Biostatistics (2000), **1**, 1, *pp*. 35–47
Printed in Great Britain

Linear regression analysis of censored medical costs

D. Y. LIN

Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA
danyu@biostat.washington.edu

SUMMARY

This paper deals with the problem of linear regression for medical cost data when some study subjects are not followed for the full duration of interest so that their total costs are unknown. Standard survival analysis techniques are ill-suited to this type of censoring. The familiar normal equations for the least-squares estimation are modified in several ways to properly account for the incompleteness of the data. The resulting estimators are shown to be consistent and asymptotically normal with easily estimated variance–covariance matrices. The proposed methodology can be used when the cost database contains only the total costs for those with complete follow-up. More efficient estimators are available when the cost data are recorded in multiple time intervals. A study on the medical cost for ovarian cancer is presented.

Keywords: Censoring; Cost analysis; Economic evaluation; Health economics; Incomplete data; Medical care; Survival analysis.

1. INTRODUCTION

The escalating cost of health care has posed serious challenges in the United States and other countries. The cost of intensive chemotherapy with peripheral stem cell support for immune function in stage II breast cancer approaches $100 000 in some treatment centers, and this cost must be balanced against the potential expense of salvage therapy for later recurrent disease. The initial costs of screening for various forms of cancer are reasonably well understood, but the exact costs of long-term care in this and other chronic diseases are largely unknown. Accurate understanding of the costs associated with alternative therapies can lead to substantial cost savings. Clinical trial data as well as patient series data from medical centers, disease registries and insurance companies present excellent opportunities to evaluate the cost of medical care. Many multi-center clinical trials groups, such as the Eastern Cooperative Oncology Group, have formed special research teams to study medical cost so that the clinical benefit of newer therapies can be evaluated in light of cost.

Despite the tremendous interest in the evaluation of medical cost, there has been little progress in the development of formal statistical methods for such evaluation. The main difficulty lies in the incompleteness of the available data. In long-term clinical or observational studies to collect cost data, it is inevitable that some patients are not followed until the endpoint of interest so that their medical costs are not fully observed. This phenomenon is referred to as censoring, which is well known for survival time data.

Statistical methods for handling censoring in survival data have been well developed. Due to the similarity between censored medical costs and censored survival times, a number of authors have suggested that standard survival analysis methods, such as the Kaplan–Meier estimator and Cox regression, be used to analyse censored medical costs. As pointed out by Lin *et al*. (1997), however, this strategy is false

because the inherent patient heterogeneity with respect to cost accumulation entails that the cumulative cost at the censoring time is positively correlated with the cumulative cost at the endpoint of interest even if the underlying censoring mechanism is purely random.

Lin *et al.* (1997) developed non-parametric methods for estimating the mean total cost based on censored data. Several unpublished reports have provided refinements of and alternatives to Lin *et al.*'s estimators. All these efforts, however, are confined to the one-sample problem. Currently, there does not exist any valid regression method for assessing the effects of covariates (e.g. therapies, insurance plans and patients characteristics) on medical cost with censored data. The regression methodology would be particularly valuable in identifying cost-effective intervention/prevention programs. A more specific application of such a methodology is to devise risk-adjusted payment systems for the Medicare or insurance companies which would reduce the incentives for hospitals to use unnecessarily expensive therapies and at the same time would avoid penalizing the hospitals that serve patients requiring more intensive care.

In the next section, we develop simple and valid methods for fitting linear regression models to censored cost data. In Section 3, we evaluate the performance of these methods by Monte Carlo simulation. In Section 4, we illustrate the proposed methods with the ovarian cancer study initially reported by Lin *et al.* (1997). Most of the technical material is relegated to the Appendix.

2. METHODS

2.1. Basic ideas

Suppose that one is interested in the total medical cost over the time period $[0, \tau]$, which may be 1 year or 50 years. For technical reasons, τ is no larger than the overall length of the study; without making stringent parametric assumptions, it would not be possible to estimate the total cost over $[0, \tau]$ if no subject were followed for τ years. Naturally, there is no further accumulation of medical cost after death. Thus, the total cost over $[0, \tau]$ is the same as the cumulative cost at $T^* = \min(T, \tau)$, where T is the survival time. If one is interested in the lifetime cost, then τ has to be larger than the support of the distribution of T.

Let Y be the cumulative cost at τ or T^*, and let Z be a $p \times 1$ vector of covariates whose effects on Y are of interest. It is convenient to relate Y to Z through the linear regression model

$$Y = \beta' Z + \epsilon, \tag{1}$$

where β is a $p \times 1$ vector of unknown regression parameters, and ϵ is a zero-mean error term with an unspecified distribution. We set the first component of Z to 1 so that the first component of β corresponds to the intercept.

As mentioned in Section 1, survival time and medical cost are normally subject to right censoring. Let C be the censoring time. Write $X = \min(T, C)$, $\delta = I(C \geq T)$, and $\delta^* = I(C \geq T^*)$, where $I(\cdot)$ is the indicator function. Clearly, Y is known if and only if $\delta^* = 1$, whereas T is known if and only if $\delta = 1$. A subject whose survival time is censored at or after τ has a complete observation on the medical cost over $[0, \tau]$, while a subject whose survival time is censored before τ has an incomplete observation. Suppose that we have a random sample of size n. For $i = 1, \ldots, n$, the variables corresponding to the ith subject are indexed by the subscript i.

In this subsection, we assume that censoring occurs in a completely random fashion such that C is independent of all other random variables. This assumption holds if, for example, all the censoring is caused by study termination, which is the so-called administrative censoring. Let $G(t) = \Pr(C \geq t)$, and let $\hat{G}(t)$ be the Kaplan–Meier estimator of $G(t)$ based on the data $(X_i, \bar{\delta}_i)$ $(i = 1, \ldots, n)$, where $\bar{\delta}_i = 1 - \delta_i$. We consider more general censoring mechanisms in the next subsection.

If no survival times are censored before τ, then β may simply be estimated by the least-squares normal

equation

$$\sum_{i=1}^{n}(Y_i - \beta' Z_i)Z_i = 0. \tag{2}$$

In practice, there is some censoring before τ. As mentioned above, Y_i is known if and only if $\delta_i^* = 1$. Because $E\{\delta_i^*/G(T_i^*)\} = 1$, it seems natural to modify (2) as

$$\sum_{i=1}^{n}\frac{\delta_i^*}{G(T_i^*)}(Y_i - \beta' Z_i)Z_i = 0. \tag{3}$$

Only the subjects with complete cost data contribute non-zero terms to the summation in (3), but their contributions are weighted inversely by their probabilities of inclusion. Thus, the left-hand sides of (2) and (3) have the same expectation, which is zero. Since G is unknown but can be consistently estimated by the Kaplan–Meier estimator $\hat{G}$, we replace G in (3) with $\hat{G}$ to yield the following estimating equation for β

$$\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{G}(T_i^*)}(Y_i - \beta' Z_i)Z_i = 0, \tag{4}$$

which has a closed-form solution

$$\hat{\beta} = \left\{\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{G}(T_i^*)}Z_i^{\otimes 2}\right\}^{-1}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{G}(T_i^*)}Y_i Z_i. \tag{5}$$

Here and in the sequel, we use the notation: $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa'$. If $\delta_i^* = 1$ for all i, then (5) reduces to the ordinary least-squares estimator.

Remark. The above idea of weighting the complete observations by their inversed probabilities of inclusion was originated by Horvitz and Thompson (1952) in the context of sample surveys. The adaptation of this idea to the setting of censored survival data was initially considered by Koul *et al.* (1981), and later on by Robins and Rotnitzky (1992) and Lin and Ying (1993). Recently, Zhao and Tsiatis (1997) applied this idea to the problem of quality adjusted survival time.

We show in the Appendix that $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ converges in distribution to a p-variate zero-mean normal random vector with a covariance matrix which can be consistently estimated by $\hat{A}^{-1}\hat{B}^{-1}\hat{A}^{-1}$, where

$$\hat{A} = n^{-1}\sum_{i=1}^{n}Z_i^{\otimes 2}, \tag{6}$$

$$\hat{B} = n^{-1}\sum_{i=1}^{n}\left[\frac{\delta_i^*(Y_i - \hat{\beta}' Z_i)Z_i}{\hat{G}(T_i^*)} + \bar{\delta}_i Q(X_i) - \sum_{j=1}^{n}\frac{\bar{\delta}_j I(X_j \le X_i)Q(X_j)}{\sum_{l=1}^{n} I(X_l \ge X_j)}\right]^{\otimes 2}, \tag{7}$$

and

$$Q(t) = \sum_{i=1}^{n}\frac{I(T_i^* > t)\delta_i^*(Y_i - \hat{\beta}' Z_i)Z_i}{\hat{G}(T_i^*)}\Bigg/\sum_{j=1}^{n} I(X_j \ge t). \tag{8}$$

Note that neither $\hat{\beta}$ nor its covariance matrix estimator involves the cost data from the subjects whose survival times are censored before τ.

2.2. Covariate-dependent censoring

It is possible to allow censoring to depend on measured covariates. If the covariates are all discrete with a limited number of values, then one may stratify the sample according to the covariate values and replace $\hat{G}$ in Section 2.1 with the stratum-specific Kaplan–Meier estimators; otherwise, it is convenient to formulate the effects of covariates on censoring through the proportional hazards model. Both of these two situations are encompassed by the following stratified proportional hazards model (Cox, 1972)

$$\lambda(t|V, W) = e^{\gamma' W(t)} \lambda_V(t), \tag{9}$$

where V represents the stratification variables and W the rest of the covariates, $\lambda(t|V, W)$ is the conditional hazard function of C given V and W, $\lambda_V(\cdot)$ is an unspecified baseline hazard function for stratum V, and γ is a set of unknown regression parameters. Naturally, V and W may include part of Z. We assume that C is independent of all other random variables conditional on (V, W). Without this assumption, there would be a non-identifiability problem.

Let V_i and W_i be the values of V and W for the ith subject. The parameters in model (9) are estimated from the data $(X_i, \bar{\delta}_i, V_i, W_i)$ $(i = 1, \ldots, n)$ by the partial likelihood principle (Cox, 1975). The validity of the model can be verified by a number of existing methods; see Lin *et al.* (1993). Under model (9), the probability that Y_i is known, i.e. $\delta_i^* = 1$, is estimated by

$$\hat{G}(T_i^*|V_i, W_i) = \exp\left\{ -\sum_{j=1}^{n} \frac{\bar{\delta}_j I(V_j = V_i, X_j < T_i^*) e^{\hat{\gamma}' W_i(X_j)}}{S^{(0)}(X_j; V_i, \hat{\gamma})} \right\},$$

where $\hat{\gamma}$ is the maximum partial likelihood estimator of γ (Lin *et al.* 1994). Here and in the sequel, we adopt the notation:

$$S^{(\rho)}(t; V, \gamma) = \sum_{i=1}^{n} I(V_i = V, X_i \geq t)\, e^{\gamma' W_i(t)} W_i^{\otimes \rho}(t), \qquad \rho = 0, 1, 2.$$

The replacement of $\hat{G}(T_i^*)$ in (4) with $\hat{G}(T_i^*|V_i, W_i)$ yields the estimating equation

$$\sum_{i=1}^{n} \frac{\delta_i^*}{\hat{G}(T_i^*|V_i, W_i)} (Y_i - \beta' Z_i) Z_i = 0, \tag{10}$$

which again has an explicit solution

$$\hat{\beta} = \left\{ \sum_{i=1}^{n} \frac{\delta_i^*}{\hat{G}(T_i^*|V_i, W_i)} Z_i^{\otimes 2} \right\}^{-1} \sum_{i=1}^{n} \frac{\delta_i^*}{\hat{G}(T_i^*|V_i, W_i)} Y_i Z_i. \tag{11}$$

The asymptotic properties of $\hat{\beta}$ are described at the end of the next subsection.

2.3. Multiple time intervals

The estimators developed in the previous two subsections may be inefficient when there is heavy censoring because the cost data from the subjects whose survival times are censored prior to τ are not used at all. In many applications, including the ovarian cancer study to be presented in Section 4, the costs are recorded in certain time intervals, e.g. every month or every year, in which case it is possible to obtain

more efficient estimators. The availability of the cost data in multiple time intervals also offers the opportunity to assess how the effects of covariates change over time. The idea of partitioning the entire study period into several time intervals to improve efficiency was initially explored by Lin *et al.* (1997) in the one-sample case, though they handled censoring with a very different approach.

Suppose that the entire time period of interest $[0, \tau]$ is divided into K intervals by $0 \equiv t_0 < t_1 < \cdots < t_{K-1} < t_K \equiv \tau$. For the ith subject, let Y_{ki} denote the cost incurred over the time interval $(t_{k-1}, t_k]$. The initial cost at $t = 0$ is included in the first time interval. We specify a linear regression model for each of the K intervals:

$$Y_{ki} = \beta_k' Z_i + \epsilon_{ki}, \qquad k = 1, \ldots, K; \qquad i = 1, \ldots, n, \tag{12}$$

where β_k $(k = 1, \ldots, K)$ are $p \times 1$ vectors of unknown regression parameters, and the error terms ϵ_{ki}s are assumed to be independent among different subjects but allowed to be correlated within the same subject. This is a semiparametric marginal model for repeated measures in that only the marginal mean structure is modelled. By summing both sides of (12) over k, we obtain

$$Y_i = \beta' Z_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $Y_i = \sum_{k=1}^K Y_{ki}$, $\beta = \sum_{k=1}^K \beta_k$, and $\epsilon_i = \sum_{k=1}^K \epsilon_{ki}$. This is the same as model (1). Neither model (1) nor (12) requires specification of the relationship between survival time and cost.

Define $T_{ki}^* = \min(T_i, t_k)$, and $\delta_{ki}^* = I(C_i \geq T_{ki}^*)$. Clearly, Y_{ki} is known if and only if $\delta_{ki}^* = 1$. Mimicking equation (10), we propose the following estimating equation for β_k

$$\sum_{i=1}^n \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} (Y_{ki} - \beta_k' Z_i) Z_i = 0, \tag{13}$$

which has an explicit solution

$$\hat{\beta}_k = \left\{ \sum_{i=1}^n \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} Z_i^{\otimes 2} \right\}^{-1} \sum_{i=1}^n \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} Y_{ki} Z_i. \tag{14}$$

The corresponding estimator of β is $\hat{\beta} = \sum_{k=1}^K \hat{\beta}_k$, or

$$\hat{\beta} = \sum_{k=1}^K \left[\left\{ \sum_{i=1}^n \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} Z_i^{\otimes 2} \right\}^{-1} \sum_{i=1}^n \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} Y_{ki} Z_i \right]. \tag{15}$$

This estimator shares the spirit of the generalized estimating equations for repeated measures (Liang and Zeger, 1986).

A subject will contribute a non-zero term to the left-hand side of (13) if he/she has complete cost data in the kth interval. In other words, a subject whose survival time is censored in the $(k+1)$th interval contributes his/her cost data from the first k time intervals to the estimation of β. By contrast, a subject whose survival time is censored before τ does not contribute any cost information to (10). Thus, (15) is expected to be more efficient than (11).

We prove in the Appendix that $n^{\frac{1}{2}}(\hat{\beta}_1 - \beta_1, \ldots, \hat{\beta}_K - \beta_K)$ is asymptotically zero-mean normal and that the limiting covariance matrix between $n^{\frac{1}{2}}(\hat{\beta}_k - \beta_k)$ and $n^{\frac{1}{2}}(\hat{\beta}_l - \beta_l)$ $(k, l = 1, \ldots, K)$ can be consistently estimated by $\hat{A}^{-1} \hat{B}_{kl} \hat{A}^{-1}$, where $\hat{A}$ is given in (6), $\hat{B}_{kl} = n^{-1} \sum_{i=1}^n \hat{\xi}_{ki} \hat{\xi}_{li}'$, and

$$\hat{\xi}_{ki} = \frac{\delta_{ki}^* (Y_{ki} - \hat{\beta}_k' Z_i) Z_i}{\hat{G}(T_{ki}^* | V_i, W_i)} + \bar{\delta}_i D_{ki}(X_i) - \sum_{j=1}^n \frac{\bar{\delta}_j I(V_j = V_i, X_j \leq X_i) e^{\hat{\gamma}' W_i(X_j)} D_{ki}(X_j)}{S^{(0)}(X_j; V_i, \hat{\gamma})}. \tag{16}$$

In expression (16),

$$D_{ki}(t) = Q_k(t; V_i) + R_k \hat{\Omega}^{-1}\{W_i(t) - S^{(1)}(t; V_i, \hat{\gamma})/S^{(0)}(t; V_i, \hat{\gamma})\},$$

where

$$Q_k(t; V) = \sum_{i=1}^{n} \frac{I(V_i = V, T_{ki}^* > t) e^{\hat{\gamma}' W_i(t)} \delta_{ki}^* (Y_{ki} - \hat{\beta}_k' Z_i) Z_i}{\hat{G}(T_{ki}^* | V_i, W_i) S^{(0)}(t; V_i, \hat{\gamma})}, \tag{17}$$

$$R_k = n^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^* (Y_{ki} - \hat{\beta}_k' Z_i) Z_i H'(T_{ki}^*; V_i, W_i)}{\hat{G}(T_{ki}^* | V_i, W_i)}, \tag{18}$$

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} \bar{\delta}_i \left\{ \frac{S^{(2)}(X_i; V_i, \hat{\gamma})}{S^{(0)}(X_i; V_i, \hat{\gamma})} - \frac{S^{(1)}(X_i; V_i, \hat{\gamma})^{\otimes 2}}{S^{(0)}(X_i; V_i, \hat{\gamma})^2} \right\},$$

and

$$H(t; V, W) = \sum_{i=1}^{n} \bar{\delta}_i I(V_i = V, X_i < t) e^{\hat{\gamma}' W(X_i)} \left\{ W(X_i) - \frac{S^{(1)}(X_i; V_i, \hat{\gamma})}{S^{(0)}(X_i; V_i, \hat{\gamma})} \right\} \Big/ S^{(0)}(X_i; V_i, \hat{\gamma}).$$

It follows that $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ is asymptotically zero-mean normal with a covariance matrix which can be consistently estimated by $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$, where $\hat{B} = \sum_{k=1}^{K}\sum_{l=1}^{K} \hat{B}_{kl}$.

The above asymptotic results also apply to the estimator given in (11) since Section 2.2 is a special case of this subsection with $K = 1$. To calculate the variance estimator for (11), we set $K = 1$ and replace T_{ki}^*, δ_{ki}^*, Y_{ki} and $\hat{\beta}_k$ in (16)–(18) with T_i^*, δ_i^*, Y_i and $\hat{\beta}$, respectively.

If censoring occurs in a completely random fashion, we set $V_i = 1$ and $W_i = 0$ $(i = 1, \ldots, n)$, and replace $\hat{G}(T_{ki}^* | V_i, W_i)$ in (13)–(15) with the Kaplan–Meier estimator $\hat{G}(T_{ki}^*)$ $(k = 1, \ldots, K;\ i = 1, \ldots, n)$. Specifically, (15) becomes

$$\hat{\beta} = \sum_{k=1}^{K} \left[\left\{ \sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^*)} Z_i^{\otimes 2} \right\}^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^*)} Y_{ki} Z_i \right]. \tag{19}$$

The aforementioned asymptotic results continue to hold, but with

$$\hat{\xi}_{ki} = \frac{\delta_{ki}^* (Y_{ki} - \hat{\beta}_k' Z_i) Z_i}{\hat{G}(T_{ki}^*)} + \bar{\delta}_i Q_k(X_i) - \sum_{j=1}^{n} \frac{\bar{\delta}_j I(X_j \le X_i) Q_k(X_j)}{\sum_{l=1}^{n} I(X_l \ge X_j)},$$

where

$$Q_k(t) = \sum_{i=1}^{n} \frac{I(T_{ki}^* > t) \delta_{ki}^* (Y_{ki} - \hat{\beta}_k' Z_i) Z_i}{\hat{G}(T_{ki}^*)} \Big/ \sum_{j=1}^{n} I(X_j \ge t).$$

If we further set $K = 1$, then $\hat{B}_{kl}$ and $\hat{B}$ reduce to (7).

3. SIMULATION STUDIES

Monte Carlo simulation was conducted to assess the operating characteristics of the proposed methods in practical settings. The simulation scheme was a modification of that of Lin *et al.* (1997). To be specific, the survival times were generated from two distributions: uniform on (0, 10) years and exponential with

Table 1. *Summary statistics for the simulation studies*

			Single interval ($K = 1$)				Multiple intervals ($K = 10$)			
T	c	n	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
Uniform	20	100	0.0037	0.405	0.392	0.941	0.0005	0.378	0.368	0.942
		200	−0.0003	0.286	0.280	0.942	−0.0006	0.267	0.263	0.943
		500	−0.0031	0.180	0.178	0.948	−0.0024	0.168	0.167	0.948
	15	100	0.0039	0.458	0.431	0.930	0.0025	0.403	0.388	0.939
		200	0.0011	0.321	0.310	0.935	−0.0002	0.285	0.278	0.940
		500	−0.0013	0.201	0.198	0.947	−0.0020	0.179	0.177	0.949
Exponential	20	100	0.0018	0.453	0.440	0.935	0.0023	0.409	0.400	0.940
		200	0.0050	0.316	0.314	0.946	0.0036	0.288	0.285	0.947
		500	−0.0015	0.200	0.199	0.946	−0.0011	0.181	0.181	0.950
	15	100	−0.0017	0.526	0.500	0.925	0.0007	0.435	0.422	0.936
		200	0.0065	0.368	0.358	0.937	0.0047	0.306	0.302	0.944
		500	−0.0014	0.228	0.227	0.948	−0.0015	0.192	0.192	0.949

Note: Bias is the mean of $\hat{\beta}$ minus β; SE is the standard error of $\hat{\beta}$; SEE is the mean of the standard error estimator for $\hat{\beta}$; CP is the coverage probability of the 95% confidence interval for β. Each entry is based on 10 000 replicates.

a mean of six years. The censoring times were generated from the uniform$(0, c)$ distribution. We set $c = 20$ and 15 years, resulting in censoring probabilities of approximately 25% and 35% under the uniform survival time distribution and 30% and 40% under the exponential distribution. The 10-year cost was considered to be the quantity of main interest.

We postulated U-shaped sample paths for the costs. Specifically, the entire time period of interest $[0, 10]$ was divided into 10 1-year intervals. Within each interval, there was a baseline cost of uniform $(0, 1)$. In addition, there were a diagnostic cost of uniform $(0, 1)$ at $t = 0$ and a uniform $(0, 1)$ cost in the final year of life. The covariate was a treatment indicator with $\frac{n}{2}$ subjects in each of the two groups. The treatment assignment was independent of the costs so that the regression coefficient for the treatment indicator is 0.

Table 1 summarizes the results of simulation studies on the estimation of treatment difference based on estimator (5) and estimator (19) with $K = 10$. Both estimators appear to be virtually unbiased. In general, the standard error estimators adequately reflect the true variations of the parameter estimators and the associated confidence intervals have reasonable coverage probabilities. Making use of the cost data in multiple time intervals not only enhances the efficiency of the estimation but also improves the accuracy of the asymptotic approximation in small samples.

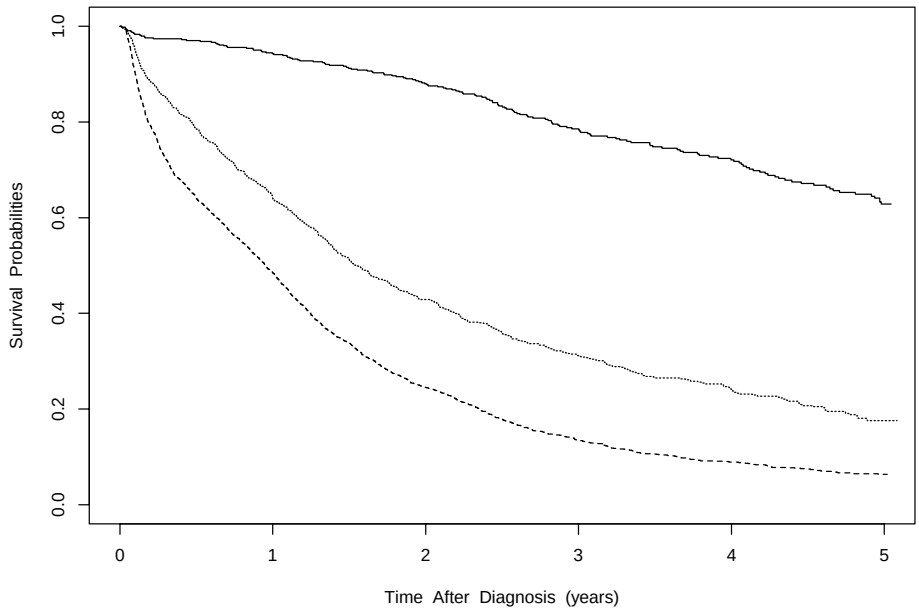

Fig. 1. Kaplan–Meier estimates of the survival probabilities for epithelial ovarian cancer patients: solid curve, local stage; dotted curve, regional stage; dashed curve, distant stage.

4. OVARIAN CANCER STUDY

In this section, we use the linked SEER–Medicare database (Potosky *et al*., 1993) to study the medical cost for epithelial ovarian cancer among the Medicare enrollees in the United States. We consider the 3550 Medicare beneficiaries over the age of 65 who were diagnosed with ovarian cancer from 1984 to 1989. Among them, 540, 836 and 2174 subjects were diagnosed with local, regional and distant stages, respectively. The data on mortality and monthly medical costs were collected during the period of 1984 to 1990. From a public health point of view, it is important to assess how the stage at diagnosis affects the future survival and medical cost.

The survival times and medical costs are censored on the patients who were still alive at the end of 1990. The censoring times as measured from the times of diagnosis vary substantially among the subjects as the diagnoses staggered over a period of 6 years. Because censoring was solely caused by the limited study duration, it is reasonable to assume that censoring is independent of all other random variables.

Figure 1 shows the Kaplan–Meier estimates of the survival probabilities separated by the three disease stages at diagnosis. Clearly, the patients with less aggressive disease have better survival experiences. The 5-year survival probabilities are approximately 70%, 20% and 10% for the local, regional and distant stages, respectively.

To evaluate how the stage of disease affects the 5-year post-diagnosis cost, we fit model (12) with $n = 3550$, $K = 60$ and three covariates: $Z_i = (1, 0, 0)'$, $(1, 1, 0)'$ or $(1, 0, 1)'$ if the ith patient was diagnosed with the local, regional or distant stages, respectively. By the definition of the covariates, the

Table 2. *Regression estimates for the 5-year cost in ovarian cancer*

(a) Using monthly costs

Parameter	Estimate	St. error	Est/SE	95% Confidence Interval
β_1^*	32229	1241	25.97	(29797, 34662)
$\beta_2^\dagger$	5972	1683	3.55	(2673, 9271)
$\beta_3^\ddagger$	4527	1402	3.23	(1779, 7275)

(b) Using 5-year costs

Parameter	Estimate	St. error	Est/SE	95% Confidence interval
β_1	33675	1689	19.93	(30365, 36986)
β_2	6515	2182	2.99	(2239, 10790)
β_3	3614	1832	1.97	(24, 7205)

(c) Naive complete-cases analysis

Parameter	Estimate	St. error	Est/SE	95% Confidence interval
β_1	34590	1640	21.09	(31375, 37805)
β_2	3393	1969	1.72	(−466, 7253)
β_3	736	1764	0.42	(−2723, 4194)

* Local stage.

† Difference of regional stage from local stage.

‡ Difference of distant stage from local stage.

first component of each β_k corresponds to the local stage, and the second and third components correspond to the differences of the regional and distant stages from the local stage. Table 2(a) summarizes the results for the overall regression parameters $\beta = (\beta_1, \beta_2, \beta_3)'$, while Figure 2 displays the cumulative costs for the three stages based on individual $\hat{\beta}_k$ $(k = 1, \ldots, 60)$.

As is evident from Figure 2, the medical cost for the local stage is lower than those of the other 2 stages in the first 2 years after diagnosis, but the opposite is true in later years. This phenomenon is mainly due to the fact that the local-stage patients survived longer. The average 5-year cost for a local-stage patient is slightly over \$32 000, while those of the regional- and distant-stage patients are about \$5 000 higher. The differences are statistically significant.

To demonstrate the efficiency gain of using multiple time intervals over using a single interval, we also apply the method of §2.1 to the total 5-year costs. The results are shown in Table 2(b). A comparison of Table 2(a) and (b) shows that the use of the monthly cost data yields considerable variance reduction.

The results in Table 2(c) are obtained by applying the ordinary least-squares method to the cases with complete data on the 5-year cost. Such an analysis is biased towards the costs of the patients with shorter survival times because longer survival times are more likely to be censored. The estimates in Table 2(c) differ appreciably from those of Table 2(a) and (b).

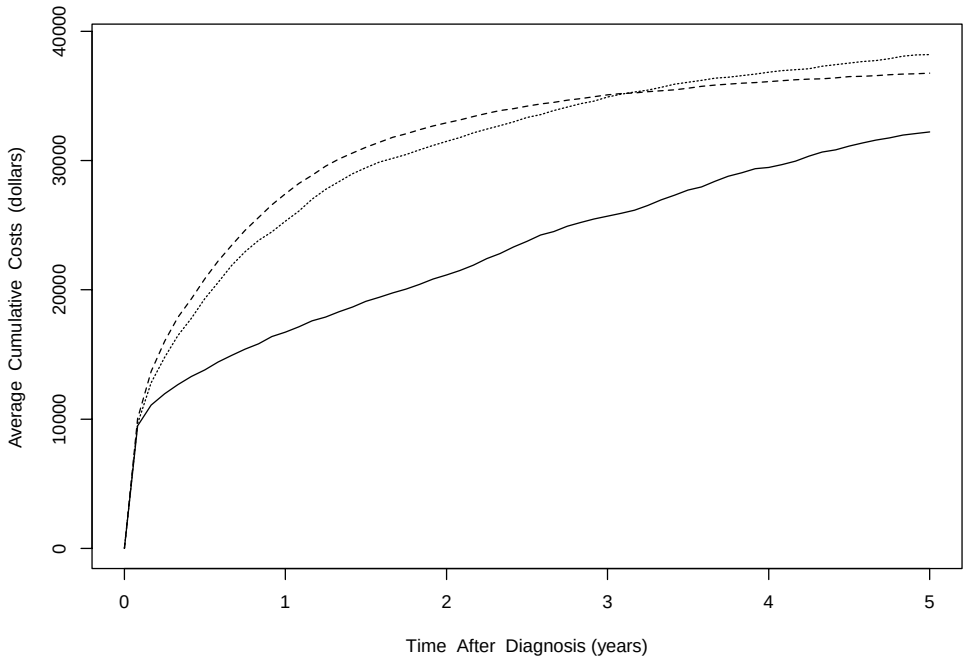

Fig. 2. Estimates of the average cumulative costs for epithelial ovarian cancer patients: solid curve, local stage; dotted curve, regional stage; dashed curve, distant stage.

5. REMARKS

It might seem natural to apply the existing regression methods in the survival analysis literature, such as those based on the Cox proportional hazards and accelerated failure time models, to censored medical costs by treating $(\tilde{Y}_i, \delta_i^*, Z_i)$ $(i = 1, \ldots, n)$ as censored survival data, where $\tilde{Y}_i$ is the cumulative cost at $\min(T_i^*, C_i)$, i.e. the minimum of the cumulative costs at T^* and C_i. As mentioned in Section 1, this approach is invalid because the cumulative cost at T^* is positively correlated with the cumulative cost at C even if T^* and C are independent. This phenomenon of dependent censoring is caused by the fact that the patients are heterogeneous such that those who accumulate cost at higher rates over time tend to generate higher cumulative costs at all time points, including T^* and C, as compared with those with lower accumulation rates.

The approach taken in this paper allows arbitrary censoring patterns, whereas that of Lin *et al.* (1997) requires censoring to occur only at the cut-points $t_1, \ldots, t_K$. In the one-sample case, the proposed estimators do not reduce to those of Lin *et al.* (1997). The main motivation for developing the regression methods is to handle a large number of continuous and discrete covariates. The ovarian cancer example given in Section 4 did not demonstrate the full power of the proposed regression methods because the available database does not contain any continuous covariates. The new methods are currently being applied to several ongoing clinical trials. The results of those applications will be reported in medical journals.

This paper focuses on the actual total medical cost, which acknowledges the fact that there is no further accumulation of cost after death. This quantity is highly important to public health and insurance

industries. Because longer survival time tends to be associated with higher medical cost, different intervention/prevention programs should be compared not only with respect to medical cost but also with respect to survival time. In fact, the cost-effectiveness of a new program relative to the standard is normally measured by the increase in mean medical cost divided by the increase in mean survival time (Patrick and Erickson, 1993).

ACKNOWLEDGEMENTS

The author is grateful to the referees for their careful and speedy reviews of this paper, and to Drs Ruth Etzioni, Paula Diehr and Sean Sullivan for helpful discussions on related topics. This research was supported by the National Institutes of Health.

APPENDIX

Proofs of asymptotic results

In this appendix, we prove the asymptotic theory stated in Section 2.3. The theory given in Section 2.2 is a special case with $K = 1$, while that of Section 2.1 is a further special case with $K = 1$ and $V_i = 1$ and $W_i = 0$ for all $i = 1, \ldots, n$.

The left-hand side of (13) can be written as $U_k(\beta_k) = U_{k1}(\beta_k) + U_{k2}(\beta_k)$, where

$$U_{k1}(\beta_k) = \sum_{i=1}^{n} \frac{\delta_{ki}^*}{G(T_{ki}^*|V_i, W_i)}(Y_{ki} - \beta_k' Z_i)Z_i, \tag{A1}$$

$$U_{k2}(\beta_k) = \sum_{i=1}^{n} \frac{G(T_{ki}^*|V_i, W_i) - \hat{G}(T_{ki}^*|V_i, W_i)}{G(T_{ki}^*|V_i, W_i)\,\hat{G}(T_{ki}^*|V_i, W_i)}\delta_{ki}^*(Y_{ki} - \beta_k' Z_i)Z_i.$$

Because $E(\delta_{ki}^*|V_i, W_i, Y_{ki}, Z_i, T_{ki}^*) = E(\delta_{ki}^*|V_i, W_i, T_{ki}^*) = G(T_{ki}^*|V_i, W_i)$, the term $U_{k1}(\beta_k)$ consists of n independent zero-mean random vectors.

By a slight extension of the results of Lin *et al.* (1994),

$$\frac{n^{\frac{1}{2}}\{G(t|V, W) - \hat{G}(t|V, W)\}}{G(t|V, W)} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^t \frac{I(V_i = V)\, e^{\gamma' W(x)}\, dM_i(x)}{s^{(0)}(x; V)}$$

$$+h'(t; V, W)\, \Omega^{-1}\, n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^\infty \{W_i(x) - \overline{w}(x; V_i)\}\, dM_i(x) + o_p(1),$$

where $s^{(\rho)}(t; V) = \lim_{n\to\infty} n^{-1} s^{(\rho)}(t; V, \gamma)$ $(\rho = 0, 1, 2)$, $\overline{w}(t; V) = s^{(1)}(t; V)/s^{(0)}(t; V)$,

$$h(t; V, W) = \int_0^t e^{\gamma' W(x)}\, \{W(x) - \overline{w}(x; V)\}\lambda_V(x)dx,$$

$$\Omega = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \bar{\delta}_i \left\{ \frac{s^{(2)}(X_i; V_i)}{s^{(0)}(X_i; V_i)} - \overline{w}^{\otimes 2}(X_i; V_i) \right\},$$

$$M_i(t) = \bar{\delta}_i I(X_i \le t) - \int_0^t I(X_i \ge x)\, e^{\gamma' W_i(x)} \lambda_{V_i}(x)dx.$$

Thus,

$$n^{-\frac{1}{2}} U_{k2}(\beta_k) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^\infty \tilde{Q}_k(t; V_i) dM_i(t)$$

$$+\tilde{R}_k \Omega^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^\infty \{W_i(t) - \overline{w}(t; V_i)\} dM_i(t) + o_p(1),$$

where

$$\tilde{Q}_k(t; V) = n^{-1} \sum_{i=1}^{n} \frac{I(V_i = V, T_{ki}^* > t) e^{\gamma' W_i(t)} \delta_{ki}^* (Y_{ki} - \beta_k' Z_i) Z_i}{\hat{G}(T_{ki}^* | V_i, W_i) s^{(0)}(t; V_i)},$$

$$\tilde{R}_k = n^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^* (Y_{ki} - \beta_k' Z_i) Z_i h'(T_{ki}^*; V_i, W_i)}{\hat{G}(T_{ki}^* | V_i, W_i)}.$$

The law of large numbers, together with the consistency of $\hat{G}$, implies that $\tilde{Q}_k(t; V)$ and $\tilde{R}_k$ converge to well-defined limits, say $q_k(t; V)$ and r_k. Therefore,

$$n^{-\frac{1}{2}} U_{k2}(\beta_k) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^\infty \left[q_k(t; V_i) + r_k \Omega^{-1} \{W_i(t) - \overline{w}(t; V_i)\} \right] dM_i(t) + o_p(1). \tag{A2}$$

Combining (A2) with (A1), we have $n^{-\frac{1}{2}} U_k(\beta_k) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \xi_{ki} + o_p(1)$, where

$$\xi_{ki} = \frac{\delta_{ki}^*}{G(T_{ki}^* | V_i, W_i)} (Y_{ki} - \beta_k' Z_i) Z_i + \int_0^\infty \left[q_k(t; V_i) + r_k \Omega^{-1} \{W_i(t) - \overline{w}(t; V_i)\} \right] dM_i(t).$$

Because $(\xi_{1i}, \ldots, \xi_{Ki})$ $(i = 1, \ldots, n)$ are n independent zero-mean random matrices, the multivariate central limit theorem implies that $n^{-\frac{1}{2}} \{U_1(\beta_1), \ldots, U_K(\beta_K)\}$ converges in distribution to a zero-mean normal random matrix. The limiting covariance matrix between $n^{-\frac{1}{2}} U_k(\beta_k)$ and $n^{-\frac{1}{2}} U_l(\beta_l)$ is $B_{kl} \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \xi_{ki} \xi_{li}'$ $(k, l = 1, \ldots, K)$.

By the Taylor series expansion, $n^{\frac{1}{2}} (\hat{\beta}_k - \beta_k) = \tilde{A}_k^{-1} n^{-\frac{1}{2}} U_k(\beta_k)$, where

$$\tilde{A}_k = n^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{G}(T_{ki}^* | V_i, W_i)} Z_i^{\otimes 2},$$

which converges in probability to $A \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} Z_i^{\otimes 2}$. It then follows from the aforementioned asymptotic normality of $n^{-\frac{1}{2}} \{U_1(\beta_1), \ldots, U_K(\beta_K)\}$ that $n^{\frac{1}{2}} (\hat{\beta}_1 - \beta_1, \ldots, \hat{\beta}_K - \beta_K)$ converges in distribution to a zero-mean normal random matrix and the limiting covariance matrix between $n^{\frac{1}{2}} (\hat{\beta}_k - \beta_k)$ and $n^{\frac{1}{2}} (\hat{\beta}_l - \beta_l)$ is $A^{-1} B_{kl} A^{-1}$. This convergence of distribution implies the consistency of $\hat{\beta}_k$ $(k = 1, \ldots, K)$. Replacing all the unknown quantities in ξ_{ki} with their respective sample estimators, we obtain $\hat{\xi}_{ki}$ given in (16). The consistency of $\hat{B}_{kl}$ for B_{kl} $(k, l = 1, \ldots, K)$ follows from the law of large numbers, together with the consistency of $\hat{G}$, $\hat{\gamma}$ and $\hat{\beta}_k$ $(k = 1, \ldots, K)$.

REFERENCES

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B **34**, 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

HORVITZ, D. G. AND THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

KOUL, H., SUSARLA, V. AND VAN RYZIN, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics* **9**, 1276–1288.

LIANG, K.-Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIN, D. Y., ETZIONI, R., FEUER, E. J. AND WAX, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53**, 419–434.

LIN, D. Y., FLEMING, T. R. AND WEI, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.

LIN, D. Y., WEI, L. J. AND YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

LIN, D. Y. AND YING, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80**, 573–581.

PATRICK, D. L. AND ERICKSON, P. (1993). *Health Status and Health Policy: Allocating Resources to Health Care*, pp. 52–53. New York: Oxford University Press.

POTOSKY, A. L., RILEY, G. F., LUBITZ, J. D., MENTNECH, R. M. AND KESSLER, L. G. (1993). Potential for cancer related health services research using a linked Medicare-tumor registry data base. *Medical Care* **31**, 732–747.

ROBINS, J. M. AND ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, Eds N. P. Jewell, K. Dietz and V. T. Farewell, pp. 297–331. Boston, MA: Birkhäuser.

ZHAO, H. AND TSIATIS, A. A. (1997). A consistent estimator for the distribution of quality-adjusted survival time. *Biometrika* **84**, 339–348.

[*Received May 14, 1999. Revised August 19, 1999*]