Published in final edited form as:

Stat Biosci. 2015 October ; 7(2): 367–378. doi:10.1007/s12561-015-9126-8.

On Huberized calibration regression for censored medical cost data

Brent A. Johnson

Department of Biostatistics & Computational Biology, University of Rochester, 601 Elmwood Ave. Box 630, Rochester, NY 14642 U.S.A., Tel.: +011-585-273-1869, Fax: +011-585-273-1031, brent.johnson@urmc.rochester.edu

Abstract

Authors have observed that the distribution of medical expenditures has features that do not lend it to parametric modeling and can present significant challenges for least-squares-type estimators, even on a logarithmic scale. In this note, we discuss caveats and extensions of coefficient estimation in the bivariate accelerated lifetime model of medical cost and survival time on covariates. We consider the setting where medical cost is observed only when the event occurs and potential right-censoring of the event time induces a dependent censoring mechanism on cost. We adopt Huang's (2002) estimation framework using the weighted log-rank estimating equations and investigate his proposal for robust mark-scale coefficient estimation. Due to modeling restrictions on the joint distribution of survival time and cost, we conclude that his robust mark-scale coefficient estimator would benefit from a time-scale adjustment. We use basic principles from robust estimation to define a new weighted marked process that subsequently leads to a new time-corrected robust regression calibration estimator. Our simulation studies illustrate how the proposed estimator has desirable operating characteristics, including reduced sensitivity to extreme values in the cost distribution, smaller finite sample bias and variance than earlier proposals. We illustrate the method in an analysis of lifetime medical cost data from a lung cancer study conducted by the Southwest Oncology Group.

Keywords

Accelerated failure time model; Dependent censoring; Marked point processes; Robust regression; Weighted log-rank regression

1 Introduction

Most institutions that operate on a budget are keenly interested in managing limited resources, and this includes medical institutions whose charge is to identify cost-effective devices, treatments, and procedures. In order to keep premium costs low, for example, insurers may like to know which treatments cost less money over the lifetime of the insured but still have the same efficacy. In the United States of America, we anticipate that the cost of medical treatments will be scrutinized even more closely after the Affordable Care Act was signed into law in March 2010, which further underscores the importance and relevance of reliable statistical methods for the analysis of medical cost data. It is well-known that if medical cost is only observed at the failure or event time of interest which is itself subject to

right-censoring, then the cost data is subject to induced dependent censoring and ordinary survival analysis methods (cf. Cox and Oakes, 1984; Fleming and Harrington, 1991; Kalbfleisch and Prentice, 2002) cannot, in general, be used to estimate quantiles from the cost distribution (Lin et al, 1997; Zhao and Tsiatis, 1997; Huang and Louis, 1998; Bang and Tsiatis, 2000; Strawderman, 2002), test for differences in cost between two independent samples or estimate associations between cost and covariates (Lin, 2000; Bang and Tsiatis, 2002; Huang and Lovato, 2002; Huang, 2002). Although the focus of this paper is regression for censored medical costs, an incomplete list of other methodological contributions in the analysis of marked endpoints include applications to quality-adjusted survival analysis (Zhao and Tsiatis, 1997; Huang and Louis, 1998), HIV vaccine efficacy trials (Gilbert et al, 2008), and the analysis of clinical endpoints (e.g. immunology, non-AIDS defining illnesses, HIV resistance) measured at the time of virologic failure in therapeutic HIV trials (Gilbert et al, 2004, 2008; Sun et al, 2009). Hence, the scientific interest in marked endpoints extends beyond the specific application to medical costs discussed here.

Let V be a mark variable of interest observed at the event time T and $\mathbf{X}$ a d -vector of independent regressors. The goal is to understand the association between V and $\mathbf{X}$ when the observed data are $(W, U, \mathbf{X})$, $W = VI(T - C)$, $U = T \wedge C$, $I(T - C)$, C is a censoring random variable assume to be conditionally independent of (V, T) given $\mathbf{X}$, and $a \wedge b = \min(a, b)$. To this end, consider the bivariate accelerated lifetime model (Cox and Oakes, 1984; Kalbfleisch and Prentice, 2002),

$$\begin{pmatrix} \log T \\ \log V \end{pmatrix} = \mathbf{X}' \begin{pmatrix} \eta \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_T \\ \varepsilon_V \end{pmatrix}, \quad (1)$$

where β and η are unknown vectors of regression coefficients on the mark- and time-scale, respectively, and the error $\varepsilon = (\varepsilon_T, \varepsilon_V)'$ follows an unknown bivariate error distribution function F_ε . Lin (2000) proposed a weighted least-squares estimator for β where the weight is zero for censored observations and, for uncensored observations, the inverse conditional probability of censoring exceeding the follow-up time given covariates while Bang and Tsiatis (2002) extended the technique to median and quantile regression. Alternatively, Huang (2002) proposed a semi-rank based estimator built on the weighted log-rank estimating equations (Tsiatis, 1990; Wei et al, 1990; Ying, 1993). Briefly, Huang showed that one may draw mark-scale inference by adopting the weighted log-rank estimating functions but replacing the usual time-scale counting process $N(z, \eta) = I(\log U - \mathbf{X}'\eta \leq z, \quad = 1)$ with the the marked process,

$$N^v(z, \theta) = (\log V - \mathbf{X}'\beta)N(z, \eta). \quad (2)$$

Using (2), Huang proposed a mark-scale coefficient estimator for β defined as the solution of a calibrated linear system of equations, whereby consistent estimation of the regression coefficients η in the survival model appropriately rescales or recalibrates estimation of the mark model parameters β in (1).

Because medical cost data can have extremely long tails in practice (Bang and Tsiatis, 2002), even on the log-scale, we investigated robust extensions of the calibration estimator

that would be resistant to extreme outliers in the mark-space. As suggested by Huang (2002), a natural remedy appears to be the modified marked process,

$$N^v(z, \theta) = \psi(\log V - \mathbf{X}'\beta)N(z, \eta), \quad (3)$$

where $\psi(\cdot)$ pertains to Huber's ψ function (Huber, 1964). Intuitively, if we rewrite $\psi(\log V - \mathbf{X}'\beta)$ in (3) as the weighted residual $\phi(\beta)(\log V - \mathbf{X}'\beta)$ and our goal is to reduce sensitivity to extreme observations in the mark-space, then the objective is to assign small weights $\phi(\beta)$ to observations with large residuals. In a classic theory of robust M -estimators, locally optimal weights are defined by comparing the normalized residuals to a referent distribution, e.g. standard normal distribution (cf. Hogg, 1979). The extension to (3) is non-trivial because the mark-scale residuals cannot be normalized naïvely due to dependent censoring and the joint model in (1). To illustrate the issue, we display in panel (a) of Fig. 1 the Huber function ψ operating on the mark-scale errors over time, where $E[\varepsilon_V|\varepsilon_T = z]$ is assumed to follow a sinusoidal function in time. Without knowing the value of time ε_T , we have no idea whether the magnitude of ε_V is extreme or not. However, if the mark-scale errors ε_V can be viewed alongside the time-scale errors ε_T and centered appropriately, then one can identify outliers in the mark-space as illustrated in panel (b).

To overcome these challenges, we propose the weighted marked process

$$N^v(z, \theta) = \phi(z, \beta)(\log V - \mathbf{X}'\beta)N(z, \eta), \quad (4)$$

where $\phi(z, \beta)$ is a time-dependent weight function depending on β and potentially other nuisance parameters. Note, if $E[\varepsilon_V|\varepsilon_T = z]$ were constant, then the marked processes in (3) and (4) would be equivalent with $\phi(z) = \psi(z)/z$. In general, (3) is a special case of (4) but not vice-versa because the weight ϕ in (4) is time-dependent. Finally, under suitable regularity conditions on the weight $\phi(z, \beta)$, the construction of marked process in (4) allows us to borrow large sample results from Huang (2002) which guarantee the consistency and asymptotic normality of the resulting coefficient estimator. Thus, these results will be stated but the reader will be referred elsewhere for technical details.

2 Methods

2.1 Notation, Background, Assumptions

Following the notation in Sec. 1, let the observed data be $(W_i, U_i, \mathbf{X}_i)$, $i = 1, \dots, n$ and assume that ε_i , $i = 1, \dots, n$, are mutually independent and identically distributed according to distribution law F_ε . Define the at-risk, counting and marked processes,

$R_i(z, \eta) = I(\log U_i - \mathbf{X}_i'\eta \geq z)$, $N_i(z, \eta)$ and $N_i^v(z, \theta)$, respectively, for $i = 1, \dots, n$, and the filtration

$$\mathcal{F}_z = \sigma \left\{ N_i(u, \eta), N_i^{(0)}(u, \eta), V_i \cdot N_i(u, \eta), \mathbf{X}_i, u \leq z, i = 1, \dots, n \right\},$$

where $N_i^{(0)}(z, \eta) = I(\log U_i - \mathbf{X}_i'\eta \leq z, \Delta_i = 0)$. The weighted log-rank estimating function for η is,

$$\Psi_{\eta}(\theta) = \sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta) \} dN_i(z, \eta),$$

$\bar{\mathbf{X}}(z, \eta) = \sum_{i=1}^n \mathbf{X}_i R_i(z, \eta) / \sum_{i=1}^n R_i(z, \eta)$, ϕ is a $\mathcal{F}_z$ -measurable weight function satisfying regularity conditions (Ying, 1993). Substituting the marked process $N_i^{\nu}(z, \theta)$ from (2) for $N_i(z, \eta)$ in $\Psi_{\eta}(\theta)$ leads to Huang's (2002) estimating function for β :

$$\Psi_{\beta}(\theta) = \sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta) \} dN_i^{\nu}(z, \theta) = \sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta) \} (\log V_i - \mathbf{X}_i' \beta) dN_i(z, \eta). \quad (5)$$

Huang's (2002) estimator $\tilde{\theta} = (\tilde{\eta}', \tilde{\beta}')'$ is the solution to the $2d$ system of equations, $0 = \Psi_{\theta}(\tilde{\theta})$, $\Psi_{\theta} = (\Psi_{\eta}', \Psi_{\beta}')'$. Due to the fact that $\Psi_{\eta}(\theta)$ does not depend on β , the calibration estimation procedure consists of two steps: solving the weighted log-rank system of estimating equations, $0 = \Psi_{\eta}(\tilde{\theta})$, for $\tilde{\eta}$ followed by solving $0 = \Psi_{\beta}((\tilde{\eta}', \tilde{\beta}')')$. Although Huang (2002) noted that the weight function ϕ could be defined more generally as a function of both the mark- and time-scale parameters, he showed that adopting ϕ in (5) as a function of only the time-scale parameters leads to a locally optimal coefficient estimator on the mark-scale when the mark- and time-scale errors are independent, for example. Below, we show how to construct weight functions that improve the efficiency and robustness of the mark-scale calibration regression estimator.

2.2 Time-Adjusted Coefficient Estimation

In light of the joint lifetime model in (1), we consider the partly linear model for the conditional expectation of the mark variable given time,

$$E[\log V_i | \mathbf{X}_i, \varepsilon_{T,i} = z] = \nu(z) + \mathbf{X}_i' \beta, \quad (6)$$

for some unspecified function $\nu(z)$ and implies that $E[\varepsilon_{V,i} | \mathbf{X}_i, \varepsilon_{T,i} = z] = \nu(z)$. Hence, define the new weighted marked process,

$$N_i^{\nu}(z, \theta) = \varphi_i(z, \beta) (\log V_i - \mathbf{X}_i' \beta) N_i(z, \eta), \quad (7)$$

$$\varphi_i(z, \beta) = \psi\{Z_i(z)\} / Z_i(z)$$

$$Z_i(z) = \frac{\log V_i - \nu(z) - \mathbf{X}_i' \beta}{\tau(z)},$$

where $\tau^2(z) = \text{var}[\log V_i - \nu(z) - \mathbf{X}_i' \beta | \mathbf{X}_i, \varepsilon_{T,i} = z]$. Now, substituting the weighted marked process in (7) for $N_i^{\nu}(z, \theta)$ in $\Psi_{\beta}(\theta)$ from (5) leads to the proposed estimating function,

$$\Psi_{\beta, \varphi}(\theta) = \sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta) \} \varphi_i(z, \beta) (\log V_i - \mathbf{X}_i' \beta) dN_i(z, \eta).$$

We define $\hat{\theta}_{\varphi} = (\hat{\eta}', \hat{\beta}'_{\varphi})'$ and $\hat{\beta}_{\varphi}$ is the solution to the calibrated system of equations, $0 = \Psi_{\beta, \varphi}((\hat{\eta}', \hat{\beta}')')$ and η was defined earlier.

Assume that $v(z)$ and $\tau(z)$ are known and $\phi_i(z, \beta)$ are known, $\mathcal{F}_z$ -measurable functions for every $i = 1, \dots, n$. Then, each of $\phi(z, \eta)$, $\{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta) \}$, $\phi_i(z, \beta)$, and $(\log V_i - \mathbf{X}_i' \beta)$ are $\mathcal{F}_z$ -measurable and one can use standard martingale theory to establish the asymptotic properties (cf. Fleming and Harrington, 1991; Kalbfleisch and Prentice, 2002). Under Conditions (A)–(E) in Huang (2002, Appendix), $\hat{\theta}_{\varphi}$ is a consistent estimator for θ_0 and $\sqrt{n}(\hat{\theta}_{\varphi} - \theta_0)$ converges in distribution to a mean-zero normal vector with covariance $\Lambda^{-1} \mathbf{B} (\Lambda^{-1})'$, where Λ is the asymptotic slope matrix of $n^{-1} \Psi_{\theta}(\theta_0)$ and $\mathbf{B}$ is the asymptotic covariance of $n^{-1/2} \Psi_{\theta}(\theta_0)$. The proof of this result is an extension of Huang (2002) as long $\phi_i(z, \beta)$ is $\mathcal{F}_z$ -adapted and satisfies suitable regularity conditions. In practice $v(z)$ and $\tau(z)$ in $\phi_i(z, \beta)$ are unknown and must be estimated from the data; in Section 3 below, we outline one estimation strategy.

Variance estimation is challenging because the asymptotic slope matrix Λ depends on the underlying hazard functions of the mark- and time-scale errors in (1). However, one may adopt a variety of established methods for drawing statistical inference from non-smooth estimating functions including Huang's (2002) inverse numerical differentiation, numerically estimating Λ by perturbing the score function (cf. Strawderman, 2005), or the multiplier bootstrap (Kosorok, 2008, Ch. 10.2); in this paper, we adopt a multiplier bootstrap procedure.

3 Algorithm

First, suppose that $\tau(z) \equiv \tau$ and the weights $\phi_i(z, \beta_0)$, $i = 1, \dots, n$, are fixed and known. Then the solution to the system of equations, $0 = \Psi_{\beta, \varphi}((\eta_0', \beta_0')')$, is:

$$\tilde{\beta}_{\varphi} = \left[\sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta_0) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta_0) \} \varphi_i(z, \beta_0) \mathbf{X}_i' dN_i(z, \eta_0) \right]^{-1} \times \sum_{i=1}^n \int_{-\infty}^{\infty} \phi(z, \eta_0) \{ \mathbf{X}_i - \bar{\mathbf{X}}(z, \eta_0) \} \varphi_i(z, \beta_0) \log V_i dN_i(z, \eta_0). \quad (8)$$

This closed-form solution reduces to Huang's (2002) estimator when $\phi_i(z, \beta) = 1$ for $i = 1, \dots, n$, and for all z . We use (8) to construct an iteratively reweighted linear solution, which we now describe.

Compute the time-scale coefficient estimate $\hat{\eta}$ and set $\tilde{\beta}_{\varphi, [0]}$ to Huang's (2002) linear estimate as the initial value to the iterative procedure. Stratify the observations into J groups $\{S_1, \dots, S_J\}$ by quantiles of the fitted time-scale residuals. Compute the estimate of the time-

varying intercept using the uncensored mark-scale residuals in the j -th stratum and the current iterate, $\tilde{\beta}_{\varphi,[k]}: \tilde{\nu}_{i,[k]} = |S_j|^{-1} \sum_{i \in S_j} \Delta_i(\log V_i - \mathbf{X}_i' \tilde{\beta}_{\varphi,[k]})$, $\forall i \in S_j, j = 1, \dots, J$, where $|A|$ is the cardinality of the set A . The use of uncensored residuals to estimate secondary nuisance parameters in an accelerated lifetime model was first proposed by Heller (2007) for a related problem. Then, for each of the uncensored observations, compute the weights

$$\tilde{\varphi}_i(z, \tilde{\beta}_{\varphi,[k]}) = \frac{\psi(\tilde{Z}_{i,[k]})}{\tilde{Z}_{i,[k]}}, \quad \tilde{Z}_{i,[k]} = \frac{\log V_i - \tilde{\nu}_{i,[k]} - \mathbf{X}_i' \tilde{\beta}_{\varphi,[k]}}{\tilde{\tau}_{[k]}},$$

where $\tilde{\tau}_{[k]}$ is the median absolute deviation (MAD) from the median of the time-adjusted, uncensored residuals, $\log V_i - \tilde{\nu}_{i,[k]} - \mathbf{X}_i' \tilde{\beta}_{\varphi,[k]}$, divided by 0.6745. Compute the weighted mark-scale coefficient estimate by solving the linear system in (8) using the updated weights $\tilde{\varphi}_{i,[k]}$, $i = 1, \dots, n$. The procedure continues until convergence or until a maximum number of iterations is reached. Although the iterative procedure is not guaranteed to converge, our own limited experience suggests the algorithm works well in practice and often converges in fewer than 7 iterations using a tolerance of 10^{-6} . The final estimate from the iterative procedure is called $\hat{\beta}_{\varphi}$.

In the more general case where $\tau(z) \neq \tau$, we propose replacing $\tilde{\tau}_{[k]}$ in the above algorithm with a stratified estimate. A simple extension of the above estimator for the time-varying intercept is to estimate $\tau(z)$ using the strata-specific MAD divided by 0.6745 (Hogg, 1979). With sufficient data, one could more accurately account for censoring by augmenting the naïve strata-specific estimator with weights, e.g. inversely weighting observations by the probability of censoring. Alternatively, one could use the nonparametric maximum likelihood estimate Huang and Louis (1998) of the joint error distribution to estimate both $v(z)$ and $\tau(z)$. Note that the existence and identifiability of $v(z)$ is non-trivial and implies, for example, that the support of the time-scale residual exceeds the support of the censored residual (cf. Ritov, 1990). Estimating time-varying second- and higher-order moments requires large samples for stable estimation and, for the purposes of illustrating the method in finite samples, we implemented the procedure where $\tau(z) = \tau$ in our simulation studies below.

4 Simulation Studies

We simulated data according to joint lifetime model in (1) with $d = 2$ independent covariates X_1 and X_2 : X_1 is Bernoulli with equal probability on -1 and 1 and X_2 is $\text{Un}(-1, 1)$. The coefficients are $\eta = (1, 1)'$ and $\beta = (1, 0.5)'$, and the errors $\varepsilon = (\varepsilon_T, \varepsilon_V)'$ are independent and identically distributed bivariate normal with $E(\varepsilon_T) = E(\varepsilon_V) = 0$, $\text{var}(\varepsilon_T) = \text{var}(\varepsilon_V) = 1$ and $\text{cov}(\varepsilon_T, \varepsilon_V) = 0.5$. The censoring distribution was $\text{Un}(0, v)$ where v was chosen to yield 20% and 30% censoring. The Huberized estimator used $K = 5$ strata when normalizing the mark-scale residuals while the resampling procedure used 500 realizations of exponential random variables with unit rate. The results from 500 Monte Carlo data sets for the bivariate normal error distribution are displayed in Table 1.

In Table 1, we find that finite sample bias is approximately the same between the two estimators but that Huang's linear estimator has somewhat better precision than the robust estimator. With one exception, we generally found the log-rank estimators had smaller variance than the Gehan-type counterparts for both the linear and robust versions of the coefficient estimators. The resampling method worked well for all estimators, although we observed considerable bias in small samples which resulted in coverage probabilities that were too small. Again, the results in our table suggests that the exact same resampling plan led to better inference with the log-rank weight function. Finally, we conducted a simulation study with $n = 1000$ subjects to illustrate the scalability of the statistical methods. When the sample size is this large, the point and variance estimators performed their best in terms of finite sample bias, variance, and empirical coverage probability.

In the second simulation scenario, we wanted to investigate the scenario where a subject's survival time lies in the middle of the time distribution but the cost associated with the participant is extraordinarily large. So, we repeated the above experiment except that the errors follow a bivariate contaminated distribution: contaminated on the mark-scale with probability π , but uncontaminated on the time-scale. In particular, a contamination indicator ζ was simulated Bernoulli(π) and then a random error ε was generated from a the bivariate normal distribution with mean $(1 - \zeta) + 10\zeta$ and covariance $(1 - \zeta)\Omega + I_2\zeta$, where Ω is the same covariance used in the first scenario and I_2 is the 2-dimensional identity matrix. We investigated two contamination probabilities, $\pi \in \{0.01, 0.03\}$, and set the censoring distribution to $\text{Un}(0, 6)$, which resulted in 20% censored observations, on average. The simulation results from 1000 Monte Carlo data sets are presented in Table 2.

When the contamination probability is 1%, the magnitude of the bias for Huang's linear estimator is somewhat larger than the robust estimator and about twice as large as the robust estimator when 3% of the observations are contaminated. When 1% of the observations are contaminated, on average, the variance of the robust estimator is always less than the variance of Huang's estimator. When 3% of the observations are contaminated, the standard error of the Huberized estimator is about 24% smaller than Huang's estimator. In general, the log-rank weight led to more precision than the Gehan weight, which agrees with the results in Table 1. In the presence of contaminated marks, the resampling scheme worked well for the robust estimator but not so for the linear estimator. Despite this observation, the confidence intervals appeared to cover the true value at the nominal level for both the linear and robust estimators.

5 Analysis of the SWOG 9509 Data

We present a reanalysis of the lung cancer data from the Southwest Oncology Group Study 9509 (Huang, 2002). Briefly, this was a randomized clinical study of Paclitaxel plus Carboplatin versus Vinorelbine plus Cisplatin in untreated patients with advanced cell lung cancer. Resource utilization was monitored as part of the study and collected at 3 months, 6 months, and every 6 months after that until 24 months. Cost was assigned to each medical resource using standardized procedures and recalibrated back to 1998 US dollars for consistency with the date of the study. Here and in Huang (2002), lifetime medical cost is recorded in thousands of US dollars. Three baseline covariates are considered in the data

analysis: the treatment indicator with Vinorelbine plus Cisplatin as the reference, lactate dehydrogenase normality indicator, and age. The data for 398 study participants are included in the analysis. Our analysis differs from Huang (2002) only in that age is standardised to have mean zero and unit variance. The analytic results are reported in Table 3. Standard errors are based on 1000 bootstrap resamples and 95% confidence intervals are constructed using the percentile method. We used two different number of strata $K \in \{5, 20\}$ to investigate the sensitivity of the the robust estimator.

On the whole, the results from Huang's linear estimator and the robust estimator agree. As stated in Huang (2002), lifetime medical cost was significantly higher in the Paclitaxel plus Carboplatin group as compared to the Vinorelbine plus Cisplatin group even though the survival was not significantly different. The one minor difference was that our treatment effects were somewhat larger than what was presented earlier: when $K = 5$, 0.621 and 0.507 versus 0.405 and 0.338 for the Gehan and log-rank weights, respectively. This is roughly a 35% increase in the treatment effects and could be an indication of left-skewed cost data after the logarithmic transformation. We reran the analysis with $K = 20$ strata and still found that treatment effect sizes from the robust estimates were about 24% larger than with the linear estimator. Finally, although the treatment effect sizes were larger for the robust estimator, so were their confidence intervals. The length of the confidence interval for treatment effect was 0.58 and 0.51 units for the robust and linear estimators, respectively, using the Gehan weight function with $K = 5$ that is about 12.5% longer. We observed the same phenomenon regarding interval length using the log-rank weight function as well.

6 Discussion

Authors have noted peculiarities in the cost distribution that do not lend themselves to parametric modeling (Lin et al, 1997; Huang, 2009). The fact that the cumulative medical cost distribution typically has exceptionally long right tails is representative of a small percent of the population who use a disproportionately large amount of resources. In some cases, the scientific interest may be focused on just the high resource utilizers. But when the scientific interest lies in describing associations between risk factors and medical cost for the typical individual in the middle of the population, such skewness and extreme observations can have an adverse effect on parameter estimates and statistical inference if not properly addressed (Bang and Tsiatis, 2002).

A competing approach to the method presented here is robust inverse probability weighted (IPW) estimators for censored medical cost data (e.g. Bang and Tsiatis, 2002). In addition to a completely different estimation strategy, the statistical model is different fundamentally. Huang's (2002) framework models the joint conditional distribution of time and cost given covariates whereas the competing approach models the conditional distribution of medical cost given covariates. Because the conditional moments of medical cost given covariates may not be identified, it is common to introduce a user-defined but arbitrary upper time-limit which induces identifiability (cf. Zhao and Tsiatis, 1997). The upper time-limit is used to define a new time-restricted outcome and, subsequently, one draws statistical inference on the new time-restricted scale. Drawing inference on a time-restricted endpoint may or may not be interesting scientifically and one must consider the specific application carefully.

Occasionally, the user-defined time-limit is large enough such that there is little practical difference between the unrestricted and time-restricted endpoints. When the time-restricted outcome is desirable, the IPW estimator and its doubly robust extensions are relatively straightforward to implement and computationally simpler than rank-based methods. For a lengthier description of technical and practical issues between the competing approaches, we refer the reader to elsewhere in the literature (Huang, 2009).

Finally, the asymptotic variance of the cost-scale coefficient estimators depend on the hazard functions of the residuals on both the cost- and time-scales. So, when the true error distribution is contaminated as in Sect. 4, the asymptotic variance is a function of the contamination probability and the contaminated error distribution. Then, locally optimal weight functions $\phi_i(z, \beta)$ could be chosen to minimize the asymptotic variance under particular distributional assumptions. In contrast, the stratified weighting scheme proposed here was based on general principles from a theory of robust M -estimation. Formal analytic calculations could provide insight into scenarios that would favor alternative weighting schemes to the scheme presented here but such investigations were beyond the scope of the current manuscript.

Acknowledgements

The author thanks Drs. John Crowley, Mary Redman, and Ms. Kari Chansky for graciously sharing and assisting with the SWOG data. The author gratefully acknowledges helpful editorial comments from an associate editor and an anonymous referee and numerous insightful conversations with Professor Yijian Huang on this and related topics over the last several years.

References

- Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika*. 2000; 87:329–343.
- Bang H, Tsiatis AA. Median regression with censored cost data. *Biometrics*. 2002; 58:643–649. [PubMed: 12229999]
- Cox, DR.; Oakes, D. *Analysis of Survival Data*. London: Chapman and Hall; 1984.
- Fleming, TR.; Harrington, DP. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
- Gilbert PB, McKeague IW, Sun Y. Tests for comparing mark-specific hazards and cumulative incidence functions. *Lifetime Data Analysis*. 2004; 10:5–28. [PubMed: 15130048]
- Gilbert PB, McKeague IW, Sun Y. The 2-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. *Biostatistics*. 2008; 9:263–276. [PubMed: 17704528]
- Heller G. Smoothed rank regression with censored data. *Journal of the American Statistical Association*. 2007; 102(478):552–559.
- Hogg RV. Statistical robustness: One view of its use in applications today. *The American Statistician*. 1979; 33:108–115.
- Huang Y. Calibrated regression of censored lifetime medical cost. *Journal of the American Statistical Association*. 2002; 97:318–327.
- Huang Y. Cost analysis with censored data. *Medical Care*. 2009; 47:S115–S119. [PubMed: 19536024]
- Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85:785–798.
- Huang Y, Lovato L. Tests for lifetime utility or cost via calibrating survival time. *Statistica Sinica*. 2002; 12:707–723.
- Huber PJ. Robust estimation of a location parameter. *Ann Math Statist*. 1964; 35:813–828.
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley; 2002.

- Kosorok, M. Introduction to Empirical Processes and Semiparametric Inference. New York: Springer; 2008.
- Lin DY. Linear regression analysis of censored medical costs. *Biostatistics*. 2000; 1:35–47. [PubMed: 12933524]
- Lin DY, Feuer EJ, Etzioni R. Estimating medical costs from incomplete follow-up data. *Biometrics*. 1997; 53:419–434. [PubMed: 9192444]
- Ritov Y. Estimation in a linear regression model with censored data. *The Annals of Statistics*. 1990; 18(1):303–328.
- Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association*. 2002; 95:1192–1208.
- Strawderman RL. The accelerated gap times model. *Biometrika*. 2005; 92:647–666.
- Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. *Annals of Statistics*. 2009; 37:394–426. [PubMed: 19838313]
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann Statist*. 1990; 18:354–372.
- Wei LJ, Ying Z, Lin DY. Regression analysis of censored survival data based on rank tests. *Biometrika*. 1990; 77:845–851.
- Ying Z. A large sample study of rank estimation for censored regression data. *Annals of Statistics*. 1993; 21:977–984.
- Zhao H, Tsiatis AA. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*. 1997; 84:339–348.

Figure 1.
Huber's ψ acting on mark-scale residuals ε_V when $E[\varepsilon_V|\varepsilon_T = z]$ is a sinusoidal function in z .

Table 1

Simulation results for a bivariate Gaussian distribution. SE refers to the median absolute deviation of the parameter estimates divided by 0.6745, SEE refers to the median of the standard error estimates, and ECP is the empirical coverage probability. Table entries for Bias, SE, and SEE are multiplied by 1000 and ECP displayed as a percentage.

Method	n	% Censored	Parameter	Gehan				Log-rank			
				Bias	SE	SEE	ECP	Bias	SE	SEE	ECP
Linear	50	20%	β_1	3	181	164	92.4	4	171	150	92.9
			β_2	-5	338	280	87.8	-1	292	258	90.4
Robust			β_1	3	181	176	93.4	3	175	160	93.7
			β_2	-3	360	302	89.7	1	319	273	91.4
Linear		30%	β_1	7	193	169	91.4	5	177	152	92.4
			β_2	16	338	289	89.2	22	296	266	90.6
Robust			β_1	5	196	182	92.8	4	182	162	92.6
			β_2	17	388	311	91.6	24	313	279	93.6
Linear	100	20%	β_1	-2	126	123	93.6	-8	98	109	93.6
			β_2	-1	213	209	93.8	-2	193	187	94.4
Robust			β_1	-6	136	130	94.4	-14	109	115	93.2
			β_2	-4	220	222	95.4	0	205	197	94.2
Linear		30%	β_1	-1	146	126	91.8	2	129	113	92.2
			β_2	15	222	213	92.6	10	208	191	93.0
Robust			β_1	0	140	132	92.6	3	130	118	93.2
			β_2	14	228	225	92.4	9	208	202	93.6
Linear	1000	20%	β_1	1	38	40	95.1	1	32	35	94.3
			β_2	2	73	69	94.3	-1	62	61	93.9
Robust			β_1	1	40	41	94.7	0	34	37	94.9
			β_2	1	77	71	94.9	-2	61	63	93.9
Linear		30%	β_1	0	40	41	94.6	-1	34	36	94.8
			β_2	1	78	71	93.6	-1	63	62	93.4
Robust			β_1	1	42	43	94.2	-1	36	38	95.6

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Method	<i>n</i>	% Censored	Parameter	Gehan			Log-rank		
				Bias	SE	SEE	Bias	SE	SEE
			β_2	0	79	73	94.0	63	65
							−3		94.2

Table 2

Simulation results for a contaminated error distribution. SE refers to the median absolute deviation of the parameter estimates divided by 0.6745, SEE refers to the median of the standard error estimates, and ECP is the empirical coverage probability. Table entries for Bias, SE, and SEE are multiplied by 1000 and ECP displayed as a percentage.

Method	% Contaminated	n	Parameter	Gehan				Log-rank			
				Bias	SE	SEE	ECP	Bias	SE	SEE	ECP
Linear	1%	40	β_1	-1	240	205	92.0	-1	208	190	92.7
			β_2	4	250	207	93.2	4	238	196	93.9
Robust			β_1	-3	220	206	90.9	-3	197	187	91.6
			β_2	0	221	212	93.3	0	211	189	93.1
Linear		60	β_1	1	207	175	92.5	2	188	162	93.1
			β_2	-17	210	176	94.5	-10	183	162	94.3
Robust			β_1	0	179	166	92.7	0	160	152	92.9
			β_2	-10	188	170	94.1	-3	154	154	93.4
Linear	3%	40	β_1	8	331	273	93.2	7	298	260	94.1
			β_2	12	323	282	95.1	12	305	269	95.0
Robust			β_1	1	237	222	92.2	-1	196	196	92.6
			β_2	2	236	224	94.5	0	214	202	94.4
Linear		60	β_1	-6	301	255	95.2	-5	274	244	94.5
			β_2	-23	277	252	95.8	-12	258	238	95.6
Robust			β_1	-2	186	177	92.5	-1	165	157	93.8
			β_2	-12	191	182	95.2	-4	159	160	94.3

Table 3

Data Analysis for the SWOG Lung Cancer Study

Method	K	Variable	Gehan		Log-rank	
			Est. (SE)	95% CI	Est. (SE)	95% CI
Linear	-	Treatment	0.405 (0.134)	[0.140, 0.652]	0.338 (0.121)	[0.109, 0.582]
		LDH	0.164 (0.138)	[-0.094, 0.446]	0.141 (0.121)	[-0.117, 0.366]
		Age	-0.069 (0.066)	[-0.202, 0.066]	-0.050 (0.056)	[-0.168, 0.058]
Robust	5	Treatment	0.621 (0.153)	[0.274, 0.853]	0.507 (0.141)	[0.227, 0.772]
		LDH	0.199 (0.170)	[-0.110, 0.541]	0.106 (0.144)	[-0.176, 0.394]
		Age	-0.120 (0.077)	[-0.256, 0.055]	-0.116 (0.067)	[-0.238, 0.028]
20		Treatment	0.539 (0.159)	[0.227, 0.840]	0.434 (0.145)	[0.191, 0.759]
		LDH	0.086 (0.195)	[-0.118, 0.534]	0.153 (0.153)	[-0.173, 0.394]
		Age	-0.104 (0.080)	[-0.259, 0.061]	-0.091 (0.068)	[-0.235, 0.039]