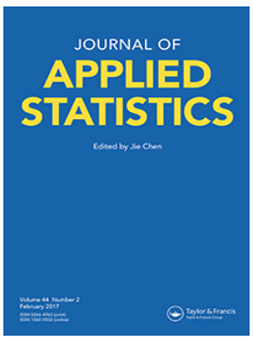

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

The weighted general linear model for longitudinal medical cost data – an application in colorectal cancer

Y. T. Hwang, C. H. Huang, W. L. Yeh & Y. D. Shen

To cite this article: Y. T. Hwang, C. H. Huang, W. L. Yeh & Y. D. Shen (2017) The weighted general linear model for longitudinal medical cost data – an application in colorectal cancer, Journal of Applied Statistics, 44:2, 288-307, DOI: 10.1080/02664763.2016.1169255

To link to this article: https://doi.org/10.1080/02664763.2016.1169255

JOURNAL OF APPLIED STATISTICS, 2017
VOL. 44, NO. 2, 288–307
http://dx.doi.org/10.1080/02664763.2016.1169255

Taylor & Francis
Taylor & Francis Group

The weighted general linear model for longitudinal medical cost data – an application in colorectal cancer

Y. T. Hwang[a], C. H. Huang[a], W. L. Yeh[a] and Y. D. Shen[b]

[a]Department of Statistics, National Taipei University, Taipei, Taiwan; [b]Department of Ophthalmology, Taipei Medical University, Taipei, Taiwan

ABSTRACT

Identifying cost-effective decisions that can take into account of medical cost and health outcome is an important issue under very limited resources. Analyzing medical costs has been challenged owing to skewness of cost distributions, heterogeneity across samples and censoring. When censoring is due to administrative reasons, the total cost might be related to the survival time since longer survivals are likely to be censored and the corresponding total cost will be censored as well. This paper uses the general linear model for the longitudinal data to model the repeated medical cost data and the weighted estimating equation is used to find more accurate estimates for the parameter. Furthermore, the asymptotic properties for the proposed model are discussed. Simulations are used to evaluate the performance of estimators under various scenarios. Finally, the proposed model is implemented on the data extracted from National Health Insurance database for patients with the colorectal cancer.

ARTICLE HISTORY
Received 20 December 2013
Accepted 18 March 2016

KEYWORDS
General linear model; inverse probability weighted method; medical cost; longitudinaldata; proportional hazards model

AMS SUBJECT CLASSIFICATION
C23; C24

1. Introduction

Increasing cost in health care becomes great financial burdens in many industrialized countries. Identifying cost-effective decisions that can take into account of medical cost and health outcome is an important issue under very limited resources. Analyzing medical costs has been challenged owing to skewness of cost distributions, heterogeneity across samples and censoring [3].

When the sample is censored exogenously, the total cost over a period such as the survival time (T) could be assumed independent of T. Nonetheless, owing to some administrative reasons, the study might be terminated at a pre-specified date τ. In turn, the total cost might be related to T since longer survivals are likely to be censored and the corresponding total cost will be censored as well. The estimates constructed using only the uncensored cases are likely to be biased [3,16,19].

Estimators for the total cost that adjust for the effect of censoring have been discussed. The standard survival analysis techniques such as the Kaplan–Meier curve and the log rank test can be applied to handle the censoring problem in estimating the total cost [9,10,13,26].

CONTACT Y. T. Hwang hwangyt@gm.ntpu.edu.tw Department of Statistics, National Taipei University, 67, Sec. 3 Ming Sheng E. Rd, Taipei, Taiwan

However, this implementation might be incorrect since the total cost can be positively correlated with the survival time [17,19]. An alternative way is to find proper weights for observations when constructing the estimator for the total cost. Dividing the entire study period of interest into several intervals, Lin *et al.* [19] derived 4 weighted sample mean estimates, where where the weight can be either the Kaplan–Meier estimator for the probability of dying or for the probability of surviving to the start of each interval conditioning on surviving to the start of the interval. On the contrary, using only the complete case for total costs, Bang and Tsiatis [2] proposed a weighted average total cost where the weight is the inverse of probability of a subject not being censored. The former estimators are more efficient when costs in subintervals are available, while the later estimator is recommended when only total costs are available on each patient [25,27].

To assess the effect of covariates, Baser *et al.*[3] and Lin [16–18] proposed regression models. The construction of the regression models depends also upon the availability of data. When only the total cost is available, Lin [17] used the proportional cox model to model the censored total cost, where the survival time is replaced by the total cost. Besides censoring, the total cost might be right-skewed and heterogoeneous and have a large proportion of zero values. Liu *et al.* [22] proposed a flexible two-part random effect model which includes the logistic model and the generalized gamma distribution and is used to model the probability of positive cost and the medical cost, where the distribution of the medical cost is assumed to be generalized gamma distributions. Chen *et al.* [6] suggested using the generalized linear mixed model (GLMM), where the working variance and covariance matrix is assumed to be a smooth function of the mean function known as the penalized spline function and is used to model the heteroscedasticity. Besides using GLMM, Locatelli and Marazzi [24] added the duration in the GLMM and the jointly estimated the coefficients with the parametric survival model. Furthermore, to be able to adjust for the incompleteness, the total cost estimate proposed by Lin *et al.* [19] is re-formulated by plugging in the estimated coefficients from the parametric models.

When the cost is collected per subperiod over the entire study period of interest, Lin [16] specified a linear regression model for each subperiod using the same predictors, where the error terms from the same subject are assumed to be correlated, and the average total cost was then estimated by summing the cost for each subinterval and the unknown parameters are estimated by the generalized estimating equation that is weighted by the inverse of the probability of a subject not being censored. Following similar model constructions and estimation techniques, Lin [16] extended the linear model to the generalized linear model to model a more complex relationship between cumulative costs and time. Instead of modeling a regression model for each subinterval, Baser *et al.* [3] treated costs for the subperiods as a panel data and proposed using the random intercept model to take into account of possible correlation between subperiods, where the observation is weighted by the inverse of the probability of a subject not being censored.

The variance and covariance structure in the random intercept model is a compound symmetric structure which assumes that constant variances for the panel and constant correlations among panels. To have a more general association among the panel, in this paper, we suggests using the general linear model for the longitudinal data to model the medical cost and the estimates of the parameters are derived based on the weighted estimating equation, where the weight equals again the probability of a patient being uncensored. The asymptotic properties of the corresponding estimators for the proposed model are derived

in Section 2.2. Section 3 evaluates the performance of proposed parameter estimators using Monte Carlo simulation. Section 4 illustrates the usage of these estimators using a real data obtained from National Health Insurance (NHI) Research Database in Taiwan. Discussions are given in Section 5.

2. Methods

2.1. Preliminaries

Let X_i and C_i denote the survival time and censoring time of the ith patient, $i = 1, 2, \ldots, n$, and be assumed to be independent. For the right-censored data, we are only able to observe the smallest time of X_i and C_i. Let the observable data be denoted as $T_i = \min(X_i, C_i)$ and $\delta_i = I[X_i \leq C_i]$, where $I[A]$ denotes the indicator of the event A.

The research duration is assumed to be fixed and be denoted as $[0, \tau)$, where τ is a pre-specified constant. Let the duration be partitioned into K intervals, where the kth interval is denoted as $[t_k, t_{k+1})$ and $t_1 = 0$, $t_{K+1} = \tau$. Let $T_i^* = \min(X_i, \tau)$, $T_{ik}^* = \min(T_i^*, t_k)$, and $\delta_{ik}^* = I[C_i \geq T_{ik}^*]$, $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n$. Let Y_{ik} denote the accumulative cost in the kth interval for the ith subject. Owing to censoring, some $\tilde{Y}_{ik}$ may be incompletely observed. Obviously, Y_{ik} is completely observed if and only if $\delta_{ik}^* = 1$. However, when $T_i < \tau$ and $T_i \in (t_k, t_{k+1})$ for some k, then the cost beyond the kth interval is not available. Thus, adopting the cost setting in [16], let $\tilde{Y}_{ij}$ denote the observable cost in the jth interval for the ith patient, where $\tilde{Y}_{ij} = 0$ for all $j > k$ if $\delta_{ij}^* = 1$, $\tilde{Y}_{ij} = \cdot$ for all $j > k$ if $\delta_{ij}^* = 0$ and otherwise $\tilde{Y}_{ij} = Y_{ij}$.

Suppose the medical cost is completely observable. In turn, the linear model for the repeated measurements can be used to identify potential significant covariates and be defined as

$$Y_{ik} = \beta_0 + \beta_1 Z_{ik,1} + \cdots + \beta_{p-1} Z_{ik,p-1} + \epsilon_{ik}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K, \tag{1}$$

where $Z_{ik,j}, j = 1, 2, \ldots, p-1$ are the covariate observed at the kth interval for the ith subject. For simplicity, let $\boldsymbol{Y}_i' = (Y_{i1}, Y_{i2}, \ldots, Y_{iK})$ denote the medical cost for the ith subject and the corresponding covariates be denoted as

$$\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{Z}_{i1}' \\ \boldsymbol{Z}_{i2}' \\ \vdots \\ \boldsymbol{Z}_{iK}' \end{pmatrix} = \begin{pmatrix} 1 & Z_{i11} & Z_{i12} & \cdots & Z_{i1,p-1} \\ 1 & Z_{i21} & Z_{i22} & \cdots & Z_{i2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{iK1} & Z_{iK2} & \cdots & Z_{iK,p-1} \end{pmatrix},$$

where $\boldsymbol{Z}_{ij}' = (1, Z_{ij1}, Z_{ij2}, \ldots, Z_{ij,p-1})$. In turn, Equation (1) can be expressed as

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ and $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iK})$ is the measurement error. The distribution of $\boldsymbol{\epsilon}_i$ is assumed to be multivariate normal with mean $\mathbf{0}$ and variance–covariance matrix $\boldsymbol{\Sigma}$.

Given the observed cost data $\boldsymbol{y}_i$, $i = 1, 2, \ldots, n$, the log-likelihood function is

$$l = -\frac{Kn}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})\right\}. \tag{2}$$

Differentiating l with respect to $\boldsymbol{\beta}$ yields

$$\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta}) = 0. \tag{3}$$

The solution of $\partial l/\partial\boldsymbol{\beta} = 0$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$. When $\boldsymbol{\Sigma}$ is completely specified, a closed form solution for the estimator of $\boldsymbol{\beta}$ can be obtained as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i\right). \tag{4}$$

If $\boldsymbol{\Sigma}$ is unknown, the estimator $\hat{\boldsymbol{\beta}}$ and an estimator of $\boldsymbol{\Sigma}$ can be also obtained from the likelihood function (see [11]). Since the MLE of $\boldsymbol{\Sigma}$ is a biased estimator, Fitzmaurice *et al.* [11] suggested using the restricted likelihood function

$$l = -\frac{Kn}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})\right\} - \frac{1}{2}\log\left|\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\Sigma}\boldsymbol{Z}_i\right|$$

to obtain the estimator of $\boldsymbol{\Sigma}$, which is denoted as $\hat{\boldsymbol{\Sigma}}$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}\left(\boldsymbol{\beta}, \left(\sum_{i=1}^{n}\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i\right)^{-1}\right)$$

(see [14]). The estimated covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\boldsymbol{Z}_i'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Z}_i\right)^{-1}.$$

When $\boldsymbol{\Sigma}$ is completely unspecified and the number of partitions increases, the number of parameters in $\boldsymbol{\Sigma}$ can increase dramatically. For the repeated measures, there exist certain characteristics in the association among the repeated measures. In turn, $\boldsymbol{\Sigma}$ may be characterized into some specified parametric structures such as the compound symmetry covariance structure, first-order autoregressive covariance structure, Toeplitz covariance structure, etc. (see [11]). An appropriate structure is often determined by the likelihood ratio statistics or Akaike information criterion (AIC).

2.2. Proposed models

Considering only the random intercept model, which is a special case of Equation (1), Baser *et al.* [3] suggested using the inverse probability weight method (IPWM) to modify

the influence of censoring. Analogously, based on the IPWM, the following derives the estimators for Equation (1).

To derive the weighted estimators for Equation (1), the estimates using only the observable data are discussed first as follows. Let the design matrix for the observable cost be denoted as $\tilde{\boldsymbol{Z}}_i = \boldsymbol{S}_i \boldsymbol{Z}_i$, where $\boldsymbol{S}_i$ is a $K \times K$ diagonal matrix and the kth diagonal element equals 1 when the cost for the kth interval for the ith subject is observable and 0 otherwise. Let $\tilde{\boldsymbol{y}}_i' = (\tilde{y}_{i1}, \tilde{y}_{i2}, \ldots, \tilde{y}_{iK})$ denote the observed medical cost for the ith subject. Based on only observable data, the model defined in Equation (1) becomes

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{Z}}_i \boldsymbol{\beta} + \tilde{\boldsymbol{e}}_i, \quad i = 1, \ldots, n, \tag{5}$$

where $\tilde{\boldsymbol{e}}_i$ has the multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix $\tilde{\boldsymbol{\Sigma}}$. To use this model, censoring is assumed to be not associated with the total cost.

By using Equation (3), the estimator of $\boldsymbol{\beta}$ is obtained and denoted as

$$\hat{\boldsymbol{\beta}}^C = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \hat{\tilde{\boldsymbol{\Sigma}}}^{-1} \tilde{\boldsymbol{Z}}_i \right)^{-1} \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \hat{\tilde{\boldsymbol{\Sigma}}}^{-1} \tilde{\boldsymbol{y}}_i \right) \tag{6}$$

and $\hat{\tilde{\boldsymbol{\Sigma}}}$. The asymptotic distribution of $\hat{\boldsymbol{\beta}}^C$ is the multivariate normal with mean $\boldsymbol{\beta}$ and an estimated asymptotic variance covariance matrix $\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^C) = (\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \hat{\tilde{\boldsymbol{\Sigma}}}^{-1} \tilde{\boldsymbol{Z}}_i)^{-1}$.

To take into account of censoring, the estimating equation is modified as follows. Since $E[\delta_{ik}^*/G(T_{ik}^*)] = 1$, Equation (3) is modified as

$$\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{Z}}_i \boldsymbol{\beta}) = \mathbf{0}, \tag{7}$$

where $\boldsymbol{W}_i$ is a $K \times K$ diagonal matrix with the kth diagonal element being $\delta_{ik}^*/\hat{G}(T_{ik}^*)$ and $\hat{G}(t)$ is an estimated censoring distribution as defined in Equation (10).

To derive the estimator of $\boldsymbol{\beta}$, we first assume $\tilde{\boldsymbol{\Sigma}}$ is known. An explicit solution of the estimator of $\boldsymbol{\beta}$ for Equation (7) can be obtained as

$$\hat{\boldsymbol{\beta}}^W = \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{Z}}_i \right)^{-1} \left(\sum_{i=1}^{n} \tilde{\boldsymbol{Z}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{y}}_i \right). \tag{8}$$

When $\tilde{\boldsymbol{\Sigma}}^{-1}$ is unknown, Gourieroux *et al.* [12] and Liang and Zeger [15] have shown that the resulting equation is asymptotically as efficient as if $\tilde{\boldsymbol{\Sigma}}^{-1}$ is known. Let $\hat{\tilde{\boldsymbol{\Sigma}}}$ be an estimator of $\tilde{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{B}}$ be an estimator of $\boldsymbol{B}$ as defined in Equation (13), which is obtained by replacing the unknown quantities in Equation (13) by the following estimators: $\hat{w}_{1i,j} = \delta_{ij}^*/\hat{G}(T_{ij}^* | \boldsymbol{V}_i)$,

$$\hat{w}_{2i,j} = \frac{\delta_{ij}^*}{n} \sum_{k=1}^{n} \frac{\delta_k^C I[T_k \le T_{ij}^*] \, \mathrm{e}^{\hat{\boldsymbol{\gamma}} \boldsymbol{V}_i(T_{ij}^*)}}{S^{(0)}(T_k, \hat{\boldsymbol{\gamma}})}$$

$$- \frac{\delta_{ij}^*}{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\delta_l^C I[T_k \ge T_l] I[T_l \le T_{ij}^*] (\mathrm{e}^{\hat{\boldsymbol{\gamma}} \boldsymbol{V}_i(T_l)})^2}{n (S^{(0)}(T_l, \hat{\boldsymbol{\gamma}}))^2} + \hat{h}'(T_{ij}^*; \boldsymbol{V}_i) \hat{\Lambda}^{-1} \frac{\delta_{ij}^*}{n} \sum_{k=1}^{n}$$

$$\times \left\{ \delta_k^C I[T_k \le T_{ij}^*](\boldsymbol{V}_i(T_k) - \bar{S}(T_k, \hat{\boldsymbol{\gamma}})) \left[1 - \frac{e^{\hat{\boldsymbol{\gamma}} \boldsymbol{V}_i(T_k)} \sum_{l=1}^n I[T_l \ge T_k]}{nS^{(0)}(T_k, \hat{\boldsymbol{\gamma}})} \right] \right\},$$

$$\hat{h}'(T_{ij}^*; \boldsymbol{V}_i) = \frac{\delta_i^C (e^{\boldsymbol{\gamma} \boldsymbol{V}_i(T_{ij}^*)})^2 (\boldsymbol{V}_i(T_{ij}^*) - \bar{S}(T_{ij}^*, \hat{\boldsymbol{\gamma}}))}{nS^{(0)}(T_{ij}^*, \hat{\boldsymbol{\gamma}})},$$

$$\hat{\Lambda} = \frac{1}{n} \sum_{k=1}^n \delta_k^C \left[\frac{S^{(2)}(T_k, \hat{\boldsymbol{\gamma}})}{S^{(0)}(T_k, \hat{\boldsymbol{\gamma}})} - (\bar{S}(T_k, \hat{\boldsymbol{\gamma}}))^{\otimes 2} \right].$$

Also, let $\hat{\boldsymbol{A}}_n$ denote an estimator of $\boldsymbol{A}_n$ as defined in Equation (A6), which is obtained by replacing $\boldsymbol{\Sigma}$ by $\hat{\hat{\boldsymbol{\Sigma}}}$ in Equation (A6). Appendix 2 shows that $n^{1/2}(\hat{\boldsymbol{\beta}}^W - \boldsymbol{\beta})$ converges in distribution to a p-variate zero-mean normal random vector with a covariance matrix that can be consistently estimated by $\hat{\boldsymbol{A}}_n^{-1} \hat{\boldsymbol{B}} \hat{\boldsymbol{A}}_n^{-1}$.

3. Simulations

The performance of model estimates is evaluated by Monte Carlo simulations. Both the survival time and the yearly cost are generated. The settings for generating the survival time is taken from [16,19]. The survival time is generated from two distributions. The first distribution is the exponential distribution with mean $\mu = 6$ and the second distribution assumes the Weibull distribution with shape $= 2$ and scale $= \sqrt{6}$. The censoring time is generated only from the exponential distribution with mean c, where c is determined according to the probability of censoring. Three censoring situations are considered. Specifically, under the exponential survival assumption, c equals 25, 15, 8, which correspond to the probability of censoring being 27%, 36% and 45%. Under the Weibull survival assumption, c equals 10, 7, 5 and results in the probability of censoring being 29%, 40% and 55%.

To have a mean response curve similar to the pattern displayed in Figure 5 , the piecewise linear model is considered and defined as

$$Y_{ik} = \beta_0 + \beta_1 Z_i + \beta_2 t_{ik} + \beta_3 (t_{ik} - 2)_+ + u_i + e_{ik}, \quad i = 1, 2, \ldots, n; \ \ k = 1, 2, \ldots, 10,$$

where Z_i is a dummy variable for the treatment, $(t)_+ = \max(0, t)$, u_i is the random normal intercept with mean 0 and variance σ_u^2 and e_{ik} is the random error having a normal distribution with mean 0 and variance $\sigma_e^2 = 1$. Assume $\sigma_u^2 = 0.25, 1, 4$ and $\beta_0 = 0.5$, $\beta_1 = 1$ and $\beta_2 = -1$, $\beta_3 = 1.5$. Also, for a random intercept model, there exists an intraclass correlation coefficient (ICC) among panels, where ICC $= \rho = \sigma_u^2/(\sigma_e^2 + \sigma_u^2)$ and under the preceding settings, ρ equals 0.2, 0.5 and 0.8. The variance and covariance matrix is a compound symmetric structure, that is,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_e^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_e^2 + \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_e^2 + \sigma_u^2 \end{bmatrix}.$$

The detailed simulation setting for cost is described in Tables A1 and A2.

The performance of each estimator is evaluated based on the bias, the standard error of estimator (SSE), the mean of standard error of estimator (SEE) and the 95% coverage probability (CP), which are computed from 10,000 simulated samples. Estimates of models are obtained for four different methods:

Method I: $\mathbf{\Sigma}$ is a diagonal matrix and no adjustment is set for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}^{\mathrm{IC}}$.
Method II: $\mathbf{\Sigma}$ is a diagonal matrix and IPWM adjustment is applied for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}^{\mathrm{IW}}$.
Method III: Dependence is considered and no adjustment is set for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}^{\mathrm{C}}$.
Method IV: Dependence is considered and IPWM adjustment is applied for the censoring bias. The estimate of β_1 is denoted as $\hat{\beta}^{\mathrm{W}}$.

Figure 1(a) displays the bias ($\times 10^{-3}$) of the estimates in terms of $P[C \leq X] = 0.27, 0.36$ and 0.45 assuming the exponential survival distribution and $\rho = 0.8$. The difference in bias is rather small. In particular, the bias reduces as n increases. The estimates obtained from the independent covariance structure have slightly larger bias. When $P[C \leq X]$ is small, the difference in bias between different methods is small, whereas for larger $P[C \leq X]$, a relative large difference is displayed. The performance in terms of SSE and SEE is given in Figure 1(b) and 1 (c). Obviously, both indices decline as the sample size increases but the difference in SSE and SEE is small. Nevertheless, both indices increase as $P[C \leq X]$ increases. The estimate $\hat{\beta}^{\mathrm{IW}}$ has the largest SSE and SEE. Finally, Figure 1(d) provides CP

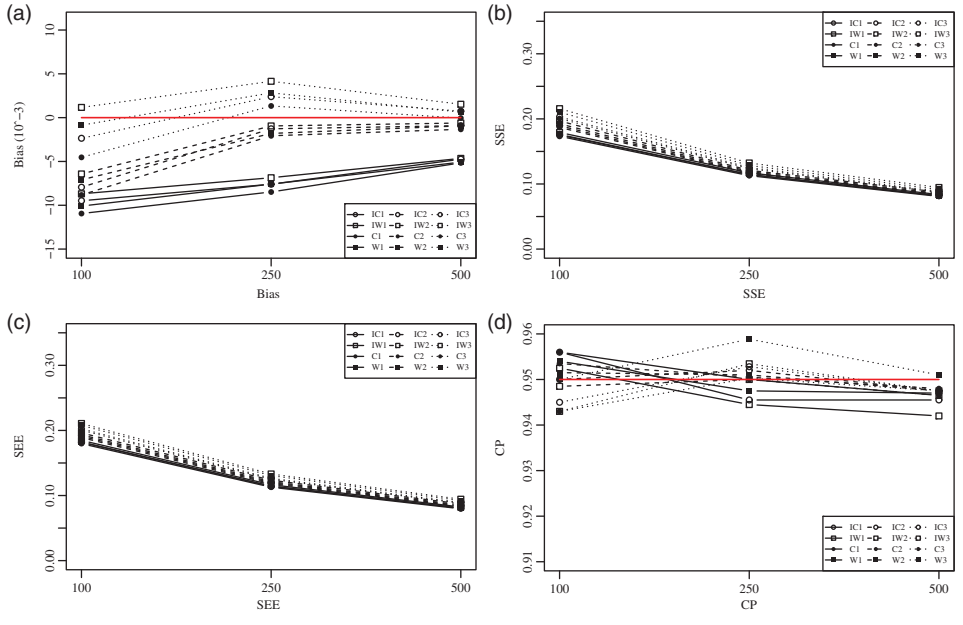

Figure 1. Performance of estimators in terms of $P[C \leq X] = 0.27, 0, 36, 0.45$ under the exponential survival time and $\rho = 0.8$, where IC1, IC2 and IC3 correspond to estimates for $P[C \leq X] = 0.27, 0.36, 0.45$ and the others are defined similarly.

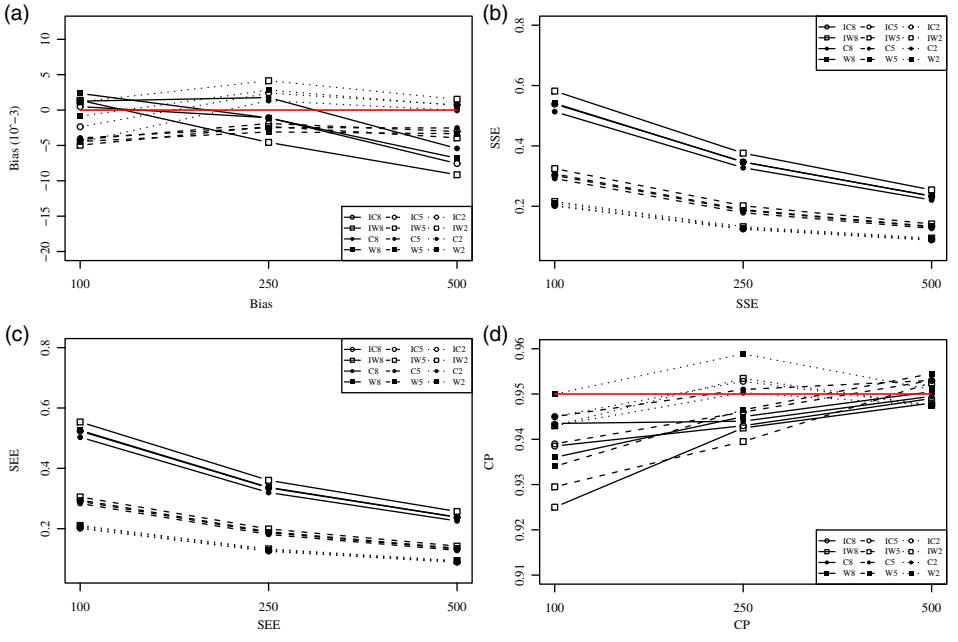

Figure 2. Performance of estimators under $\rho = 0.2, 0.5, 0.8$, $P[C \leq X] = 0.45$ and the exponential survival time, where IC2, IC5 and IC8 correspond to estimates for $\rho = 0.2, 0.5, 0.8$ and the others are defined similarly.

of the estimates with respect to the sample size and $P[C \leq X]$. As the sample size increases, CP is closer to 0.95. In particular, when the dependent covariance structure is incorporated and $P[C \leq X]$ is small, CP is closer to 0.95.

Figure 2(a) displays the bias ($\times 10^{-3}$) of the estimates in terms of ρ assuming the exponential survival distribution and $P[C \leq X] = 0.45$. The bias of estimates is again very small. When ρ is moderate, bias is very robust with respect to n. When ρ is large, the bias increases slightly when $n = 500$. However, ρ has stronger impact on SSE and SEE as given in Figure 2(b) and 2 (c). Larger ρ yields larger SSE and SEE, but the difference in SSE and SEE reduces as n increases. Figure 2(d) presents CP with respect to the sample size and ρ. Again, when n is small, CP is smaller than the desired confidence level (95%). Furthermore, increasing ρ yields CP slightly closer to the desired level, especially when n is small. Overall, estimates using dependent covariance structures have better performance. The detailed performance in terms of four indices with respect to n, ρ and $P[C \leq X]$ assuming the exponential survival time is given in Table A1.

Given $\rho = 0.8$ and the Weibull survival time, Figure 3(a) displays the performance in terms of bias with respect to n and $P[C \leq X]$. The bias is again small but is slightly larger than that under the exponential survival time. As n increases, the bias reduces slightly. Also, when n is small, $P[C \leq X]$ influences the bias and larger $P[C \leq X]$ has larger bias. Furthermore, the estimates using the independent covariance structure have slightly larger bias when n is small and $P[C \leq X]$ is large. Figure 3(b) and 3 (c) display SSE and SEE in terms of n and $P[C \leq X]$. Again, n has a strong impact on SSE and SEE, whereas $P[C \leq X]$ has very mild impact on SSE and SEE. Among different methods, SSE and SEE derived

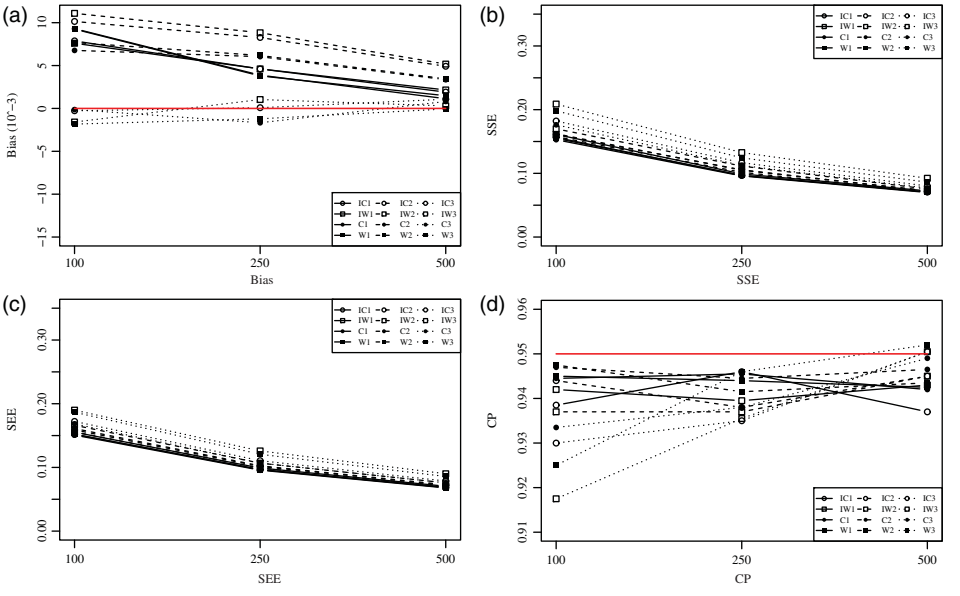

Figure 3. Performance of estimators under $P[C \leq X] = 0.29, 0.40, 0.55$, $\rho = 0.8$ and the Weibull survival time, where IC1, IC2 and IC3 correspond to estimates for $P[C \leq X] = 0.29, 0.40, 0.55$ and the others are defined similarly.

from Method I are larger than those from other methods. In particular, when $P[C \leq X]$ is large, the difference is very evident. Finally, Figure 3(d) shows the result for CP. CP is again influenced by n. As n increases, CP is much closer to the desired confidence level. Also, when n is small, $P[C \leq X]$ also influences CP. That is, CP is closer to the desired confidence level when $P[C \leq X]$ is small. Overall, estimates with dependent covariance structures have better performance.

Figure 4(a) displays the bias when $P[C \leq X] = 0.55$ and the Weibull survival time for $\rho = 0.2, 0.5, 0.8$. The bias is slightly larger than that for the Weibull survival time. n and ρ has some influences on bias. Specifically, when n is small and ρ is large, the estimates that incorporate the independent covariance structure have larger bias. The difference in bias of estimates becomes small when $n = 500$. Figure 4(b) and 4 (c) provide the tendency of SSE and SEE in terms of n for $\rho = 0.2, 0.5, 0.8$. The performance in terms of SEE and SSE with respect to n and ρ is similar to the Weibull survival time. Figure 4(d) shows the influence of n and ρ on CP. The ability in preserving CP increases as n and ρ increases. In particular, when Methods I and II are used, the ability in preserving CP reduces. The detailed performance in terms of four indices with respect to n, ρ and $P[C \leq X]$ assuming the Weibull survival time is given in Table A2. The performance of other estimates is similar to that of β_1 (data not shown).

4. Application to medical cost for colorectal cancer

The medical cost for colorectal cancer extracted from NHI database is used to illustrate the feasibility of the proposed model. Hospitals in Taiwan were categorized into the regional hospital, the teaching hospital and the academic medical center. Normally, in Taiwan, the

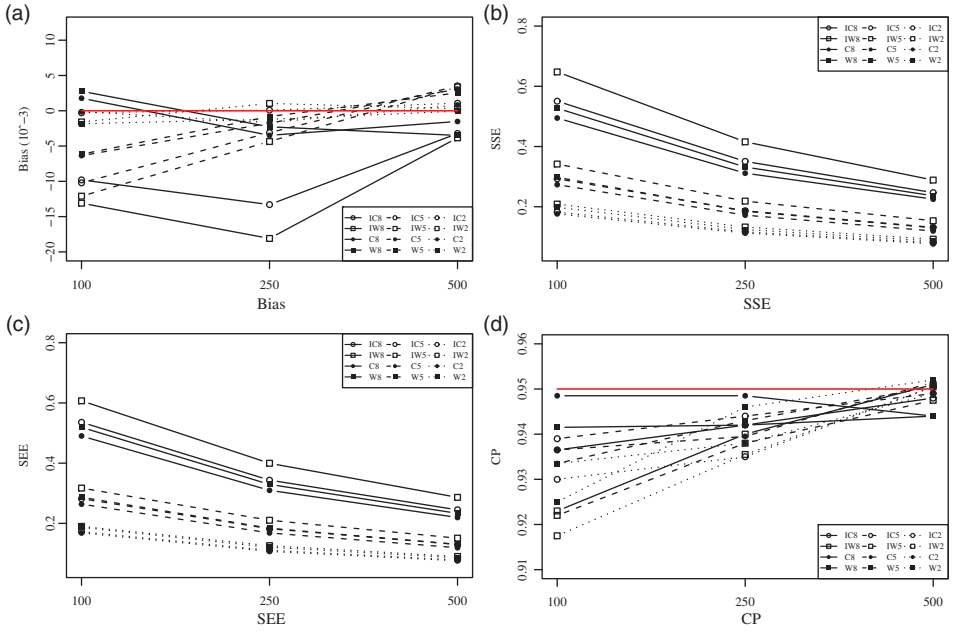

Figure 4. Performance of estimators under $\rho = 0.2, 0.5, 0.8$, $P[C \leq X] = 0.55$ and the Weibull survival time, where IC2, IC5 and IC8 correspond to estimates for $\rho = 0.2, 0.5, 0.8$ and the others are defined similarly.

academic medical center charges higher and gains more support from NHI. The purpose of the analysis is to investigate the potential influential factors that would effect the five-year medical cost for treating a colorectal patient.

The patient diagnosed of having colorectal cancer in 1999–2000 was included in the study. The inclusion criteria was set to extract patients whose ICD9 codes in inpatient records contained '155' in 1999–2000. To avoid including the recurrent patients, we excluded patients whose ICD9 codes in inpatient records contained '155' in 1998, since clinically, a patient who had colorectal cancer would require inpatient medical care and be usually followed constantly. Furthermore, since gender was an important factor, subjects with inconsistent gender informations in the medical records were excluded. In addition, the mortality data were not available in this study. Nonetheless, since it is required to be insured with NHI for all residents in Taiwan, when an insured withdraws, the possibility that the insured decreases is high, especially for the cancer patient. Thus, in this paper, the date that patients withdrew from NHI was treated as a proxy for patient's date of death. Under such an assumption, subjects were also excluded when the medical records existed after the proxy date of death. Overall, there were 7646 eligible patients.

The medical cost for 7646 eligible patients was extracted from the inpatient and outpatient medical records from 2001 to 2006. The yearly accumulative medical cost was computed for 5 consecutive years. The survival time was defined as the time from diagnosis to death. The survival time and medical costs were censored for patients who were still alive at the end of 2006. The censoring was sorely caused by the limited study duration. The overall censoring proportion was 44.7%.

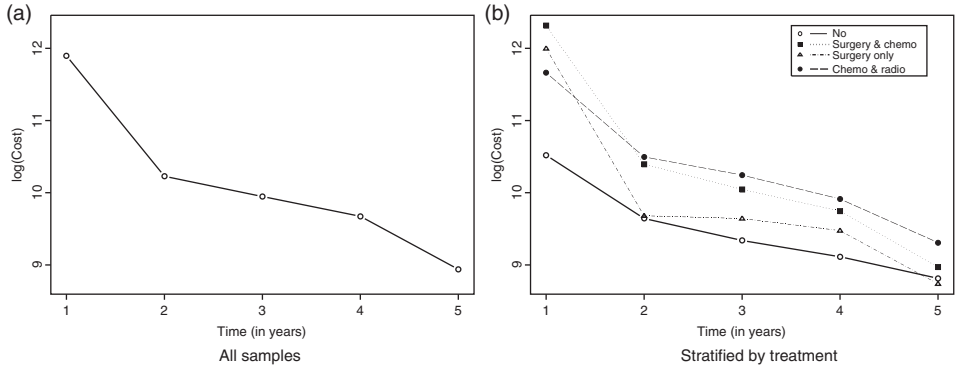

Figure 5. Mean response curve for medical cost.

Table 1. AIC and log likelihood values for the fourth degrees of polynomial mean model with various covariance structures.

		Unweighted model			Weighted model		
	df	AIC	$-2\log L$	G^2	AIC	$-2\log L$	G^2
ANTE(1)[a]	9	64,846[b]	64,826	–	60,063[b]	60,043	–
TOEPH[c]	9	65,066	65,048	222	60,359	60,339	296
ARH(1)[d]	6	65,089	65,075	249	60,386	60,372	328
CSH[e]	6	65,769	65,757	931	X	X	
Independence	1	72,624	72,622	7796	68,254	68,252	8209

Notes: [a]ANTE(1) is the ante-dependent covariance matrix. [b]Model has a smaller AIC value. [c]TOEPH is the heterogeneous Toeplit covariance matrix. [d]ARH is the heterogeneous AR(1) covariance matrix. [e]CSH is the heterogeneous compound symmetry covariance matrix. X, model does not converge.

Three basic demographic variables were considered including sex, age and area of residence. Age was categorized into 6 groups, under 35 years of age, 36–45, 46–55, 56–65, 65–75 and over 76. Areas of residence in Taiwan consist of Taipei city, northern region, central region, southern region and Kaohsiung. In particular, Taipei is the capital and Kaohsiung is the main metropolitan area in the southern Taiwan. Three disease-related information were extracted including the Charlson index (see [5]), stage of cancer and treatment. Four types of treatment combinations were discussed including without any treatment, surgery only, surgery with chemotherapy only, and surgery with chemotherapy and radiotherapy.

Figure 5 displays the mean response curve for the medical cost. Owing to the cost for the initial treatment, the medical cost for the first-year cost was much higher than that for the following years. To find an appropriate covariance pattern structure for the mean cost model, the fourth degrees of polynomial mean models with a random intercept controlling for sex, age group, residential areas, type of hospitals, the level of the hospital, cancer stage and the type of treatments, was used. Table 1 lists AIC and log likelihood values for models under various covariance–covariance pattern assumptions. The the ante-dependence structure (ANTE(1)), whose ijth element is $\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$, has the smallest AIC value for both the unweighted and weighted model. The ANTE(1) is used to establish the final mean model. The log likelihood value is obtained and is listed in Table 2. Based on the likelihood ratio test, the cubic model is selected.

Table 2. Log likelihood values for various degrees of polynomial mean models under the ante-dependent covariance structure.

Model	Linear	Quadratic	Cubic	Fourth degree
Linear	–	610	2057	2073
Quadratic	–	–	1447	1463
Cubic	–	–	–	16[a]

Note: [a]LR test is not rejected at $\alpha = 0.05$.

Table 3. Parameter estimates for the colorectal cancer data.

	Unweighted model			Weighted model		
Variable	$\hat{\beta}$	SE	*p* Value	$\hat{\beta}$	SE	*p* Value
Sex						
Female vs. Male	−0.0887	0.0179	< .0001	−0.0819	0.0182	< .0001
Age						
35 – vs. 76+	−0.0175	0.0573	.7602	0.1236	0.0601	.0396
36 – 45 vs. 76+	0.0200	0.0398	.6158	0.0888	0.0409	.0299
46 – 55 vs. 76+	−0.0042	0.0314	.8943	0.0515	0.0320	.1075
56 – 65 vs. 76+	−0.0231	0.0270	.3928	0.0195	0.0275	.4771
66 – 75 vs. 76+	−0.0685	0.0239	.0042	−0.0442	0.0242	.0681
Hospital private vs. public	−0.0193	0.0206	.3475	−0.0271	0.0208	.1942
Region Central vs. Taipei	−0.0860	0.0261	.0010	−0.0826	0.0265	.0018
Kaohsiung vs. Taipei	−0.2108	0.0269	< .0001	−0.2067	0.0274	< .0001
Northern vs. Taipei	−0.2244	0.0277	< .0001	−0.2314	0.0281	< .0001
Southern vs. Taipei	−0.0417	0.0297	.1599	−0.0291	0.0301	.3346
Level regional vs. Medical	−0.1490	0.0221	< .0001	−0.1359	0.0224	< .0001
Teaching vs. medical	−0.5160	0.0273	< .0001	−0.4717	0.0281	< .0001
Stage II vs. I	0.4054	0.0252	< .0001	0.3702	0.0253	< .0001
III vs. I	0.6191	0.0322	< .0001	0.5639	0.0319	< .0001
IV vs. I	0.7569	0.0274	< .0001	0.7023	0.0273	< .0001
Comorbidity	0.0835	0.0066	< .0001	0.0738	0.0066	< .0001
Treatment						
Surgery vs. no	3.7478	0.4727	< .0001	3.7789	0.4010	< .0001
Surgery and chemo vs. no	3.0543	6.7400	< .0001	3.0191	0.3777	< .0001
Chemo and radio vs. no	0.8964	1.7800	.0751	0.8841	0.4387	.0439
Time						
Time	−4.2170	0.6678	< .0001	−4.2025	0.5419	< .0001
$Time^2$	1.3468	0.2589	< .0001	1.3852	0.2056	< .0001
$Time^3$	−0.1383	0.0293	< .0001	−0.1450	0.0228	< .0001
Time × Treatment						
Time × surgery	−3.3441	0.7121	< .0001	−3.6238	0.5974	< .0001
Time × surgery and chemo	−1.9857	−2.9000	.0037	−2.2035	0.5634	< .0001
Time × chemo and radio	−0.0142	−0.0200	.9850	−0.1751	0.6557	.7894
$Time^2$ × Treatment						
$Time^2$ × surgery	−0.1007	0.0311	.0012	1.1042	0.2266	< .0000
$Time^2$ × surgery and chemo	−0.0548	−1.8200	.0680	0.6193	0.2139	.0038
$Time^2$ × chemo and radio	−0.0004	−0.0100	.9899	0.0170	0.2504	.9460
$Time^3$ × treatment						
$Time^3$ × surgery	1.0183	0.2752	.0002	−0.1085	0.0251	< .0001
$Time^3$ × surgery and chemo	0.5528	2.0900	.0369	−0.0609	0.0237	.0104
$Time^3$ × chemo and radio	−0.0126	−0.0400	.9658	−0.0020	0.0279	.9417

Table 3 provides the estimates of coefficients for the unweighted method and the weighted method. Estimates obtained from both methods for controlling variables except age were similar. Controlling for other factors, males spend significantly more. Patients who resided in Taipei city had significantly higher averaged cost than those who resided in other areas. The largest difference in estimates appeared between Taipei city and the

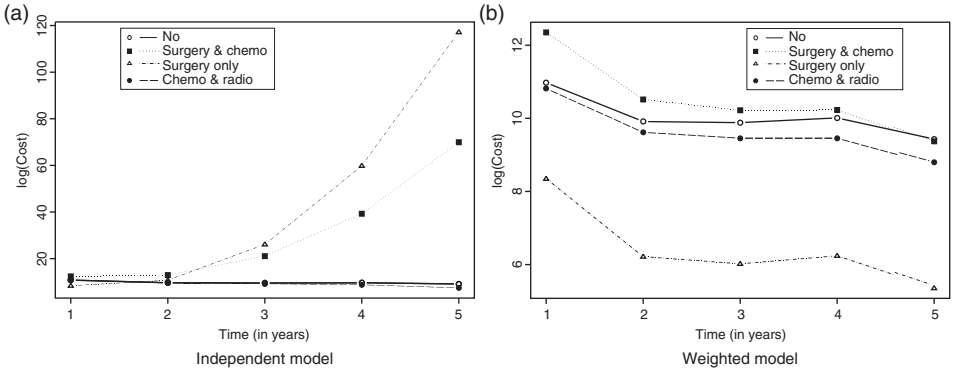

Figure 6. Predicted medical cost for the reference group.

northern area. As expected, patients who were treated in the medical center would spend more than those patients who were treated in the regional hospitals and teaching hospitals. Furthermore, patients, whose cancer stage was more advanced and who had more comorbidity, had higher average medical cost.

The other explanatory variables showed different impact on the cost. The weighted method showed that as compared to patients with 76 years of age and older, younger patients (under 35 years old and 36–45 years of age) spend significantly more medical expenditure, whereas the unweighted method only discovered that patients who were 76 years old and higher would spend significantly more than those who were 66–75 years of age. A different time trend and time by treatment interaction were found. The estimates of the third degree for the unweighted method were positive, whereas those for the weighted method were negative. Furthermore, the magnitudes for the unweighted method were much larger.

To be more precise, Figure 6(a)–6(b) display the predicted cost derived from both models for the reference group. The unweighted method provided an increasing trend in cost, which is very different from the mean response curve shown in Figure 5(b). Furthermore, the cost for patients having only surgery was much higher than that for others and increased exponentially with time and unexpectedly, the estimated cost for patients with no further treatment and with chemotherapy and radiotherapy was constant over the five years. On the contrary, the predicted cost for the weighted method decreased as time increased, which is much coincide with the mean response curve displayed in Figure 5(a) and the estimated cost for patients with only surgery or without any further treatment would be lower than those who have other treatments as expected.

5. Discussion and conclusion

Since censoring in the cost data is not the same as that for the survival data, the usual methodology for analyzing survival data can not be implemented directly. Taking the advantage that the cost data are normally sequentially recorded, this paper modifies the estimating equation in the general linear model for the longitudinal data to analyze the cost data. This modification takes into account of the sampling bias by considering a

weight defined as the inverse of the probability of a patient not being censored. Furthermore, we derive the asymptotic properties of model estimators and use simulations to demonstrate the feasibility of the proposed models. The simulation shows that estimates with proper covariance data structure and proper adjustment for the selection bias have better performance.

When cost data can be divided into subintervals, Lin [16,18] specified a set of linear regression models for the cost in each subinterval and the average of the total cost is estimated by summing up the regression coefficients. Although the model establishment allows some correlations within the same subject, but the estimating procedures in Lin [16,18] do not take into account of these correlations. Based on the proposed model, the relationship between time and cost can be modeled by selecting the appropriate trend model. By weighting the observations by IPW, Baser *et al.* [3] proposed using the random intercept model to model the panel cost data. Under the random intercept model, the variance of covariance of cost among the panel is the compound symmetric structure. The estimating procedure for the proposed model adapted from Equation (13) in [16] is more general and takes into account a more general association structure for the panel.

A parametric models for covariance structure is often preferred when analyzing longitudinal data since it involves fewer parameters and can be estimated more accurately. The more the complexity of covariance structure, the less the efficiency of estimators. However, the gains in efficiency may be often modest and well be outweighed by the potential loss of consistency when the parametric assumption for the covariance structure is wrongly specified (see [8]).

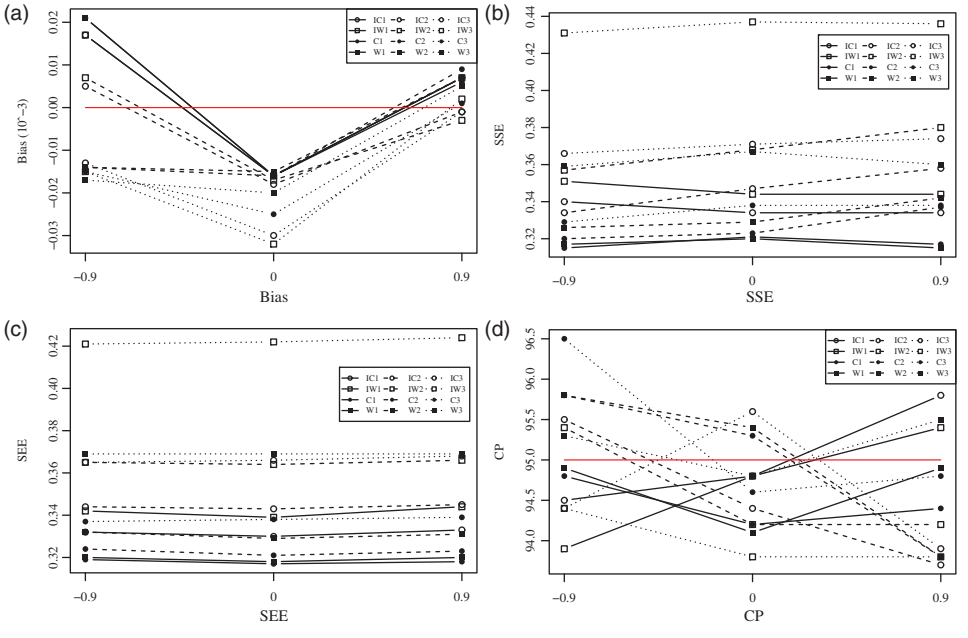

Figure 7. Performance of estimators in terms of skewness $\alpha = -0.9, 0, 0.9$, $\rho = 0.5$ and $n = 250$, the exponential survival time and a skewed normal random intercept, where IC1, IC2 and IC3 correspond to estimates for $P[C \leq X] = 0.29, 0.40, 0.55$ and the others are defined similarly.

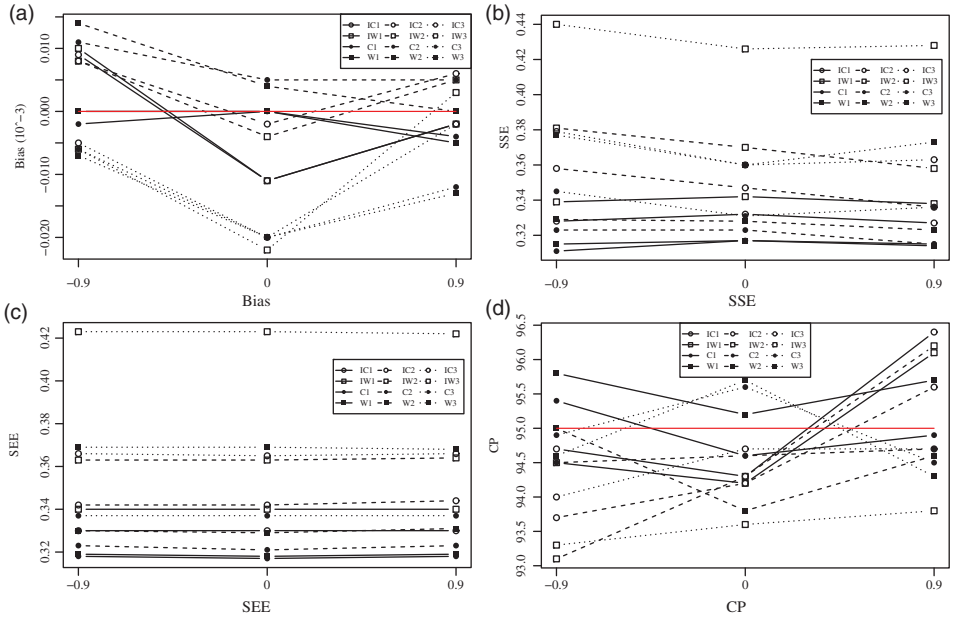

Figure 8. Performance of estimators in terms of skewness $\alpha = -0.9, 0, 0.9$, $\rho = 0.5$ and $n = 250$, the exponential survival time and a skewed normal random error, where IC1, IC2 and IC3 correspond to estimates for $P[C \leq X] = 0.29, 0.4, 0.55$ and the others are defined similarly.

To understand the robustness of our proposed model in terms of the normal assumption, additional simulations are performed. That is, we assume that the distribution of the random intercept and random error is the skewed normal distribution (see [1]). Varying the skewness, Figure 7 provides the performance of $\hat{\beta}$ in terms of four indices assuming the distribution of random intercept is a skewed normal distribution. Bias again is small. when the correct covariance structure and proper weights are implemented, SSE and SEE are relatively small and are robust with respect to skewness. However, when the distribution is left-skewed, CP might be slightly away from the desired level as the probability of censoring increases. Figure 8 provides the performance of $\hat{\beta}$ in terms of four indices assuming the distribution of the random error is a skewed normal distribution. Bias again is small. SSE and SEE are relatively small when proper adjustments are implemented and are robust with respect to skewness. When the skewness increases, the estimate with an appropriate adjustment is much closer to the nominal CP level. Depending upon the skewness, CP for the estimate constructed from the independent situation might be overestimated or underestimated the nominal level depending upon the probability of censoring.

IPWM is a conceptually simple way to adjust for incomplete data. However, it is inefficient and sensitive to the choice of weighting method. A series of papers that proposed to improve IPWM over the last decade have been published (see [4,28–30]). Implementing different imputing methods may improve the efficiency of estimators for the proposed model. In addition, Liu *et al.* [23] and Liu [21] proposed the joint model for the mixed model and proportional hazards model that included a shared random effect to incorporate the association between medical costs and survival may be considered.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] A. Azzalini and A. Capitanio, *The Skew-Normal and Related Families*, IMS Monographs series, Cambridge University Press, New York, 2014.

[2] H. Bang and A.A. Tsiatis, *Estimating medical costs with censored data*, Biometrika 87 (2000), pp. 329–343.

[3] O. Baser, J.C. Gardiner, C.J. Bradley, H. Yuce, and C. Given, *Longitudinal analysis of censored medical cost data*, Health Econ. 15 (2006), pp. 513–525.

[4] J.R. Carpenter, M.G. Kenward, and S. Vansteelandt, *A comparison of multiple imputation and doubly robust estimation for analyses with missing data*, J. Roy. Statist. Soc. Ser. A 169 (2006), pp. 571–584.

[5] M.E. Charlson, P. Pompei, and K.L. Ales, and C.R. MacKenzie, *A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation*, J Chronic Dis 40 (1987), pp. 373–383.

[6] J. Chen, L. Liu, D. Zhang, and Y.C. Shih, *A flexible model for the mean and variance functions, with application to medical cost data*, Stat. Med. 32 (2013), pp. 4306–4318.

[7] D.R. Cox, *Partial likelihood*, Biometrika 62 (1975), pp. 269–276.

[8] P.J. Diggle, P. Heagerty, K.Y. Liang, and S.L. Zeger, *Analysis of Longitudinal data*, 2nd ed., Oxford, Great Britian, 2002.

[9] R.D. Etzioni, E.J. Feuer, S.D. Sullivan, D. Lin, C. Hu, and S.D. Ramsey, *On the use of survival analysis techniques to estimate medical care costs*, J. Health Econ. (1999), pp. 365–380.

[10] P. Fenn, A. McGurie, B. Phillips, M. Backhouse, and D. Jones, *The analysis of censored treatment cost data in economic evaluation*, Med. Care 33 (1995), pp. 851–863.

[11] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware, *Applied Longitudinal Analysis*, 2nd ed., John Wiley and Sons, Inc., New York, 2011.

[12] C. Gourieroux, A. Monfort, and A. Trognon, *Psuedo maximum likelihood methods: theory*, Econometrica 52 (1984), pp. 681–700.

[13] R.A. Hiatt, C.P. Quesenberry, J.V. Selby, B.H. Fireman, and A. Knight, *The cost of acquired immunodeficiency syndrome in North California: The experience of a large prepaid health plan*, Arch. Intern. Med. 150 (1990), pp. 833–838.

[14] R.V. Hogg, J.W. McKean, and A.T. Craig, *Introduction to Mathematical Statistics*, 7th ed., Pearson, Taipei, 2013.

[15] K.-Y. Liang and S.L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika 73 (1986), pp. 13–22.

[16] D.Y. Lin, *Linear regression analysis of censored medical costs*, Biostatistics 1 (2000), pp. 35–47.

[17] D.Y. Lin, *Proportional means regression for censored medical costs*, Biometrics 56 (2000), pp. 775–778.

[18] D.Y. Lin, *Regression analysis of incomplete medical cost data*, Stat. Med. 22 (2003), pp. 1181–1200.

[19] D.Y. Lin, E.J. Feuer, R. Etzioni, and Y. Wax, *Estimating medical costs from incomplete follow-up data*, Biometrics 53 (1997), pp. 419–434.

[20] D.Y. Lin, T.R. Fleming, and L.J. Wei, *Confidence bands for survival curves under the proportional hazards model*, Biometrika 81 (1994), pp. 73–81.

[21] L. Liu, *Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data*, Stat. Med. 28 (2009), pp. 972–986.

[22] L. Liu, R.L. Strawderman, M.E. Cowen, and Y.C. Shih, *A flexible two-part random effects model for correlated medical costs*, J. Health Econ. 29 (2010), pp. 110–123.

[23] L. Liu, R.A. Wolfe, and J.D. Kalbfleish, *A shared random effects model for censored medical costs and mortality*, Stat. Med. 26 (2007), pp. 139–155.

[24] I. Locatelli and A. Marazzi, *Robust parametric indirect estimates of the expected cost of a hospital stay with covariates and censored data*, Stat. Med. 32 (2013), pp. 2457–2466.
[25] A. O'Hagan and J.W. Stevens, *On estimators of medical costs with censored data*, J. Health Econ. 23 (2004), pp. 615–625.
[26] C.P. Quesenberry, B. Fireman, R.A. Hiatt, and J.V. Selby, *A survival analysis of hospitalization among patients with acquired immunodeficiency syndrome*, Am. J. Public Health 79 (1989), pp. 1643–1647.
[27] M. Raikou and A. McGuire, *Estimating medical care costs under conditions of censoring*, J. Health Econ. 23 (2004), pp. 443–470.
[28] D.B. Robins, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
[29] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Analysis of semiparametric regression models for repeated outcomes in the presence of missing data*, J. Amer. Statist. Assoc. 90 (1995), pp. 106–121.
[30] D.O. Scharfstein, A. Rotnizky, and J.M. Robins, *Adjusting for nonignorable drop-out using semiparametric nonresponse models (with comments)*, J. Amer. Statist. Assoc. 94 (1999), pp. 1096–1120.

Appendix 1. Estimating the censoring distribution

When the censoring time depends on covariates, the proportional hazards model can be used to formulate the effects of covariates. Define the proportional hazards model as

$$\lambda(t \mid \boldsymbol{V}) = \lambda_0(t)\, \mathrm{e}^{\boldsymbol{\gamma}' \boldsymbol{V}(t)}, \tag{A1}$$

where $\boldsymbol{V}(t)$ is a subset of covariates $\boldsymbol{Z}$, $\lambda_0(t)$ is an unspecified baseline hazard function of C and $\lambda(t \mid \boldsymbol{V})$ is the conditional hazard function of C given $\boldsymbol{V}$. To avoid the identifiability problem, we assume C is independent of all other random variables conditioning on $\boldsymbol{V}$.

The parameters in Equation (A1) can be estimated by the partial likelihood function [7]. Let the estimators be denoted as $\hat{\boldsymbol{\gamma}}$ and also let $\delta_i^C = 1 - \delta_i$. Furthermore, to ease the representations, we introduce the notations $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$ for any vector $\boldsymbol{a}$ and

$$S^{(\rho)}(t; \boldsymbol{\gamma}) = \sum_{i=1}^{n} I[T_i \geq t]\, \mathrm{e}^{\boldsymbol{\gamma}' \boldsymbol{V}_i(t)} \boldsymbol{V}_i^{\otimes \rho}(t), \quad \rho = 0, 1, 2.$$

$$\bar{S}(t, \boldsymbol{\gamma}) = \frac{S^{(1)}(t, \boldsymbol{\gamma})}{S^{(0)}(t, \boldsymbol{\gamma})}.$$

Let $G(t) = P[C_i \geq t]$. Under Equation (9), an estimator for G is given as

$$\hat{G}(t \mid \boldsymbol{V}) = \exp\left\{ - \sum_{j=1}^{n} \frac{\delta_j^C I[T_j < t]\, \mathrm{e}^{\hat{\boldsymbol{\gamma}}' \boldsymbol{V}(T_j)}}{S^{(0)}(T_j; \hat{\boldsymbol{\gamma}})} \right\}. \tag{A2}$$

Appendix 2. Proofs of asymptotic results

To derive the asymptotic properties of $\hat{\boldsymbol{\beta}}^W$, we first rewrite the ith diagonal element of $\boldsymbol{W}_i$ as

$$\frac{\delta_{ik}^*}{\hat{G}(t \mid \boldsymbol{V})} = \frac{\delta_{ik}^*}{G(t \mid \boldsymbol{V})} + \frac{\delta_{ik}^* (G(t \mid \boldsymbol{V}) - \hat{G}(t \mid \boldsymbol{V}))}{\hat{G}(t \mid \boldsymbol{V})}. \tag{A3}$$

Modifying (2.1) in [20], we can show that the second term in the right-hand side of Equation (A3) is asymptotically equivalent to

$$\frac{\delta_{ik}^*}{n} \sum_{i=1}^{n} \int_0^t \frac{\mathrm{e}^{\boldsymbol{\gamma}' \boldsymbol{V}_i(t)}}{s^{(0)}(x)} \mathrm{d}M_i(x) + h'(t; \boldsymbol{V}) \boldsymbol{\Lambda}^{-1} \frac{\delta_{ik}^*}{n} \sum_{i=1}^{n} \int_0^t \boldsymbol{V}_i(x) - \bar{s}(x)\, \mathrm{d}M_i(x) + o_p(1), \tag{A4}$$

where $s^{(\rho)}(t) = \lim_{n\to\infty} n^{-1}S^{(\rho)}(t;\boldsymbol{\gamma})$, $\bar{s}(t) = s^{(1)}(t)/s^{(0)}(t)$,

$$h(t;\boldsymbol{V}) = \int_0^t \mathrm{e}^{\boldsymbol{\gamma}'\boldsymbol{V}(t)}(\boldsymbol{V}(x) - \bar{s}(x))\lambda_0(x)\,\mathrm{d}x,$$

$$\boldsymbol{\Lambda} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\delta_i^C\left\{\frac{s^{(2)}(T_i)}{s^{(0)}(T_i)} - \bar{s}^{\otimes 2}(T_i)\right\},$$

$$M_i(t) = \delta_i^C I[T_i \le t] - \int_0^t I[T_i \ge x]\,\mathrm{e}^{\boldsymbol{\gamma}'\boldsymbol{V}_i(x)}\lambda_0(x)\,\mathrm{d}x.$$

By the law of large numbers, when $t = T_{ik}^*$, Equation (A4) converges in probability to a well-defined limits, say $w_{2i,k}$.

Let $\boldsymbol{W}_i = \boldsymbol{W}_{1i} + \boldsymbol{W}_{2i}$ where $\boldsymbol{W}_{1i}$ and $\boldsymbol{W}_{2i}$ are the diagonal matrices with the jth diagonal element being $\delta_{ij}^*/G(T_{ij}^*|\boldsymbol{V})$ and $w_{2i,j}$, respectively. The left-hand side of Equation (7) can be rewritten as $\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{U}_{1n}(\boldsymbol{\beta}) + \boldsymbol{U}_{2n}(\boldsymbol{\beta})$, where

$$\boldsymbol{U}_{1n}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_{1i}\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta}),$$

$$\boldsymbol{U}_{2n}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_{2i}\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{Z}}_i\boldsymbol{\beta}).$$

Based on the consistence of $\hat{G}$ and the Delta method, $n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta})$ has an asymptotic multivariate normal distribution with zero-mean and covariance matrix

$$\boldsymbol{B} = \sum_{i=1}^{n}\tilde{\boldsymbol{Z}}_i'(\boldsymbol{W}_{1i} + \boldsymbol{W}_{2i})\tilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{W}_{1i} + \boldsymbol{W}_{2i})'\tilde{\boldsymbol{Z}}_i. \tag{A5}$$

From the Delta method, $n^{1/2}(\hat{\boldsymbol{\beta}}^W - \boldsymbol{\beta}) = \boldsymbol{A}_n^{-1}n^{-1/2}\boldsymbol{U}_n(\boldsymbol{\beta})$, where

$$\boldsymbol{A}_n = n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{Z}}_i'\boldsymbol{W}_i\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{Z}}_i, \tag{A6}$$

has an asymptotic multivariate normal distribution with zero-mean and covariance matrix $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$, where $\boldsymbol{A}_n$ converges in probability to $\boldsymbol{A}$.

Appendix 3. Detailed simulation results

Table A1. Performance of adjusted cost estimator under the exponential survival.

			$P[C \le X] = 0.27$				$P[C \le X] = 0.36$				$P[C \le X] = 0.45$			
n	ρ	Indices	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$
100	0.8	Bias	0.010	0.011	0.005	0.006	−0.000	−0.001	0.000	−0.001	0.000	0.001	0.001	0.002
		SSE	0.514	0.530	0.489	0.498	0.526	0.554	0.498	0.516	0.539	0.582	0.513	0.541
		SEE	0.487	0.500	0.467	0.476	0.503	0.524	0.483	0.498	0.524	0.554	0.503	0.526
		CP	93.4	93.1	94.1	94.0	93.1	92.8	93.6	93.4	93.9	92.5	94.4	93.6
	0.5	Bias	0.008	0.008	0.009	0.008	−0.005	−0.006	−0.006	−0.007	−0.004	−0.005	−0.004	−0.004
		SSE	0.278	0.286	0.268	0.274	0.284	0.296	0.275	0.283	0.302	0.325	0.292	0.307
		SEE	0.269	0.275	0.262	0.266	0.278	0.288	0.271	0.278	0.291	0.305	0.284	0.295
		CP	93.7	93.6	94.3	93.7	94.4	93.9	94.4	94.2	93.9	93.0	94.5	93.4
	0.2	Bias	−0.009	−0.009	−0.011	−0.010	−0.008	−0.006	−0.009	−0.007	−0.002	0.001	−0.005	−0.001
		SSE	0.175	0.179	0.173	0.176	0.190	0.197	0.187	0.192	0.203	0.216	0.201	0.210
		SEE	0.182	0.185	0.180	0.182	0.191	0.196	0.189	0.193	0.203	0.211	0.200	0.208
		CP	95.6	95.3	95.6	95.4	95.0	94.9	95.4	95.2	94.5	94.3	94.3	95.0
250	0.8	Bias	0.004	0.003	0.003	0.003	0.011	0.013	0.008	0.010	−0.001	−0.005	0.002	−0.001
		SSE	0.316	0.327	0.300	0.306	0.322	0.340	0.304	0.315	0.347	0.376	0.328	0.346
		SEE	0.312	0.322	0.297	0.304	0.323	0.338	0.307	0.318	0.336	0.360	0.320	0.337
		CP	94.9	94.8	95.1	95.3	94.5	94.6	94.8	94.9	94.3	94.3	94.4	94.5
	0.5	Bias	−0.004	−0.004	−0.004	−0.004	0.006	0.007	0.005	0.006	−0.002	−0.002	−0.002	−0.003
		SSE	0.173	0.178	0.168	0.170	0.182	0.191	0.176	0.180	0.186	0.202	0.178	0.189
		SEE	0.172	0.176	0.166	0.169	0.178	0.186	0.173	0.178	0.187	0.199	0.181	0.190
		CP	94.7	94.7	95.0	94.8	94.4	94.2	94.5	94.8	94.6	94.0	95.1	94.7
	0.2	Bias	−0.008	−0.007	−0.008	−0.008	−0.001	−0.001	−0.002	−0.002	0.002	0.004	0.001	0.003
		SSE	0.114	0.117	0.113	0.115	0.117	0.121	0.116	0.119	0.125	0.132	0.124	0.129
		SEE	0.114	0.117	0.113	0.114	0.120	0.123	0.118	0.121	0.126	0.133	0.125	0.131
		CP	94.6	94.5	95.0	94.8	95.2	95.0	95.1	95.1	95.3	95.3	95.0	95.9
500	0.8	Bias	−0.006	−0.006	−0.006	−0.006	0.001	0.001	0.002	0.001	−0.008	−0.009	−0.005	−0.007
		SSE	0.225	0.233	0.212	0.217	0.234	0.246	0.222	0.229	0.233	0.254	0.221	0.234
		SEE	0.221	0.228	0.210	0.214	0.228	0.240	0.217	0.224	0.238	0.257	0.226	0.239
		CP	94.6	94.7	94.7	94.9	94.0	93.6	93.8	93.9	94.9	94.8	95.0	95.1
	0.5	Bias	0.005	0.004	0.004	0.004	0.003	0.004	0.003	0.004	−0.003	−0.004	−0.003	−0.003
		SSE	0.123	0.128	0.118	0.120	0.130	0.136	0.126	0.130	0.131	0.141	0.127	0.133
		SEE	0.122	0.125	0.118	0.120	0.127	0.132	0.122	0.126	0.133	0.142	0.128	0.135
		CP	94.8	94.5	95.2	95.1	94.1	94.1	94.3	94.2	95.3	95.3	95.3	95.5
	0.2	Bias	−0.005	−0.005	−0.005	−0.005	−0.001	−0.001	−0.001	−0.001	0.001	0.002	−0.000	0.001
		SSE	0.082	0.084	0.081	0.082	0.084	0.088	0.083	0.086	0.089	0.095	0.087	0.092
		SEE	0.081	0.082	0.080	0.081	0.084	0.087	0.083	0.085	0.089	0.094	0.088	0.092
		CP	94.6	94.2	94.7	94.7	94.8	94.7	94.8	94.8	94.8	94.8	94.8	95.1

Notes: SSE is the sampling standard error of estimator. SEE is the sampling mean of standard error estimator. CP is the coverage probability of the 95% confidence interval.

Table A2. Performance of adjusted cost estimator under the Weibull survival.

			$P[C \le X] = 0.29$				$P[C \le X] = 0.40$				$P[C \le X] = 0.55$			
n	ρ	Indices	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$	$\hat{\beta}^{IC}$	$\hat{\beta}^{IW}$	$\hat{\beta}^{C}$	$\hat{\beta}^{W}$
100	0.8	Bias	0.003	0.003	−0.004	−0.005	0.003	0.003	−0.000	−0.002	−0.010	−0.013	0.002	0.003
		SSE	0.504	0.522	0.490	0.493	0.514	0.550	0.485	0.494	0.551	0.648	0.495	0.527
		SEE	0.491	0.506	0.481	0.484	0.507	0.537	0.480	0.489	0.536	0.607	0.490	0.519
		CP	93.6	93.7	94.4	94.4	93.9	93.8	94.7	95.0	93.7	92.3	94.9	94.2
	0.5	Bias	−0.012	−0.012	−0.007	−0.007	−0.006	−0.005	−0.002	0.001	−0.010	−0.012	−0.006	−0.006
		SSE	0.260	0.268	0.251	0.252	0.273	0.290	0.255	0.261	0.293	0.342	0.273	0.298
		SEE	0.255	0.262	0.247	0.249	0.263	0.277	0.251	0.258	0.282	0.317	0.264	0.287
		CP	94.3	94.0	94.4	94.5	94.0	94.1	94.4	94.4	93.9	92.2	93.7	93.4
	0.2	Bias	0.008	0.008	0.009	0.009	0.010	0.011	0.007	0.008	−0.000	−0.002	−0.000	−0.002
		SSE	0.156	0.161	0.153	0.155	0.161	0.170	0.158	0.162	0.182	0.209	0.176	0.198
		SEE	0.152	0.155	0.151	0.152	0.159	0.166	0.156	0.161	0.172	0.190	0.168	0.187
		CP	93.9	94.2	94.5	94.5	94.4	93.7	94.7	94.8	93.0	91.8	93.4	92.5
250	0.8	Bias	0.004	0.004	0.006	0.005	0.006	0.007	0.000	0.000	−0.013	−0.018	−0.003	−0.002
		SSE	0.332	0.343	0.316	0.318	0.329	0.353	0.306	0.311	0.351	0.416	0.311	0.332
		SEE	0.315	0.326	0.306	0.308	0.326	0.347	0.305	0.311	0.344	0.400	0.310	0.329
		CP	94.1	94.3	94.2	94.2	93.9	94.3	95.0	94.8	94.2	94.0	94.9	94.2
	0.5	Bias	−0.002	−0.002	−0.004	−0.005	0.008	0.008	0.009	0.009	−0.003	−0.004	−0.002	−0.001
		SSE	0.169	0.174	0.164	0.165	0.172	0.183	0.162	0.166	0.187	0.219	0.172	0.185
		SEE	0.164	0.169	0.159	0.160	0.171	0.182	0.162	0.166	0.182	0.210	0.168	0.184
		CP	94.7	94.7	93.8	94.1	94.9	95.1	94.6	95.1	94.4	93.8	94.0	94.3
	0.2	Bias	0.005	0.005	0.004	0.004	0.008	0.009	0.006	0.006	0.000	0.001	−0.002	−0.001
		SSE	0.097	0.099	0.096	0.097	0.106	0.111	0.102	0.105	0.117	0.133	0.112	0.124
		SEE	0.097	0.099	0.096	0.097	0.102	0.107	0.099	0.103	0.111	0.126	0.107	0.121
		CP	94.6	94.0	94.6	94.4	93.8	93.7	94.5	94.2	93.5	93.6	93.8	94.6
500	0.8	Bias	0.010	0.011	0.007	0.007	0.001	0.001	0.002	0.002	−0.003	−0.004	−0.002	−0.003
		SSE	0.224	0.231	0.216	0.218	0.231	0.247	0.220	0.222	0.247	0.289	0.225	0.236
		SEE	0.223	0.231	0.217	0.218	0.231	0.246	0.216	0.220	0.245	0.286	0.220	0.233
		CP	94.9	95.1	95.6	95.4	94.0	94.3	95.0	95.1	94.8	95.1	94.4	94.4
	0.5	Bias	−0.003	−0.003	−0.002	−0.002	0.003	0.004	0.002	0.003	0.004	0.003	0.003	0.003
		SSE	0.117	0.120	0.113	0.114	0.121	0.129	0.113	0.116	0.131	0.154	0.119	0.130
		SEE	0.117	0.120	0.113	0.113	0.122	0.129	0.115	0.117	0.130	0.151	0.119	0.130
		CP	94.5	94.2	94.4	94.1	95.2	95.2	95.5	95.8	94.9	94.8	95.2	95.1
	0.2	Bias	0.002	0.002	0.001	0.001	0.005	0.005	0.003	0.003	0.001	0.000	0.001	−0.000
		SSE	0.071	0.072	0.070	0.071	0.072	0.075	0.071	0.073	0.080	0.092	0.077	0.086
		SEE	0.069	0.070	0.068	0.068	0.072	0.076	0.070	0.073	0.079	0.090	0.076	0.086
		CP	93.7	94.3	94.2	94.3	94.5	94.5	94.7	94.4	95.1	95.1	94.9	95.2

Notes: SSE is the sampling standard error of estimator. SEE is the sampling mean of standard error estimator. CP is the coverage probability of the 95% confidence interval.