2006

Longitudinal analysis of censored medical cost data

Onur Baser
Thomson-Medstat

Joseph C. Gardiner
Michigan State University College of Law

Cathy J. Bradley
Virginia Commonwealth University

Huseyin Yuce
CUNY New York City College of Technology

Charles Given
Michigan State University

HEALTH ECONOMICS
Health Econ. **15**: 513–525 (2006)
Published online 3 January 2006 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.1087

Longitudinal analysis of censored medical cost data

Onur Başer[a,*], Joseph C. Gardiner[b], Cathy J. Bradley[c], Hüseyin Yüce[d] and Charles Given[e]
[a] *Thomson-Medstat, Ann Arbor, USA*
[b] *Division of Biostatistics, Department of Epidemiology, Michigan State University, USA*
[c] *Department of Health Administration, Virginia Commonwealth University, Richmond, USA*
[d] *Department of Mathematics, Florida International University, USA*
[e] *Department of Family Practice, Michigan State University, USA*

Summary

This paper applies the inverse probability weighted (IPW) least-squares method to estimate the effects of treatment on total medical cost, subject to censoring, in a panel-data setting. IPW pooled ordinary-least squares (POLS) and IPW random effects (RE) models are used. Because total medical cost might not be independent of survival time under administrative censoring, unweighted POLS and RE cannot be used with censored data, to assess the effects of certain explanatory variables. Even under the violation of this independency, IPW estimation gives consistent asymptotic normal coefficients with easily computable standard errors. A traditional and robust form of the Hausman test can be used to compare weighted and unweighted least squares estimators. The methods are applied to a sample of 201 Medicare beneficiaries diagnosed with lung cancer between 1994 and 1997. Copyright © 2006 John Wiley & Sons, Ltd.

JEL classification: C23; I1

Keywords censoring; longitudinal analysis; inverse probability weighted estimation; pooled OLS; random effect

Introduction

Rising health care expenditures in many industrialized countries has spurred the development of methods for analysis of medical costs in conjunction with evaluation of health outcomes. Challenges in analyzing cost data include addressing skewness in cost distributions, heterogeneity across samples and more challenging, complexities due to censoring.

Ordinary least squares estimation (OLS) can be used to analyze cost data under exogenous censoring. With exogenous censoring, once covariates have been selected, the total cost Y over the period T (survival time) are assumed independent. With longitudinal data, administrative censoring is due to study termination when, for instance, the analyst chooses a closing date for data collection. In these circumstances, OLS is not appropriate because total costs and survival time are likely to be associated. Because longer survival times and their associated costs are more likely to be censored, estimates of cost based only on the uncensored cases are biased towards patients with shorter survival times.

In this paper, we apply an inverse probability weighted (IPW) least squares method to assess the effects of covariates (e.g. patient and clinical characteristics) on medical cost with censored data. In our application to costs in patients

*Correspondence to: Thomson-Medstat, 777 Eisenhower Parkway, Ann Arbor, MI 48108, USA.
E-mail: onur.baser@thomson.com

Received 15 December 2004
Accepted 25 October 2005

diagnosed with lung cancer, we aim to observe the effects of treatment on average cost per subperiod (e.g. monthly, quarterly) over a circumscribed window of observation. In particular, our method examines how various treatment regimens (e.g. surgery only, chemotherapy, radiation and combinations of surgery, chemotherapy, and radiation) affects the cost of lung cancer care per month over the two years of initial diagnosis.

The IPW least squares method has a long history in statistics [1–6]. Our work is strongly influenced by the more general framework that develops the asymptotic properties of the IPW M-estimator for variable probability samples [7, 8]. IPW least squares produces consistent asymptotically normal coefficients with easily computable standard errors, even under violation of the exogenous censoring assumption.

Other published applications of IPW estimation included Lin [9, 10] Jain and Strawderman [11] and Willan *et al*. [12]. Lin [9, 10] developed a method to estimate the mean cost conditional on covariates from data subject to censoring. Jain and Strawderman [11] extend Lin's method to implement inverse probability of censoring weighted estimation in a hazard regression model for the conditional distribution of life time cost given covariates. Both of these methods analyze the cross-sectional data. Willan *et al*. [12] proposed an extension of Lin's methods to allow for longitudinal structure for cost effectiveness analysis. In particular, the researchers use seemingly unrelated regression (SUR) equations when comparing two groups in a cost effectiveness analysis. In contrast, our model controls for both continuous and several categorical time dependent variables, whereas Willan *et al*. [12] have only one categorical time dependent variable.

Another popular approach in the health services literature is to create a measure of cost per-individual per-month from longitudinal data on expenditures over the period of cost accumulation. Our proposed method is based on panel data. Thus, it differs from that of Lin [9,10] and Jain and Strawderman [11]. Using data gathered over time from the same cross sectional units is useful for several reasons. First, it allows us to examine dynamic relationships, which is not possible with a single cross section. Second, the panel data structure extends Lin's method to accommodate covariates that are time dependent.

The panel data model is conceptually different from the SUR model. The errors, for example, are homoskedastic and serially independent both within and between individuals. In SUR models the errors are allowed to be contemporarily correlated and heteroskedastic between individuals. If there are a large number of independent individuals observed (more than 500) for a few time periods (less than 30), it is not possible to estimate different individual slopes for all the exogeneous variables. Panel data are not subject to these restrictions.

None of the previous work using IPW methods offers a test to compare their methodology with potential bias methods. As a secondary contribution, we show how to apply a traditional and robust form of the Hausman test [13] to determine if systematic differences are present between OLS and IPW least squares methods. This allows us to determine whether bias introduced by applying OLS on the uncensored data leads to statistically significant differences in the coefficients.

We first introduce IPW pooled ordinary least squares (POLS) and IPW random effects (RE) models. The choice between the two models is dependent upon the presence of unobserved heterogeneity in the data. The next section describes the proposed Hausman type of test. We demonstrate our methods for assessing covariate effects on costs using data from Medicare claim files for a sample of patients diagnosed with lung cancer. Further, we present the detail of application. The final section summarizes our findings. All technical details are presented in Appendix A.

General framework

Suppose that we are interested in the total medical cost over period $[0, L]$. If there are data on cost and explanatory variables at multiple intervals such as months or years, they fit naturally into a panel format. Let the entire time period of interest be divided into G intervals: $0 = t_0 < t_1 < \cdots < t_G = L$. Since there is no further medical expense after death, the total cost over $(t_{g-1}, t_g]$ is the same as the cost incurred up to $T_g^* = \min(T, t_g)$, where T is the survival time. The distribution of T is assumed to be continuous from 0 to L.

Survival time and medical cost may be subject to right censoring and therefore are not always fully observable. Censoring of cost occurs when a patient's follow-up time is less than t_G, and the patient is alive at the time of censoring. Because no

further expense is incurred after death, for all observed deaths the total costs are known.

One advantage of dividing the total period into intervals is that we can consider the ith individual as uncensored in the gth interval $(t_{g-1}, t_g]$ whenever the censoring time C exceeds the minimum of T and t_g. Therefore, some individuals regarded as censored in studies which we do not partition the period of interest can be considered uncensored in some intervals during the period of interest.

For the ith individual let $T_i^* = \min(T_i, L)$, $Z_i = \min(T_i^*, C_i)$ and $s_{ig} = I(C_i \geq T_{ig}^*)$, where $I(.)$ is the indicator function. Therefore, cost in the gth interval is censored if $s_{ig} = 0$.

Let y_{ig} be the medical cost (or log-transformed cost) for the ith individual in the interval $(t_{g-1}, t_g]$. If there is an initial cost at $t = 0$, we include that cost in first time interval. The following situations arise:

(a) $Z_i \geq t_g$: Here the patient survives beyond t_g and is not censored by time t_g. Therefore y_{ig} is observed.
(b) $Z_i \geq t_{g-1}$ *and* $s_{ig} = 1$: If death occurs in $(t_{g-1}, t_g]$ then T_i is observed. The period costs y_{ig} in $(t_{g-1}, T]$ is observed. If death does not occur in $(t_{g-1}, t_g]$ then we are back to (a).
(c) $Z_i \geq t_{g-1}$ *and* $s_{ig} = 0$: Here censoring occurs in $(t_{g-1}, t_g]$ and the cost y_{ig} is censored.
(d) $Z_i < t_{g-1}$: Either death or censoring precedes t_{g-1}. Therefore, the cost y_{ig} in $(t_{g-1}, t_g]$ is either zero (if death had occurred) or is censored.

(a)–(d) captures all possibilities. For example, if our study is 12 months and costs are assessed monthly, then an observed death in month 1 would mean that $y_{it} (\geq 0)$ is observed, and $y_{i2} = \cdots = y_{i12} = 0$ for the next 11 months. According to our model we will use all the y_{ig}'s as long as $s_{ig} = 1$. This is true for the POLS estimator $\hat{\boldsymbol{\beta}}_{up}$ and IPW POLS estimator $\hat{\boldsymbol{\beta}}_{wp}$ that we describe next. The asymptotic theory of the estimators $\hat{\boldsymbol{\beta}}_{up}$ and $\hat{\boldsymbol{\beta}}_{wp}$ still hold true, although one could claim that we could do better modelling the zero-cost observations. We will mention this again after introducing our random effects model.

Pooled ordinary least squares estimation (POLS)

The properties of POLS under exogenous censoring can be summarized as follows. Assume the usual linear model for independent identically distributed cross-sections: for each i

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i \quad i = 1, 2, \ldots, N$$

where $\mathbf{X}_i = (\mathbf{x}_{i1}', \mathbf{x}_{i2}', \ldots, \mathbf{x}_{iG}')'$ is $G \times K$ matrix of explanatory variables, $\boldsymbol{\beta}$ is the $K \times 1$ vector of unknown regressions parameters, $\mathbf{u}_i$ is $G \times 1$ vector of unobservables whose distribution is unspecified. Let $\mathbf{S}_i$ be a $G \times G$ matrix whose gth diagonal element $s_{ig} = 1$ if $(\mathbf{x}_{ig}, y_{ig})$ is observed, zero otherwise. Generally, we have an unbalanced panel. We can define our explanatory variables and response variables for the selected sample as $\tilde{\mathbf{X}}_i = \mathbf{S}_i\mathbf{X}_i$, $\tilde{\mathbf{y}}_i = \mathbf{S}_i\mathbf{y}_i$.

Assumption 1.

(i) $E(\mathbf{u}_i|\mathbf{X}_i) = 0$;
(ii) $E(\mathbf{u}_i|\mathbf{X}_i) = E(\mathbf{u}_i|\mathbf{X}_i, \mathbf{S}_i)$;
(iii) $E(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)$ has rank K.

Under assumption 1, the unweighted POLS estimator $\hat{\boldsymbol{\beta}}_{up}$ of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{up} = \left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'\tilde{\mathbf{y}}_i\right)$$

is consistent, asymptotically normal with its asymptotic robust variance matrix estimated by

$$\hat{\mathbf{V}}_{up} = \hat{\mathbf{A}}_{up}^{-1}\hat{\mathbf{B}}_{up}\hat{\mathbf{A}}_{up}^{-1}/N$$

where

$$\hat{\mathbf{A}}_{up} = \left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\right)$$

$$\hat{\mathbf{B}}_{up} = N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i'(\tilde{\mathbf{u}}_i)(\tilde{\mathbf{u}}_i)'\tilde{\mathbf{X}}_i$$

and $\tilde{\mathbf{u}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}_{up}$.

Assumption 1(ii) is the key exogenous censoring assumption underlying the validity of the unweighted POLS estimator from the censored sample. This assumption is not true in the estimation of medical cost from administratively censored data, because assumption 1(ii) entails, for all g

$$E(y_{ig}|\mathbf{x}_{ig}, s_{ig}) = E(y_{ig}|\mathbf{x}_{ig}) \tag{1}$$

Under administrative censoring, although C_i and y_{ig} are independent, y_{ig} and T_i could be correlated. We will see that IPW least squares estimation produces a consistent, asymptotically normal estimator of $\boldsymbol{\beta}$ even when (1) does not hold, but

under the following assumptions. Suppose that T and C are independent given $\mathbf{x}$.

Assumption 1′.

(i) $E(\mathbf{X}_i'\mathbf{u}_i) = 0$;
(ii) $E(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)$ has rank K;
(iii) $\mathbf{x}_{ig}$ and y_{ig} are *ignorable* in the censoring equation, that is

$$P(s_{ig} = 1|\mathbf{x}_{ig}, y_{ig}, T_i) \overset{A}{=} P(C_i \geq T_{ig}^*|\mathbf{x}_{ig}, y_{ig}, T_i) \overset{B}{=} P(C_i \geq \min(t_{ig}, L)|\mathbf{x}_{ig}, y_{ig}, T_i) \overset{C}{=} P(s_{ig} = 1|T_i)$$

A is the definition of $s_{ig} = 1$; B is the definition of T_{ig}^*; C is a consequence of the assumption that the censoring time C_i is independent of $(\mathbf{x}_{ig}, y_{ig}, T_i)$ and L is constant. For similar formats see for example Lin [9]. We observe T_i whenever T_i is uncensored, we observe C_i when $s_i = 0$. A weaker assumption would be that the censoring time C_i is independent of (y_{ig}, T_i) given $\mathbf{x}_{ig}$. The censoring probability is then $P(s_{ig} = 1|\mathbf{x}_{ig}, y_{ig}, T_i) = P(s_{ig} = 1|T_i)$.

Another advantage of weighting the observations, other than solving the censoring problem, is that we derive consistency with the weaker assumption 1′(i) rather than assumption 1(i). Assumption 1′(ii) is the appropriate rank condition. Assumption 1′(iii) requires that the censoring probability is observable when $s_{ig} = 1$.

Under Assumption 1′ the IPW POLS estimator is, $\hat{\boldsymbol{\beta}}_{wp}$

$$\hat{\boldsymbol{\beta}}_{wp} = \hat{\mathbf{A}}_{wp}^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{X}}_i'\hat{\mathbf{y}}_i\right)$$

where

$$\hat{\mathbf{A}}_{wp} = \left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{X}}_i'\hat{\mathbf{X}}_i\right)$$

$\hat{\mathbf{X}}_i = \mathbf{S}_i\mathbf{P}_i^{-1}\mathbf{X}_i$, $\hat{\mathbf{y}}_i = \mathbf{S}_i\mathbf{P}_i^{-1}\mathbf{y}_i$, and $\mathbf{P}_i$ is $G \times G$ diagonal matrix in which the gth diagonal element is $\sqrt{p_{ig}}$ where

$$p_{ig} = P(C_i \geq T_{ig}^*|T_i) = p(T_{ig}^*) \tag{2}$$

and $p(t) = P[C_i \geq t]$. Then, $\hat{\boldsymbol{\beta}}_{wp}$ is consistent, asymptotically normal and its asymptotic robust variance matrix is estimated by

$$\hat{\mathbf{V}}_{wp} = \hat{\mathbf{A}}_{wp}^{-1}\hat{\mathbf{B}}_{wp}\hat{\mathbf{A}}_{wp}^{-1}/N \tag{3}$$

where

$$\hat{\mathbf{B}}_{wp} = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{X}}_i'(\underline{\hat{\mathbf{u}}}_i)(\underline{\hat{\mathbf{u}}}_i)'\hat{\mathbf{X}}_i$$

and $\underline{\hat{\mathbf{u}}}_i = \hat{\mathbf{y}}_i - \hat{\mathbf{X}}_i\hat{\boldsymbol{\beta}}_{wp}$.

Each observation of $(\mathbf{y}_i, \mathbf{x}_i)$ is weighted by the inverse probability of appearing in the sample. Assumption 1′(iii) requires the function $p(t)$ to be known, so $\hat{\boldsymbol{\beta}}_{wp}$ is computable from observed data.

The estimated covariance matrix in (3) is the White heteroskedasticity-robust covariance matrix [14] applied to all variables for observation i in the gth interval and weighted by the inverse probability of appearing in the sample. Hence, under our assumptions censoring can be handled fairly easily because most standard statistics software programs compute a heteroskedasticity-robust covariance matrix.

Usually the sampling probability function, p_{ig}, is unknown and needs to be estimated. Assume a parametric form $p(t, \theta)$ for $p(t)$ is known except for the unknown θ. Let $s_i = I(C_i \geq T_i^*)$. Using the sample, $\{(Z_i, \bar{s}_i) : i = 1, \ldots, N\}$ where $\bar{s}_i = 1 - s_i$, we construct a consistent estimator $\hat{p}(t) = p(t, \hat{\theta})$ of $p(t)$. Then,

$$\hat{p}_{ig} = \hat{p}(T_{ig}^*, \hat{\theta}), \quad i = 1, \ldots, N; \ g = 1, \ldots, G \tag{4}$$

Application of Lemma 4.3 in [15] shows that if p_{ig} in (2) is replaced by $\hat{p}_{ig}$, under the conditions in which the uniform weak law of large numbers can be applied, then $\hat{\boldsymbol{\beta}}_{wp}$ consistently estimates $\boldsymbol{\beta}$. Except where censoring is exogenous, one should adjust the variance matrix in (3) to account for the first stage estimation of censoring probabilities. The adjusted variance matrix is given in (A7) in Appendix A.

Random effects model

Panel data usually provide researchers with a large number of data points that increase the degrees of freedom and reduce collinearity among explanatory variables. It also provides a way to resolve or reduce the magnitude of an econometric problem that often arises in empirical studies, namely, omitted variables that are correlated with explanatory variables. One has greater flexibility in controlling for the effects of unobserved variables by using information on both the intertemporal dynamics and the individuality of the entities being investigated [16].

Let us first investigate assumptions under which the random effects estimator is consistent under exogenous censoring. The model is the unobserved effects model for any i and all G time periods

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i \tag{5}$$

where $\mathbf{X}_i$ is $G \times K$, $\boldsymbol{\beta}$ is $K \times 1$, and $\mathbf{v}_i$ is the vector of composite errors, $\alpha_i\mathbf{j}_G + \mathbf{u}_i$, where α_i is the unobserved heterogeneity and $\mathbf{j}_G$ is $G \times 1$ vector with all entries equal to 1.

Assumption 2.

(i) $E(\mathbf{v}_i|\mathbf{X}_i) = 0$;
(ii) $E(\mathbf{v}_i|\mathbf{X}_i) = E(\mathbf{v}_i|\mathbf{X}_i, \mathbf{S}_i)$;
(iii) rank $E(\mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{X}_i) = K$;
(iv) $E(\mathbf{v}_i\mathbf{v}_i'|\mathbf{X}_i, \mathbf{S}_i) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \mathbf{R}^{-1}(\mathbf{R}')^{-1}$. Assuming $\boldsymbol{\Omega}$ is positive definite, $\mathbf{R}$ can be taken as the unique $G \times G$ lower triangular, nonsingular matrix with positive diagonal elements.

As with the POLS, a random effect analysis, puts α_i into the error term and imposes more restrictive assumptions. The random effect approach exploits the serial correlation in the composite error in a generalized least squares (GLS) framework. In order to ensure feasible GLS is consistent under exogenous censoring, we need assumption 2(i)–(iv).

Typically, we would assume that $\boldsymbol{\Omega}$ has the standard random effects form. This standard random effect form is $\boldsymbol{\Omega} = \sigma_u^2\mathbf{I}_G + \sigma_\alpha^2\mathbf{j}_G\mathbf{j}_G'$, where $E(u_{ig}^2) = \sigma_u^2$, $E(\alpha_i^2) = \sigma_\alpha^2$, $\mathbf{I}_G$ is $G \times G$ identity matrix and $\mathbf{j}_G\mathbf{j}_G'$ is the $G \times G$ matrix with unity in every element. There is a simple analytical form for $\mathbf{R}$ when $\boldsymbol{\Omega}$ has the random effect form. To see this, define $z_g = \{[(g\sigma_\alpha^2 + \sigma_u^2)]/[(g+1)\sigma_\alpha^2\sigma_u^2 + \sigma_u^4]\}^{1/2}$ for $g = 1, 2, \ldots, G$ and $z_0 = [1/(\sigma_\alpha^2 + \sigma_u^2)]^{1/2}$. Then $\mathbf{R}$ can be written as:

$$\begin{pmatrix} z_{G-1} & 0 & \cdots & \\ -\frac{\sigma_\alpha^2 z_{G-1}}{(G-1)\sigma_\alpha^2+\sigma_u^2} & z_{G-2} & 0 & \cdots \\ \vdots & -\frac{\sigma_\alpha^2 z_{G-2}}{(G-2)\sigma_\alpha^2+\sigma_u^2} & \ddots & \\ & \vdots & & z_0 \end{pmatrix}$$

However, this standard random effect form assumption on $\boldsymbol{\Omega}$ is not necessary for the following theoretical development. We can transform Equation (5) to

$$\mathbf{y}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{v}_i^*$$

where $\mathbf{y}_i^* = \mathbf{R}\mathbf{y}_i, \mathbf{X}_i^* = \mathbf{R}\mathbf{X}_i$ and $\mathbf{v}_i^* = \mathbf{R}\mathbf{v}_i$.

The reason why we choose $\mathbf{R}$ as a lower triangular matrix is due to the attrition problem. Note that $(\mathbf{x}_{ig}, y_{ig})$ is observed if and only if $(\mathbf{x}_{ig}, y_{ig})$ and $(\mathbf{x}_{is}, y_{is})$ are observed, $s < g$. Therefore because $\mathbf{R}$ is lower triangular, $(\mathbf{x}_{ig}^*, y_{ig}^*)$ is observed if and only if $(\mathbf{x}_{ig}, y_{ig})$ is observed. Then $\mathbf{S}_i\mathbf{X}_i^*$ is observed. This would not be true if we do not choose $\mathbf{R}$ lower triangular, or if we have other patterns of missing data.

Using this set-up, we obtain the unweighted GLS estimator of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{ur} = \left(N^{-1}\sum_{i=1}^N \mathbf{X}_i^{*\prime}\mathbf{S}_i\mathbf{X}_i^*\right)^{-1}\left(N^{-1}\sum_{i=1}^N \mathbf{X}_i^{*\prime}\mathbf{S}_i y_i^*\right) \tag{6}$$

Obtaining GLS requires knowing $\boldsymbol{\Omega}$ up to scale. In feasible GLS (FGLS) estimation, we replace the unknown matrix $\boldsymbol{\Omega}$ with a consistent estimator and get asymptotic properties that are identical to those of the GLS estimator. For example, under the standard random effects form assumption, we can replace σ_α^2 and σ_u^2 with their consistent estimators, respectively,

$$\hat{\sigma}_\alpha^2 = \frac{1}{[NG(G-1)/2 - K]}\sum_{i=1}^N\sum_{g=1}^{G-1}\sum_{s=g+1}^{G}\tilde{u}_{ig}\tilde{u}_{is} \tag{7}$$

$$\hat{\sigma}_u^2 = \left(\frac{1}{[NG-K]}\sum_{i=1}^N\sum_{g=1}^{G}\tilde{u}_{ig}^2\right) - \hat{\sigma}_\alpha^2 \tag{8}$$

where $\tilde{u}_{ig}$ is the estimated ith POLS residual at the gth interval.

This estimator is feasible and the consistency of $\hat{\boldsymbol{\beta}}_{ur}$ follows under assumption 2(i)–(iv). Explicitly, by the usual law of large numbers argument, and by using Equation (6)

$$p\lim\hat{\boldsymbol{\beta}}_{ur} = [E(\mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{X}_i)]^{-1}E(\mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{y}_i) = \boldsymbol{\beta}$$

To obtain the asymptotic variance of $\hat{\boldsymbol{\beta}}_{ur}$, let $\mathbf{A}_{ur} = E(\mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{X}_i)$, and write

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{ur} - \boldsymbol{\beta}) = \mathbf{A}_{ur}^{-1}\left(N^{-1/2}\sum_{i=1}^N \mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{v}_i\right) + o_p(1) \tag{9}$$

The asymptotic variance of the bracketed term in (9) is $E(\mathbf{X}_i^{*\prime}\mathbf{S}_i\mathbf{R}\mathbf{v}_i\mathbf{v}_i'\mathbf{R}'\mathbf{S}_i\mathbf{X}_i^*)$. Under assumption 2(iv) this reduces to $\mathbf{A}_{ur}$. This shows that the asymptotic variance of the LHS of Equation (9) can be estimated by

$$\hat{\mathbf{V}}_{ur} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{X}_i \right)^{-1} \tag{10}$$

assuming that we know $\mathbf{R}$. Otherwise, assuming the standard form of $\boldsymbol{\Omega}$ and the derived form $\mathbf{R}$, (7) and (8) produce an estimate of $\mathbf{R}$.

Correlation between the survival times and medical costs would violate exogenous censoring assumption 2(ii), making $\hat{\boldsymbol{\beta}}_{ur}$ inconsistent. Inverse probability weighted estimation produces consistent and $\sqrt{N}$ asymptotically normal estimators even under violation of the assumption 2(ii) if the following assumptions hold.

Assumption 2′.

(i) $E(\mathbf{X}_i^{*\prime}\mathbf{v}_i^*) = 0$;
(ii) $E(\mathbf{X}_i^{*\prime}\mathbf{X}_i^*)$ has rank K;
(iii) $\mathbf{x}_{ig}$ and y_{ig} are *ignorable* in the selection equation, that is,

$$\begin{aligned} P(s_{ig} = 1|\mathbf{X}_i, y_i, T_i) &= P(s_{ig} = 1|T_i) \\ &= P(C_i \geq T_{ig}^*|T_i) \end{aligned}$$

As in the case of POLS, another advantage of weighting the observations, other than solving the censoring problem is that we derive consistency with the weaker assumption 2′(i) rather than assumption 2(i). Assumption 2′(ii) is the appropriate rank condition. In terms of conditioning set, assumption 2′(iii) is much stronger than the one presented under POLS section. Write $\tilde{\mathbf{S}}_i = \mathbf{S}_i\mathbf{P}_i^{-1}$.

Using this set-up, IPW RE estimator is

$$\hat{\boldsymbol{\beta}}_{wr} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i^{*\prime}\tilde{\mathbf{S}}_i\mathbf{X}_i^* \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i^{*\prime}\tilde{\mathbf{S}}_i\mathbf{y}_i^* \right)$$

We can estimate $\mathbf{R}$ by using IPW POLS residuals in Equations (7) and (8), assuming the selection probabilities in $\mathbf{P}_i$ are known or can be estimated. This makes the estimator feasible. To derive the consistency of $\hat{\boldsymbol{\beta}}_{wr}$ write

$$\hat{\boldsymbol{\beta}}_{wr} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{y}_i \right)$$

By the usual law of large numbers argument

$$p\lim \hat{\boldsymbol{\beta}}_{wr} = [E(\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i)]^{-1} E(\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{y}_i)$$

But the usual iterated expectations argument gives

$$\begin{aligned} E(\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i) &= E[\mathbf{X}_i'\mathbf{R}'E(\tilde{\mathbf{S}}_i|\mathbf{X}_i, \mathbf{y}_i)\mathbf{R}\mathbf{X}_i] \\ &= E[\mathbf{X}_i'\mathbf{R}'\mathbf{R}\mathbf{X}_i] \\ &= E[\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i] \end{aligned}$$

Essentially the same argument gives $E(\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{y}_i) = E[\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{y}_i]$. Therefore, under the assumption 2(i) and obvious rank condition assumption rank $E(\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i) = K$

$$p\lim \hat{\boldsymbol{\beta}}_{wr} = E[\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i]^{-1} E[\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{y}_i] = \boldsymbol{\beta}$$

To obtain the asymptotic variance of $\hat{\boldsymbol{\beta}}_{wr}$ let $\mathbf{A}_{wr} = E(\mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i)$, and write

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wr} - \boldsymbol{\beta}) = \mathbf{A}_{wr}^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{v}_i \right) + o_p(1)$$

Then

$$A\,\mathrm{var}[\sqrt{N}(\hat{\boldsymbol{\beta}}_{wr} - \boldsymbol{\beta})] = \mathbf{A}_{wr}^{-1}\mathbf{B}_{wr}\mathbf{A}_{wr}^{-1}$$

where $\mathbf{B}_{wr} = E(\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{v}_i\mathbf{v}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i)$. Both $\mathbf{A}_{wr}$ and $\mathbf{B}_{wr}$ can be consistently estimated, and there are no simplifications even under all the assumptions of the random effects model in the population. The estimated asymptotic variance of IPW RE estimator is, therefore,

$$\begin{aligned} \hat{\mathbf{V}}_{wr} = &\left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i \right)^{-1} \\ &\times \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\hat{\underline{\mathbf{y}}}_i\hat{\underline{\mathbf{y}}}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i \right) \\ &\times \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i \right)^{-1} \end{aligned} \tag{11}$$

where $\hat{\underline{\mathbf{y}}}_i = \mathbf{y}_i^* - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}_{wr}$.

As in the case of POLS, except when the censoring is exogeneous, $\hat{\mathbf{V}}_{wr}$ is unadjusted, because the estimation of $\mathbf{P}_i$ at the first stage has not been accounted for. The adjusted variance matrix can be obtained by applying the results from Appendix A. Usually, the adjustment for estimation at the first step has little effect on the asymptotic standard errors.

As mentioned previously, our cost vector $\mathbf{y}_i$ may include zero components. With RE model, the linear transformation $\mathbf{S}_i\mathbf{R}$ is applied to the vector $\mathbf{y}_i$ in the unweighted case, and $\mathbf{S}_i\mathbf{P}_i^{-1}\mathbf{R}$ is used in the weighted case. In both situations, it is very

unlikely in practice to have many zeros values in the final analysis vector $\mathbf{S}_i\mathbf{R}\mathbf{y}_i$ or $\mathbf{S}_i\mathbf{P}_i^{-1}\mathbf{R}\mathbf{y}_i$. The asymptotic theory of the estimators $\hat{\boldsymbol{\beta}}_{ur}$ and $\hat{\boldsymbol{\beta}}_{wr}$ is still valid. However, this does not say anything about how good the model fit would be.

The cost of treating patients who have died usually accelerate as the patient gets closer to death. Since our model allows for estimating cost in a period among those who died relative to those survived, we only need to add a time by death status interaction term. Costs of those who die in a period can be compared with costs of those who survived that period. These same groups can also be compared for the period before death. To model this we need: (i) intercept, (ii) period indicator, (iii) interaction of period and death indicators.

Weighted or unweighted estimator?

It has been shown that the unweighted estimator is no less efficient than the weighted estimator under homoskedasticity and exogenous censoring [7]. For a linear regression model, the Gauss–Markov Theorem for independent observations implies that the OLS estimator is the best linear unbiased estimator. It is better than any other weighted estimator, which is linear and unbiased.

Because the unweighted estimator is inconsistent when the censoring scheme is not exogenous and the weighted estimator is consistent with or without exogenous censoring, we can apply a Hausman test [13] to determine exogeneity of censoring. The traditional form of Hausman statistics can be used under the homoskedasticity assumption. We can state this assumption for the POLS estimator as follows:

$$E(\tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'\tilde{\mathbf{X}}_i) = \sigma_0^2 E(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i) \tag{12}$$

When Equation (12) holds, estimation of the unweighted POLS variance estimator is simplified further

$$\hat{\mathbf{V}}_{up} = \hat{\sigma}^2\hat{\mathbf{A}}_{up}^{-1} \tag{13}$$

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

In general form, the Hausman test statistic can be stated as

$$\mathbf{H} = (\hat{\boldsymbol{\theta}}_w - \hat{\boldsymbol{\theta}}_u)'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}_u)$$

The distribution of $\mathbf{H}$ under the null hypothesis is chi-square with K degrees of freedom. For weighted and unweighted POLS, choose $\hat{\boldsymbol{\theta}}_w$, $\hat{\boldsymbol{\theta}}_u$ as $\hat{\boldsymbol{\beta}}_{wp}$, $\hat{\boldsymbol{\beta}}_{up}$, respectively. $\hat{\mathbf{V}} \equiv \hat{\mathbf{V}}_w - \hat{\mathbf{V}}_u$, where $\hat{\mathbf{V}}_w$ is given by (3) and $\hat{\mathbf{V}}_u$ by (13) under the homoskedasticity assumption.

For the RE model, $\hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{\theta}}_u$ are $\hat{\boldsymbol{\beta}}_{wr}$ and $\hat{\boldsymbol{\beta}}_{ur}$, respectively. $\hat{\mathbf{V}}_w$ is given by (11) and $\hat{\mathbf{V}}_u$ by (10). In many cases we may want to use a Hausman test when the homoskedasticity assumption is violated. This requires a robust form that replaces $\hat{\mathbf{V}}$ for POLS estimation by

$$(\hat{\mathbf{A}}_{wp}^{-1} \mid -\hat{\mathbf{A}}_{up}^{-1})\left(N^{-1}\sum_{i=1}^{N}\sum_{g=1}^{G}\hat{\mathbf{e}}_{ig}\hat{\mathbf{e}}_{ig}'\right)(\hat{\mathbf{A}}_{wp}^{-1} \mid -\hat{\mathbf{A}}_{up}^{-1})'/N$$

where $(.|.)$ denotes the augmented matrix obtained by appending two matrices and $\hat{\mathbf{e}}_{ig} = (\hat{w}_{ig}\hat{u}_{ig}\mathbf{x}_{ig}', s_{ig}\tilde{u}_{ig}\mathbf{x}_{ig}')'$. $\hat{u}_{ig}$ and $\tilde{u}_{ig}$ are the residuals after weighted and unweighted POLS estimation. For RE estimation, we replace $\hat{\mathbf{V}}$ by

$$(\hat{\mathbf{A}}_{wr}^{-1} \mid -\hat{\mathbf{A}}_{ur}^{-1})\left(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{e}}_i\tilde{\mathbf{e}}_i'\right)(\hat{\mathbf{A}}_{wr}^{-1} \mid -\hat{\mathbf{A}}_{ur}^{-1})'/N \tag{14}$$

where $\tilde{\mathbf{e}}_i = (\mathbf{X}_i'\mathbf{R}'\tilde{\mathbf{S}}_i\mathbf{R}\mathbf{X}_i\hat{\mathbf{v}}_i, \mathbf{X}_i'\mathbf{R}'\mathbf{S}_i\mathbf{R}\mathbf{X}_i\tilde{\mathbf{v}}_i)'$, and $\hat{\mathbf{v}}_i, \tilde{\mathbf{v}}_i$ are the residuals after weighted and unweighted RE estimation.

If the Hausman test indicates rejection, then the assumption of exogenous censoring is violated, and the unweighted estimators are inconsistent. A failure to reject means the coefficients from unweighted and weighted estimators are not systematically different. The typical response is to conclude that the exogeneity assumption holds and therefore, we should use OLS estimates. Unfortunately due to the low power of the Hausman test we might commit a Type II error. Therefore, it is recommended that the results from both estimations be presented.

The lung cancer study

Data

The data set is derived from a broader study of health care cost, utilization and physical health function in a cohort of newly diagnosed elderly lung cancer patients recruited from several Michigan oncology clinics during 1994 through 1997. For our application, we use data from 201 (out of 223) Medicare beneficiaries age 65 or older who agreed to participate in this study. We excluded 22

patients because their demographic data was missing. Detailed cost data were obtained from Medicare claim files for each patient for a 2-year period following diagnosis. Payments by Medicare were used as a proxy for direct Medicare costs (as opposed to billed charges).

Patient demographic data were obtained through interviews. Physical function 3 months prior to diagnosis was measured by the short form SF-36. The physical function subscale of the SF-36 [17] is a 10 item measure of patients ability to perform a series of ordered activities including lifting, bending, stooping, and carrying packages of a given weight, walking different distances, and climbing stairs and performing self care activities such as dressing and bathing oneself. The scale is a weighted sum score with 100 representing high level of functioning and lower scores indicating persons who are less able to perform physical activities. Comorbid conditions were assessed using questions from the Aging and Health in America Survey (1996), which documents 15

Table 1. Summary statistics from the lung cancer study

Variable	Variable description	Mean ($N = 4335$)
Total cost	Total medicare payments	\$2620
	$\sum$ (Inpatient, Outpatient, Provider)	(\$7175)
Age	Patient's age within two weeks of	71.97
	initiating either radiation or chemotherapy	(4.85)
Physical functioning	Three months prior to diagnosis	72.10
	using the subscale from the SF-36	(27.30)
Symptoms	A count of all symptoms	10.87
		(4.99)
Comorbidity	= 1 if patient's comorbid	0.65
	conditions are three or more	(0.48)
Late stage	= 1 if patient's disease	0.64
	stage is regional, distant or invasive	(0.48)
White	= 1 if patient's race is white	0.93
	stage is regional, distant or invasive	(0.27)
Male	= 1 if patient's gender is male	0.59
	stage is regional, distant or invasive	(0.49)
Pays all	= 1 if insurance coverage pays all expenses	0.40
		(0.49)
Pays more	= 1 if insurance coverage requires minor expenses	0.48
		(0.50)
Pays little	= 1 if insurance coverage requires many expenses	0.09
		(0.29)
Pays none	= 1 if many services are not covered	0.03
		(0.17)
No treatment	= 1 if patient received	0.8104
	no treatment	(0.3920)
Surgery	= 1 if patient received	0.01407
	surgery only	(0.1178)
Surgery & Chemo	= 1 if patient received	0.0005
	surgery and chemotherapy	(0.2147)
Surgery & Radiation	= 1 if patient received	0.0002
	surgery and radiation	(0.0151)
Surgery & Chemo & Radiation	= 1 if patient received	0.0018
	surgery, chemotherapy and radiation	(0.0429)
Chemo & Radiation	= 1 if patient received	0.0209
	chemotherapy and radiation	(0.1434)
Chemotherapy	= 1 if patient received	0.0911
	chemotherapy only	(0.2878)
Radiation	= 1 if patient received	0.0609
	radiation only	(0.2392)

Standard deviations are in parentheses.

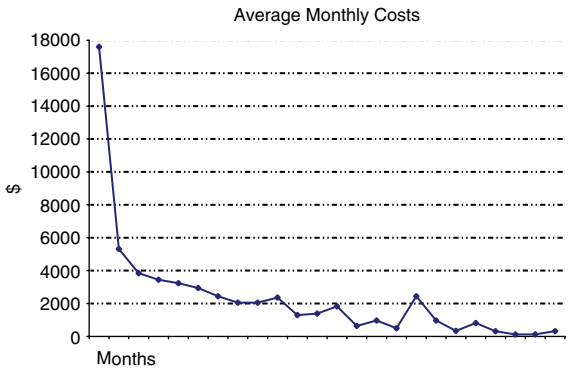

Figure 1. The distribution of average monthly cost values for uncensored cases

diseases and health problems other than lung cancer. Disease stage was determined by the Tumor Nodes and Metastasis (TNM) staging system of the American Joint Committee on Cancer (AJCC) using the pathological data obtained from audit of patients' medical records. Table 1 shows the summary statistics for each variable as well as short description of the variables.

A patient's medical cost was regarded as censored if the patient was alive at the end of 1997 and if follow up was less than two years. Because censoring is solely due to the limit of study duration, it is reasonable to assume that censoring is independent of all other random variables.

Figure 1 shows the distribution of average monthly costs for uncensored cases. Expenditure shows a spike in the first month after diagnosis due to surgery. Interventions such as surgery and radiation incur large costs within the first few months of following diagnosis, whereas chemotherapy, which is less costly, may be administered over a much longer time.

Regression analysis

Two analyses were performed to examine how patient- and treatment-related variables explain total medical cost for older persons newly diagnosed with lung cancer. In particular, we are interested in how various treatment regimens (e.g. surgery only, chemotherapy, radiation, and combinations of surgery, chemotherapy, and radiation) affected the total cost of lung cancer care. Total medical cost is the expenditure incurred from initiation of treatment until death or for a 2-year period, whichever comes first. Monthly expenditures were derived for this period. Following Manning and Mullahy [18] the cost estimates satisfied conditions for which an OLS-based model for log-transformed dependent variable was appropriate. One of the disadvantages of log-transformed models are zero cost observations. We did not have zero-cost as long as the individual was alive. Since we are considering total cost, some cost would be observed even there is no treatment. We did have a zero cost issue, if an individual died. In our sample, however, only 5 patients out of 201 died during the study period. We assumed that these patients had a cost of $1 per month, following the month death, so that when we transformed cost into its natural log, the cost per month would then be 0. If the percentage of patients died was higher, we would have used generalized linear model (GLM) approach suggested by Manning *et al.* [19]. Table 2 shows the results of the regression analysis for correlates of the total cost.

Because the population may have a different distribution in different periods we allowed the intercept to differ across different months. These are the time dependent factors. The first month after diagnosis was the base month and dummy variables were added for all other months. The estimated coefficients were all negative and statistically significant ($p < 0.05$). (These results are not shown).

The control variables include time independent covariates such as gender, race, comorbid conditions, stage of cancer and physical functions and time dependent covariates such as age and treatment-related variables. We divided treatment into seven categories: no treatment, radiation only, chemotherapy only, surgery and radiation, surgery and chemotherapy, chemotherapy and radiation, and finally surgery, chemotherapy and radiations. The latter was chosen as the reference group.

In our sample, all subjects are enrolled in Medicare and thus insurance payer is exogenous. Researchers using data from subjects that have other forms of insurance (or are uninsured) may want to include insurance payer as an explanatory variable.

Disease severity, as measured by cancer stage, had a statistically significant effect under both IPW RE and IPW POLS models. Regional stage decreased total cost of care almost 68% according to IPW POLS and 41% according to IPW RE compared to *in situ* or local stage cancer. On average, expenses for patients who had no

Table 2. Estimates of the log transformed total medical cost

	POLS	IPWPOLS	RE	IPWRE
Variable	$n = 4335$	$n = 4335$	$n = 4335$	$n = 4335$
Age	−0.0044	−0.0038	−0.033	−0.05342
	(0.0183)	(0.0184)	(0.01911)	(0.02196)
Physical functioning	0.0031	0.0024	0.0019	0.0065
	(0.0035)	(0.0035)	(0.0039)	(0.0042)
Symptoms	−0.0013	−0.0027	−0.0004	0.006
	(0.0191)	(0.0193)	(0.0205)	(0.0228)
Comorbidity	0.2053	0.2293	0.0204	0.3271
	(0.1989)	(0.1996)	(0.2054)	(0.2205)
Late stage	−1.1153	−1.1887	−1.0626	−1.1212
	(0.2049)**	(0.2086)**	(0.2040)**	(0.2301)**
White	−0.1955	−0.2289	−0.0306	−0.1614
	(0.3812)	(0.3752)	(0.4511)	(0.4803)
Male	−0.1843	−0.1846	−0.1691	−0.1299
	(0.1915)	(0.1946)	(0.1944)	(0.2144)
Pays more	−0.1746	−0.1783	−0.1243	−0.0471
	(0.1993)	(0.2004)	(0.3651)	(0.2277)
Pays little	−0.1817	−0.1492	−0.2696	−0.1446
	(0.3717)**	(0.3863)**	(0.3651)**	(0.4358)**
Pays none	−0.4758	−0.5136	−0.5501	−0.9925
	(0.3911)	(0.4073)	(0.3889)	(0.5359)
Surgery	5.7724	5.7976	5.3409	5.4178
	(0.2084)	(0.2098)	(0.2678)	(0.2691)
Surgery & Chemo	4.8626	4.8787	5.0051	5.1912
	(0.7230)	(0.7176)	(0.5007)	(0.4546)
Surgery & Chemo & Radiation	6.3000	6.3184	3.8124	3.8299
	(0.4585)	(0.4523)	(0.2067)	(0.20695)
Surgery & Radiation	5.7719	5.7822	5.7458	5.8515
	(0.2959)	(0.2989)	(0.3572)	(0.3705)
Chemo & Radiation	5.7356	5.7774	5.4732	5.5766
	(0.1936)	(0.1937)	(0.2273)	(0.2299)
Chemotheraphy	5.2680	5.3257	4.7609	4.8458
	(0.1768)**	(0.1792)**	(0.1787)	(0.1832)
Radiation	4.8976	4.9337	4.6567	4.7324
	(0.1804)**	(0.1807)**	(0.1879)*	(0.1877)*
Constant	5.2925	5.2426	5.6180	5.7946
	(1.6023)**	(1.5999)**	(1.7307)**	(1.9492)**
Monthly dummies	Yes	Yes	Yes	Yes
R-squared	0.7179	0.7305	0.7512	0.7564

Robust standard errors in parentheses *significant at 5%; **significant at 1%.

treatment were almost 99% less than for the patients who had surgery chemotherapy and radiation according to the IPW POLS and 4.46 times greater according to IPW RE models. A person who received radiation only had decreased the total medical cost relative to the average cost for persons with surgery plus adjuvant therapies. The estimates with respect to IPW POLS and IPW RE are 72 and 49%.

The Hausman test comparing the POLS and IPW POLS, and RE and IPW RE models, suggest that the exogenous censoring assumption is not violated. Thus, coefficients from weighted and unweighted estimations are both consistent.

Conclusion

Measurement of treatment cost is especially important in the evaluation of medical interventions, in the analysis of clinical trials, and in social

experiments. However, because cost records are incomplete, it is difficult to estimate cost accurately. Current statistical methods that would be applicable to administrative data, which is often censored, are under-developed.

One advantage of cost data is that they often fit naturally into a panel data format. This paper estimates medical cost per patient as a linear function of time varying covariates over a time interval $[0, L]$ following diagnosis. This interval is divided into G periods, so a panel structure arises. Censoring (in some periods) occurs when, for a given patient, the follow-up time is smaller than L and smaller than the survival time of the patient. The IPW least squares method was applied to longitudinal data to illustrate how possible censoring bias can be removed. The main motivation for developing the method is to handle a large number of continuous and discrete covariates.

We analyzed POLS and RE models and examined their statistical properties under censoring. Without exogenous censoring, the usual POLS and RE estimators are inconsistent. Generally, censoring is not exogenous because per-period medical cost may not be independent of survival time and the later is not independent of whether or not censoring occurs. To correct for censoring bias, we propose using IPW estimators, either in a pooled OLS or in a random effects framework. IPW estimators are consistent and $\sqrt{N}$ asymptotically normal. We also derived these estimators' first stage adjusted variance matrix.

Since unweighted POLS and RE estimators are consistent under exogenous censoring and more efficient under the homoskedasticity assumption, the Hausman test can be used to compare the systematic differences in coefficients between weighted and unweighted estimators. This test can be use to ascertain whether the exogenous censoring assumption is violated and whether the censoring bias creates statistically meaningful differences in the coefficients. We also derived and applied robust forms of the Hausman test in case the homoskedasticity assumption is violated.

Although it does not demonstrate the full power of the IPW least squares method, the lung cancer study demonstrated our proposed regression methods and test statistics. We fail to reject the hypothesis that the exogenous censoring assumption is violated. In order to see that this assumption was not violated in the lung cancer example, we needed to apply IPW estimation. Thus while the censoring bias created by applying POLS or RE on complete observations does not produce statistically different results than IPW POLS and IPE RE produce, though the latter two do correct for possible censoring bias.

One of the problems with medical cost data is fraction of zeros. This is especially dominant when we analyze subcategory costs such as inpatient costs. In order to deal with skewness generalized linear models (GLM) are proposed by several authors [18–20]. Especially with cost-per-individual per month analysis, those without the disease will have even higher fraction of zeroes than they do for annual data. Further, the subperiod data on positive expenditures will be even more skewed than is the case for annual data. Here, we are faced with the robustness-efficiency trade-off, which is very common in econometrics. Our analysis is probably more efficient with monthly data, but less robust than an analysis that uses just annual data. Although we focus on log-transformed OLS on modeling medical cost data, the present framework can be adapted to GLM models to deal with both zeroes and skewness assuming that the correct link function is known or is estimable.

Acknowledgements

The authors thank Jeffrey M. Wooldridge, participants of North America Econometric Society Meetings and two anonymous referees for their 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, and the Agency for Healthcare Research and Quality under Grant HS14206.

Appendix A: Derivation of IPW POLS Variance Matrix adjusted to first stage estimation of censoring probabilities

Let $\hat{\hat{\boldsymbol{\beta}}}$ be IPW POLS estimator

$$\hat{\hat{\boldsymbol{\beta}}}_{wp} = \left(N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig} \mathbf{x}_{ig} \mathbf{x}'_{ig}}{\hat{p}_{ig}} \right)^{-1} \times \left(N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig} \mathbf{x}_{ig} y_{ig}}{\hat{p}_{ig}} \right) \quad \text{(A1)}$$

where $\hat{p}_{ig}$ is defined in Equation (4) in the main text. It is convenient to express $\hat{p}_{ig}$ as $p(T^*_{ig}-,\hat{\theta})$, where $\hat{\theta}$ is the vector of estimated parameters that appear in the first stage estimation. As mentioned in the text, consistency of $\hat{\hat{\boldsymbol{\beta}}}_{wp}$ be easily read off from (A1) by using Lemma 4.3 in Newey and McFadden [15] under usual assumption. In the application here, we need to obtain the asymptotic variance of $\sqrt{n}(\hat{\hat{\boldsymbol{\beta}}}_{wp}-\boldsymbol{\beta}_{wp})$ when p_{ig}'s are estimated in the preliminary stage.

By substituting for y_{ig}, (A1) can be re-written as

$$\sqrt{n}(\hat{\hat{\boldsymbol{\beta}}}_{wp}-\boldsymbol{\beta}_{wp})=\left(N^{-1}\sum_{i=1}^{N}\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}\mathbf{x}'_{ig}}{p(T^*_{ig}-,\hat{\theta})}\right)^{-1}\times\left(N^{-1/2}\sum_{i=1}^{N}\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}u_{ig}}{p(T^*_{ig}-,\hat{\theta})}\right) \quad \text{(A2)}$$

Applying the uniform law of large numbers (Lemma 4.3 of Newey and McFadden [15]) shows that the first term on the right-hand side of (A2) converges to

$$E\left(\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}\mathbf{x}'_{ig}}{p(T^*_{ig}-,\theta_0)}\right)=E\left(\sum_{g=1}^{G}\mathbf{x}_{ig}\mathbf{x}'_{ig}\right)=\mathbf{A}_w \quad \text{(A3)}$$

where θ_0 is the true parameter. Standard maximum likelihood estimation is used to estimate θ_0 by $\hat{\theta}$. The part of the likelihood of (Z_i,s_i) that is relevant for estimation of θ has the form $\{p(Z_i,\theta)\}^{s_i}\{g(Z_i,\theta)\}^{1-s_i}\{p(t_G,\theta)\}[T\wedge C>t_g]$ where $g(t,\theta)$ is a density for C_i. We need to assume that $p(t_G,\theta)>0$ and that $\theta\to g(.,\theta)$ to fulfill all regularity conditions needed for maximum likelihood estimation of θ.

Note that $z_i\in(t_{g-1},t_g]$ and $s_i=1$ is equivalent to $[C_i\geq T^*_{ig}][t_{g-1}\leq T^*_{ig}<t_g]=s_{ig}I_g(T^*_{ig})$, whereas $z_i\in(t_{g-1},t_g]$ and $s_i=0$ is equivalent to $[t_{g-1}\leq C_i<T^*_{ig}]=(1-s_{ig})I_g(C_i)$, where $I_g(t)=[t_{g-1}\leq t<t_g]$. To include the interval $t\geq t_G$, define the indicator $I_{G+1}(t)=[t\geq t_G]$. Then the derivative with respect to θ of the aforementioned log-likelihood can be written

$$\sum_{i=1}^{G}\left\{s_{ig}I_g(T^*_{ig})\frac{\nabla_\theta p(T^*_{ig},\theta)}{p(T^*_{ig},\theta)}+(1-s_{ig})I_g(C_i)\frac{\nabla_\theta g(C_i,\theta)}{g(C_i,\theta)}+(1/G)I_{G+1}(T_i\wedge C_i)\frac{\nabla_\theta p(L,\theta)}{p(L,\theta)}\right\}=\sum_{i=1}^{G}d_{ig}(\theta)$$

The estimator $\hat{\theta}$ is a solution $\sum_{i=1}^{N}\sum_{i=1}^{G}d_{ig}(\theta)=0$. Consistency of $\hat{\theta}$ follows from the standard regularity conditions on the function $\theta\to g(.,\theta)$ for maximum likelihood estimation of θ.

Let θ_0 be the true parameter. Note that $d_{ig}(\theta_0)$ is a $q\times 1$ vector. Using a Taylor expansion of $\sum_{i=1}^{N}\sum_{i=1}^{G}d_{ig}(\hat{\theta})=0$ at θ_0, one can show

$$\sqrt{n}(\hat{\theta}-\theta_0)=J^{-1}(\theta_0)\left(\sqrt{N}\sum_{i=1}^{N}\sum_{g=1}^{G}d_{ig}(\theta_0)\right)+o_p(1) \quad \text{(A4)}$$

where $J(\theta_0)=-E(\sum_{g=1}^{G}d_{ig}(\theta_0))(\sum_{g=1}^{G}d'_{ig}(\theta_0))$ is a $q\times q$ matrix.

By using the standard asymptotic representation of a maximum likelihood estimator based on the information matrix equality and (A4), we can write the second term on the right-hand side of (A2) to get

$$N^{-1/2}\sum_{i=1}^{N}\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}u'_{ig}}{p(T^*_{ig}-,\hat{\theta})}=N^{-1/2}\sum_{i=1}^{N}\{\mathbf{k}_i-D(\theta_0)J^{-1}(\theta_0)\mathbf{d}_i(\theta_0)\mathbf{j}_G\}+o_p(1) \quad \text{(A5)}$$

where $\mathbf{d}_i(\theta_0)=[d_{i1}(\theta_0),\ldots,d_{iG}(\theta_0)]$ is a $q\times G$ matrix, $\mathbf{j}_G$ is a $G\times 1$ vector of 1's, and $D(\theta_0)=E(\sum_{i=1}^{N}\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}u_{ig}}{(p(T^*_{ig},\tilde{\theta}))^2}(\nabla_\theta p(T^*_{ig},\theta_0))')$ is a $K\times q$ matrix, and $\mathbf{k}_i=\sum_{g=1}^{G}\frac{s_{ig}\mathbf{x}_{ig}u'_{ig}}{p(T^*_{ig}-,\theta_0)}$ is a $K\times 1$ vector.

Combining the terms, (A2) can be re-written as

$$\sqrt{n}(\hat{\hat{\boldsymbol{\beta}}}_{wp}-\boldsymbol{\beta}_{wp})=\mathbf{A}_w^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{e}_i\right)+o_p(1) \quad \text{(A6)}$$

where $\mathbf{e}_i=(\mathbf{k}_i-D(\theta_0)J^{-1}(\theta_0)\mathbf{d}_i(\theta_0)\mathbf{j}_G)$. A direct calculation of the variance matrix of the right-hand side of (A6) shows that $\sqrt{n}(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_w)$ has asymptotic variance $\mathbf{V}_{wa}=\mathbf{A}_w^{-1}\mathbf{F}_w\mathbf{A}_w^{-1}$ where $\mathbf{F}_w=E(\mathbf{k}_i\mathbf{k}'_i)-D(\theta_0)J^{-1}(\theta_0)D'(\theta_0)$.

Under exogenous censoring (see assumption 1) we get $D(\theta_0)=0$ and so the asymptotic variance reduces to $\mathbf{V}_{wa}=\mathbf{A}_w^{-1}E(\mathbf{k}_i\mathbf{k}'_i)\mathbf{A}_w^{-1}$ which is the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}_{wp}$ in the main text (see (3)). This is the variance if the censoring probabilities were known. In general, the difference $\mathbf{V}_{wa}-\mathbf{V}_{wu}$ is negative definite which makes the asymptotic variance after first stage estimation no larger than that if the first stage was ignored.

References

1. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite population. *J Am Stat Assoc* 1952; **47**: 663–685.
2. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology-Methodological Issues*, Jewell H, Dietz K, Farewell V (eds). 1992; 297–331.
3. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
4. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc* 1995; **90**: 122–129.
5. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc* 1997; **82**: 387–394.
6. Horowitz JL, Manski CF. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *J Econ* 1998; **84**: 37–58.
7. Wooldridge JM. Asymptotic properties of weighted M-estimators for variable probability sampling. *Econometrica* 1999; **6**: 1385–1406.
8. Wooldridge JM. Asymptotic properties of weighted M-estimator for standard stratified samples. *Econometric Theory* 2001; **17**: 451–470.
9. Lin DY. Linear regression analysis of censored medical cost. *Biostatistics* 2000; **1**: 35–47.
10. Lin DY. Proportional means regression for censored medical cost. *Biometrics* 2000; **56**: 775–778.
11. Jain AK, Strawderman RL. Flexible hazard regression modeling for medical cost. *Biostatistics* 2002; **3**: 101–118.
12. Willan AR, Lin DY, Manca A. Regression methods for cost-effectiveness analysis with censored data. *Stat Med* 2005; **24**: 131–145.
13. Hausman JA. Specification tests in econometrics. *Econometrica* 1978; **46**: 1251–1271.
14. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; **48**: 817–838.
15. Newey WK, McFadden D. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, Engle RF, Mc Fadden D (eds). 1994; 2111–2245.
16. Hsiao C. *Analysis of Panel Data*. The University Press: Cambridge, 1999.
17. Ware Jr. JE, Sherbourne CD. The MOS 36item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473–483.
18. Manning WG, Mullahy H. Estimating log models: to transform or not to transform? *J Health Econ* 2001; **20**: 461–494.
19. Manning WG, Basu A, Mullahy H. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ* 2005; **24**: 465–488.
20. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ* 1999; **18**: 153–171.

