean Stat. Soc., 2017, 46(2): 298–307.

[16] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe[J]. J. Amer. Stat. Assoc., 1952, 47(260): 663–685.

[17] Brown B M, Wang Y G. Induced smoothing for rank regression with censored survival times[J]. Stat. Med., 2007, 26: 828–836.

加速失效时间模型下关于广义病例队列抽样功效计算的一个注记

石跃勇[1,3], 曹永秀[2], 焦雨领[2], 余吉昌[2]

(1.中国地质大学(武汉)经济管理学院, 湖北 武汉 430074)

(2.中南财经政法大学统计与数学学院, 湖北 武汉 430073)

(3.中国地质大学(武汉)资源环境经济研究中心, 湖北 武汉 430074)

摘 要: 本文在加速失效时间模型下研究了广义病例队列抽样的功效计算问题. 利用光滑加权Gehan估计方程方法估计了未知回归参数, 研究了固定预算下广义病例队列抽样的功效计算. 模拟研究和实际数据分析评估了提出方法在有限样本下的表现.

关 键 词: 加速失效时间模型; 广义病例队列抽样; 诱导光滑; 功效计算

MR(2010)**主题分类号**: 62D05; 62J99; 62N01 **中图分类号**: O212.1