Longitudinal Analysis
of Censored Medical Cost Data

Onur Başer, a Joseph C. Gardiner, b Cathy J. Bradley, c
Huseyin Yuce, d Charles Given e

a The MEDSTAT Group, 777 Eisenhower Parkway, Ann Arbor, MI 48108
b Division of Biostatistics, Department of Epidemiology, Michigan State University, East Lansing, MI 48824
c Department of Medicine, Michigan State University, East Lansing, MI 48824
d Department of Mathematics, Michigan State University, East Lansing, MI 48824
e Department of Family Practice, Michigan State University, East Lansing, MI 48824
Abstract:

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 can not 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.

Key words: Censoring, Longitudinal Analysis, Inverse Probability Weighted Estimation, Pooled OLS, Random Effect
1 Introduction

A significant concern in many industrialized countries, the rising health care has triggered several responses. One is the search for cost effective therapies which requires proper analysis of medical cost data.

The analysis of medical-cost data can be challenging because of the complexities introduced by censoring. Previous methods to address these analytic complications hinged on the assumption of homogeneously distributed data—a rare occurrence in most empirical applications. Lin (2000a,b) developed a methodologically sound means for cost estimation using censored cross-section data. Another approach that has gained some popularity in the health-services literature is to creates per-member-per-month measure for the dependent variable by dividing expenditures during the window period by time in the window period (Etzioni et al., 1999). Lin (2000a) introduced a partitioned estimation technique appropriate for time independent covariates. We extend this work by proposing a technique appropriate for both time independent and dependent covariates.

Ordinary least squares (OLS) can be used to analyze cost data with exogenous censoring. With exogenous censoring, once covariates are selected (e.g. patient characteristics), total cost is independent of censoring and survival time. Administrative censoring, on the other hand, is caused by study termination when, for instance, the analyst chooses a closing date for data collection. OLS is not appropriate under these conditions because total cost and survival time are likely to be associated. Since longer survival times and their associated costs are more likely to be censored, estimates of cost based only on the un-
censored cases are biased toward 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. IPW has a long history in statistics (Horvitz et al., 1952; Robins et al., 1992, 1995; Robins & Rotnitzky, 1995; Rosenbaum, 1997; Horowitz et al., 1998). Our work is strongly influenced by the more general framework that develops the asymptotic properties of the IPW M-estimator for variable probability samples (Wooldridge, 1999, 2001). IPW produces consistent asymptotically normal coefficients with easily computable standard errors even under the violation of the exogenous censoring assumption.

We show how to apply a traditional and robust form of the Hausman test (Hausman, 1978) 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 cases only leads to statistically significant differences on the coefficients.

Expanding on methods developed by Lin (2000a) to account for time independent as well as time dependent covariates, 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. 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. Technical details are presented in the Appendix.
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: \(0 = t_0 < t_1 < ... < t_G = L\). Since there is no further medical expense after death, the total cost over \((t_{g-1}, t_g]\) is the same as the cumulative cost at \(T_g^* = \min(T, t_g)\), 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 \(i\)-th individual as uncensored in the \(g\)-th interval whenever the censoring time \(C\) exceeds the minimum of \(T\) and \(t_g\). Therefore, some individuals regarded as censored in the studies which do not partition the period of interest can be considered uncensored in some intervals during the period of interest.

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

Let \(y_{ig}\) be the medical cost per unit of time, i.e. it is the total medical (or log transformed) cost for the \(i\)-th individual for the interval \((t_{g-1}, t_g]\). If there is
initial cost at \( t = 0 \), we include that cost in the first time interval.

### 2.1 Pooled Ordinary Least Squares Estimation

The properties of POLS for the linear data under exogenous censoring can be summarized as follows. Assume that the model is the usual linear model for i.i.d cross-sections: for any \( i \),

\[
y_i = X_i \beta + u_i \quad i = 1, 2, ..., N
\]

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

**Assumption 1 :**

(i) \( E(u_i|X_i) = 0 \)

(ii) \( E(u_i|X_i) = E(u_i|X_i, S_i) \)

(iii) \( E(\tilde{X}_i \tilde{X}_i') \) has rank \( K \)

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

\[
\hat{\beta}_{up} = (N^{-1} \sum_{i=1}^{N} \tilde{X}_i \tilde{X}_i')^{-1}(N^{-1} \sum_{i=1}^{N} \tilde{X}_i' \tilde{y}_i);
\]
is consistent, asymptotically normal with its asymptotic robust variance matrix estimated by

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

where

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

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

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

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

$$E(y_{ig}|x_{ig}, s_{ig}) = E(y_{ig}|x_{ig}). \quad (1)$$

Under administrative censoring, although $C_i$ and $y_{ig}$ are independent, $y_{ig}$ and $T_i$ could be correlated. We will see that IPW estimation produces a consistent and $\sqrt{N}$ asymptotically normal estimator of $\beta$ even under the violation of equation (1), under the following assumptions. Suppose that $T$ and $C$ are independent given $x$.

Assumption 1’:

(i) $E(X'u_i) = 0$
(ii) $E(\hat{X}_i\hat{\beta}_i)$ has rank $K$

(iii) $x_{ig}$ and $y_{ig}$ are *ignorable* in the selection equation, that is,

$$P(s_{ig} = 1|x_{ig}, y_{ig}, T_i) = P(s_{ig} = 1|T_i) = P(C_i \geq T^*_{ig}|T_i).$$

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 selection probability is observable when $s_{ig} = 1$.

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

$$\hat{\beta}_{wp} = (N^{-1} \sum_{i=1}^{N} \hat{X}_i \hat{y}_i),$$

where

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

$$\hat{X}_i = S_i \hat{P}_i^{-1} X_i, \hat{y}_i = S_i \hat{P}_i^{-1} y_i,$$

and $P_i$ is $G \times G$ diagonal matrix in which the $g$-th diagonal element is $\sqrt{p_{ig}}$ where

$$p_{ig} = P(C_i \geq T^*_{ig}|T_i) = p(T^*_{ig}),$$

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

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

where
\[ \hat{B}_{wp} = N^{-1} \sum_{i=1}^{N} \hat{X}_i'(\hat{u}_i)(\hat{u}_i)'\hat{X}_i \]

and \( \hat{u}_i = \hat{y}_i - \hat{X}_i\hat{\beta}_{wp} \).

Each observation of \((y_i, x_i)\) is weighted by the inverse probability of appearing in the sample. Assumption \(1'(iii)\) requires the function \(p(t)\) to be known, so \( \hat{\beta}_{wp} \) is computable from observed data.

The estimated covariance matrix in (3) is the White heteroskedasticity-robust covariance matrix (White, 1980) applied to all variables for observation \(i\) in the \(g\)-th 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 \(\theta\). Let \(s_i = I(C_i \geq T_i^*)\). Using the sample, \(\{(Z_i, s_i) : i = 1, ..., N\}\) where \(s_i = 1 - s_i\), we construct a consistent estimator \(\hat{p}(t) = p(t, \hat{\theta})\) of \(p(t)\). Then,

\[ \hat{p}_{ig} = \hat{p}(T_i^*, \hat{\theta}) \quad i = 1, ..., N; g = 1, ..., G. \] (4)

Lemma 4.3 in (Newey and McFadden, 1994) 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 \(\hat{\beta}_{wp}\) consistently estimates \(\beta\). Except where censoring is exogenous, one should also adjust the variance matrix in (3) to the first stage estimation of the censoring probabilities. The adjusted variance matrix is given in (A7) in the appendix.
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 (Hsiao, 1999).

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,

\[
y_i = X_i \beta + v_i, \quad (5)
\]

where \( X_i \) is \( G \times K \), \( \beta \) is \( K \times 1 \), and \( v_i \) is the vector of composite errors, \( \alpha_i j_G + u_i \), where \( \alpha_i \) is the unobserved heterogeneity and \( j_G \) is \( G \times 1 \) vector with all entries equal to 1.

Assumption 2:

(i) \( E(v_i|X_i) = 0 \)

(ii) \( E(v_i|X_i) = E(v_i|X_i, S_i) \)

(iii) \( \text{rank } E(X_i' R S_i R X_i) = K \)
(iv) \( E(v_i v_i' | X_i, S_i) = \Omega \), where \( \Omega = R^{-1}(R')^{-1} \). Assuming \( \Omega \) is positive definite, \( R \) 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 \( \alpha_i \) 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 \( \Omega \) has the standard random effect form. This standard random effect form is \( \Omega = \sigma^2_u I_G + \sigma^2_\alpha j_G j_G' \), where \( E(u^2_{ig}) = \sigma^2_u \), \( E(\alpha^2_i) = \sigma^2_\alpha \), \( I_G \) is \( G \times G \) identity matrix and \( j_G j_G' \) is the \( G \times G \) matrix with unity in every element. There is a simple analytical form for \( R \) when \( \Omega \) has the random effect form. To see this, define \( z_g = \left\{ (g \sigma^2_\alpha + \sigma^2_u) / [(g+1)\sigma^2_\alpha \sigma^2_u + \sigma^4_u] \right\}^{1/2} \) for \( g = 1, 2, ..., G \) and \( z_0 = [1/(\sigma^2_\alpha + \sigma^2_u)]^{1/2} \). Then \( R \) can be written as:

\[
\begin{pmatrix}
    z_G & 0 & \cdots \\
    -\frac{\sigma^2_u z_{G-1}}{(G-1)\sigma^2_\alpha + \sigma^2_u} & z_{G-2} & 0 & \cdots \\
    \vdots & -\frac{\sigma^2_u z_{G-2}}{(G-2)\sigma^2_\alpha + \sigma^2_u} & \ddots & \ddots \\
    \vdots & \vdots & \ddots & z_0
\end{pmatrix}
\]

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

\[
y^*_i = X^*_i \beta + v^*_i;
\]

where \( y^*_i = R y_i, X^*_i = RX_i \) and \( v^*_i = R v_i \).
The reason why we choose $R$ as a lower triangular matrix is due to the attrition problem. Note that $(x_{ig}, y_{ig})$ is observed if and only if $(x_{ig}, y_{ig})$ and $(x_{is}, y_{is})$ are observed, $s < g$. Therefore because $R$ is lower triangular, $(x_{ig}^*, y_{ig}^*)$ is observed if and only if $(x_{ig}, y_{ig})$ is observed. Then $S_i X_i^*$ is observed. This would not be true if we do not choose $R$ lower triangular, or if we have other missing data.

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

$$
\hat{\beta}_{ur} = \left( N^{-1} \sum_{i=1}^{N} X_i^t S_i X_i^* \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} X_i^t S_i y_i^* \right).
$$

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

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

$$
\hat{\sigma}^2_u = \frac{1}{NG - K} \sum_{i=1}^{N} \sum_{g=1}^{G} \tilde{u}_{ig}^2 - \hat{\sigma}^2_\alpha,
$$

where $\tilde{u}_{ig}$ is the estimated $i$th POLS residual at the $g$-th interval.

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

$$
\text{plim} \hat{\beta}_{ur} = \left[ E(X_i^t R^t S_i R X_i) \right]^{-1} E(X_i^t R^t S_i R y_i) = \beta.
$$

To obtain the asymptotic variance of $\hat{\beta}_{ur}$, let $A_{ur} = E(X_i^t R^t S_i R X_i)$, and write
\[
\sqrt{N}(\hat{\beta}_{ur} - \beta) = A^{-1}_{ur} \left( N^{-1/2} \sum_{i=1}^{N} X_i' R'S_i Rv_i \right) + o_p(1). \tag{9}
\]

The asymptotic variance of the bracketed term in (9) is \( E(X_i^s'S_i Rv_i v_i'R'S_i X_i^s) \). Under assumption 2(iv) this reduces to \( A_{ur} \). This shows that the asymptotic variance of the LHS of equation (9) can be estimated by

\[
\hat{V}_{ur} = \left( N^{-1} \sum_{i=1}^{N} X_i' R'S_i R X_i \right)^{-1}, \tag{10}
\]

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

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

**Assumption 2’** :

(i) \( E(X_i^s'v_i^s) = 0 \)

(ii) \( E(X_i^s'X_i^s) \) has rank \( K \)

(iii) \( x_{ig} \) and \( y_{ig} \) are *ignorable* in the selection equation, that is,

\[
P(s_{ig} = 1|X_i, y_i, T_i) = P(s_{ig} = 1|T_i) = P(C_i \geq T_{ig}^*|T_i).
\]

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 $\tilde{S}_i = S_i P_i^{-1}$.

Using this set-up, IPW RE estimator is

$$\hat{\beta}_{wr} = \left( N^{-1} \sum_{i=1}^{N} X_i' \tilde{S}_i X_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} X_i' \tilde{S}_i y_i \right).$$

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

$$\hat{\beta}_{wr} = \left( N^{-1} \sum_{i=1}^{N} X_i' R' \tilde{S}_i R X_i \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} X_i' R' \tilde{S}_i R y_i \right).$$

By the usual law of large numbers argument,

$$\text{plim} \hat{\beta}_{wr} = \left[ E(X_i' R' \tilde{S}_i R X_i) \right]^{-1} E(X_i' R' \tilde{S}_i R y_i).$$

But the usual iterated expectations argument gives

$$E(X_i' R' \tilde{S}_i R X_i) = E[X_i' R' E(\tilde{S}_i | X_i, y_i) R X_i]$$
$$= E[X_i' R' R X_i]$$
$$= E[X_i' \Omega^{-1} X_i].$$

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

$$\text{plim} \hat{\beta}_{wr} = E[X_i' \Omega^{-1} X_i]^{-1} E[X_i' \Omega^{-1} y_i] = \beta.$$

To obtain the asymptotic variance of $\hat{\beta}_{wr}$ let $A_{wr} = E(X_i' \Omega^{-1} X_i)$, and write
\[
\sqrt{N}(\hat{\beta}_{wr} - \beta) = A_{wr}^{-1} \left( N^{-1} \sum_{i=1}^{N} X'_i \hat{R}' \hat{S}_i R v_i \right) + o_p(1).
\]

Then

\[
A_{\text{var}}[\sqrt{N}(\hat{\beta}_{wr} - \beta)] = A_{wr}^{-1} B_{wr} A_{wr}^{-1},
\]

where \( B_{wr} = E(X'_i \hat{R}' \hat{S}_i R v_i \hat{v}'_i \hat{R}' \hat{S}_i \hat{R} x_i) \). Both \( A_{wr} \) and \( B_{wr} \) 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,

\[
\hat{V}_{wr} = \left( N^{-1} \sum_{i=1}^{N} X'_i \hat{R}' \hat{S}_i \hat{R} x_i \right)^{-1} \times \left( N^{-1} \sum_{i=1}^{N} X'_i \hat{R}' \hat{S}_i \hat{R} \hat{v}'_i \hat{R}' \hat{S}_i \hat{R} x_i \right) \times \left( N^{-1} \sum_{i=1}^{N} X'_i \hat{R}' \hat{S}_i \hat{R} x_i \right)^{-1}
\]

(11)

where \( \hat{v}_i = y_i^* - X_i' \hat{\beta}_{wr} \).

As in the case of POLS, except when the censoring is exogeneous, \( \hat{V}_{wr} \) is unadjusted, because the estimation of \( P_i \) at the first stage has not been accounted for. The adjusted variance matrix can obtain by applying the results from the appendix \(^1\).

Usually, the adjustment for estimation at the first step has little effect on the asymptotic standard errors.

\(^1\) Let \( r_{ij} \) be the element of \( i \)th row and \( j \)th column of the matrix \( R \) and let \( x_{ig} = \sum_{j=1}^{g} x_{ij} r_{gj} \), \( y_{ij} = \sum_{j=1}^{g} y_{ij} r_{gj} \), \( u_{ij} = \sum_{j=1}^{g} u_{ij} r_{gj} \) for \( g = 1, \ldots, G \) in appendix equation (A1).
3 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 (Wooldridge, 1999). For a linear regression model, the Gauss-Markov Theorem for independent observations imply that the OLS estimator is the best linear unbiased estimator. It is better than a 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 (Hausman, 1978) 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:

\[
E(\hat{\mathbf{X}}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\hat{\mathbf{X}}_i) = \sigma_0^2 E(\hat{\mathbf{X}}_i'\hat{\mathbf{X}}_i). \tag{12}
\]

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 \(\sigma_0^2\).

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

\[
\mathbf{H} = (\hat{\mathbf{\theta}}_w - \hat{\mathbf{\theta}}_u)' \hat{\mathbf{V}}^{-1} (\hat{\mathbf{\theta}}_w - \hat{\mathbf{\theta}}_u).
\]
The distribution of $H$ under the null hypothesis is $\chi^2(K)$. For weighted and unweighted POLS, choose $\hat{\theta}_w, \hat{\theta}_u$ as $\hat{\beta}_{wp}, \hat{\beta}_{up}$, respectively. $\hat{V} \equiv \hat{V}_w - \hat{V}_u$, where $\hat{V}_w$ is defined in equation (3) and $\hat{V}_u$ is defined in equation (13) under the homoskedasticity assumption.

For the RE model, $\hat{\theta}_w$ and $\hat{\theta}_u$ are $\hat{\beta}_{wr}$ and $\hat{\beta}_{ur}$ respectively. $\hat{V}_w$ is as in equation (11) and $\hat{V}_u$ is as in equation (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 $\hat{V}$ for the POLS estimation:

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

where $\hat{e}_{ig} = (\hat{w}_{ig} \hat{u}_{ig} x_{ig}', s_{ig} \hat{u}_{ig} x_{ig}')'$. $\hat{u}_{ig}$ and $\hat{u}_{ig}$ are the residuals after weighted and unweighted POLS estimation. For RE estimation,

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

(14)

where $\tilde{e}_{i} = (X_i' \tilde{R} \tilde{S}_i \tilde{R} X_i \tilde{v}_i, X_i' \tilde{R} S_i \tilde{S} X_i \tilde{v}_i)'$, and $\tilde{v}_i, \tilde{v}_i$ are the residuals after weighted and unweighted RE estimation.

If the Hausman test indicates rejection, then the exogenous censoring assumption 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, we may be committing a Type II error, so results from both estimations should be presented.
4 The Lung Cancer Study

4.1 The Data

The sample consists of data from 201 Medicare beneficiaries over age 65 who were diagnosed with lung cancer from 1994 through 1997.

Medicare claim files for each patient for two years following diagnosis were obtained. These files contained monthly costs and treatments provided. Payments by Medicare were used as a proxy for direct Medicare costs as opposed to billed charges.

Patient information (such as age, sex, race) was obtained through interviews. In addition, we collected data on patients’ physical functions three months prior to diagnosis as measured by the short form 36 (SF-36). Comorbid conditions were assessed using questions from the Aging and Health in America Survey (1996), which documents 15 diseases and health problems other than lung cancer. Disease stage was determined by 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.

The medical costs are censored for patients alive at the end of 1997 and when patient follow-up is less than two years. Because censoring is solely caused by the limited study duration, it is reasonable to assume that censoring is independent of all other random variables.

The distribution of average monthly cost values for uncensored cases is given in
Figure 1. It shows that medical care expenditures for lung cancer patients spike in the first month after diagnosis, during the surgical period. Interventions such as surgery and radiation incur large costs within the first couple of months; whereas chemotherapy, which is less costly, may be administered over a much longer time.

4.2 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. Total medical cost is the expenditure incurred from initiation of treatment until death or during two years, whichever comes first. Following Manning et al. (2001), the cost estimates satisfied the conditions in which an OLS-based model with a long-transformed dependent variable is suitable.

Table 1 shows the results of the regression analysis predicting the total cost of care.

Because the population may have a different distribution in different periods, therefore we have allowed the intercept to differ across months. We chose the first month after diagnosis as the base month and included dummy variables for all but the first month after diagnosis. The coefficients were all negative and statistically significant \((p < .05)\). Since our main concern is to determine the effect of patient- and treatment-related variables on total medical cost, in the interest of brevity, we did not present the coefficients of monthly dummy variables.

The control variables are age, gender, race, comorbid conditions, stage of can-
cer, physical function, 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. Our time independent variables are gender, race, comorbid conditions, stage of cancer and physical function.

Disease severity measure, or stage, has statistically significant effects under IPW RE and IPW POLS. Regional stage decreased total cost of care almost 68% according to IPW POLS and 41% according to IP WRE compared to in situ or local stage cancer. On average, expenses for patients who had no treatment were almost 99% less than the patients who had surgery, chemotherapy and raditation according to the IPW POLS and IPW RE models. A person who received radiation only decreased the total medical cost relative to mean 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, in comparison with POLS and IPW POLS, RE and IPW RE models, suggests that the exogenous censoring assumption is not violated. Thus, coefficients from weighted and unweighted estimations are both consistent.

5 Conclusion

Measurement of treatment cost is especially important in the evaluation of medical intervention, in the analysis of clinical trials, and in social experiment. However, since cost records are incomplete, it is difficult to estimate
cost accurately. Currently, statistical methods the 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 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. Usual POLS and RE estimations create an inconsistent estimator without exogenous censoring. 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, produce consistent and \(\sqrt{N}\) asymptotically normal estimators. We also derived estimators’ first stage adjusted variance covariance 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. The derived test shows whether the exogenous censoring assumption is violated and whether the censoring bias
creates statistically sound differences in the coefficients. We also derived and applied the robust forms of the Hausman test in case the homoscedasticity 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 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.

Acknowledgements

The authors thank Jeffrey M. Wooldridge and participants of North America Econometric Society Meetings 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, the Agency for Healthcare Research & Quality under grant HS09514.
Appendix

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

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

$$\hat{\beta}_{wp} = \left( N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig}x_{ig}'y_{ig}}{\hat{p}_{ig}} \right)^{-1} \left( N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig}x_{ig}'y_{ig}}{\hat{p}_{ig}} \right)$$  \hspace{1cm} (A1)

where $\hat{p}_{ig}$ is defined in equation (4) in the main text. It is convenient to express $\hat{p}_{ig}$ as $p(T^*_g - \hat{\theta})$, where $\hat{\theta}$ is the vector of parameters that appear in the first stage estimation. As mentioned in the text, consistency of $\hat{\beta}_{wp}$ be easily read off from (A1) by using Lemma 4.3 in Newey and McFadden (1994) under usual assumption. In the application here, we need to obtain the asymptotic variance of $\sqrt{n}(\hat{\beta}_{wp} - \beta_{wp})$ when $p_{ig}$’s are estimated in the preliminary stage.

By the usual substitution, (A1) can be re-written as

$$\sqrt{n}(\hat{\beta}_{wp} - \beta_{wp}) = \left( N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig}x_{ig}'y_{ig}}{p(T^*_g - \hat{\theta})} \right)^{-1} \left( N^{-1/2} \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig}x_{ig}'y_{ig}}{p(T^*_g - \hat{\theta})} \right)$$  \hspace{1cm} (A2)

Applying the law of large numbers and standard uniform convergence results, for example Newey & McFadden (1994); the first term on the RHS of the (A2) converges to

$$E \left( \sum_{g=1}^{G} \frac{s_{ig}x_{ig}'x_{ig}'}{p(T^*_g - \hat{\theta})} \right) = E \left( \sum_{g=1}^{G} x_{ig}'x_{ig}' \right) = A_w.$$  \hspace{1cm} (A3)

The part of the likelihood of $(Z_i, s_i)$ that is relevant for estimation of $\theta$ 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

23
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 \( \theta \).

Note that \( z_i \epsilon(t_{g-1}, t_g) \) 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 \( z_i \epsilon(t_{g-1}, t_g) \) 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 \( \theta \) of the aforementioned log-likelihood can be written

\[
\sum_{i=1}^G \left\{ s_{ig}I_g(T_{ig}^*) \frac{\nabla_\theta p(T_{ig}^*, \theta)}{p(T_{ig}^*, \theta)} + (1 - s_{ig})I_g(C_i) \frac{\nabla_\theta g(C_i, \theta)}{g(C_i, \theta)} + \frac{1}{G}I_{G+1}(T_i \wedge C_i) \frac{\nabla_\theta p(L, \theta)}{p(L, \theta)} \right\} = \sum_{i=1}^G d_{ig}(\theta).
\]

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

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

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

where \( J(\theta_0) = -E \left( \left( \sum_{g=1}^G d_{ig}(\theta_0) \right) \left( \sum_{g=1}^G d_{ig}^*(\theta_0) \right) \right) \) 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

\[
N^{-1/2} \sum_{i=1}^N \sum_{g=1}^G s_{ig}X_{ig}u_{ig}' = N^{-1/2} \sum_{i=1}^N \{ k_i - D(\theta_0)J^{-1}(\theta_0) d_i(\theta_0) j_G \} + o_p(1), \quad (A5)
\]
where \( \mathbf{d}_i(\theta_0) = [d_{i1}(\theta_0), ..., d_{iG}(\theta_0)] \) is a \( q \times G \) matrix, \( \mathbf{j}_G \) is a \( G \times 1 \) vector of 1’s, and \( D(\theta_0) = E \left( \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{s_{ig}x_{ig}u_{ig}}{p(T_{ig}^*, \theta_0)^2} (\nabla \varphi(T_{ig}^*, \theta_0))^\prime \right) \) is a \( K \times q \) matrix.

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

\[
\sqrt{n}(\hat{\beta}_w - \beta_w) = \mathbf{A}_w^{-1}(N^{-1/2} \sum_{i=1}^{N} \mathbf{e}_i) + \text{op}(1). \tag{A6}
\]

where \( \mathbf{e}_i = (\mathbf{k}_i - D(\theta_0)J^{-1}(\theta_0)\mathbf{d}_i(\theta_0)\mathbf{j}_G) \).

The variance matrix adjusted for first stage estimation of censoring probabilities is,

\[
\sqrt{n}(\hat{\beta} - \beta_w) = \mathbf{A}_w^{-1} \mathbf{F}_w \mathbf{A}_w^{-1}, \tag{A7}
\]

where \( \mathbf{F}_w = E(\mathbf{e}_i\mathbf{e}_i') \).

One important implication is to see how “adjusted” and “unadjusted” variance matrix is equal when \( T_{ig}^*, C_i, \mathbf{u}_i \) is independent given \( \mathbf{x}_i \), i.e. exogenous censoring. This can be easily seen because \( D(\theta_0) = 0 \) under exogenous censoring, so the variance estimator given (A7) is equal to the one given in (3) in the main text.
References


Figure 1. The distribution of average monthly cost values for uncensored cases.
Table 1. Estimates of the log transformed total medical cost by POLS, IPWPOLS, RE and IPWRE Methods

<table>
<thead>
<tr>
<th>Variable</th>
<th>POLS</th>
<th>IPWPOLS</th>
<th>RE</th>
<th>IPWRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=4335</td>
<td>n=4335</td>
<td>n=4335</td>
<td>n=4335</td>
<td>n=4335</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.00863</td>
<td>-0.00808</td>
<td>-0.01212</td>
<td>-0.01215</td>
</tr>
<tr>
<td>(0.01762)</td>
<td>(0.01772)</td>
<td>(0.01928)</td>
<td>(0.01929)</td>
<td></td>
</tr>
<tr>
<td>PHYSICAL FUNCTIONING</td>
<td>0.00385</td>
<td>0.00391</td>
<td>0.00197</td>
<td>0.00196</td>
</tr>
<tr>
<td>(0.00359)</td>
<td>(0.00360)</td>
<td>(0.00403)</td>
<td>(0.00402)</td>
<td></td>
</tr>
<tr>
<td>SYMPTOMS</td>
<td>-0.00821</td>
<td>-0.00975</td>
<td>0.00059</td>
<td>0.00059</td>
</tr>
<tr>
<td>(0.01824)</td>
<td>(0.01827)</td>
<td>(0.02027)</td>
<td>(0.02027)</td>
<td></td>
</tr>
<tr>
<td>COMORBIDITY</td>
<td>0.20258</td>
<td>0.2284</td>
<td>0.05969</td>
<td>0.05922</td>
</tr>
<tr>
<td>(0.19960)</td>
<td>(0.20049)</td>
<td>(0.21900)</td>
<td>(0.21911)</td>
<td></td>
</tr>
<tr>
<td>LATE STAGE</td>
<td>-1.13064</td>
<td>-1.163218</td>
<td>-0.87603</td>
<td>-0.87591</td>
</tr>
<tr>
<td>(0.20322)**</td>
<td>(0.20700)**</td>
<td>(0.20708)**</td>
<td>(0.20738)**</td>
<td></td>
</tr>
<tr>
<td>WHITE</td>
<td>-0.18628</td>
<td>-0.21972</td>
<td>-0.02838</td>
<td>-0.02733</td>
</tr>
<tr>
<td>(0.37348)</td>
<td>(0.36675)</td>
<td>(0.44775)</td>
<td>(0.44751)</td>
<td></td>
</tr>
<tr>
<td>MALE</td>
<td>-0.18636</td>
<td>-0.18668</td>
<td>-0.19706</td>
<td>-0.19739</td>
</tr>
<tr>
<td>(0.19241)</td>
<td>(0.19570)</td>
<td>(0.20178)</td>
<td>(0.20188)</td>
<td></td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th></th>
<th>SURG. ONLY</th>
<th>SURG. CHEMO</th>
<th>NO TREATMENT</th>
<th>SURG.&amp; RAD</th>
<th>CHE.&amp; RAD</th>
<th>CHE. ONLY</th>
<th>RAD. ONLY</th>
<th>CONSTANT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.46294</td>
<td>-1.42077</td>
<td>-6.23854</td>
<td>-0.47386</td>
<td>-0.50907</td>
<td>-0.94254</td>
<td>-1.32588</td>
<td>11.73381</td>
</tr>
<tr>
<td></td>
<td>(0.41896)</td>
<td>(0.77451)</td>
<td>(0.43146)**</td>
<td>(0.40151)</td>
<td>(0.39625)</td>
<td>(0.42480)**</td>
<td>(0.39032)**</td>
<td>(1.32801)**</td>
</tr>
<tr>
<td></td>
<td>-0.45149</td>
<td>-1.41767</td>
<td>-6.25017</td>
<td>-0.40439</td>
<td>-0.48517</td>
<td>-0.91301</td>
<td>-1.30260</td>
<td>11.69306</td>
</tr>
<tr>
<td></td>
<td>(0.42208)</td>
<td>(0.40439)</td>
<td>(0.42569)**</td>
<td>(0.40439)</td>
<td>(0.39748)</td>
<td>(0.42584)**</td>
<td>(0.39143)**</td>
<td>(1.34687)**</td>
</tr>
<tr>
<td></td>
<td>0.18319</td>
<td>-0.38542</td>
<td>-5.38061</td>
<td>0.35893</td>
<td>0.17230</td>
<td>-0.41751</td>
<td>-0.610530</td>
<td>4.37906</td>
</tr>
<tr>
<td></td>
<td>(0.34567)</td>
<td>(0.62645)</td>
<td>(0.36390)**</td>
<td>(0.33968)</td>
<td>(0.32340)</td>
<td>(0.35917)</td>
<td>(0.32367)*</td>
<td>4.373564</td>
</tr>
<tr>
<td></td>
<td>0.18039</td>
<td>-0.38861</td>
<td>-5.38382</td>
<td>0.35619</td>
<td>0.168981</td>
<td>-0.420476</td>
<td>-0.613489</td>
<td>4.373564</td>
</tr>
<tr>
<td></td>
<td>(0.34584)</td>
<td>(0.62669)</td>
<td>(0.36408)**</td>
<td>(0.33979)</td>
<td>(0.32363)</td>
<td>(0.35932)</td>
<td>(0.32389)*</td>
<td>(0.61394)**</td>
</tr>
</tbody>
</table>

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