Estimation from Censored Medical Cost Data

Baser, Onur and Gardiner, Joseph C and Bradley, Cathy J and Given, Charles W

2004

Estimation from Censored Medical Cost Data

Onur Başer, Ph.D., Joseph C. Gardiner, Ph.D.,
Cathy J. Bradley, Ph.D., and **Charles W. Given, Ph.D.**

Summary

This paper applies the inverse probability weighted least-squares method to predict total medical cost in the presence of censored data. Since survival time and medical costs may be subject to right censoring and therefore are not always observable, the ordinary least-squares approach cannot be used to assess the effects of explanatory variables. We demonstrate how inverse probability weighted least-squares estimation provides consistent asymptotic normal coefficients with easily computable standard errors. In addition, to assess the effect of censoring on coefficients, we develop a test comparing ordinary least-squares and inverse probability weighted least-squares estimators. We demonstrate the methods developed by applying them to the estimation of cancer costs using Medicare claims data.

Key words: Censoring; Inverse probability weighted estimation; Two-stage estimation; Exogenous censoring; Costs.

1. Introduction

Statistical methods applicable to the estimation of cost from Medicare or Medicaid claim files that are often censored, are not well developed. Censoring can be due to incomplete data ascertainment or the longitudinal nature of data collection. The average total cost for a group of patients has been estimated in one of three ways: (1) by estimating the sample mean of observed costs from all cases, (2) by estimating the sample mean of uncensored subjects only, and (3) by using modifications of standard survival analysis techniques. For various reasons, these methods yield biased estimators. The sample mean from all subjects creates a downward bias because it does not account for the costs incurred after the time of censoring. The sample mean from uncensored subjects is biased toward the costs for subjects with shorter survival times since longer survival times are likely to be censored (Lin et al. 1997; Bang and Tsiatis, 2000).

Application of methods such as Kaplan-Meier estimation or Cox regression on costs is not valid if subjects accumulate costs with different rate functions over time. Survival analysis techniques assume independence between the cost at the survival time and the cost at the censoring time. In fact the two are generally positively correlated. To address this dependency, Lin et al. (1997) proposed a partitioned estimator to assess average costs. This method partitions the entire time period of interest into a number of smaller intervals and calculates for each interval an average cost and a product-limit estimate of survival. The sum of the product of these two components is called the "product-limit sampling average estimator" of total cost from the sample. Because of its rather complicated formula, the computation of the variance estimator is a challenging programming exercise, which has led to the bootstrap method to obtain variance estimates in applications (Sloan et al., 1999). Bang and Tsiatis

(2000) extended this method and proposed a partitioned estimator whose asymptotic distribution does not depend on the choice of partition or the discreteness of the censoring times. If we are interested in conditional average costs where the mean cost is adjusted for several factors, these estimation methods assume some homogeneity in the medical cost data in the sense that they are independent of patient characteristics or the type of treatment. Because variables such as patient's age, disease severity and comorbid conditions can influence cost, these variables should be accounted for in the estimation, using for example, regression-based methods (Lin, 2000).

In this paper, the inverse probability weighted (IPW) least squares method is used to assess the effects of covariates (e.g., patient and treatment characteristics) on medical costs with censored data. IPW estimation has a long history in statistics (Horvitz and Thompson, 1952). More recent developments and applications of the IPW method are presented in several works (Horowitz and Manski, 1998; Robins and Rotnitzky, 1995, 1992; Robins et al., 1995; Rosenbaum, 1997). The method is ideally suited to estimation from non-random samples, which might arise due to censoring or by the sampling strategy used. We demonstrate how IPW estimation produces consistent estimators with a covariance matrix that can be calculated by most commercial statistics software programs. We also develop a test to compare IPW least squares and ordinary least squares methods (OLS) estimators.

This paper is organized as follows. The first section outlines IPW least squares as applied to censored medical cost data, including the statistical properties of the estimation. We then introduce a Hausman type test to compare the estimators calculated by using IPW least squares and OLS over uncensored data. The third section describes an application of our methods to the estimation of cancer costs. STATA (version 7.0) is used for all estimations. The last section presents our conclusions.

2. IPW Least Squares

Suppose we are interested in the total medical cost over period $[0, L]$. Since there is no further medical expense after death, the total cost over $[0, L]$ is the same as the cumulative cost at $T^* = \min\,(T, L)$, where T is the survival time. The distribution of T is assumed to be continuous from 0 to L.

Assume that in the population of interest

$$y = \boldsymbol{x}\boldsymbol{\beta} + u\,, \tag{1}$$

where y, $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are respectively the cumulative cost (or transformed cost) at T^*, a $1 \times K$ vector of explanatory variables, a $K \times 1$ vector of unknown regression parameters, and u is the unobservable random disturbance or error, with an unspecified distribution. The first component of $\boldsymbol{x}$ is set to 1 so that the first component of $\boldsymbol{\beta}$ represents the intercept.

Assume that

$$E(\boldsymbol{x}'u) = 0\,. \tag{2}$$

Under random sampling from the population, equation (2) is the crucial assumption in obtaining consistency of the OLS estimator of $\boldsymbol{\beta}$ in (1). With (2) and the rank assumption rank $E(\boldsymbol{x}'\boldsymbol{x}) = K$, the OLS estimator using a random sample will be consistent for $\boldsymbol{\beta}$. Assuming $E(u) = 0$ alone does not guarantee consistency.

Survival time and medical cost may be subject to right censoring and therefore are not always fully observable. Cost censoring occurs when a subject's follow-up time is less than L, and the patient is alive at the time of censoring. Since no further expense is incurred after death, whether death occurs before or after L is immaterial for the cost estimation. Let C be the time of censoring.

Let $Z = \min\,(C, T^*)$, $s = I(C \geq T^*)$, where $I(.)$ is the indicator function of the displayed event. Assume T and C are independent given $\boldsymbol{x}$.

Assumption 1:
(i) $T^*, y, \boldsymbol{x}$ are observed when $s = 1$,
(ii) y can be ignored in the selection equation, conditional on $\boldsymbol{x}$:

$$P(s = 1 \mid \boldsymbol{x}, y) = P(s = 1 \mid \boldsymbol{x}) = P(C \geq T^* \mid \boldsymbol{x}) = P(C \geq T^*)\,.$$

When censoring is due to early study termination, the censoring time C is always observed along with the subject characteristics $\boldsymbol{x}$. The cost y is observed provided censoring has not occurred before the death time T or the time horizon L, that is, provided $s = 1$. Assumption 1 (ii) implies that the likelihood of cost observation (given $\boldsymbol{x}$) does not depend on the level of cost. This applies if C is independent of $(y, T, \boldsymbol{x})$. Note that y and T could be dependent. The last part of assumption 1 (ii) could be dropped by allowing $P(C \geq T^* \mid \boldsymbol{x})$ to depend on some of the covariates $\boldsymbol{x}$. We refer to $P(C \geq T^*)$ as the selection probability.

Suppose we have a random sample $\{(\boldsymbol{x}_i,\ y_i,\ s_i)$: $i = 1, 2, \ldots, N\}$ from the population to estimate $\boldsymbol{\beta}$. The underlying model is

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i\,, \tag{3}$$

with $E(\boldsymbol{x}_i'u_i) = 0$.

The IPW least square estimators, $\hat{\beta}_w = \left(\sum_{i=1}^{N} w_i\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{N} w_i\boldsymbol{x}_i'y_i\right)$ solves the minimization problem,

$$\min_{\boldsymbol{\beta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} w_i(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2\,, \tag{4}$$

where $w_i = (s_i/P(C_i \geq T_i^*))$ and $\boldsymbol{\Theta}$ is the parameter space of $\boldsymbol{\beta}$. Under assumption 1 and equation (2), $\hat{\boldsymbol{\beta}}_w$ is consistent asymptotically with asymptotic variance estimated by $V(\hat{\boldsymbol{\beta}}_w) = \hat{\boldsymbol{A}}_w^{-1}\hat{\boldsymbol{B}}_w\hat{\boldsymbol{A}}_w^{-1}/N$, where

$$\hat{\boldsymbol{A}}_w = N^{-1}\sum_{i=1}^{N} w_i\boldsymbol{x}_i'\boldsymbol{x}_i\,, \tag{5}$$

$$\hat{\boldsymbol{B}}_w = N^{-1}\sum_{i=1}^{N} w_i^2\hat{u}_i^2\boldsymbol{x}_i'\boldsymbol{x}_i\,, \tag{6}$$

and $\hat{u}_i = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_w$ are the residuals after IPW least squares estimation (Wooldridge, 1999).

The objective function in (4) weights each observation $(y_i, \boldsymbol{x}_i)$ by the inverse probability of its appearing in the sample. Observations for which $s_i = 0$ are not used in the optimization problem. Since $(y_i, \boldsymbol{x}_i)$ is observed only when $s_i = 1$, $\hat{\boldsymbol{\beta}}_w$ is computable from the observed data provided that $P(C_i \geq T_i^*)$ is known. If unknown, we need to replace $P(C_i \geq T_i^*)$ with a consistent estimator, an issue we address shortly.

Note that neither $\hat{\boldsymbol{\beta}}_w$ nor its covariance matrix estimator use the censored observations. In addition the estimated covariance matrix is the (White, 1980) heteroscedasticity-robust covariance matrix estimator, obtained by applying the weight $\sqrt{w_i}$ to all variables in the i-th observation. Heteroscedasticity-robust standard errors after the weighted regression provide the estimated asymptotic standard errors. Censoring can then be handled easily using a heteroscedasticity-robust covariance matrix.

Another advantage of this weighting scheme is that we can derive consistency of $\hat{\boldsymbol{\beta}}_w$ under the much weaker assumption (2) rather than $E(u \mid \boldsymbol{x}) = 0$. From assumption 1, $E(w \mid \boldsymbol{x}, y) = 1$ and so

$$E(w\boldsymbol{x}'u) = E(E(w \mid \boldsymbol{x}, y)\, \boldsymbol{x}'u) = E(\boldsymbol{x}'u) = 0\,. \tag{7}$$

So far, we have assumed that the selection probability $P(C \geq T^*)$ is known. In practice, this will be unknown and therefore to operationalize $\hat{\boldsymbol{\beta}}_w$, we need a suitable estimator for this probability. Suppose censoring is not covariate dependent, and define $p(t) = P(C > t)$. For simplification of variance, assume a parametric form $p(t, \theta)$ for $p(t)$ is known except for the unknown θ. A nonparametric version has been addressed by Lin (Lin, 2000). Using the sample, $\{(Z_i, \bar{s}_i):\ i = 1, 2, \ldots, N\}$ where $\bar{s}_i = 1 - s_i$, we construct an estimator $\hat{p}(t) = p(t, \hat{\theta})$ of $p(t)$. In particular, we assumed C_i is the response variable, T_i^* is the censoring variable. Then the unknown weights w_i are estimated by

$$\hat{w}_i = \frac{s_i}{\hat{p}(T_i^*-)}\,, \tag{8}$$

where $p(t-) = P(C \geq t)$. To have all quantities properly defined, we impose the mild restriction, $p(L) > 0$.

Under standard regularity conditions,[1] a two step IPW least squares estimator that uses $\hat{w}_i$ instead of w_i in equation (4) consistently estimates $\hat{\boldsymbol{\beta}}$ (Newey and McFadden, 1994).

Formally, we define the two-step IPW least squares estimator $\tilde{\boldsymbol{\beta}}_w$ by

$$\tilde{\boldsymbol{\beta}}_w = \left(\sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' y_i \right) \tag{9}$$

which is the solution to the minimization problem

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} \hat{w}_i (y_i - \boldsymbol{x}_i \tilde{\boldsymbol{\beta}})^2 . \tag{10}$$

Then $\tilde{\boldsymbol{\beta}}_w$ is asymptotically normally distributed with estimated variance

$$\tilde{\boldsymbol{V}}_w \equiv V(\tilde{\boldsymbol{\beta}}_w) = \tilde{\boldsymbol{A}}_w^{-1} \tilde{\boldsymbol{B}}_w \tilde{\boldsymbol{A}}_w^{-1} / N , \tag{11}$$

where,

$$\tilde{\boldsymbol{A}}_w = N^{-1} \sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' \boldsymbol{x}_i , \tag{12}$$

$$\tilde{\boldsymbol{B}}_w = N^{-1} \sum_{i=1}^{N} \hat{w}_i^2 \tilde{u}_i^2 \boldsymbol{x}_i' \boldsymbol{x}_i, \tag{13}$$

where $\tilde{u}_i = y_i - \boldsymbol{x}_i \tilde{\boldsymbol{\beta}}_w$ are the residuals.

Note that our estimator is identical in form to that proposed by Lin (2000). However, because we will estimate a parametric form of $p(t)$, instead of Kaplan-Meier estimation, our weight $\hat{w}_i$ is different. This leads to some simplification in computation of variance estimators. We will also provide both first stage estimation adjusted and unadjusted variance matrices.

Clearly the variance expression in (11) does not adjust for the estimation of the w_i. In the appendix, we show that the use of the estimate of w_i in the second step yields a variance estimator that is asymptotically equivalent to that estimated with known w_i under the much stronger assumption that given $\boldsymbol{x}$, (y, T, C) are independent. However, unless we have short interval cost values, such as monthly or weekly, we would expect (y, T) to be correlated. In this case $\tilde{\boldsymbol{V}}_w$ in (11) has to be adjusted for estimation of w_i. In practice, it has been found that this adjustment has little effect on the asymptotic standard errors. However, as shown in the appendix, using the estimated selection probability will produce smaller standard errors than those given by (11). By ignoring this adjustment for simplicity, inference based on (11) would be conservative. This is somewhat unusual for two-step estimation problems, where the prevailing wisdom is that larger standard errors occur by adjusting standard errors for a first stage estimation.

3. Comparison of the IPW and Unweighted Estimators

The OLS estimator for cases with complete data, called the unweighted estimator, $\tilde{\boldsymbol{\beta}}_u$ solves

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} s_i (y_i - \boldsymbol{x}_i \boldsymbol{\beta}_i)^2 . \tag{14}$$

It is well-known that selection under exogenous censoring does not cause problems if we impose the stronger assumption, $E(u \mid \boldsymbol{x}) = 0$. By exogenous censoring we mean $E(u \mid \boldsymbol{x}, s) = 0$ in equation (1). This follows from $E(u \mid \boldsymbol{x}) = 0$ under the assumption that y and (T, C) are independent given $\boldsymbol{x}$. With

[1] The conditions in which the uniform weak law of large numbers can be applied. For details; see Theorem 12.1 in Wooldridge, 2002). Lemma 4.3 in (Newey and McFadden, 1994) shows that if w_i is replaced with a consistent estimator, the convergence is still valid.

exogenous censoring, once covariates (e.g. patient and clinical characteristics) are selected, total cost is independent of censoring and survival times. Then $\tilde{\boldsymbol{\beta}}_u$ is consistent and asymptotically normally distributed and the usual variance matrix estimator $V(\tilde{\boldsymbol{\beta}}_u) = \tilde{\boldsymbol{A}}_u^{-1}\tilde{\boldsymbol{B}}_u\tilde{\boldsymbol{A}}_u^{-1}/N$ is consistent, where

$$\tilde{\boldsymbol{A}}_u = N^{-1}\sum_{i=1}^{N} s_i\boldsymbol{x}_i'\boldsymbol{x}_i\,, \tag{15}$$

$$\tilde{\boldsymbol{B}}_u = N^{-1}\sum_{i=1}^{N} s_i\breve{u}_i^2\boldsymbol{x}_i'\boldsymbol{x}_i\,, \tag{16}$$

$\breve{u}_i = y_i - \boldsymbol{x}_i\tilde{\boldsymbol{\beta}}_u$ are the residuals after OLS estimation of uncensored sample.

If censoring is exogenous, then unweighted and weighted estimators are both consistent. In such a case, theory suggests that an unweighted estimator is more efficient under conditional homoscedasticity (Wooldridge, 1999). That states the variance of the unobservable error conditional on the explanatory variables is constant.

Because the unweighted estimator is inconsistent under the violation of exogenous censoring and the weighted estimator is consistent with or without exogenous censoring we can apply a Hausman test to determine the exogeneity of censoring (Hausman, 1978).

The traditional form of Hausman statistics can be used under the assumption of conditional homoscedasticity. We can state this assumption as follows: For the selected sample,

$$E(s_iu_i^2\boldsymbol{x}_i'\boldsymbol{x}_i) = \sigma_0^2E(s_i\boldsymbol{x}_i'\boldsymbol{x}_i). \tag{17}$$

When equation (17) holds, the unweighted least squares variance estimator is

$$\tilde{\boldsymbol{V}}_u \equiv V(\tilde{\boldsymbol{\beta}}_u) = \tilde{\sigma}^2\left(N^{-1}\sum_{i=1}^{N} s_i\boldsymbol{x}_i'\boldsymbol{x}_i\right)^{-1} \tag{18}$$

provided we have a consistent estimator of $\tilde{\sigma}^2$ of σ_0^2.

In general form, the Hausman test statistic can be written as:

$$H = (\tilde{\boldsymbol{\beta}}_w - \tilde{\boldsymbol{\beta}}_u)'\,\tilde{\boldsymbol{V}}^{-1}(\tilde{\boldsymbol{\beta}}_w - \tilde{\boldsymbol{\beta}}_u)\,, \tag{19}$$

where $\tilde{\boldsymbol{V}} \equiv \tilde{\boldsymbol{V}}_w - \tilde{\boldsymbol{V}}_u$, with $\tilde{\boldsymbol{V}}_w$ is defined in equation (11) and $\tilde{\boldsymbol{V}}_u$ is defined in equation (18).

In many cases, we may want to use a Hausman test when the homoscedasticity assumption is violated. Homoscedasticity fails if (17) does not hold. This requires a robust form that replaces $\tilde{\boldsymbol{V}}$ by

$$(\tilde{\boldsymbol{A}}_w^{-1} \mid -\tilde{\boldsymbol{A}}_u^{-1})\left(N^{-1}\sum_{i=1}^{N}\tilde{\boldsymbol{e}}_i\tilde{\boldsymbol{e}}_i'\right)(\tilde{\boldsymbol{A}}_w^{-1} \mid -\tilde{\boldsymbol{A}}_u^{-1})'/N\,, \tag{20}$$

$(.\mid.)$ is used to denote the appending of matrices side by side and $\tilde{\boldsymbol{e}}_i = \left(\hat{w}_i\tilde{u}_i\boldsymbol{x}_i', s_i\breve{u}_i\boldsymbol{x}_i'\right)'$. Here $\tilde{u}_i$ and $\breve{u}_i$ are the residuals after IPW least squares and OLS estimations of the selected sample respectively, and $\hat{w}_i$, $\tilde{\boldsymbol{A}}_w$, $\tilde{\boldsymbol{A}}_u$ are defined in equations (8), (12) and (15).

Under the null hypothesis, censoring is exogenous, $H \overset{a}{\sim} \chi^2(K)$. If we reject this hypothesis, exogenous censoring assumption is violated and OLS estimation using complete cases does not produce consistent estimators, so IPW least squares method should be used. If we fail to reject the hypothesis, the typical response is to conclude that the exogeneity assumption holds and OLS estimates are consistent. Unfortunately, we may be committing a Type II error by failing to reject the null hypothesis when it is false. Therefore, we recommend that in applications results be reported from both estimation procedures.

If heteroscedasticity is present, the OLS estimator is no longer the best linear unbiased estimator. Many tests for heteroscedasticity have been suggested over the years. The two common ones are Breusch–Pagan (1980) and White (1980) tests. Breusch–Pagan test assumes that heteroscedasticity is expressed through a linear function of explanatory variables. White's test adds the squares and cross products of all of the independent variables into the variance-estimation equation intending to test for forms of heteroscedasticity that invalidates the usual OLS standard errors and test statistics. It is possible to obtain more efficient estimators than those obtained using OLS, when the form of heteroscedasticity is known. In practice, we rarely know how the variance depends on a particular independent variable in a simple form.

Exogenous censoring occurs when cost, survival time, and censoring time are independent given explanatory variables. When censoring is due to early study termination, censoring time is independent of the others. Cost and survival time, however, are expected to be correlated even after we control for certain explanatory variables (e.g. patient and treatment characteristics). This occurs especially when we have long interval costs, such as a year or more. If we have short interval cost values, such as monthly or weekly, investigators can use the simple and more efficient method (OLS, for example), if the conditional homoscedasticity assumption is satisfied. However, failing to reject the proposed Hausman test does not guarantee the assumption of exogenous censoring. The best approach is to present both IPWLS and OLS estimates. We can summarize the decisions depending on the outcome of the Hausman and heteroscedasticity tests as follows:

Hausman	Heteroscedasticity	
	Present	Not Present
Reject	IPWLS	IPWLS
Fail to Reject	IPWLS	OLS, IPWLS

The methods described are easily applied using standard commercial statistical software programs. The traditional Hausman test is part of most statistical programs (for example, SAS, STATA), but the robust form of the Hausman test requires programming. We can use an alternative approach, the regression-based Hausman test, for easy computation of the robust form (Wooldridge 1990). Since the Hausman test compares systematic differences in the coefficients, if we regress the dependent variable on weighted and unweighted explanatory variables and the coefficients are not different between the two, then the *F* test for the coefficients on weighted explanatory variables should result in an insignificant value. It can be shown that the statistics obtained from this procedure are asymptotically equivalent to Hausman statistics that compare the weighted and unweighted estimators.

4. Application to Cancer Treatment Cost

4.1 Data

From 1994 through 1997, 773 patients with incident cases of lung, prostate, colon and breast cancer were recruited from 24 Michigan community hospitals and their affiliated oncology units. Each patient provided written consent for researchers to acquire his or her Medicare claim files.

We obtained Medicare claim files for the two years following cancer diagnosis. The files included any reimbursement claims for inpatient or outpatient care, physician provider services (including laboratory tests and diagnostics, mammography, radiation, and intravenously chemotherapy), home health care, and/or skilled nursing facilities.

Total cost is the sum of these costs. Medicare payments were used as a proxy for direct medical care costs rather than billed charges. Medicare reimbursements formulas are designed to reflect an underlying pattern of resource use, whereas charges inflate actual cost. Charges were adjusted for inflation to 1997 prices by using the National Medical Care Price Index, 1994–1997. The costs of prescription drugs, unpaid caregiver services, and the services paid by other insurers or out-of-pocket were not included.

Surgical procedures were identified by the International Classification of Diseases version 9 (ICD-9) and Current Procedural Terminology (CPT) Codes. We used all ICD-9 and CPT codes avaliable in the inpatient, outpatient, and physician supplier files to identify chemotherapy and radiation. These data were coded as dichotomous variables with yes/no categories for comparison purposes.

Physical function three months prior to diagnosis was assessed using the subscale from the Short Form (SF)-36 (Ware et al., 2000). The 10-item subscale asks questions about such activities as lifting heavy objects, participating in strenuous sports, climbing stairs, walking various distances, and ability to bathe and dress. Patients were also asked if during the past two weeks they had experienced any of 33 symptoms. A count of all symptoms was summed for each patient.

Comorbid conditions were assessed with an instrument from the Aging and Health in America Study, a national survey that asks patients to indicate whether a health professional has ever told them they have one of 15 problems. The total number of positive responses was summed for each patient and sorted into one of two categories: $0-2$, and $3+$. A comparison of patient reports of comorbid conditions with medical record audits indicates that patients are able to recall other diagnosed illnesses (Katz et al., 1996) and restricting the categories for comorbid conditions does not result in lost predictive power (Newschaffer et al., 1996).

Disease stage was determined using the American Joint Committee on Cancer (AJCC) Tumor Nodes Metastasis (TNM) staging system which was applied to pathological data obtained from an audit of patients' medical records.

4.2 Descriptive analysis

Table 1 shows the variable definitions and summary statistics. We had complete data for 541 subjects and incomplete data for 232 subjects. So, approximately 31 percent of the sample had censored data.

As shown in Table 1, the patient sample can be described as white and in their early seventies for both censored and uncensored subjects. One third of the subjects were diagnosed with late stage disease. Most subjects had three or fewer comorbidities and experienced some level of symptoms related to cancer treatment. The patient sample is high functioning in terms of physical health.

The next six rows of Table 1 show the categorical variables related to treatment types. For censored and uncensored subjects, except radiation we have similar percentages for the patients. Thirty percent of the patients received radiation only in the complete subjects whereas 17% of incomplete subjects had radiation.

The distribution of our sample in terms of cancer site is in the last four rows of Table 1. The percentage of breast and lung cancer were nearly equivalent for censored and uncensored subjects. The difference is subtler in prostate and colon cancer patients. About 31% of complete subjects were diagnosed with prostate cancer whereas 26% of incomplete subjects had prostate cancer. The percentages are 16 and 22 respectively for colon cancer with the complete and incomplete cases.

The dependent variable, total Medicare payments two years following diagnosis is shown in the last row of Table 1. Considering the mean alone, we find that the total cost of all care is \$ 60,429 for the two years following a lung-cancer diagnosis for complete cases and \$ 55,877 for incomplete cases.

4.3 Regression analysis

Our aim is to determine how independent variables (e.g., gender, comorbidity, etc) predict total medical cost of cancer in the two years following diagnosis after we account for the bias introduced by censoring.

We find that the cost distribution is skewed to the right, so we transformed the cost variable to a log scale. We started with the log-scale residuals from a generalized linear model with a logarithmic link function and found that the log-scale residuals are dense at the tails. Following Manning and Mullahy (2001) we adopted an OLS-based model with a log-transformed dependent variable. Heteroscedasticity is present according to both Breusch–Pagan and White tests.

Table 2 shows the result of the regression analysis predicting total cost of care for the two years following a cancer diagnosis. The first column of Table 2 shows the unweighted regression coeffi-

Table 1 Summary statistics from the cancer study.

Variables	Variable Description	Mean	
		Uncensored Cases ($n = 541$)	Censored Cases ($n = 232$)
		Mean	Mean
Age	Patient's Age	72.51 (4.93)	72.25 (5.09)
Physical Functioning	Patient's physical functioning	81.18 (24.08)	80.84 (23.99)
Symptoms	A count of all symptoms	8.32 (4.66)	7.81 (4.26)
Comorbidity	= 1 if patient's comorbid conditions are three or more	0.27 ($n = 146$)	0.28 ($n = 66$)
Late Stage	= 1 if patient's disease stage is regional,distant or invasive	0.31 ($n = 170$)	0.31 ($n = 72$)
Surgery	= 1 if patient received surgery only	0.22 ($n = 122$)	0.34 ($n = 79$)
Surgery & Chemo	= 1 if patient received surgery and chemotherapy	0.08 ($n = 46$)	0.10 ($n = 23$)
Surgery & Radiation	= 1 if patient received surgery and radiation	0.16 ($n = 85$)	0.15 ($n = 34$)
Surgery & Chemo & Radiation	= 1 if patient received surgery, chemotherapy and radiation	0.09 ($n = 48$)	0.10 ($n = 24$)
Chemo & Radiation	= 1 if patient received chemotherapy and radiation	0.13 ($n = 70$)	0.13 ($n = 30$
Chemotherapy	= 1 if patient received chemotherapy only	0.02 ($n = 9$)	0.01 ($n = 2$)
Radiation	= 1 if patient received radiation only	0.30 ($n = 161$)	0.17 ($n = 40$)
Lung	= 1 if patient has lung cancer	0.27 ($n = 143$)	0.25 ($n = 58$)
Prostate	= 1 if patient has prostate cancer	0.31 ($n = 170$)	0.26 ($n = 61$)
Colon	= 1 if patient has colon cancer	0.16 ($n = 85$)	0.22 ($n = 50$)
Breast	= 1 if patient has breast cancer	0.27 ($n = 143$)	0.27 ($n = 63$)
Total Cost	Total Cost	\$ 60,429 (\$ 69,138)	\$ 55,877 (\$ 60,502)

For continuous variables, standard deviations are in parentheses; for categorical variables number of cases are in parentheses.

Table 2 Correlates of log transformed total medical cost by OLS and IPW Least Squares Methods, $N = 541$.

Variable	Coefficient	
	OLS	IPWLS
Age	0.013 (0.008)	0.014 (0.008)
Physical Functioning	−0.007 (0.002)**	−0.007 (0.002)**
Symptoms	0.001 (0.010)	−0.001 (0.011)
Late Stage	0.204 (0.110)	0.188 (0.111)
Comorbidity	0.087 (0.092)	0.085 (0.092)
Surgery Only	−0.175 (0.138)	−0.168 (0.140)
Surgery & Chemo	−0.009 (0.167)	0.017 (0.168)
Surgery & Radiation	−0.342 (0.152)*	−0.343 (0.156)*
Chemo & Radiation	−0.442 (0.186)*	−0.425 (0.193)*
Chemo Only	−0.722 (0.382)	−0.680 (0.402)
Radiation Only	−0.921 (0.152)**	−0.934 (0.155)**
Prostate	−0.759 (0.139)**	−0.751 (0.142)**
Colon	−0.249 (0.147)	−0.244 (0.149)
Breast	−1.243 (0.138)**	−1.225 (0.142)**
Intercept	10.957 (0.690)**	10.909 (0.687)**
R-squared	0.34	0.34

Notes: Robust standard errors in parentheses. * significant at 5%; ** significant at 1%
Omitted categories: In situ/local stage, lung cancer, no treatment.

cients, while the second column shows weighted regression coefficients. The reference group for treatment modalities is surgery plus adjuvant therapies and the reference group for site of cancer is lung.

Variables that reach statistical significance $(p < 0.05)$ include physical function, type of cancer (except colon), surgery and radiation, radiation only, and chemotherapy and radiation. Ten additional points in patients' prior physical function score decreases total medical cost by 0.7% according to the unweighted estimation and the IPW least square estimation.

According to unweighted estimation breast cancer patients and prostate cancer patients cost 71.1% and 53.2% less than lung cancer patients respectively. These estimates are 70.6% and 52.8% in IPW least square estimation.

Whether or not a person receives radiation or chemotherapy separately or in combination significantly decreased the total medical cost relative to the mean costs for persons receiving surgery plus adjuvant therapies. The estimates with respect to the unweighted and weighted least squares are: for radiation only, 60.1% and 60.7% and for chemotherapy and radiation 35.7% and 34.6%. The cost of surgery and radiation therapy is 28.9% less than the surgery plus adjuvant therapies. Both models explain 34% of the variability in total costs at two years following diagnosis.

The difference between OLS and IPW least squares estimators are statistically different. F statistics from regression based Hausman test suggests that OLS and IPW estimates are significantly different $(F(14, 513) = 8.25; (p < 0.05))$. We reject the hypothesis that the censoring scheme is exogenous. In this case unweighted estimators are inconsistent.

While the magnitude of Hausman statistics determines statistical significance of the coefficients, the actual difference in the estimated coefficients determines what we might call practical significance. An estimate can be statistically significant without being especially large. A statistically significant value without being practically significant often occurs when we are working with large samples. A consistency test rejects the null hypothesis with the probability approaching one as the sample size grows whenever the alternative is true. Therefore, a discussion of the practical significance along with the statistical significance of the estimates is appropriate. The results from OLS and IPWLS in our cancer cost analysis suggest that the difference is important statistically but not practically.

5. Conclusions

This paper brings together theoretical work from the econometrics and biostatistics literatures to estimate censored cost data using a weighted estimation methodology. As an example we estimated censored medical costs using OLS and IPW least-squares estimation techniques.

One limitation should be discussed. The exact asymptotic variances adjustment for the first stage estimation should be made, therefore if the interest lies on marginally insignificant variables, they should be interpreted with caution since with adjustment, they may become statistically significant.

The IPW least squares method solves for inconsistencies in the coefficients caused by censoring. The method is easily applicable using most statistical software programs. Under the assumption that selection can be ignored, the inverse probability weighting scheme identifies the population parameters. The method introduced handles large numbers of continuous and discrete explanatory variables for any mean type regressions.

The application of the method is a two-step estimation process. In the first step, we estimate selection probabilities by using the maximum likelihood method, where we assumed the density function is piecewise-constant over relevant time intervals by reversing the role of censoring and survival time. In the second step, we estimate heteroscedastic robust OLS on the uncensored dataset where each variable is weighted with the inverse of the square root of the estimated selection probabilities from the first stage. We showed that first stage adjusted covariance matrix is asymptotically equivalent to unadjusted covariance matrix under exogenous censoring assumption. We also proved that violation of this assumption yields a first stage adjusted variance matrix which is always as large as the unadjusted covariance matrix. This fact is important in practice. Most of the time investigators are interested in the effect of a certain variable (for example, treatment) on the dependent variable. If this effect is significant by IPW estimation without calculating a rather complicated adjusted variance matrix, it is known in advance that the effect will be significant even if we compute the adjusted variance matrix.

We also developed a test to compare the coefficients estimated by the IPW least squares and by OLS. This test, combined with common heteroskedasticity tests, can be used to assess efficiency improvement between two models. Specifically, if we reject the null hypothesis that the sampling scheme is exogenous, IPW least squares method should be used because OLS yields inconsistent estimates. Failing to reject the null hypothesis could be used to support unweighted estimation under conditional homoscedasticity.

We also applied the proposed method to an inception cohort of patients newly diagnosed with cancer. Our findings support that unweighted estimation yields inconsistent estimator due to censoring bias. IPW least square estimation method removes that bias.

Appendix

Derivation of the Variance Formula for the Two Stage IPW Estimator

From (9) and (3) write

$$\sqrt{N}\,(\tilde{\boldsymbol{\beta}}_w - \boldsymbol{\beta}) = \left(N^{-1}\sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' \boldsymbol{x}_i\right)^{-1} \left(N^{-1/2}\sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' u_i\right), \tag{A1}$$

where $\hat{w}_i = \dfrac{s_i}{\hat{p}(T_i^*-)}$. The censoring variable C_i is assumed independent of (y_i, T_i) given the covariate $\boldsymbol{x}_i$. We consider a parametric estimator $\hat{p}_t = p(t, \hat{\theta})$ for $p(t, \theta) = P[C > t \mid \theta]$ where θ is q-dimensional. The estimator $\hat{\theta}$ of θ is obtained via maximum likelihood based on the sample $\{(Z_i, \overline{s}_i)$: $i = 1, 2, \ldots., N\}$. The likelihood can be derived by considering three cases: (1) C is observed so that $C < L$ and $T > C$, or (2) T is observed so that $T < L$ and $T < C$, or (3) the limit L is reached and neither T nor C are observed, that is $T \geq L, C \geq L$. Then the likelihood for the i-th subject can be expressed as $\{S(Z_i)\ g(Z_i, \theta)\}^{1-s_i}\{p(T_i, \theta)\, f(T_i)\}^{s_i[T_i<L]}\ \{p(L, \theta)\ S(L)\}^{[T_i \geq L, C_i \geq L]}$ where S is the survival distribution and f the density of T, and g the density of C. The derivative with respect to θ of the log-likelihood is the $q \times 1$ vector

$$\begin{aligned} d_i &= (1 - s_i)\ \nabla_\theta\, g(Z_i, \theta)/g(Z_i, \theta) + [T_i < L]\ s_i \nabla_\theta\, p(Z_i, \theta)/p(Z_i, \theta) \\ &\quad + [T_i \geq L, C_i \geq L]\ \nabla_\theta\, p(L, \theta)/p(L, \theta) \\ &= \ (1 - s_i)\, \nabla_\theta\, g(Z_i, \theta)/g(Z_i, \theta) + s_i p(Z_i, \theta)/p(Z_i, \theta)\ . \end{aligned}$$

Then $\hat{\theta}$ is obtained as a solution to $\sum_{i=1}^{N} d_i(\theta) = 0$. Under the usual regularity conditions,

$$\sqrt{N}\,(\hat{\theta} - \theta) = -J^{-1}(\theta) \left(N^{-1/2}\sum_{i=1}^{N} d_i(\theta)\right) + o_p(1)\,, \tag{A2}$$

where $J(\theta) = E\{\dot{d}_i(\theta)\} = -E\{d_i(\theta)\ d_i'(\theta)\}$ is a $q \times q$ matrix assumed to be positive definite. We expand the second term in (A1) as

$$\begin{aligned} N^{-1/2}\sum_{i=1}^{N} s_i \boldsymbol{x}_i' u_i / p(T_i^*, \hat{\theta}) = {} & N^{-1/2}\sum_{i=1}^{N} s_i \boldsymbol{x}_i' u_i / p(T_i^*, \theta) \\ & - \left(N^{-1}\sum_{i=1}^{N} s_i \boldsymbol{x}_i' u_i / (p(T_i^*, \tilde{\theta}))^2\ \nabla_\theta'\, p(T_i^*, \tilde{\theta})\right) \sqrt{N}\,(\hat{\theta} - \theta)\,, \end{aligned} \tag{A3}$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and θ. Again by standard arguments the term in parenthesis on the right hand side converges to $E((s_i \boldsymbol{x}_i' u_i / (p(T_i^*, \theta))^2)\ \nabla_\theta'\, p(T_i^*, \theta)) = E\big((\boldsymbol{x}_i' u_i / p(T_i^*, \theta))\ \nabla_\theta'\, p(T_i^*, \theta)\big) = D(\theta)$, a $K \times q$ matrix. Define $k_i = s_i \boldsymbol{x}_i' u_i / p(T_i^*, \theta)$ and note that $D(\theta) = E(k_i d_i')$. From (A2) and (A3)

$$N^{-1/2}\sum_{i=1}^{N} s_i \boldsymbol{x}_i' u_i / p(T_i^*, \hat{\theta}) = N^{-1/2}\sum_{i=1}^{N} \{k_i - D(\theta)\ J^{-1}(\theta)\ d_i(\theta)\} + o_p(1)\,. \tag{A4}$$

By the law of large numbers, the consistency of $\hat{\theta}$ and standard convergence results (Newey and McFadden, 1994), the matrix in (A1) converges in probability to $E(s_i \boldsymbol{x}_i' \boldsymbol{x}_i / p(T_i^*, \theta)) = E(\boldsymbol{x}_i' \boldsymbol{x}_i) = A$. Hence from (A1) and (A4) we obtain

$$\sqrt{N}\,(\tilde{\boldsymbol{\beta}}_w - \boldsymbol{\beta}) = A^{-1} N^{-1/2} \sum_{i=1}^{N} \{k_i - D(\theta)\, J^{-1}(\theta)\, d_i(\theta)\} + o_p(1)\,. \tag{A5}$$

Application of the central limit theorem to (A5) reveals that $\sqrt{N}\,(\tilde{\boldsymbol{\beta}}_w - \boldsymbol{\beta})$ converges in distribution to a K-variate normal with mean vector 0 and covariance matrix V, where $V = A^{-1} B A^{-1}$ and $B = E(k_i k_i') - D(\theta)\, J^{-1}(\theta)\, D'(\theta)$. Had we ignored the estimation of θ assuming $p(t, \theta)$ were known, this variance $V_0 = A^{-1} E(k_i k_i') A^{-1}$. Since $V_0 - V$ is non negative definite, estimation of the variance of $\tilde{\boldsymbol{\beta}}_w$ that ignores first stage estimation is conservative. Also, if given $\boldsymbol{x}$, (y, T, C) are independent then under $E(u \mid \boldsymbol{x}) = 0$ we get $D(\theta) = 0$ and so $V_0 = V$. Hence use of an estimate of the selection probability in the second stage yields a variance estimator that is asymptotically equivalent to that estimated with known selection probability.

Acknowledgements The authors thank Jeffrey M. Wooldridge for his helpful comments on various versions of this paper. This research was supported in part by the National Institute of Nursing Research, the National Cancer Institute under grant NR 1915-06, the Agency for Healthcare Research Quality under grant HS09514.

References

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

Breusch, T. S. and Pagan, A. R. (1980). The LM test and its application to model specification in econometrics. *Review of Economic Studies* **47**, 239–254.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251–1271.

Horowitz, J. L. and Manski, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics* **84**, 37–58.

Horvitz, D. and Thompson D. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association* **47**, 663–685.

Katz, J. N., Chung, L. C., Sango, O., Fossel, A H. and Bates, D. W. (1996). Can comorbidity be measured by questionnaire rather than medical record review? *Medical Care* **34**, 73–84.

Lancaster, T. (1992). *Econometric Analysis of Transition Data*. Cambridge University Press, USA. pages 176–178.

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

Lin, D. Y. (2000). Linear regression analysis of censored medical cost. *Biostatistics* **1**, 35–47.

Manning, W. G. and Mullahy, H. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics* **20**, 461–494.

Newey, W. K., McFadden, D. (1994). Large sample estimation and hypothesis testing, in: R. F. Engle and D. McFadden, eds. *Handbook of Econometrics* **4**, 2111–2245.

Newschaffer, C. J., Penberthy, L. and Desch, C. L. (1996). The effect of age and comorbidity in the treatment of elderly women with nonmetastatic breast cancer. *Archives of Internal Medicine* **156**, 85–90.

Robins, J. M., Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers, in: H. Jewell, K. Dietz and V. Farewell, eds. *AIDS Epidemiology-Methodological Issues*. pages 297–331.

Robins, J. M., Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rosenbaum, P. R. (1997). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 387–394.

Sloan, J. A., Cha, S. S., Wagner, J. L., Alberts, S. R. and Lindman J. (1999). Analyzing oncology patient health care costs using the SAS system. *SUGI-24*, Paper 284.

Ware, J., Snow, K. K. and Kosinski, M. (2000). *Health Survey. Manual and Interpretation Guide*. Lincoln, RI, Quality Metric, Inc.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Wooldridge, J. M. (1990). An encompassing approach to conditional mean tests with applications to testing nonnested hypotheses. *Journal of Econometrics* **6**, 1385–1406.

Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability sampling. *Econometrica* **6**, 1385–1406.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA, MIT Press, USA, page 347.