This is to certify that the

dissertation entitled

ESTIMATION FROM CENSORED MEDICAL COST DATA

presented by

ONUR BASER

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in ECONOMICS

Major professor

Date May 9, 2002

ESTIMATION FROM CENSORED MEDICAL COST DATA

By

ONUR BAŞER

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Economics

2002

ABSTRACT

ESTIMATION FROM CENSORED MEDICAL COST DATA

By

ONUR BAŞER

Health care inflation is a concern in many industrialized countries. One response is search for cost effective therapies which requires proper analysis of treatment cost data. Common problem with medical cost data is censoring and statistical properties of estimating medical cost from a censored data is not well developed. In my thesis, I propose two method, one with an extension to panel data setting, to estimate medical cost from censored data.

First chapter applies the inverse probability weighted least-squares method to predict censored total medical cost. Since survival time and medical costs may be subject to right censoring and therefore are not always observable, the ordinary least-squares approach cannot be used to assess the effects of certain explanatory variables. Inverse probability weighted least-squares estimation provides consistent asymptotic normal coefficients with easily computable standard errors. A test is derived to compare the differences between the coefficients estimated by the ordinary least-squares approach and the inverse probability weighted least-squares estimation. A study on the medical cost of lung cancer is used as an application of the method.

Second chapter applies the inverse probability weighted (IPW) least-squares method to predict total medical cost from panel data subject to censoring. Specifically, IPW pooled ordinary-least squares(POLS) and IPW random effects(RE) models are used. Because total medical cost is not independent of the survival time under administrative censoring, unweighted POLS and RE cannot be used with uncensored data, to assess the effects of certain explanatory variables. IPW estimation gives consistent asymptotic normal coefficients with easily computable standard errors. A traditional

and robust form of Hausman test can be used to compare the coefficients estimated by weighted and unweighted estimation methods. The method developed in this paper are applied to lung cancer cost data.

In the third chapter, a method for testing and correcting for sample selection bias for cross-sectional data is proposed. Specifically, this paper provides a systematic treatment of the correction for nonrandom sample selection of medical cost data where the selection rule is described by a censored regression model. We show that the population parameters are identified, and provide straightforward $\sqrt{N}$-consistent and asymptotically normal estimation methods under the assumption that the selection rule is governed by a censored Tobit Model. A study on the medical cost of lung cancer is used as an application of the method.

For my wife, Deniz and my brother, Erdem

Table of Contents

List of Tables

List of Figures

Chapter 1: Estimation of Censored Medical Cost Data

Introduction

The rising cost of health care is a concern in many industrialized countries. One response is the search for cost-effective therapies, which requires proper analysis of treatment cost data. Cost-effectiveness analysis (CEA) involves estimating the net, or incremental, costs and effects of an intervention. Treatment costs and health outcomes are compared with some alternative, which might be the care that would be given if the interventions were not used at all. (Gold et al., 1996). The first step in this important process is the estimation of costs. Costs can be estimated from a variety of sources, including Medicare claims files. However, censoring is a common problem with these administrative data.

Statistical methods applicable to censored cost are not well developed. Nevertheless, the average total cost for a group of patients has been estimated in one of three ways: (1) by estimating the sample mean of observed costs from all cases, (2) by estimating the sample mean the uncensored subjects only, and (3) by using modifications of standard survival analysis techniques (Lin, 1997). All of these methods yield biased estimators. The sample mean from all cases creates a downward bias because it does not account for the costs incurred after censoring. The sample mean from uncensored subjects is biased toward the costs for patients with a shorter survival time since a longer survival time is likely to be censored.

Standard survival analysis on costs is not valid if patients accumulate costs with different rate functions over time. This technique assumes independence

between the cost at the survival time and the cost at the censoring time, whereas, the two are generally positively correlated. To adjust for this dependency, Lin et al. (1997) proposed a "partitioned estimator" method to assess average costs. This method partitions the entire time period of interest into a number of smaller intervals and calculates average cost and product-limit estimates for each interval. The sum of the product of these two components becomes "product-limit sampling average estimator" of total cost for the sample. An application by Sloan et al. (1999) to health care costs of patients in a oncology clinical trial found Lin's variance estimation to be an arduous and untenable numerical programming exercise. Instead, they used a straightforward application of the bootstrap method to obtain variance estimates.

If we are interested in conditional average costs, all three estimation methods incorrectly assume some homogeneity in the medical cost data in the sense that they are independent of patient characteristics or the type of treatment. Since cost can depend on patients' age, disease stage, comorbid conditions, symptoms, type of treatments received,etc., estimation should account for these control variables. Multivariable regression analysis is required but no such method is available using standard software programs. The two approaches developed by Lin (2000a, 2000b) require a high level of computer programming and have not been emprically tested.

In this chapter, the inverse probability weighted (IPW) least squares method is used 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 and Thomson (1952), Robins and Rotnisky (1992,1995), Robin, Rotzinky and Zhao (1995), Horowitz and Manski (1998), Rosenbaum (1987) and Hirano, Imbens, and Ridder (2000)). In more general framework, Wooldridge (1999, 2001) examined the asymptotic properties of the IPW M-estimator for variable probability samples and standard stratified samples and Wooldridge(2002b) provides an overview of IPW

M-estimation for cross-section applications.

IPW estimation produces consistent estimators with a covariance matrix that can be calculated by most commercial statistics software programs. We also developed a test to compare the coefficients estimated by IPW least squares and ordinary least squares methods (OLS).

This chapter is organized as follows. The first section outlines IPW least squares as applied to censored medical cost data, including the statistical properties of the estimation and step-by-step procedures for implementation. The next section introduces a Hausman type of test to compare the estimators calculated by using IPW least squares and OLS over uncensored data. The third section describes an application of our methods to a study on the medical cost in lung cancer patient. The last section presents our conclusions.

1.1 IPW least squares

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

Assume that in the population of interest

$$y = \boldsymbol{x}\boldsymbol{\beta} + u, \tag{1.1}$$

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

component of β represents the intercept.

Assume that

$$E(\boldsymbol{x}'u) = 0. \tag{1.2}$$

Under random sampling from the population, equation (1.2) is the crucial assumption in obtaining consistency of the OLS estimator of β in (1.1). With (1.2) and the rank assumption E($\boldsymbol{x}'\boldsymbol{x}$) =K, the OLS estimator using a random sample will be consistent for β . The assumption that u is a zero-mean error term does not guarantee consistency.

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

Let $Z = min(T, C)$, $s = I(C \geq T)$, and $s^* = I(C \geq T^*)$, where $I(.)$ is the indicator function. There are two types of censoring: time censoring if $s = 0$ that is, $T > C$, and cost censoring if $s^* = 0$, that is, $min(T, L) > C$. A generic element from the population can be denoted $(y, \boldsymbol{x}, s^*)$. Suppose that T and C are independent given $\boldsymbol{x}$.

Assumption 1

(i) $\boldsymbol{x}$ and C are always observed, and T^*, y are observed when $s^* = 1$.

(ii) y can be ignored in the selection equation, conditional on $\boldsymbol{x}$:

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

Assumption 1 indicates that $\boldsymbol{x}$ is always observed and that, conditional on $\boldsymbol{x}$,

the response variable does not affect the selection probability. Part (ii) of Assumption 1 is crucial because of the requirement that the selection probability is observable when $s^* = 1$. Since we can ignore y from the selection equation, having a censored y value does not create a problem for estimating selection probabilities.

Suppose we have a random sample of size N from the population to estimate $\boldsymbol{\beta}$. Thus $\{(\boldsymbol{x_i}, y_i)\colon i = 1, 2, ..., N\}$ are treated as independent, identically distributed random variables, where $\boldsymbol{x_i}$ is $1 \times K$ and y_i is a scalar, and s_i^* is a corresponding sample selector indicator. The underlying model is then,

$$y_i = \boldsymbol{x_i}\boldsymbol{\beta} + u_i, \tag{1.3}$$

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

The IPW least square estimators, $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$, solves

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} w_i(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2, \tag{1.4}$$

where $w_i = (s_i^*/P(C_i \geq T_i^*))$. Under assumption 1 and equation (1.2), $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ is consistent with asymptotically normal distribution and the estimated asymptotic variance $V(\hat{\boldsymbol{\beta}}_{\boldsymbol{w}})=\hat{\boldsymbol{A}}_{\boldsymbol{w}}^{-1}\,\hat{\boldsymbol{B}}_{\boldsymbol{w}}\,\hat{\boldsymbol{A}}_{\boldsymbol{w}}^{-1}/N$, where

$$\hat{\boldsymbol{A}}_{\boldsymbol{w}} = N^{-1}\sum_{i=1}^{N} w_i\boldsymbol{x_i'}\boldsymbol{x_i}, \tag{1.5}$$

$$\hat{\boldsymbol{B}}_{\boldsymbol{w}} = N^{-1}\sum_{i=1}^{N} w_i^2\hat{u}_i^{\,2}\boldsymbol{x_i'}\boldsymbol{x_i}, \tag{1.6}$$

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

The objective function in (1.4) simply weights each observation $(y_i, \boldsymbol{x}_i)$ by the inverse probability of appearing in the sample, that is, observations for which $s_i^* = 0$ do not appear in the optimization problem. Assumption 1 part (ii) requires $P(C_i \geq T_i^*)$ to be known whenever $s_i^* = 1$, so $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ is computable from observed data assuming we know $P(C_i \geq T_i^*)$.

Note that neither $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ nor its covariance matrix estimator involves the incomplete observations. In addition the estimated covariance matrix is the White (1980) heteroskedasticity-consistent covariance matrix estimator applied to all variables for observation i weighted by $\sqrt{w_i}$. Note that even if there is no heteroskedasticity in the potential model (1.3), we treat the model as heteroskedastic due to censoring. Heteroskedasticity-robust standard errors after the weighted regression provide the estimated asymptotic standard errors. Censoring, then can be handled easily because most standard statistics software programs compute a heteroskedasticity-consistent covariance matrix.

Another advantage of weighting the observations, other than solving the censoring problem, is that we derive consistency with the weaker assumption (1.2) rather than $E(u|\boldsymbol{x}) = 0$. Since w is independent of $(\boldsymbol{x}, u)$,

$$E(w\boldsymbol{x}'u) = E(E(w|u, \boldsymbol{x})\boldsymbol{x}'u)) = E(E(w|y, \boldsymbol{x})\boldsymbol{x}'u)) = E(E(w|\boldsymbol{x})\boldsymbol{x}'u)) = 0. \quad (1.7)$$

The last equality follows because $E(w|\boldsymbol{x}) = 1$.

So far, it has been assumed that the sampling probability function is known. Usually, that function is unknown and needs to be estimated. We propose to estimate the sampling probability function by the product-limit estimator (Kaplan and Meier 1958), with the roles of C_i and T_i reversed (i.e., T_i censors C_i).

Assuming censoring is not covariate dependent, define $p(t) = P(C \geq t)$, and let $\hat{p}(t)$ be the product limit estimator of $p(t)$ based on the data $(Z_i, \overline{s_i})$ $(i = 1, ..., N)$,

where $\overline{s_i} = 1 - s_i$. Then,

$$\hat{w}_i = \frac{s_i^*}{\hat{p}(T_i^* -)} \quad i = 1, ..., N. \tag{1.8}$$

Under standard regularity conditions [1] two step IPW least square estimator that uses $\hat{w}_i$ instead of w_i in equation (1.4) consistently estimates $\hat{\boldsymbol{\beta}}_{\boldsymbol{w}}$ (Newey and MacFadden 1994).

Exogeneous censoring implies that

$$E(y|\boldsymbol{x}\ , T^*, C) = E(y|\boldsymbol{x}\ ). \tag{1.9}$$

It can be shown that under exogenous sampling the use of an estimate of the probabilities for the second step yields a variance estimator that is asymptotically equivalent to that estimated with known probability values. Therefore, define $\tilde{\boldsymbol{\beta}}_{\boldsymbol{w}}$, the two step IPW least squares estimator, as the solution to

$$\min_{\beta \in \Theta} \sum_{i=1}^{N} \hat{w}_i (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}_i)^2. \tag{1.10}$$

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

$$\tilde{\boldsymbol{V}}_{\boldsymbol{w}} \equiv V(\tilde{\boldsymbol{\beta}}_{\boldsymbol{w}}\ ) = \tilde{\boldsymbol{A}}_{\boldsymbol{w}}^{-1}\ \tilde{\boldsymbol{B}}_{\boldsymbol{w}}\ \tilde{\boldsymbol{A}}_{\boldsymbol{w}}^{-1}\ /N, \tag{1.11}$$

where,

$$\tilde{\boldsymbol{A}}_{\boldsymbol{w}} = N^{-1} \sum_{i=1}^{N} \hat{w}_i \boldsymbol{x}_i' \boldsymbol{x}_i, \tag{1.12}$$

[1] The conditions in which the uniform weak law of large numbers can be applied. For details; see Theorem 12.1 in Wooldridge(2000a). Lemma 4.3 in Newey and McFadden (1994) shows that if w_i is replaced with a consistent estimator, the convergence still valid.

$$\tilde{\boldsymbol{B}}_{\boldsymbol{w}} = N^{-1} \sum_{i=1}^{N} \hat{w}_i^2 \tilde{u}_i^2 \boldsymbol{x}_i' \boldsymbol{x}_i, \tag{1.13}$$

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

Under administrative censoring, for example, all censoring is caused by study termination, and C is independent of y. However, unless we have short interval cost values, such as monthly or weekly, we may expect that T^* and y are correlated. In this case,$\tilde{\boldsymbol{V}}_{\boldsymbol{w}}$ in (1.11) has to be adjusted for estimation of w_i (for adjusted covariance matrix, see Wooldridge (2002b)). The estimated covariance matrix of the two-step IPW least squares estimator is the White (1980) heteroskedasticity-consistent covariance matrix estimator applied to all variables for observation i weighted by $\sqrt{\hat{w}_i}$. Robust covariance matrix is built into most statistical programs, adjustment for the $\tilde{\boldsymbol{V}}_{\boldsymbol{w}}$ in (1.11) requires programming. In practice, it has been found that adjusting for the first-step estimators usually has little effect on the asymptotic standard errors. Moreover, Wooldridge (2002b) shows that using the estimated selection probability will produce smaller standard errors than true estimated by using a known selection probability. In other words, if we compute the asymptotic variance as if we have not estimated the probabilities, inference is conservative. Adjustment for estimation of w_i requires programming in the application. The main point of this paper is to suggest an easily applicable method. By ignoring adjustment for simplicity, we produce higher standard errors, however obtaining significant estimates using *unadjusted* standard errors leads to larger t statistics after correction. This is somewhat unusual for two-step estimation problems, where the prevailing wisdom is that larger standard errors occur by adjusting standard errors for a first stage estimation.

The steps for deriving consistent two-step IPW least squares estimators and

their *unadjusted* asymptotic variance estimators can be summarized as follows.[2]

(i) Calculate the product-limit estimator, m_i, based on data $(Z_i, 1 - s_i)$ $(i = 1, ..., N)$.

(ii) Generate $p_i = m_i$ for the cases $Z_i \leq L$; $p_i = l_i$ were l_i is the value of m_i at $Z_i = L$.

(iii) Generate weight, $w_i = s_i/p_i$ $(i = 1, ..., N)$.

(iv) Generate weighted response and explanatory variables: $y_i^* = \sqrt{w_i} y_i$, $\boldsymbol{x}_i^* = \sqrt{w_i}\, \boldsymbol{x}_i$ $(i = 1, ..., N)$.

(v) Run the OLS regression of y_i^* on $\boldsymbol{x}_i^*$ with the heteroskedasticity robust option.

Total medical cost data are typically characterized by a skewed empirical distribution of the nonzero realizations (Manning and Mullahy, 2001). The most common method for analyzing such data is logarithmic transformation of the response variable. In our estimation procedure, y_i can be chosen as the transformed dependent variable. Retransformation then can be done using the smearing estimator (Duan, 1983). The smearing estimator is the exponential of the expected response on the log-scale multiplied by the average of the exponential cost. Anderson et al. (2000) developed the heteroskedastic smearing estimator for use when the variances of the residuals are not constant.[3]

[2]STATA commands are in the Appendix

[3]In the method described above, a heteroskedastic-robust variance matrix is used. The robust variance matrix is needed because of stratification whether or not there is heteroskedasticity. Therefore, homoskedastic smearing transformation can still be chosen after robust estimation if the variance of the residuals are constant.

1.2 Comparison Of The IPW And Unweighted Estimators

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

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} s_i^*(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i)^2. \tag{1.14}$$

It is well-known that selection under exogeneous sampling does not cause problems if we impose the stronger assumption, $E(u|\boldsymbol{x}) = 0$. Then $\tilde{\boldsymbol{\beta}}_{\boldsymbol{u}}$ is consistent and asymptotically normally distributed and the usual variance matrix estimator $V(\tilde{\boldsymbol{\beta}}_{\boldsymbol{u}}) = \tilde{\boldsymbol{A}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{B}}_{\boldsymbol{u}} \tilde{\boldsymbol{A}}_{\boldsymbol{u}}^{-1} / N$ is consistent, where

$$\tilde{\boldsymbol{A}}_{\boldsymbol{u}} = N^{-1} \sum_{i=1}^{N} s_i^* \boldsymbol{x}_i' \boldsymbol{x}_i, \tag{1.15}$$

$$\tilde{\boldsymbol{B}}_{\boldsymbol{u}} = N^{-1} \sum_{i=1}^{N} s_i^* \breve{u}_i^2 \boldsymbol{x}_i' \boldsymbol{x}_i, \tag{1.16}$$

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

If equation (1.9) is satisfied, then unweighted and weighted estimators are both consistent. In such a case, theory suggests that an unweighted estimator is more efficient under conditional homoskedasticity and weighted estimator is more efficient under unknown heteroscedasticity (Wooldridge, 1999).

Because the unweighted estimator is inconsistent under the violation of equation (1.9) and the weighted estimator is consistent with or without exogeneous

sampling, we can apply a Hausman (1978) test to determine exogeneity of sampling.

The traditional form of Hausman statistics can be used under homoskedasticity assumption. We can state this assumption as follows: For the selected sample, $i = 1, 2, ..., N_0$:

$$E(u_i^2 \boldsymbol{x}_i' \boldsymbol{x}_i) = \sigma_0^2 E(\boldsymbol{x}_i' \boldsymbol{x}_i). \tag{1.17}$$

When equation (1.17) holds, the unweighted least squares variance estimator is

$$\tilde{\boldsymbol{V}}_{\boldsymbol{u}} \equiv V(\tilde{\boldsymbol{\beta}}_{\boldsymbol{u}}) = \tilde{\sigma}^2 \left(N^{-1} \sum_{i=1}^{N} s_i^* \boldsymbol{x}_i' \boldsymbol{x}_i \right)^{-1} \tag{1.18}$$

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

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

$$H = (\tilde{\boldsymbol{\beta}}_{\boldsymbol{w}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{u}})' \tilde{\boldsymbol{V}}^{-1} (\tilde{\boldsymbol{\beta}}_{\boldsymbol{w}} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{u}}), \tag{1.19}$$

where $\tilde{\boldsymbol{V}} \equiv \tilde{\boldsymbol{V}}_{\boldsymbol{w}} - \tilde{\boldsymbol{V}}_{\boldsymbol{u}}$. $\tilde{\boldsymbol{V}}_{\boldsymbol{w}}$ is defined in equation (11) and $\tilde{\boldsymbol{V}}_{\boldsymbol{u}}$ is defined in equation (18).

In many cases we may want to use a Hausman test when the homoskedasticity asssumption is violated. This requires a robust form that replaces $\tilde{\boldsymbol{V}}$ by

$$(\tilde{\boldsymbol{A}}_{\boldsymbol{w}}^{-1} \mid - \tilde{\boldsymbol{A}}_{\boldsymbol{u}}^{-1}) \left(N^{-1} \sum_{i=1}^{N} \tilde{\boldsymbol{e}}_{\boldsymbol{i}} \, \tilde{\boldsymbol{e}}_{\boldsymbol{i}}' \right) (\tilde{\boldsymbol{A}}_{\boldsymbol{w}}^{-1} \mid - \tilde{\boldsymbol{A}}_{\boldsymbol{u}}^{-1})'/N, \tag{1.20}$$

and $\tilde{\boldsymbol{e}}_{\boldsymbol{i}} = \left(\hat{w}_i \tilde{u}_i \boldsymbol{x}_{\boldsymbol{i}}' \, , s_i^* \check{u}_i \boldsymbol{x}_{\boldsymbol{i}}' \right)'$. Here $\tilde{u}_i$ and $\check{u}_i$ are the residuals after IPW least squares and OLS estimations of the selected sample, respectively. $\hat{w}_i$, $\tilde{\boldsymbol{A}}_{\boldsymbol{w}}$, $\tilde{\boldsymbol{A}}_{\boldsymbol{u}}$ are defined in equation (1.8),(1.12),(1.15) respectively.

Under the null hypothesis the sampling scheme is exogenous, $H \overset{a}{\sim} \chi_P^2$. If we

reject the hypothesis, IPW least squares method should be used. Since we have endogenous sampling, OLS estimation using complete cases is not consistent. If we fail to reject the hypothesis, the typical response is to conclude that the exogeneity assumption holds and we should use OLS estimates. Unfortunately, we may be committing a Type II error by failing to reject exogeneity assumption when it is false. Therefore, we should report results from both estimation procedures.

1.3 Application to the Lung Cancer Study

1.3.1 Data

From 1994 through 1997, 202 patients with incident cases of lung-cancer were recruited from 24 Michigan community hospitals and their affiliated oncology units. Each patient provided written consent for researchers to acquire his or her Medicare claim files; 189 patient had lung-cancer treatment.

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

Several cases had missing data. One case was missing age, and nine cases were missing stage, nineteen cases did not reported comorbid conditions, three cases were missing symptoms and six cases had no data on physical function. We first assumed a completely random distribution of missing data,that is, cases with complete data are not distinguisable from cases with incomplete data. We used mean substitution, median substitution, pairwise deletion and regression methods to complete the data set. We then assumed that complete data are different from cases with incomplete

data but a missing pattern is tractable. We used a multiple imputation method to complete the data set. None of imputation methods yielded a significant estimate in the regression model for the variables with missing values, therefore we used median substitution without loss of generality.

1.3.2 Variables

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

Age. Age is defined as continuous variable, patient's age within two weeks of initiating either radiation or chemotherapy.

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

Hospitalization. The number of inpatient Medicare claims was used to derive the number of hospitalizations for reasons other than lung cancer surgery.

Physical Function. Physical function three months prior to diagnosis was assessed using the subscale from the SF-36 (Ware et al. 2000). The 10-item subscale asks questions about such activities as lifting heavy objects, participating in

strenuous sports, climbing stairs, walking various distances, and ability to bathe and dress. Response categories are: limited a lot, limited a little, and not limited at all. Scores are standardized and range from 100 (no limitation) to 0 (severe limitation).

Symptoms. Patients were asked if during the past two weeks they had experienced any of 33 symptoms. A count of all symptoms was summed for each patient.

Death. The Office of Vital Statistics,Michigan Department of Community Health, Death Certificate Registry was used to identify the date of death.

Comorbid Conditions. Comorbid conditions were assessed with an instrument from the Aging and Health in America Study, a national survey that asks patients to indicate whether a health professional has ever told them they have one of 15 problems. The total number of positive responses was summed for each patient and sorted into one of two categories: zero to two,and three or more. A comparison of patient reports of comobid conditions with medical record audits indicates that patients are able to recall other diagnosed illnesses (Katz et al., 1996). Restricting the categories for comorbid conditions does not result in lost predictive power (Newschaffer, 1998).

Stage. Disease stage was determined using the American Joint Committee on Cancer (AJCC) Tumor Nodes & Metastasis (TNM) staging system which was applied to pathological data obtained from an audit of patients' medical records. Stage of cancer at diagnosis was collapsed into early (in situ and local) and late (regional and distant).

Gender. The value is 1 for males; 0 for females.

Race. The value is 1 for whites, 0 for blacks.

Table 1.1: Descriptive Statistics from the Lung Cancer Study

	Uncensored(n=135)		Censored(n=48)	
Variables	Mean	Std	Mean	Std
total cost	63939	41680	62877	40114
lstage	.67		.54	
lcomorbi	.64		.63	
hospitalize	.62		.54	
chemo only	.08		.06	
radiation only	.26		.25	
chemo and radiation	.36		.54	
symptoms	11.13	5.14	10.12	4.72
physical functions	73.55	26.76	71.45	28.71
age	71.96	4.85	72.68	5.201
gender	.57		.62	
white	.93		.91	
death	.53		0	

1.3.3 Descriptive Analysis

All analysis were done using STATA version 7. Table 1.1 shows the summary statistics. Nineteen cases had no treatment related to lung cancer, so we dropped these cases from the sample. Out of the remaining 183 patients, we had complete data for 135 cases and incomplete data for 48 cases. So, approximately 26 percent of the sample had cencored data.

As shown in Table 1.1, the patient sample can be described as predominantly white and in their early seventies for both censored and uncensored cases. Two thirds of the patients were diagnosed with late stage disease for complete cases whereas half of the patients with incomplete data were diagnosed with late stage disease. Eighty-three patients who have compete data were hospitalized for reasons other than lung cancer surgery while 26 censored patients were hospitalized.

Most patients had three or more comorbidities and experienced some level of symptoms related to cancer treatment. The patient sample is relatively high

functioning in terms of physical health. Fifty-seven percent of the complete cases were male relative to 62% percent of the cases in the censored data.

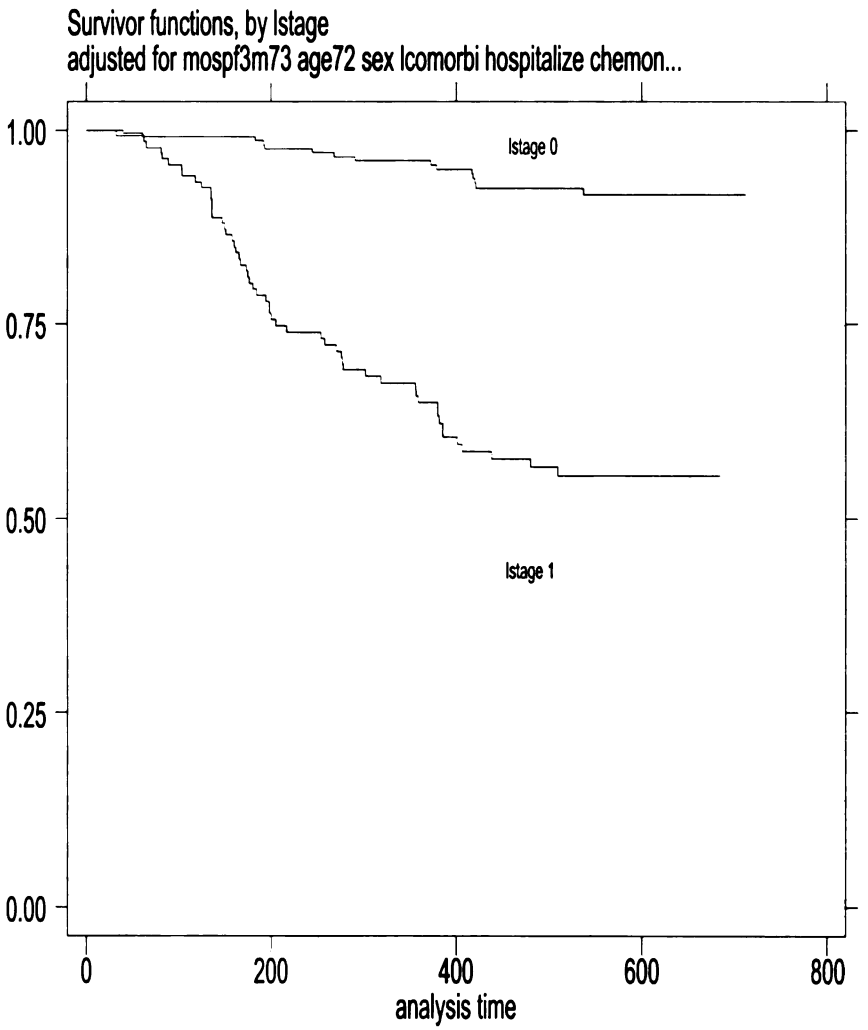
The last four rows of Table 1.1 show the categorical variables related to treatment types. For censored and uncensored cases, we have similar percentages for the patients who had radiation only or chemotherapy only, approximately 25% and 7% respectively. Twenty-four percent of the patients had surgery or surgery plus adjuvant therapy in the complete cases whereas 35% had them in incomplete cases. For chemotherapy and radiation, the ratios are 36% and 54% respectively for censored and uncensored cases.

The dependent variable, total Medicare payments two years following diagnosis is shown in the first row of Table 1.1. Considering the mean alone, we find that the total cost of all care is $63,939 for the two years following a lung-cancer diagnosis for complete cases and $62,877 for incomplete cases.

1.3.4 Survival Curves

Figure 1.1 and Figure 1.2 show the separately estimated baseline survival curves for the variable of interest after we conditioned on the explanatory variables. For each graph, we estimated a separate Cox (1972) proportional hazards model on the explanatory variables of interest so that we can compare the effects of certain variables on survival time and total medical cost conditioning on the others. As shown in Figure 1a, the patients with less aggressive disease have better survival probabilities after we control for physical health three months prior to diagnosis, age, gender,race, comorbidity conditions, hospitalization and the treatments. The chances are approximately 90% for early stage and 30% for the late stage. Interestingly, the patients who have hospitalizations for reasons other than lung cancer surgery have a 30% chance of survival , compared to 95% for those who do

Figure 1.1: Survival functions according to disease stage level



not, conditioning on the other explanatory variables(Figure 1.2). For the other categorical variables, comoribid conditions, race, and treatment types; we did not observe differences in the base line curves after we controlled for the explanatory variables.

1.3.5 Regression Analysis

Our aim is to determine how the variables age, gender, comorbid conditions, stage of cancer, symptoms, death status, physical functions, hospitalization, and treatment account for the total medical cost of lung cancer in the two years following diagnosis.

We found that costs are skewed to the right, so we transformed the cost equation to a log-linear scale. We started with the log-scale residuals from a generalized linear model with a logarithmic link function and found that the log-scaler residuals are dense at the tails. Following Manning and Mullahy (2001) we considered an OLS-based model with a log-transformed dependent variable.

Table 1.2 shows the result of the regression analysis predicting total cost of care for the two years following a lung cancer diagnosis. The first column of Table 1.2 shows the unweighted regression coefficients, while the second column shows weighted regression results. The reference groups for treatment modalities are surgery only and surgery plus adjuvant therapies.

Variables that reach statistical significance ($p < .05$) include hospitalization for reasons other than lung cancer surgery, chemotherapy only, radiation only, and chemotherapy and radiation.

Hospitalization for reasons other than lung cancer surgery increases total medical cost during the period of interest by 107% according to the unweighted estimation and by 114% according to the IPW least square estimation.

Figure 1.2: Survival functions according to hospitalization for reasons other than lung cancer surgery

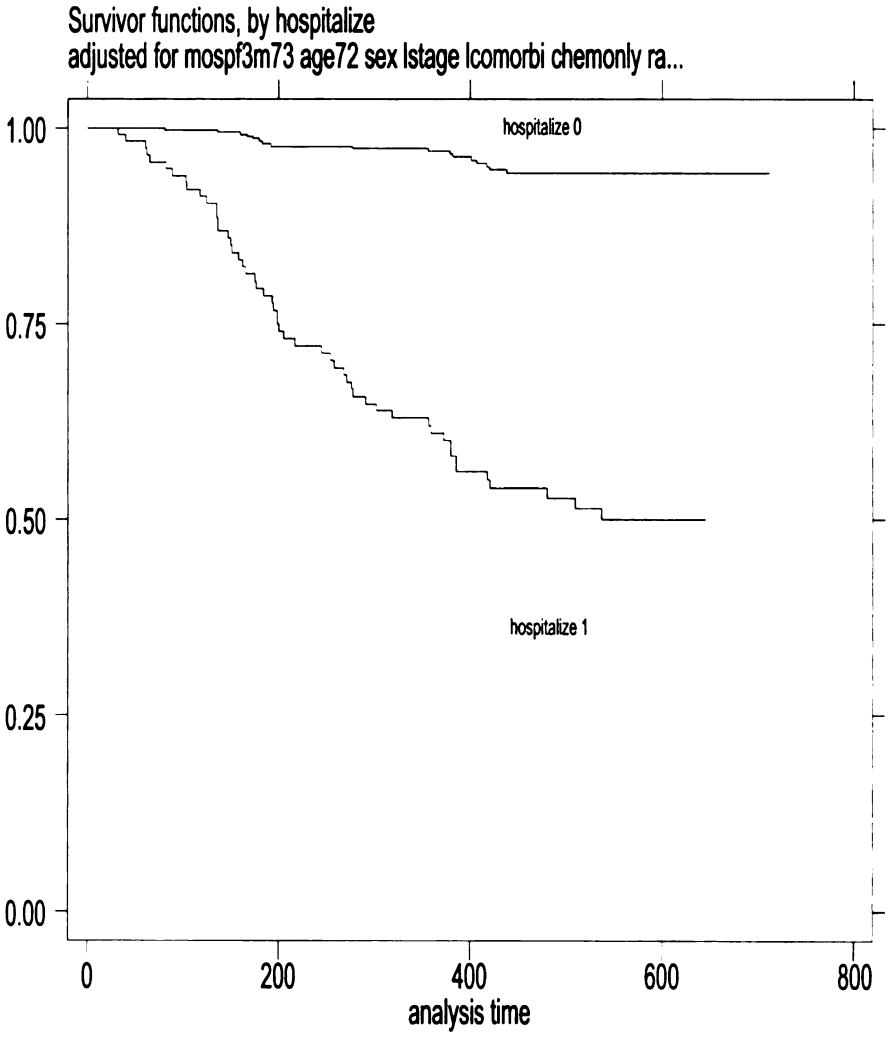


Table 1.2: Estimates of the Log(tcost) Equation by OLS and IPW

Explantory	OLS	IPW
constant	10.74	10.70
	(1.06)**	(1.04)**
hospitalize	.72	.75
	(.18)**	(.18)**
chemotherapy only	-.92	-.83
	(.29)**	(.31)**
radiation only	-.79	-.73
	(.23)**	(.22)**
chemothrepay and radiation	-.49	-.43
	(.22)*	(.21)*
Death	-.02	-.03
	(.12)	(.12)
symptoms	.004	.009
	(.014)	(.014)
physical functions	.001	.002
	(.003)	(.002)
age	-.003	-.005
	(.014)	(.013)
gender	.081	.12
	(.12)	(.12)
white	.12	.23
	(.20)	(.19)
Observations	135	135
R-squared	0.13	0.15

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

Whether or not a person receives radiation or chemotherapy separately or in combination significantly decreased the total medical cost relative to the mean costs for persons receiving surgery only or surgery plus adjuvant therapies. The estimates with respec to the unweighted and weighted least squares are: for radiation only, 120% and 105%; for chemotherapy only, 148 % and 127%; for chemotherapy and radiation, 65% and 55%.

As we demonstrated in Table 1.2, age, gender, physical functions, stage, comorbid conditions, race and death status do not have a statistically significant effect. Our models explain 13% of the variability in total costs the two years following diagnosis according to unweighted estimation and 15% according to IPW least squares.

A comparison of the weighted and unweighted estimations does not reveal significant differences, although the former statistically corrects for potential bias. The test developed in section 1.2 can be used to support this argument. We failed to reject the hypothesis that sampling scheme is exogenous. In this case, there is a chance that our unweighted estimators are consistent. Both estimators are reported in Table 1.2 and are statistically and practically the same.

Adjusted means can be calculated using the smearing estimation. These are shown below.

Method	Mean	Standard Deviation
Uncensored-Unadjusted	\$63,939	\$41,680
Unweighted Estimation	\$64,043	\$14,177
Weighted Estimation	\$64,563	\$15,850

Whether the sample selection depends on the conditioning variables, or is independent, then the weighted and unweighted estimators are consistent. Since we have evidence of exogenous sampling with the robust form of Hausman Test, we reached this conclusion. In this case, the theory suggests that the unweighted

estimator is more efficient under conditional homoskedasticity. In our model, we do not have heteroscedasticity,therefore the standard errors from the predicted means are in the expected direction.

1.4 Conclusions

Prior to Lin (2000a, 2000b) the methods of estimating censored costs incorrectly assume some homegeneity in the medical cost data in the sense that they are independent of covariates such as patient and clinical characteristics. In 2000, Lin developed a technique for estimating censored costs. However, his approach, while correct, is extraordinarily complex and not applicable using commercially available statistical software programs. Therefore no empirical tests of this model have been completed.

This paper examines the IPW least squares method to solve for inconsistencies due to censoring and is easily applicable using most statistical software programs. Under the key assumption that selection is ignorable, the inverse probability weighting scheme identifies the population parameters.The regression method introduced can handle large numbers of continuous and discrete explanatory variables.

The application of the method is a two-step estimation process where at first step, we estimate selection probabilities by using the product limit estimation where the role of censoring and survival time is reversed. At the second step, we estimate heteroskedastic robust OLS on the uncensored data set where each variable is weighted with the inverse of the square root of the estimated selection probabilities from first stage.

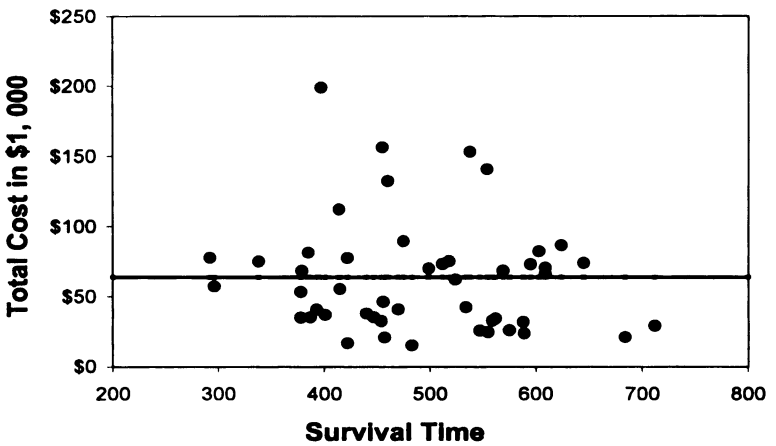
We also developed a test to compare the coefficients estimated by the IPW least squares and by OLS. This test can be used to asses efficiency improvement

between two models. Specifically, if we reject the null hypothesis that the sampling scheme is exogenous, IPW least squares method should be used because the other method yields inconsistent estimates. Failing to reject the null hypothesis could be used to support unweighted estimation under conditional homoskedasticity.

We also applied the proposed method to an inception cohort of patients newly diagnosed with lung cancer. The findings from the lung cancer study can be summarized as follows. Although lung cancer stage does not affect the total medical cost, it decreases survival time. Comorbid conditions are not significant for the estimation of total medical cost and do not effect survival time. Hospitalization for reasons other than lung cancer surgery decreases survival time and it also doubles the total medical cost during the period of interest.

Several limitations should be discussed. The lung cancer study does not demonstrate the full power of the IPW least squares method. First, the sample size is small and all of the results demonstrated in the first two sections are asymptotically valid. Second, the censored observations in the data set are relatively homogeneous. Applying OLS to the cases with complete data yields an unbiased estimator toward the cost of the patients with shorter survival time because a longer survival time is more likely to be censored. Since a longer survival time tends to be associated with higher medical cost, the cost values of the censored case should be well above the mean value for cases with complete data. Figure 1.3 shows that is not the case for data in this study. All the censored cases cluster around the mean of uncensored cases. With the available data set, where the number of observation is large and deviation between censored and uncensored observations is significant, we would see the full power of the IPW least squares method over OLS.

Figure 1.3: Total medical cost distribution among the censored cases



Third, for the exact asymptotic variances adjustment for the first stage estimation should be made. So marginally insignificant variables should be interpreted with the caution since with the adjustment they may turn out to be significant.

In conclusion, our study improves upon previous studies by propose a multivariate regression analysis that solves for inconsistencies due to censoring and a statistical test to asses the efficiency improvement between the old methods and the more easily replicatable proposed method. Furthermore, an application of lung cancer study shows how the method can be applied by using most of the statistical software programs, including step-by-step procedures.

Chapter 2: The Longitudinal Analysis of Censored Medical Cost Data

Introduction

Proper analysis of treatment cost data is more challenging than is generally recognized. A common problem is that censoring and statistical methods applicable to estimation of medical cost from censored data are not well developed.

Until recently the methods (Lin et al. 1997, Bang and Tsiatis 2000) for analyzing censored medical cost assumed homogeneity in the data, which in practice is rare. Proper analysis requires multivariate regression analysis. The two approaches developed by Lin (2000a, 2000b) require a high level of computer programming and have not been fully empirically tested.

Analysis of censored data under exogenous sampling can be done easily by using the ordinary least-squares (OLS) method for uncensored data. It produces consistent estimators which we refer as unweighted estimators throughout the paper. Exogenous sampling in the context of estimation from censored medical cost means that once explanatory variables are selected, such as patient characteristics or the type of treatments, total medical cost does not depend on the censoring time and survival time.

Under administrative censoring, that is when all censoring is caused by study termination, it is reasonable to assume that total cost is independent of censoring time. Exogenous sampling assumptions are violated; if total cost and survival time

are correlated after we condition on explanatory variables. Since longer survival time may be associated with higher medical cost, the unweighted method yields an estimator biased toward the cost of patients with shorter survival times.

In the first chapter, I applied the inverse probability weighted (IPW) least-squares method to predict total medical cost in patients with lung cancer two years after diagnosis. In more general framework, Wooldridge (1999, 2001) examined the asymptotic properties of the IPW M-estimator for variable probability samples and standard stratified samples and Wooldridge(2002b) provides an overview of IPW M-estimation for cross-section applications. IPW produces consistent asymptotically normal coefficients with easily computable standard errors under the violation of the exogenous sampling assumption. With small data sets the resulting estimator may be unstable if the censoring is heavy (Bang 2000). It is necessary to ensure that sufficient follow-up is available during the period for which we wish to compute medical costs.

In this chapter, we extend first the method described in the first chapter to handle data with extensive censoring. The method covers the partitioned estimation suggested by Lin (2000a), that estimator can be used only for time independent regressors. Our method can be applied for both time dependent and independent explanatory variables. We use weighted estimation specifically, IPW pooled ordinary-least square (POLS) and IPW random effects (RE) models. The choice between the two depends on whether unobserved heterogeneity is present. If present, IPW RE should be used, otherwise the simpler IPW POLS will produce consistent asymptotically normal coefficients.

Second, since an unweighted estimator is inconsistent when exogenous sampling is violated and the weighted estimator is consistent with or without exogenous sampling, traditional and robust form of the Hausman (1978) test will be applied to determine systematic differences in the models in a panel data setting.

The third section of the chapter describes a study on the medical cost for lung cancer that is used to demonstrate the methods. The last section presents conclusions.

2.1 General Framework

Suppose that we are interested in the total medical cost over period $[0, L]$. If data on cost and explanatory variables are available in multiple intervals such as every month or every year, we can set up the data into a panel format by dividing the entire time period of interest 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. Distribution of T is assumed to be continuous on $[0, L]$.

Survival time and medical costs 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 whenever the censoring time C exceeds the maximum T and t_g. Therefore, some individuals counted as censored in our previous work can be considered uncensored in some interval during the period of interest. The increase in the sample size allows more precise estimators and test statistics with more power.

For ith individual, let $Z_i = min(T_i, C)$, $s_i^* = I(C \geq T_i)$, and $s_{ig} = I(C \geq T_{ig}^*)$, where $I(.)$ is the indicator function. There are two types of censoring: time censoring if $s_i^* = 0$; that is, $T_i > C_i$, and cost censoring if $s_{ig} = 0$, that is,

$min(T_i, t_{ig}) > C_i$.

Let y_{ig} be the total medical (or log transformed) cost for ith 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.

2.1.1 Pooled Ordinary Least Squares (POLS) Estimation

The properties of POLS for the linear data can be summarized as follows. Assume that the model is the usual linear model for $i.i.d$ cross-sections: for any i,

$$\boldsymbol{y_i} = \boldsymbol{X_i}\,\boldsymbol{\beta} + \boldsymbol{u_i} \qquad i = 1, 2, ..., N \tag{2.1}$$

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

Assumption 1 :

(i) $E(\boldsymbol{u_i} \mid \boldsymbol{S_i}, \boldsymbol{X_i}) = E(\boldsymbol{u_i} \mid \boldsymbol{X_i}) = 0$.

(ii) $E(\tilde{\boldsymbol{X}}'_i\,\tilde{\boldsymbol{X}}_i)$ has rank K.

It is well known that under the assumption 1, the unweighted POLS estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{UP}}$ is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{UP}} = \hat{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1}\,(N^{-1}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}'_i\,\tilde{\boldsymbol{y}}_i); \tag{2.2}$$

where

$$\hat{\boldsymbol{A}}_{\boldsymbol{UP}} = (N^{-1} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{\boldsymbol{i}}' \, \tilde{\boldsymbol{X}}_{\boldsymbol{i}}) \tag{2.3}$$

on the unbalanced panel is consistent; and its asymptotic robust variance matrix is $V(\hat{\boldsymbol{\beta}}_{\boldsymbol{UP}})=\hat{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1} \, \hat{\boldsymbol{B}}_{\boldsymbol{UP}} \, \hat{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1} /N$, where

$$\hat{\boldsymbol{B}}_{\boldsymbol{UP}} = N^{-1} \sum_{i=1}^{N} \tilde{\boldsymbol{X}}_{\boldsymbol{i}}' \, (\boldsymbol{S}_{\boldsymbol{i}} \tilde{\boldsymbol{u}}_{\boldsymbol{i}}) \, (\boldsymbol{S}_{\boldsymbol{i}} \tilde{\boldsymbol{u}}_{\boldsymbol{i}})' \, \tilde{\boldsymbol{X}}_{\boldsymbol{i}} \tag{2.4}$$

and $\tilde{\boldsymbol{u}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}} \hat{\boldsymbol{\beta}}_{\boldsymbol{UP}}$ (Wooldridge, 2002).

The key exogenous sampling assumption underlying the validity of the unweighted POLS estimator on the selected sample is given in assumption 1(i). Exogenous sampling in this setup implies that

$$E(y_{ig} | \boldsymbol{x}_{\boldsymbol{ig}} \, , T_{ig}^{*}, C_i) = E(y_{ig} | \boldsymbol{x}_{\boldsymbol{ig}}) \quad g = 1, 2, ...G \quad i = 1, 2, ..., N. \tag{2.5}$$

Under administrative censoring C_i is independent of y_{ig} but we would expect that T_{ig}^{*} and y_{ig} may be correlated. Correlation increases with the length of the interval. Violation of equation (2.5) would yield an inconsistent POLS estimator.

IPW estimation produces consistent and $\sqrt{N}$ asymptotically normal estimators even under the violation of equation (2.5) with the following assumption. Suppose that T and C are independent given x.

Assumption 1' :

(i) $E(\boldsymbol{X}_{\boldsymbol{i}}' \, \boldsymbol{u}_{\boldsymbol{i}}) = 0.$

(ii) Same as assumption 1 part (ii).

(iii) $\boldsymbol{x}_{\boldsymbol{ig}} \, , y_{ig}, T_{ig}^{*}$ are observed when $s_{ig} = 1$, C_i is always observed.

(iv) $\boldsymbol{x_{ig}}$ and y_{ig} can be *ignorable* in the selection equation

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

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'(iii) simply defines when the data are observable. Part (iv) requires that the selection probability is observable when $s_{ig} = 1$.

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

$$\hat{\boldsymbol{A}}_{\boldsymbol{WP}}^{-1} (N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{X}}_{\boldsymbol{i}}' \hat{\boldsymbol{y}}_{\boldsymbol{i}}), \tag{2.6}$$

where

$$\hat{\boldsymbol{A}}_{\boldsymbol{WP}} = (N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{X}}_{\boldsymbol{i}}' \hat{\boldsymbol{X}}_{\boldsymbol{i}}), \tag{2.7}$$

$\hat{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{W_i}\, \boldsymbol{X_i}$, $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{W_i}\, \boldsymbol{y_i}$, and $\boldsymbol{W_i}$ is a $G \times G$ diagonal matrix in which the gth diagonal element is $\sqrt{w_{ig}}$ where

$$w_{ig} = s_{ig}/P(C_i \geq T_{ig}^*). \tag{2.8}$$

$\hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}$ is consistent, asymptotically normal and its asymptotic robust variance matrix is

$$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}) = \hat{\boldsymbol{A}}_{\boldsymbol{WP}}\, \hat{\boldsymbol{B}}_{\boldsymbol{WP}}\, \hat{\boldsymbol{A}}_{\boldsymbol{WP}}^{-1}/N, \tag{2.9}$$

where

$$\hat{\boldsymbol{B}}_{\boldsymbol{WP}} = N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{X}}_{\boldsymbol{i}}' \left(\boldsymbol{W}_{\boldsymbol{i}} \hat{\boldsymbol{u}}_{\boldsymbol{i}}\right) \left(\boldsymbol{W}_{\boldsymbol{i}} \hat{\boldsymbol{u}}_{\boldsymbol{i}}\right)' \hat{\boldsymbol{X}}_{\boldsymbol{i}} \tag{2.10}$$

and $\hat{\boldsymbol{u}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}} \hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}$ (Wooldridge, 1999).

Each observation of $(y_i, \boldsymbol{x}_i)$ is weighted by the inverse probability of appearing in the sample. Assumption 1' part (iv) requires $P(C_i > T_{ig}^*)$ to be known whenever $s_{ig}^* = 1$, so $\hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}$ is computable from observed data assuming we know $P(C_i > T_{ig}^*)$.

Usually the sampling probability function, w_{ig}, is unknown and needs to be estimated. We propose to estimate the unknown survivor function by the Kaplan-Meier (1958) estimator, with the roles of C and T reversed.

Define $p(t) = P(C \geq t)$, and let $\hat{p}(t)$ be the product limit estimator of $p(t)$ based on the data $(Z_i, \overline{s_i^*})$ $(i = 1, ..., N)$, where $\overline{s_i^*} = 1 - s_i^*$. Then,

$$\hat{w}_{ig} = \frac{s_{ig}}{\hat{p}(T_{ig}^*)} \qquad i = 1, ..., N; g = 1, ..., K. \tag{2.11}$$

Lemma 4.3 in Newey and McFadden (1994) shows that if w_{ig} in (8) is replaced with consistent estimator $\hat{w}_{ig}$, under the conditions in which the uniform weak law of large numbers can be applied, then $\hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}$ consistently estimates $\hat{\boldsymbol{\beta}}$ in equation (2.1).

The estimated covariance matrix in (2.9) is the White (1980) heteroskedasticity-consistent covariance matrix applied to all variables for observation i at the gth interval and weighted by $\sqrt{\hat{w}_{ig}}$. Censoring therefore can be handled fairly easily because most standard statistics software programs compute a heteroskedasticity-consistent covariance matrix.

This simplicity does not work when w_{ig} is replaced with $\hat{w}_{ig}$ for variance of IPW POLS because it should be adjusted for estimation of selection probabilities.

Fortunately, Wooldridge (2002b) shows that estimating the selection probabilities leads to a more efficient estimator than using known probabilities. In other words, if we compute the asymptotic covariance matrix as if we have no estimated probabilities and if we get significant estimators by using the easily computable matrix in (2.9), we know that they would have smaller standard errors under corrected covariance matrix calculation. This is somewhat unusual for two-step estimation problems. Estimating (2.9) by using $\hat{w}_{ig}$ instead of w_{ig} results in a conservative inference.

The steps for deriving consistent two-step IPW least-squared estimators and their *unadjusted* asymptotic variance estimators can be summarized as follows.

(i) Calculate the product-limit estimator, m_i, based on data $(Z_i, 1 - s_i^*)$ $(i = 1, ..., N)$.

(ii) Generate $p_{ig} = m_{ig}$; where m_{ig} is the value of m_i at $T_{ig}*$ and $s_{ig} = 1$ if (y_{ig}, x_{ig}) is observed at $(t_{g-1}, t_g]$.

(iii) Generate weight, $w_{ig} = s_{ig}/p_{ig}$ $(i = 1, ..., N)$.

(iv) Generate weighted response and explanatory variables: $y_{ig}^* = \sqrt{w_{ig}} y_{ig}$, $\boldsymbol{x}_{ig}^* = \sqrt{w_{ig}}\, \boldsymbol{x}_{ig}$ $(i = 1, ..., N)$.

(v) Compute the OLS regression of y_{ig}^* on $\boldsymbol{x}_{ig}^*$ with robust option.

2.1.2 Random Effect Model

Panel data usually provides the researcher with a large number of data points that increases the degrees of freedom and reduces collinearity among explanatory variables. Panel data 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. By using information on both the intertemporal dynamics and the individuality of the entities being investigated, one

is better able to control for the effects of unobserved variables (Hsiao, 1999).

Let us first investigate assumptions under which the random effects estimator is consistent under exogenous selection. The model is the unobserved effects model for any i,

$$y_{ig} = \boldsymbol{x_{ig}}\,\boldsymbol{\beta} + \alpha_i + u_{ig} \qquad g = 1, 2, ..., G, \tag{2.12}$$

where α_i is unobserved effect, $\boldsymbol{x_{ig}}$ is $1 \times K$; and $\boldsymbol{\beta}$ is the $K \times 1$ vector of interest.

We can write the model as

$$\boldsymbol{y_i} = \boldsymbol{X_i}\,\boldsymbol{\beta} + \boldsymbol{v_i} \tag{2.13}$$

by defining $\boldsymbol{y_i} = (y_{i1}, y_{i2}, ..., y_{iG})'$ and $\boldsymbol{v_i} = \alpha_i j_G + \boldsymbol{u_i}$, where j_G is the $G \times 1$ vector of ones, $\boldsymbol{u_i} = (u_{i1}, u_{i2}, ..., u_{iG})'$ and $\boldsymbol{X_i} = (\boldsymbol{x'_{i1}}, \boldsymbol{x'_{i2}}, ..., \boldsymbol{x'_{iG}})'$.

Define the variance matrix of $\boldsymbol{v_i}$ over uncensored cases as
$\boldsymbol{\Omega_{si}} = \boldsymbol{S_i}\, E(\boldsymbol{v_i}\, \boldsymbol{v'_i})\boldsymbol{S_i}$, a $G \times G$ matrix that we assume to be positive definite.

Assumption 2:

(i) $E(\boldsymbol{v_i} \,|\boldsymbol{X_i}, \boldsymbol{S_i}) = E(\boldsymbol{v_i} \,|\boldsymbol{X_i}) = 0$

(ii) $rank\ E(\tilde{\boldsymbol{X}}_i' \,\boldsymbol{\Omega}_{si}^{-1}\, \tilde{\boldsymbol{X}}_i) = K$

Under assumption 2, generalized least squares (GLS) over uncensored data is consistent and it can be shown that a consistent estimator of the RE model is

$$\tilde{\boldsymbol{\beta}}_{RE} = \tilde{\boldsymbol{A}}_{RE}^{-1} \left(N^{-1} \textstyle\sum_{i=1}^{N} \tilde{\boldsymbol{X}}_i' \,\boldsymbol{\Omega}_{si}^{-1}\, \tilde{\boldsymbol{y}}_i \right); \tag{2.14}$$

where

$$\tilde{\boldsymbol{A}}_{\boldsymbol{RE}} \;=\; N^{-1}\sum_{i=1}^{N}\tilde{\boldsymbol{X}}_{\boldsymbol{i}}'\,\boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\,\tilde{\boldsymbol{X}}_{\boldsymbol{i}}\;; \tag{2.15}$$

with variance matrix

$$V(\tilde{\boldsymbol{\beta}}_{\boldsymbol{RE}}\,) = \tilde{\boldsymbol{A}}_{\boldsymbol{RE}}^{-1}\,\tilde{\boldsymbol{B}}_{\boldsymbol{RE}}\,\tilde{\boldsymbol{A}}_{\boldsymbol{RE}}^{-1}\,/N, \tag{2.16}$$

where

$$\tilde{\boldsymbol{B}}_{\boldsymbol{RE}} \;= N^{-1}\sum_{i=1}^{N}\Big(\tilde{\boldsymbol{X}}_{\boldsymbol{i}}'\,\boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\,(\boldsymbol{S}_{\boldsymbol{i}}\,\tilde{\boldsymbol{\eta}}_{\boldsymbol{i}}\,)(\boldsymbol{S}_{\boldsymbol{i}}\,\tilde{\boldsymbol{\eta}}_{\boldsymbol{i}}\,)'\boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\,\tilde{\boldsymbol{X}}_{\boldsymbol{i}}\,\Big)\,; \tag{2.17}$$

and $\tilde{\boldsymbol{\eta}}_{\boldsymbol{i}} = \tilde{\boldsymbol{y}}_{\boldsymbol{i}} \;- \boldsymbol{X}_{\boldsymbol{i}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{RE}}$ is the vector of residual i.

As mentioned previously, correlation between survival times and medical costs violates the exogenous sampling assumption. Violation of assumption 2(i) makes $\hat{\boldsymbol{\beta}}_{\boldsymbol{RE}}$ defined in (2.14) inconsistent.

Inverse probability weighted estimation with the following assumption produces consistent and $\sqrt{N}$ asymptotically normal estimators under violation of the exogenous sampling assumption.

Define the (unconditional) weighted variance matrix of $\boldsymbol{v}_{\boldsymbol{i}}$ as
$\boldsymbol{\Omega}_{\boldsymbol{wi}} = \boldsymbol{W}_{\boldsymbol{i}}\mathrm{E}(\boldsymbol{v}_{\boldsymbol{i}}\boldsymbol{v}_{\boldsymbol{i}}')\boldsymbol{W}_{\boldsymbol{i}}$ and $\boldsymbol{T}_{\boldsymbol{i}}^{*} = (T_{i1}^{*}, T_{i2}^{*}, ...T_{iG}^{*})$.

Assumption 2':

(i) $E(\boldsymbol{X}_{\boldsymbol{i}}'\,\boldsymbol{v}_{\boldsymbol{i}}\,) = 0$

(ii) $rank\ E(\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\,\boldsymbol{\Omega}_{\boldsymbol{wi}}^{-1}\,\hat{\boldsymbol{X}}_{\boldsymbol{i}}\,) = K$

(iii) $\boldsymbol{x}_{\boldsymbol{ig}}\,, y_{ig}, T_{ig}^{*}$ are observed when $s_{ig} = 1$, C_i is always observed.

(iv) $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ can be *ignorable* in the selection equation

$$P(s_{ig}^{*}=1|\boldsymbol{x_i}\ ,\boldsymbol{y_i},C_i,\boldsymbol{T_i^{*}})=P(s_{ig}^{*}=1|C_i,T_{ig}^{*})=P(C_i\geq T_{ig}^{*}).$$

Under assumption 2' GLS is consistent, however, obtaining GLS requires knowing $\boldsymbol{\Omega_{wi}}$ up to scale. In feasible GLS (FGLS) estimation, we replace unknown matrix $\boldsymbol{\Omega_{wi}}$ with a consistent estimator and get asymptotic properties that are identical to those of the GLS estimator.

Since the sampling probability function is unknown, we can use a proposed consistent estimator of w_{ig} to find the IPWRE estimator:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{WRE}} \ = \hat{\boldsymbol{A}}_{\boldsymbol{WRE}}^{\ -1}\left(N^{-1}\textstyle\sum_{i=1}^{N}\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\,\hat{\boldsymbol{\Omega}}_{\boldsymbol{wi}}^{-1}\hat{\boldsymbol{y}}_{\boldsymbol{i}}\right); \tag{2.18}$$

where

$$\hat{\boldsymbol{A}}_{\boldsymbol{WRE}} \ = \ N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\,\hat{\boldsymbol{\Omega}}_{\boldsymbol{wi}}^{-1}\hat{\boldsymbol{X}}_{\boldsymbol{i}}\ ; \tag{2.19}$$

provided that $\hat{\boldsymbol{\Omega}}_{\boldsymbol{wi}}$ is consistent estimator of $\boldsymbol{\Omega_{wi}}$.

The unajusted robust variance of the IPWRE estimator is

$$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{WRE}}\ )=\hat{\boldsymbol{A}}_{\boldsymbol{WRE}}^{-1}\,\hat{\boldsymbol{B}}_{\boldsymbol{WRE}}\,\hat{\boldsymbol{A}}_{\boldsymbol{WRE}}^{-1}/N, \tag{2.20}$$

where

$$\hat{\boldsymbol{B}}_{\boldsymbol{WRE}}\ =N^{-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{X}}_{\boldsymbol{i}}'\ \hat{\boldsymbol{\Omega}}_{\boldsymbol{wi}}^{-1}(\hat{\boldsymbol{W}}_{\boldsymbol{i}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})(\hat{\boldsymbol{W}}_{\boldsymbol{i}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})'\hat{\boldsymbol{X}}_{\boldsymbol{i}}\ \right), \tag{2.21}$$

and $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}=\boldsymbol{y_i}\ -\boldsymbol{X_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{WRE}}$ is the vector of residual i.

$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{WRE}})$ is unadjusted because it should be adjusted for the first-stage estimation of $\boldsymbol{W_i}$. As in the IPW POLS estimation, we also know the direction. By using an unadjusted covariance matrix, we can get a conservative inference. The estimated covariance matrix in (2.20) is the White (1980) heteroskedasticity-consistent covariance matrix after random effect estimation, applied to all variables for observation i at the gth interval weighted by $\sqrt{\hat{w}_{ig}}$. Based on the results of Wooldridge (2002b), we know that any significant variables determined by this easily computable method will have more power under the adjusted variance matrix.

The steps for deriving consistent IPW RE estimators and their *unadjusted* asymptotic variance estimators are the same as the steps described earlier for IPW POLS; with the following exception:

(v) Compute the RE regression of y_{ig}^{*} on $\boldsymbol{x}_{\boldsymbol{ig}}^{*}$ with robust option.

2.2 Weighted or Unweighted Estimator?

It has been shown that the unweighted estimator is no less efficient than the weighted estimator under homoskedasticity and exogenous sampling (Wooldridge 1999, 2001). For a linear regression model, the Gauss-Markov Theorem for independent observation implies that an 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 sampling scheme is not exogenous and the weighted estimator is consistent with or without exogenous sampling, we can apply a Hausman(1978) test to determine exogeneity of sampling.

The traditional form of Hausman statistics can be used under the homoskedasticity assumption. We can state this assumption for the POLS estimator

as follows:

$$E(\tilde{\boldsymbol{X}}_i'\boldsymbol{S}_i\boldsymbol{u}_i\boldsymbol{u}_i'\boldsymbol{S}_i\tilde{\boldsymbol{X}}_i\ ) = \sigma_0^2 E(\tilde{\boldsymbol{X}}_i'\tilde{\boldsymbol{X}}_i\ ) \tag{2.22}$$

When (2.22) holds, estimation of the unweighted POLS variance estimator is simplified further:

$$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{UP}}\ ) = \hat{\sigma^2}\hat{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1}\ , \tag{2.23}$$

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

The homoskedasticity assumption under RE is

$$E(\hat{\boldsymbol{X}}_i'\ \boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\boldsymbol{v}_i'\boldsymbol{v}_i\boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\ \hat{\boldsymbol{X}}_i\ ) = E(\hat{\boldsymbol{X}}_i'\boldsymbol{\Omega}_{\boldsymbol{si}}^{-1}\ \hat{\boldsymbol{X}}_i\ ). \tag{2.24}$$

Then, the unweighted RE variance estimator becomes

$$V(\hat{\boldsymbol{\beta}}_{\boldsymbol{RE}}\ ) = \hat{\boldsymbol{A}}_{\boldsymbol{RE}}^{-1}\ . \tag{2.25}$$

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

$$\boldsymbol{H} = \left(\hat{\boldsymbol{\theta}}_{\boldsymbol{w}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{u}}\right)\hat{\boldsymbol{V}}^{-1}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{w}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{u}}\right). \tag{2.26}$$

For weighted and unweighted POLS, choose $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{u}}$ as $\hat{\boldsymbol{\beta}}_{\boldsymbol{WP}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{UP}}$, respectively. $\hat{\boldsymbol{V}} \equiv \hat{\boldsymbol{V}}_{\boldsymbol{w}} - \hat{\boldsymbol{V}}_{\boldsymbol{u}}$, where $\hat{\boldsymbol{V}}_{\boldsymbol{w}}$ is defined in equation (2.9) and $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ is defined in equation (2.23) under the homoskedasticity assumption.

For the RE model, $\hat{\boldsymbol{\theta}}_{\boldsymbol{w}}$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{u}}$ is $\hat{\boldsymbol{\beta}}_{\boldsymbol{WRE}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{RE}}$. $\hat{\boldsymbol{V}}_{\boldsymbol{w}}$ is as in equation (2.20); and $\hat{\boldsymbol{V}}_{\boldsymbol{u}}$ is as in equation (2.25).

In many cases we may want to use a Hausman test when the homoskedasticity

assumption is violated. This requires a robust form that replaces $\hat{\boldsymbol{V}}$ for the POLS estimation:

$$(\hat{\boldsymbol{A}}_{\boldsymbol{WP}}^{-1} \mid -\tilde{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1}) \left(N^{-1} \sum_{i=1}^{N} \sum_{g=1}^{G} \hat{\boldsymbol{e}}_{\boldsymbol{ig}} \, \hat{\boldsymbol{e}}_{\boldsymbol{ig}}' \right) (\hat{\boldsymbol{A}}_{\boldsymbol{WP}}^{-1} \mid -\tilde{\boldsymbol{A}}_{\boldsymbol{UP}}^{-1})'/N, \tag{2.27}$$

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

$$(\hat{\boldsymbol{A}}_{\boldsymbol{WRE}}^{-1} \mid -\tilde{\boldsymbol{A}}_{\boldsymbol{RE}}^{-1}) \left(N^{-1} \sum_{i=1}^{N} \tilde{\boldsymbol{e}}_{\boldsymbol{i}} \, \tilde{\boldsymbol{e}}_{\boldsymbol{i}}' \right) (\hat{\boldsymbol{A}}_{\boldsymbol{WRE}}^{-1} \mid -\tilde{\boldsymbol{A}}_{\boldsymbol{RE}}^{-1})'/N, \tag{2.28}$$

where $\tilde{\boldsymbol{e}}_{\boldsymbol{i}}=\left(\hat{\boldsymbol{X}}_{\boldsymbol{i}}' \, \hat{\boldsymbol{\Omega}}_{\boldsymbol{wi}}^{-1} \hat{\boldsymbol{W}}_{\boldsymbol{i}} \hat{\boldsymbol{\eta}}_{\boldsymbol{i}}, \tilde{\boldsymbol{X}}_{\boldsymbol{i}}' \, \boldsymbol{\Omega}_{\boldsymbol{si}}^{-1} \boldsymbol{S}_{\boldsymbol{i}} \tilde{\boldsymbol{\eta}}_{\boldsymbol{i}} \right)'$, and $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{i}}$ are the residuals after weighted and unweighted RE estimation (Wooldridge, 1995).

The methods described are easily applicable using standard commercial statistical software programs. The traditional Hausman test is built in to most statistical programs, but the robust form Hausman requires programming. We can use an alternative approach, the regression-based Hausman test, for easy computation of the robust form. Ruud (1984) and Wooldridge (1990) examine this issue. Since the Hausman test compares systematic differences in the coefficients, if we regress the independent variables on weighted and unweighted explanatory variables and the coefficients are not different, then the F test for the coefficients on weighted explanatory variables should result in an insignificant value. It can be shown that the statistics obtained from this procedure are asymptotically equivalent to Hausman statistics that compare the difference of weighted and unweighted estimators. To obtain the traditional form of the Hausman test:

(i) Compute the POLS(RE) regression of y_{ig} on $\boldsymbol{x}_{\boldsymbol{ig}}$ and $\boldsymbol{x}_{\boldsymbol{ig}}^{*}$ with the heteroskedasticity robust option.

(ii) Compute the F test for the coefficients of x_{ig}^{*}.

We can also obtain the traditional Hausman test statistics if we repeat (i) without the heteroskedasticity robust option.

If the Hausman test indicates rejection then the exogeneous sampling assumption is violated; and the unweighted estimator are inconsistent. A failure to reject means the coefficients from unweighted and weighted estimators are not systematically different and can be used as evidence of exogeneous sampling.

2.3 The Lung Cancer Study

2.3.1 The Data

The data are from a project entitled "Family Home Care for Cancer: A Community-Based Model from: the National Institute of Nursing Research and National Cancer Institute (grant No. NR1915-06)",which studied 202 Medicare beneficiaries over age 65 who were diagnosed with lung cancer from 1994 through 1997. Among them, 183 subjects who had some kind of treatment, whether surgery, radiation, or chemotherapy.

Medicare claim files for each patient for two years following diagnosis were obtained. These files revealed monthly cost values, treatment types, hospitalization, and death status during the 24 months. Payments by Medicare were used as a proxy for direct Medicare costs as opposed to bill charges. Costs are adjusted for inflation to 1997 prices by using the National Medicare Price Index, 1994-1997.

Patient information (such as age, sex, race) was obtained through interviews. In addition, we collected data on patients' physical function three months prior to diagnosis as measured by the short form 36 (SF-36). Comorbid conditions were assessed by 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 Commitee 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 2.1. It shows that medical care expenditures for lung cancer patients spike in the first month after diagnosis, during the surgical period. The interventions such as surgery and radiation incur large costs within the first couple of months; whereas chemotherapy may be administered over a much longer time.

2.3.2 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 and Mullahy (2001), our cost values satisfied the conditions in which an OLS-based model with a long-transformed dependent variable is suitable.

Table 2.1 shows the results of the regression analysis predicting the total cost of care. The first two columns present the regression results estimated by POLS and IPW POLS. As emphasized throughout this paper, POLS is likely to suffer from omitted variable problems.Therefore, RE and IPW RE models were estimated as an

Figure 2.1: Distribution of average monthly cost

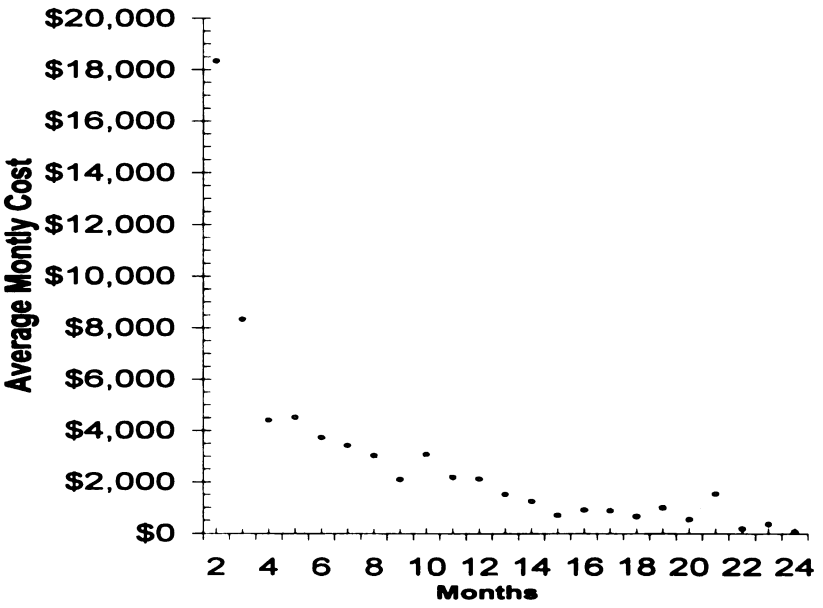


Table 2.1: Estimation of Log of Total Medical Cost from Longitudinal Data

Variables	POLS	IPWPOLS	RE	IPWRE
constant	6.67	6.47	6.80	6.23
	(1.52)**	(0.68)**	(1.57)**	(1.54)**
late stage	−1.10	−1.17	−1.02	−1.12
	(0.20)**	(0.09)**	(0.20)**	(0.20)**
late comorbidity	0.48	0.51	0.49	0.53
	(0.20)*	(0.09)**	(0.21)*	(0.19)**
hospitalize	3.61	3.71	3.50	3.45
	(0.23)**	(0.17)**	(0.21)**	(0.20)**
radiation	4.05	4.08	3.88	3.83
	(0.16)**	(0.17)**	(0.16)**	(0.16)**
$radiation_1$	1.02	1.06	0.90	0.87
	(0.18)**	(0.22)**	(0.18)**	(0.18)**
$radiation_2$	0.90	0.90	0.71	0.64
	(0.22)**	(0.21)**	(0.21)**	(0.21)**
chemothrapy	2.95	2.99	2.78	2.73
	(0.22)**	(0.18)**	(0.21)**	(0.21)**
$chemothrapy_1$	1.11	1.16	1.04	0.99
	(0.18)**	(0.21)**	(0.18)**	(0.17)**
$chemothrapy_2$	0.96	0.94	0.88	0.77
	(0.19)**	(0.19)**	(0.19)**	(0.18)**
other	2.97	2.07	2.93	2.90
	(0.21)**	(0.21)**	(0.21)**	(0.20)**
$other_1$	1.35	1.40	1.30	1.26
	(0.22)**	(0.24)**	(0.23)**	(0.22)**
$other_2$	0.65	0.64	0.57	0.49
	(0.24)**	(0.24)**	(0.24)**	(0.23)**
death	0.02	0.06	0.20	0.18
	(0.19)	(0.09)	(0.19)	(0.18)
physical functions	0.008	0.008	0.007	0.007
	(0.003)*	(0.002)**	(0.003)*	(0.03)*
age	−0.02	−0.02	−0.02	−0.01
	(0.02)	(0.01)*	(0.02)	(0.02)
sex	−0.33	−0.32	−0.30	−0.30
	(0.18)	(0.18)	(0.19)	(0.19)
white	−0.29	−0.34	−0.21	−0.27
	(0.36)	(0.17)*	(0.19)	(0.34)
observations	4000	4000	4000	4000
$r-squared$	0.61	0.59	0.60	0.61

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

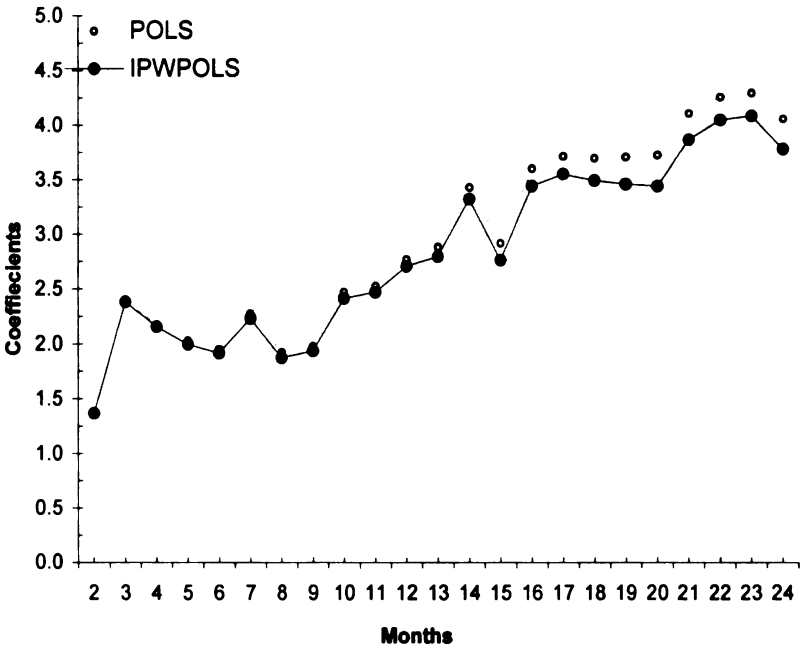
alternative way to use panel data to view the unobserved factors affecting the dependent variable. Almost all models explained 60% of the variation in total cost.

The population may have a different distribution in different periods, therefore we allow 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 ($p < .05$). Figure 2.2 shows the pattern of the absolute value of the coefficients under POLS and IPW POLS estimation. For example, after we control for patient and treatment related variables, a patient's total medical cost 4.3 less in month 24 after diagnosis than in first month after diagnosis. As shown in Table 2.1, the control variables are age, gender, race, comorbid conditions, stage of cancer,death status, physical function, and treatment-related variables. We divided treatment into four categories: no treatment, radiation only, chemothrapy only and, others which includes chemo and radiation, surgery only, and surgery plus other therapies. No treatment was chosen as the reference group. Our time independent variables are gender, race, comorbid conditions, stage of cancer and physical function.

The only variables that did not reach statistical significance under POLS and IPW POLS estimation is death. The coefficients for physical functioning and age, while statistically significant, are small in magnitude. On average, expenses for male patients were almost 31% less than for female patients. Race is also significant. The costs for whites is 34% less than black people.

Disease severity measures, such as comorbid condition and stage, have different and statistically significant effects. As shown in columns 1 and 2 of Table 2.1, having three or more comorbid conditions increase cost by almost 48% and 51%, respectively. Disease stage has a large negative effect on costs. Regional stage decreased total cost of care almost 1.1 times compared to in situ or local stage cancer according to POLS and IPWPOLS.

Figure 2.2: Distribution of absolute value of monthly dummy coefficients under POLS and IPW POLS estimation.



Hospitalization for reasons other than lung cancer surgery increases total medical cost 3.6 times (column 1) and 3.7 times (column 2) during the period of interest.

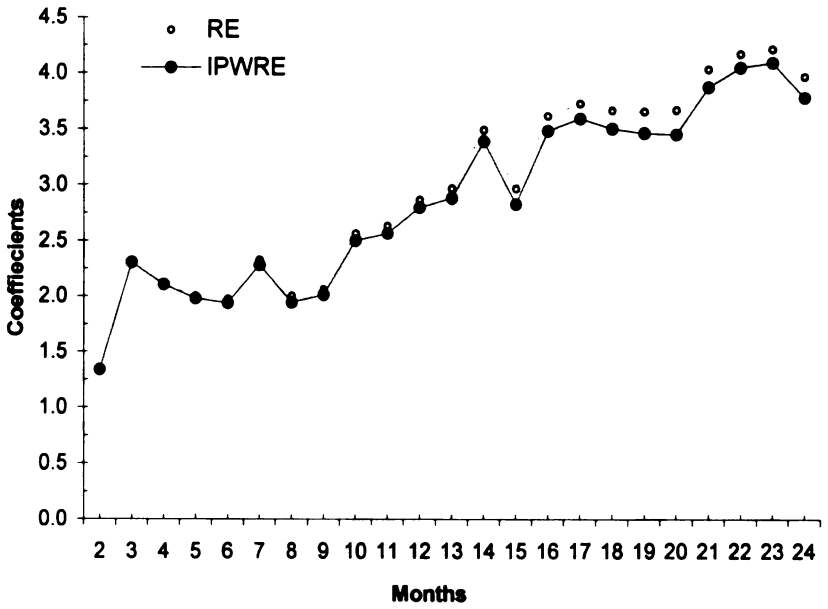
A two-period lag effect is found treatment-related variables. If a person receives radiation in a particular month; on average cost increases almost 4 times relative to the ones who have no treatment in that month according to both weighted and unweighted POLS estimation. If the same person has radiation one month prior, the effect becomes almost 5 times. To see the effect of radiation alone relative to no treatment, we need to add three coefficients. So overall, if a person receives a radiation, total cost is 6 times more than for someone who had no treatment. The effects for chemothrapy only and the others category are almost 5 times.

Note that the coefficients estimated by POLS and IPW POLS are not practically different. Since the exogenous sampling assumption is violated, however, POLS estimators are inconsistent. The traditional and robust Hausman test described in section 2.2 reject the null that sampling scheme is exogenous (pvalue is 0 for five decimal points).

Figure 2.3 shows the distribution of absolute value of monthly dummy variable coefficients when the first month after diagnosis is the base month. The variation of the coefficients between weighted and unweighted RE estimation is relatively smaller than that of POLS. As revealed in columns 3 and 4 of Table 2.1, death status, race and age are not statistically significant and physical functioning is practically insignificant. Gender has a significant effect. The cost for male patients is 29 % and 27% less than for female patients according to RE and IPW RE, respectively.

We can see differences for disease severity measured under weighted and unweighted RE models. Regional stage decreases total cost of care by 1.3 times according to IPW RE and by 1 time according to RE estimation. A patient with late comorbidity conditions paid 50% to 60% more on average; depending on the

Figure 2.3: Distribution of absolute value of monthly dummy coefficients under RE and IPW RE estimation.



estimation method.

Hospitalization other than for lung cancer surgery increases total medical costs almost 3.5 times, which is very similar to the POLS estimation.

The two-period lag effect persists under RE estimation methods. The permanent effects of radiation only, chemotherapy only, and others are 5.4, 4.5, and 4.7 times greater relative to nontreatment under unweighted RE estimation and are 5.7, 4.7, 4.8 times more under IPW RE estimation.

Since survival times are correlated with total cost, values in column 4 of Table 2.1 are the consistent estimators. Unweighted RE estimators, given in column 3, are consistent under exogenous sampling. However, both traditional and robust form of the Hausman Test as in the case of POLS reject the null hypothesis (pvalue is 0 for five decimal place). So consistent estimators are in the one in column 4.

2.4 Conclusion

The IPW least-squares method was applied to longitudinal data to illustrate how the censoring problem can be solved. The main motivation for developing regression methods is to handle a large number of continuous and discrete covariates.

POLS and RE models were analyzed and their statistical properties examined under censoring. Usual POLS and RE estimation will create an inconsistent estimator without exogenous sampling. Since survival times are correlated with total medical cost, the exogenous sampling assumption is violated. IPW estimators produce consisted and $\sqrt{N}$ asymptotically normal estimators. The method is easy to apply and can be done with most statistical software programs on the market. Step by step procedures were provided.

Since unweighted POLS and RE estimators are consistent under exogenous sampling 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. Traditional and robust forms of Hausman test described to determine between the models.

The lung cancer study, although it does not demonstrate the full power of the IPW least squares method, served as an example of how to use proposed regression methods and test statistics. We create *artificial* panel data by dividing two years after diagnosis into the months. The better estimates can be obtained if the data set originally was set in panel data format. That would be an interesting research topic to explore.

Chapter 3: Full Parametric Estimation of Censored Medical Cost

Introduction

Due to escalating cost of medical care it is important that costs of health care interventions and treatments are carefully assessed. A common problem with medical cost data is censoring since not all patients are followed until the endpoint of interest.Therefore, their medical costs are not fully observed.

The estimation of medical costs might be addressed through multivariate regression analysis. Multivariate analysis can control for patient and clinical characteristics to estimate medical cost. Using regression methodology estimating medical cost is a relatively new technique. The regression methodology would be particularly valuable in identifying cost-effective intervention programs. These intervention programs require that treatment costs are compared with alternatives that requires proper analysis of conditional means.

Although there have been several non-parametric approaches (Lin, 2000a) and a semi-parametric approach (Lin, 2000b) suggested for handling censored data, currently there is not a valid full parametric regression method for assessing the effects of covariates (e.g. patient and clinical characteristics) on censored medical costs.

Non parametric methods are often not as efficient as parametric statistics. Parametric methods are the "best practice" for estimation. They provide speed,

accuracy and flexibility to estimating processes.

This chapter suggests a full parametric method for estimating the parameters in linear structural equations when the selection rule is governed by the Tobit model. The resulting estimators are shown to be consistent and asymptotically normal. The procedure introduced in this paper involves a two stage estimation. In the first stage, the selection equation is estimated by Tobit and in the second stage, an additional variable estimated by Tobit parameters is included in the structural equation to correct for possible sample selection bias. The resulting model in this paper could be estimated by full maximum-likelihood estimation (MLE). However, the two-stage approach has the advantage of being easier to compute and is usually more robust than full MLE. The drawback to our approach is that the asymptotic variance matrix is cumbersome to estimate; although it can be done by using the general methods in Newey and McFadden (1994).

This chapter is organized as follows. Section 1 outlines the general framework to show the conditions under which the ordinary least squares (OLS) estimator using selected sample is consistent. The next section demonstrates how to apply this framework to the cases for which the selection rule is determined by the Tobit model with the specific example of estimating censored medical cost data, including statistical properties of the estimation and step-by-step procedures. Section 3 describes an application of our methods to a study on medical costs with a comparison of non-parametric method results in the first chapter. Concluding remarks are given in Section 4. The appendix contains proofs of the propositions in the main text.

3.1 General Framework

Assume that there is a population represented by the random vector $(\boldsymbol{x}, y)$ where $\boldsymbol{x}$ is a $1 \times K$ vector of explanatory variables, y is the scalar response variable.

Suppose that the population model of interest is

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + u = \boldsymbol{x}\boldsymbol{\beta} + u, \tag{3.1}$$

where we define $x_1 = 1$ and u as the error term.

Let s be a binary indicator such that $s = 1$ if $(\boldsymbol{x}, y)$ is observed and $s = 0$ otherwise. Assume that $s = h(\boldsymbol{x})$ for some non-random function $h(.)$.

Let $(\boldsymbol{x_i}, y_i)$, $i = 1, 2, ..., N$, be a random sample from the population and let $s_i = h(\boldsymbol{x_i})$. Then OLS estimator using the selected subsample can be written as

$$\hat{\boldsymbol{\beta}} = \left(N^{-1} \sum_{i=1}^{N} s_i \boldsymbol{x_i'} \boldsymbol{x_i} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i \boldsymbol{x_i'} y_i \right). \tag{3.2}$$

Theorem 1:

In model (1), assume that $E(u^2) < \infty$, $E(x_i^2) < \infty$; $i = 1, 2, ..., K$. Let $s = h(\boldsymbol{x})$ be a binary indicator for non-random function $h(.)$. Assume that $E(u|\boldsymbol{x}) = 0$ and $E(s\boldsymbol{x'x}) = K$, then the OLS estimator using the selected sample, given by equation (2) is consistent for $\boldsymbol{\beta}$. All proofs are given in the Appendix.

In the next section, we show how to apply this framework for cases where selection is determined by a censored selection variable.

3.2 Statistical Methods

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

Assume that the population model of interest is defined as (1), where $y, \boldsymbol{x}, \boldsymbol{\beta}$ are respectively a scalar representing cumulative cost (or transformed cost) at L or T, a $1 \times K$ explanatory variables (patient characteristics, treatment types, others), a $K \times 1$ unknown regression parameters.

Medical costs may be subject to right censoring and therefore are not always fully observable. Let C be the time of censoring. Suppose individuals enter the study at different times and terminal point of the study is predetermined by the researcher, so that censoring times are known when an individual is entered into the study. This form of censoring is called administrative censoring. From figure 3.1, we can see that cases subject-2 and subject-4 is subject to administrative censoring. A convenient representation of such data is to rescale each individual's starting time to 0 as described in the figure 3.2. Since we are interested in the total medical cost over period $[0, L]$; we impose a second type of censoring which we will call *artificial* censoring for some cases. From figure 3.2, cases such as subject-1 or subject-3, who are not censored due to administrative censoring, will be *artificially* censored because their duration on study is greater than L. Subject-2 is both *artificially* and administratively censored. Note that in such cases, *artificial* censoring precedes administrative censoring. Subject-5 is neither artificially nor administratively censored.

Let $s = 1(C > T^*)$, where $1(.)$ is indicator function. So y is observed when $s = 1$ (all cases except subject-4 in figure 3.2). We calculate the expected value of y

Figure 3.1: Administrative censoring when each individual has different starting time

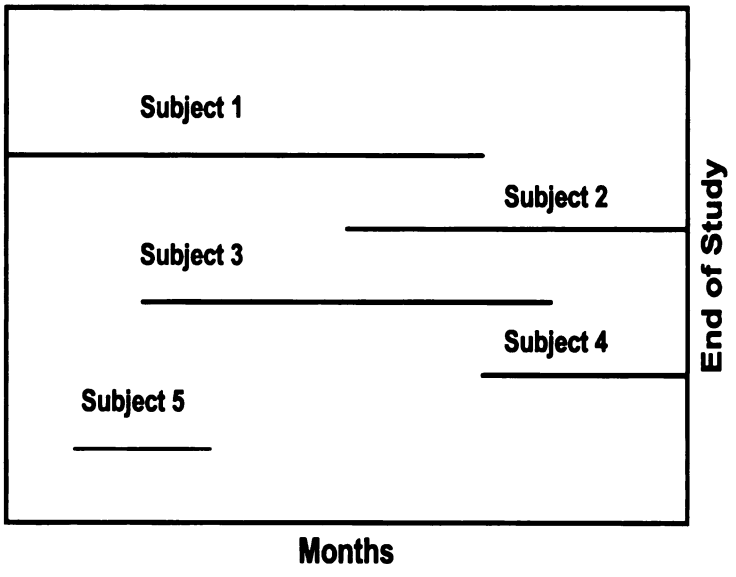
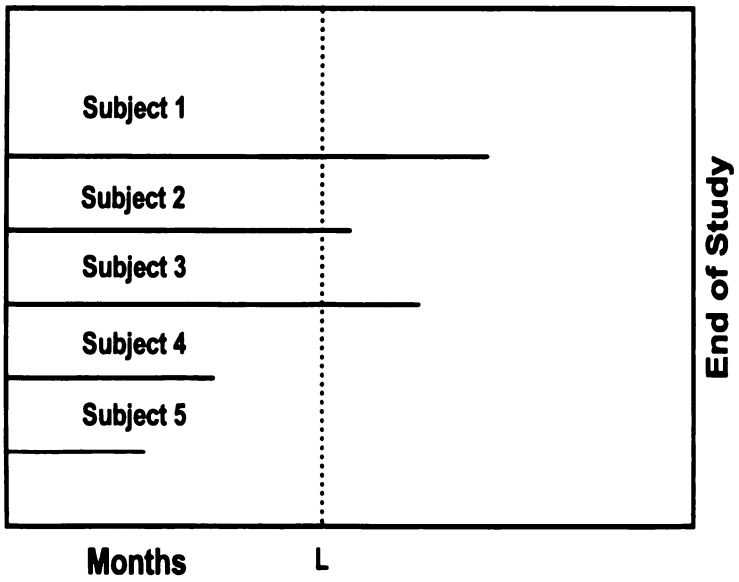


Figure 3.2: Starting time backed up to 0 for the individuals faced administrative censoring



conditional on $\boldsymbol{x}$,T^*,C; that is;

$$E(y|\boldsymbol{x}, T^*, C) = \boldsymbol{x}\boldsymbol{\beta} + E(u|\boldsymbol{x}, T^*, C) = \boldsymbol{x}\boldsymbol{\beta} + E(u|\boldsymbol{x}, T^*). \tag{3.3}$$

Since C is caused by study termination and L is determined by the research question (e.g., a week, a month, a year). Both are independent of y and $\boldsymbol{x}$. However, survival time depends on patient characteristics, treatment types ,and other factors. Let $log(T) = \boldsymbol{x}\boldsymbol{\alpha} + v$. Therefore, OLS estimation of y on $\boldsymbol{x}$ yields inconsistent estimators because $E(u|\boldsymbol{x})$ is not equal to 0.

The problem described so far can be transformed into the following statistical model:

$$y = \boldsymbol{x}\boldsymbol{\beta} + u \tag{3.4}$$

$$log(T^*) = min\,(logL, \boldsymbol{x}\boldsymbol{\alpha}+\mathrm{v}) \tag{3.5}$$

where $\boldsymbol{x}$ is always observed in the population but y is observed only when $log(T^*) < logC$.

Equation (3.4) and (3.5) are known as the censored Tobit Model (after Tobin, 1956). We refer equation (3.4) as the "structural regression equation" and equation (3.5) as the "selection equation." With the following assumption we show how to estimate $\boldsymbol{\beta}$ and its asymptotical covariance matrix consistently.

Assumption 1

(i) $\boldsymbol{x}$ is always observed in the population but y is observed only when $T^* < C$.

(ii) (u, v) is independent of $\boldsymbol{x}$ with zero mean.

(iii) $v \sim Normal(0, \tau^2)$

(iv) $E(u|v) = \rho v$

Assumption 1 part(i) defines the particular sample selection problem. y is not observable unless $T^* < C$. Part (ii) is a standard form of exogeneity of $\boldsymbol{x}$. Part (iii) is the most restrictive assumption, but it is needed to derive a conditional expectation given selected sample. Part (iv) requires linearity in the population model u on v. Under bivariate normality it always holds. However, the normality of u is not necessary.

Under assumption 1,

$$E(u|\boldsymbol{x}, T^*) = \rho E(v|\boldsymbol{x}, T^*). \tag{3.6}$$

Equation (3.3) can be written as,

$$E(y|\boldsymbol{x}, T^*, C) = \boldsymbol{x}\boldsymbol{\beta} + \rho E(v|\boldsymbol{x}, T^*). \tag{3.7}$$

If we calculate $E(v|\boldsymbol{x}, T^*)$, then from Theorem 1, we could consistently estimate $\boldsymbol{\beta}$ and ρ from the regression y on $\boldsymbol{x}$ and $E(v|\boldsymbol{x}, T^*)$.

From (3.5), for the uncensored cases, $1[logT < logL]$,

$$E(v|\boldsymbol{x}, T^*) = logT - \boldsymbol{x}\boldsymbol{\alpha} = v, \tag{3.8}$$

and for the censored cases,$1[logT \geq logL]$,

$$\begin{aligned} E(v|\boldsymbol{x}, T^*) &= E(v|\boldsymbol{x}, \boldsymbol{x}\boldsymbol{\alpha} + v \geq logL), \\ &= \tau E\left(v|\boldsymbol{x}, \frac{v}{\tau} \geq \frac{logL - \boldsymbol{x}\boldsymbol{\alpha}}{\tau}\right) \\ &= \tau\phi\left(\frac{logL - \boldsymbol{x}\boldsymbol{\alpha}}{\tau}\right) / \Phi\left(\frac{logL - \boldsymbol{x}\boldsymbol{\alpha}}{\tau}\right) \\ &= \tau\lambda\left(\frac{logL - \boldsymbol{x}\boldsymbol{\alpha}}{\tau}\right) \end{aligned} \tag{3.9}$$

where λ is the inverse Mills Ratio. Equation (3.8) and (3.9) can be written succinctly,

$$E(v|\boldsymbol{x}, T^*) = 1[logT < logL]v + 1[logT \geq logL]\tau\lambda\left(\frac{logL - \boldsymbol{x\alpha}}{\tau}\right) \tag{3.10}$$

Estimation of equation (3.5) with Tobit, replaces the unknown variables in equation (3.10) with their consistent estimators. In other words, replacing v with $\hat{v}$, residuals, τ with $\hat{\tau}$, estimated standard error, $\boldsymbol{\alpha}$ with $\hat{\boldsymbol{\alpha}}$, estimated coefficients, of the equation (3.5) after the Tobit estimation does not effect consistency of the parameters of equation(3.7), $\boldsymbol{\beta}$ and ρ. This result follows from Newey and McFadden (1994).

Therefore, for each i in the selected sample, if we define;

$$\hat{\eta}_i = 1[logT_i < logL]\hat{v}_i + 1[logT_i \geq logL]\hat{\tau}_i\lambda_i\left(\frac{logL - \boldsymbol{x_i\hat{\alpha}}}{\hat{\tau}}\right) \tag{3.11}$$

and $\hat{g}_i = (\boldsymbol{x_i}, \hat{\eta}_i)$, then OLS estimator of $\boldsymbol{\Theta} = (\boldsymbol{\beta'}, \rho')$ using selected sample can be written

$$\hat{\boldsymbol{\Theta}} = \left(N^{-1}\sum_{i=1}^{N} s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{g}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} s_i\hat{\boldsymbol{g}}_i y_i\right). \tag{3.12}$$

Theorem 2:

Under the Assumption 1, OLS estimator given equation (3.12) is consistent for $\boldsymbol{\Theta}$; and a consistent estimator of $Avar(\hat{\boldsymbol{\Theta}})$ is

$$\left(\sum_{i=1}^{N} s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{g}}_i\right)^{-1}\left(\sum_{i=1}^{N}\left(s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{e}}_i + \hat{\rho}\hat{C}\hat{r}_i\right)'\left(s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{e}}_i + \hat{\rho}\hat{C}\hat{r}_i\right)\right)\left(\sum_{i=1}^{N} s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{g}}_i\right)^{-1} \tag{3.13}$$

where $\hat{C} = \frac{1}{N}\sum_i^N s_i \hat{\boldsymbol{g}}_i'$, $\hat{e}_i = y_i - \boldsymbol{x_i}\hat{\boldsymbol{\beta}} - \hat{\rho}\hat{\eta}_i$ (for $s_i = 1$), and

$$\hat{r}_i = r_i(\hat{\boldsymbol{\gamma}}) = -\nabla_{\hat{\boldsymbol{\gamma}}}\eta_i \hat{\mathrm{H}}^{-1}\sum_i^N a_i,$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\alpha}}, \hat{\tau})$, $\nabla_{\hat{\boldsymbol{\gamma}}}\eta_i = \left(\frac{\partial n_i}{\partial \boldsymbol{\gamma_1}}(\hat{\boldsymbol{\gamma}}), \frac{\partial n_i}{\partial \boldsymbol{\gamma_2}}(\hat{\boldsymbol{\gamma}}), ..., \frac{\partial n_i}{\partial \boldsymbol{\gamma_p}}(\hat{\boldsymbol{\gamma}})\right)$, $\hat{\mathrm{H}}$ is the $P \times P$ Tobit hessian and $\hat{a}_i$ is the score of the censored tobit log likelihood for observation i valued at estimated parameters.

Since the term with the generated regressor $\hat{\eta}_i$ does not appear in the variance matrix (3.13) when $\hat{\rho} = 0$, the usual variance matrix for $\hat{\boldsymbol{\Theta}}$ is valid under homoskedasticity and the robust version is valid under heteroskedasticity. Therefore testing $\rho = 0$ is just usual t-statistics or its heteroskedastic robust version.

The steps for deriving consistent estimators in the structural equation can be summarized as follows.

(i) Estimate equation (3.5) by Tobit using all N observations. For $logT < logL$, define

$$\hat{v_{1i}} = logT - \boldsymbol{x_i}\hat{\boldsymbol{\alpha}}$$

For $logT \geq logL$, obtain

$$\hat{v_{2i}} = \hat{\tau}\lambda\left(\frac{logL - \boldsymbol{x_i}\hat{\boldsymbol{\alpha}}}{\hat{\tau}}\right)$$

(ii) Using observations for which $logT^* < logC$, estimate $\boldsymbol{\beta}, \rho$ by OLS regression

$$y_i \quad \text{on} \quad \boldsymbol{x_i}, \hat{\eta}_i \tag{3.14}$$

where $\hat{\eta}_i = 1[logT < logL](\hat{v_{1i}} + 1[logT \geq logL]\hat{v_{2i}})'$.

Equation (3.14) produces consistent, $\sqrt{N}$ asymptotically normal estimators of β and ρ under the assumption 1.

The statistic to test censoring bias is just the usual t statistics on $\hat{\eta}_i$ in regression (3.14). If it is statistically insignificant usual variance matrix in regression (3.14) can be used under homoscedasticity, and the robust version is valid under heteroscedasticity. Otherwise, standard errors should be adjusted as described in Theorem 2. In practice, it has been found adjusting for first-step estimators has usually has little effect on the asymptotical standard errors. It is not surprising to find little effect of the adjustment for modest amounts of sample selection since no correction is needed when $\rho = 0$.

3.3 Lung Cancer Study

We apply our procedure on lung cancer treatment cost data. The data set was in our study of inverse probability weighted estimation of censored medical cost data. This data set consists of an inception cohort of 183 lung cancer patients, 48 of whom are subject to administrative censoring, 65 of whom are subject to *artificial* censoring. Seventy of the cases are neither administratively nor artificially censored. The dependent variable was the *log* of total medicare payments two years following diagnosis. The exogenous variables include late disease stage (*lstage*), late comorbid conditions (*lcomorbi*), hospitalization for the reasons other than lung cancer surgery (*hospitalize*), treatment types (*chemotherapy only*, *radiation only*, *chemotherapy and radiation* with the reference group as combination of *surgery only* and *surgery plus adjuvant therapy*), *death*, *symptoms*, *physical functions*, *age*, *sex*(=1 if male), *race* (=1 if white). Table 3.1 shows the summary statistics for the administratively censored, artificially censored and uncensensored groups.

Table 3.2 shows the results of the regression analysis predicting total cost of

Table 3.1: Summary Statistics from the Lung Cancer Study

	Administratively Censored n=48		Artificially Censored n=65		Uncensored n=70	
Variable	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
total cost	62878	40115	63344	44646	64490	39075
lstage	.54 n=26		.52 n=34		.78 n=55	
lcomorbi	.62 n=30		.65 n=42		.64 n=45	
hospitalize	.54 n=26		.46 n=30		.76 n=53	
chemo only	.06 n=3		.06 n=4		.1 n=7	
radiation only	.25 n=12		.20 n=13		.31 n=22	
chemo and radiation	.33 n=16		.31 n=20		.31 n=28	
symptoms	10.12	28.71	10.68	5.19	11.56	5.11
phsical functions	71.46	28.71	75.70	27.08	71.57	26.51
age	72.68	5.21	71.91	4.72	72.01	4.99
sex	.62 n=30		.54 n=35		.61 n=43	
race	.92 n=44		.92 n=60		.94 n=66	

care for the two years following a lung cancer diagnosis. For comparison, we also obtain the estimates using OLS and using IPW procedure in the chapter 1. The results are obtained using the statistical package Stata. The robust standard errors are given in parentheses; no adjustment has been made to account for the generated regressor, $\hat{\eta}$, since there is little evidence of sample selection bias. $H_0 : \rho = 0$ cannot be rejected at even 40 percent significance level for any specification. As we showed in section 2, robust standard errors on all explanatory variables are valid when $\rho = 0$.

We found no sample selection bias with the data set in chapter 1. The results by using Procedure 3 supports that outcome. The same variables, hospitalization for reasons other than lung cancer surgery, chemotherapy only, radiation only, and chemotherapy and radiation reach statistical significance ($p < .05$). The coefficients are closer to OLS estimates relative to IPW estimators.

Hospitalization for reasons other than surgery increases the total medical cost during the period of interest by 109% according to our procedure, 114% according to IPW least squares estimation and 107% according to OLS.

The total medical cost relative to the mean costs for persons receiving surgery only or surgery plus adjuvant therapies decreased for the patients who receive radiation or chemotherapy separately or in combination. The estimates with respect to OLS, IPW least squares and our procedure are: for radiation only, 120%, 105%, 118%, for chemotherapy only 151%, 129%, and 146%, for chemotherapy and radiation, 63%, 54%, 63%.

Age, gender, physical function, stage, comorbid conditions, and race do not have a statistically significant effect in all three estimation methods.

Our procedure explained 14% of the variability in total costs for the two years following diagnosis. This value is somewhat in the middle of the values calculated by OLS (13%) and by IPW least squares (15%).

Table 3.2: Estimates of the Log(tcost) Equation by OLS, IPW and Procedure 3

Explantory Variable	OLS	IPW	Procedure 3
constant	10.74	10.70	10.73
	(1.06)	(1.04)	(1.07)
late stage	.02	-.06	.05
	(.16)	(.16)	(.17)
late comorbidity	.004	-.046	-.001
	(.132)	(.136)	(.131)
hospitalize	.72	.75	.74
	(.18)	(.18)	(.17)
chemotherapy only	-.92	-.83	-.90
	(.29)	(.31)	(.29)
radiation only	-.79	-.73	-.78
	(.23)	(.22)	(.23)
chemothrepay and radiation	-.49	-.43	-.49
	(.22)	(.21)	(.22)
symptoms	.004	.009	.005
	(.014)	(.014)	(.014)
physical functions	.001	.003	.003
	(.002)	(.003)	(.003)
age	-.003	-.005	-.003
	(.013)	(.012)	(.013)
sex	.08	.12	.09
	(.12)	(.12)	(.12)
race	.12	.23	.13
	(.20)	(.19)	(.21)
$\hat{\eta}$			.04
			(.05)
Observations	135	135	135
R-squared	0.13	0.15	0.14

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

3.4 Conclusion

This paper shows a new method for testing and correcting for sample selection bias for cross-section data under the assumption that the selection rule is governed by a censored regression model. The method is easily applicable by using standard software programs. Application the method to censored lung-cancer medical cost data illustrates its simplicity.

Several limitations should be discussed. The first and obvious one is the strong distributional assumption on the selection equation to derive a conditional expectation given a selected sample. It is possible to derive a semi-parametric extension for selection equation. If we write equation (3.3) as

$$E(y|\boldsymbol{x}, T^*, C) = \boldsymbol{x}\boldsymbol{\beta} + h(.), \tag{3.15}$$

where h is an unknown function of sample selection variables, then $h(.)$ can be estimated without specifying distributional assumption. Powell (1987), Robinson (1988), Newey (1988), Coslett (1991) offers different ways of dealing with the presence of the function $h(.)$. The second limitation is that we assume $E(u|v)$ is linear in v. We can also relax this assumption by adding quadratic terms in assumption 1 part (iv), and the formulas can be adjusted accordingly.

It will be useful to extend the methods for a longitudinal data setting. The results in section 3.2 are easily modified to panel a data setting.

APPENDICES

.1 Appendix for Chapter 1

Stata Commands

(i) stset z,failure$(1 - s)$

stset gen $m = k$

(ii) gen p=m if $Z << L$

stbase, at(L)

stset gen $l = k$

replace $p = l$ if $Z >= L$

(iii) gen w=s/p

(iv) gen $ys = sqrt(w) * x$

gen $xs = sqrt(w) * x$

(v) reg ys xs,robust

.2 Appendix for Chapter 3

Proof of Theorem 1 :

Substituting $y_i = \boldsymbol{x_i}\boldsymbol{\beta} + u_i$ into equation (3.2) gives

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} s_i \boldsymbol{x_i'} \boldsymbol{x_i} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i \boldsymbol{x_i'} \boldsymbol{u_i} \right).$$

Under $E(u|\boldsymbol{x}) = 0$, since $s_i \boldsymbol{x_i} = h(\boldsymbol{x_i})\boldsymbol{x_i}$ is just some function of $\boldsymbol{x_i}$:

$$E(s_i \boldsymbol{x_i'} \boldsymbol{u_i}) = E(E(s_i \boldsymbol{x_i'} \boldsymbol{u_i} | \boldsymbol{x_i})) = E(s_i \boldsymbol{x_i'} E(\boldsymbol{u_i} | \boldsymbol{x_i})) = 0$$

With the rank condition and the second moment conditions which is necessary to apply the law of large numbers, consistency follows.

Proof of Theorem 2 :

Substituting $y_i = \boldsymbol{x_i}\boldsymbol{\beta} + \rho\eta_i + e \equiv \boldsymbol{g_i}\Theta + e = \hat{\boldsymbol{g}}_{\boldsymbol{i}}\Theta + \rho(\eta_i - \hat{\eta}_i) + e_i$ in equation (3.12) gives

$$\sqrt{N}(\hat{\Theta} - \Theta) = \left(N^{-1} \sum_{i=1}^{N} s_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}' \hat{\boldsymbol{g}}_{\boldsymbol{i}} \right)^{-1} \left\{ N^{-\frac{1}{2}} \sum_{i=1}^{N} \rho s_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}' (\eta_i - \hat{\eta}_i) + N^{-\frac{1}{2}} \sum_{i=1}^{N} s_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}' e_i \right\} \quad (1)$$

Since each estimator is $\sqrt{N}$-consistency of each estimators for its *plim*, it can be shown that

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} s_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}' e_i = N^{-\frac{1}{2}} \sum_{i=1}^{N} s_i \boldsymbol{g_i'} e_i + op(1) \quad (2)$$

where $\boldsymbol{g_i} = (\boldsymbol{x_i}, \eta_i)$. Also, by an application of the UWLLN (see Newey and McFadden (1994, Lemma 4.3))

$$N^{-1} \sum_{i=1}^{N} s_i \hat{\boldsymbol{g}}_{\boldsymbol{i}}' \hat{\boldsymbol{g}}_{\boldsymbol{i}} \xrightarrow{p} E(si \boldsymbol{g_i'} \boldsymbol{g_i}) \equiv \boldsymbol{A} \quad (3)$$

which is nonsingular by identification assumption that $E(\boldsymbol{g}_i'\boldsymbol{g}_i|s_i = 1)$ has rank $K + 1$.

Since $E(e_i|\boldsymbol{x}_i, \eta_i) = 0$, $\hat{\boldsymbol{g}}_i$ depends on $\boldsymbol{x}_i, T_i$ and with equations (2) and (3), consistency of $\hat{\Theta}$ can be read off from equation (1).

When $\rho \neq 0$, second term at the right hand side of equation (16), contributes to the asymptotic variance of $\hat{\Theta}$. Let $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \tau)$ be a $P \times 1$ vector of unknown parameters where $P = K + 1$. Then $\hat{\boldsymbol{\gamma}}$, a $\sqrt{N}$- asymptotically normal estimator of $\boldsymbol{\gamma}$ has representation;

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = N^{-\frac{1}{2}}\boldsymbol{H}^{-1}\sum_{i=1}^{N} a_i + op(1) \tag{4}$$

where $\boldsymbol{H}$ is the $P \times P$ tobit hessian, a_i is the score of the censored tobit log-likelihood for observation i. The formulas are given in Wooldridge (2002, section 16.4).

From mean value expansion;

$$\hat{n}_i = n_i + \nabla_{\boldsymbol{\gamma}} n_i(\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \tag{5}$$

where $\nabla_{\boldsymbol{\gamma}} n_i(\boldsymbol{\gamma})$ is the $1 \times P$ gradient of $n_i(\boldsymbol{\gamma})$.

By combining equations (2) to (4), it can be shown that

$$\sqrt{N}(\hat{\Theta} - \Theta) \xrightarrow{d} \quad Normal(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}) \tag{6}$$

where $\boldsymbol{B} = Var(\boldsymbol{p}_i) = E(\boldsymbol{p}_i'\boldsymbol{p}_i)$ is defined as $\boldsymbol{p}_i = s_i\boldsymbol{g}_i'e_i + \rho s_i\boldsymbol{g}_i'r_i$ where $r_i = -\nabla_{\boldsymbol{\gamma}} n_i(\boldsymbol{\gamma})\boldsymbol{H}^{-1}\sum_{i=1}^{N} a_i$.

To estimate $Avar(\hat{\Theta}) \equiv \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}/N$, first define

$$\hat{\boldsymbol{A}} \equiv N^{-1}\sum_{i=1}^{N} s_i \hat{\boldsymbol{g}}_i'\hat{\boldsymbol{g}}_i \quad \text{and} \quad \hat{\boldsymbol{B}} \equiv N^{-1}\left(\sum_{i=1}^{N}\left(s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{e}}_i + \hat{\rho}s_i\hat{\boldsymbol{g}}_i'\hat{r}_i\right)'\left(s_i\hat{\boldsymbol{g}}_i'\hat{\boldsymbol{e}}_i + \hat{\rho}s_i\hat{\boldsymbol{g}}_i'\hat{r}_i\right)\right) \quad (7)$$

where $\hat{r}_i = r_i(\hat{\boldsymbol{\gamma}}) = -\nabla_{\boldsymbol{\gamma}}\eta_i(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{H}}^{-1}\sum_i^N \hat{a}_i$, in which gradient of $\hat{\eta}_i(\boldsymbol{\gamma})$, tobit hessian and the score of the censored tobit likelihood for observation i evaluated at $\hat{\boldsymbol{\gamma}}$; $\hat{e}_i = y_i - \boldsymbol{x_i}\hat{\boldsymbol{\beta}} - \hat{\rho}\hat{\eta}_i$ (for $s_i = 1$) and $\hat{g}_i = (\boldsymbol{x_i}, \hat{\eta}_i)$.

The asymptotic variance of $\hat{\boldsymbol{\Theta}}$ is estimated as $Avar(\hat{\Theta}) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1}/N$, and the asymptotic standard errors are obtained as the square roots of the diagonal elements of this matrix. When testing exclusion of the generated regressors, then one can take $\hat{\rho} \equiv 0$.

Bibliography

[1] Anderson, C., K. Anderson and P. Kragu-Sorensen (2000), Cost function estimation: The choice of a model to apply to dementia, Health Economics 9, 397-409.

[2] Bang, H., and A.A. Tsiatis (2000), Estimating Medical Costs Censored Data, Biometrica 87, 329-343.

[3] Cox, D.R. (1972), Regression models and life-tables(with discussions, Journal of the Royal Statistical Society Series 34, 187-220.

[4] Duan, N. (1983), Smearing Estimate: A non-parametric retransformation method, Journal of the American Statistical Association 78, 697-718.

[5] Gold, M.R., J.E. Siegel, L.B. Russel, and M.C. Weinstein (1996), Cost-effectiveness in Health and Medicine. New York: Oxford University Press.

[6] Hausman, J.A., (1978), Specification Tests in Econometrics, Econometrica 46, 1251-1271.

[7] Hirano, K., G.W. Imbens, and G. Ridder (2000), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, mimeo, UCLA Department of Economics.

[8] Horowitz, J.L. and C.F. Manski (1998), Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations, Journal of Econometrics 84, 37-58.

[9] Horvitz, D. and D. Thompson (1952), A Generalization of Sampling without replacement from a Finite Population, Journal of the American Statistical Association 47, 663-685.

[10] Hsiao, C. (1999), Analysis of Panel Data. Cambridge: The University Press.

[11] Kaplan, E.L., P. Meier (1958), Non parametric estimation incomplete observations, Journal of the American Statistical Association 53, 457-481.

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

[13] Lin, D.Y., E.J. Feuer, R. Etzioni and Y. Wax (1997), Estimating Medical Costs from Incomplete follow-up data, Biometrics 53, 419-434.

[14] Lin, D.Y. (2000a), Linear Regression Analysis of Censored Medical Cost, Biostatistics 1, 35-47.

[15] Lin, D.Y. (2000b), Proportional Means Regression for Censored Medical Cost, Biometrics 56, 775-778.

[16] Manning, W.G. and H. Mullahy (2001), Estimating log models: to transform or not to transform?, Journal of Health Economics 20, 461-494.

[17] Newey, W.K. (1988), Two Step Estimation of Sample Selection Models, manuscript, Princeton University Department of Economics.

[18] Newey, W.K., and D. Macfadden (1994), Large Sample Estimation and Hypothesis Testing, in: R.F. Engle and D. Mc Fadden, eds., Handbook of Econometrics, Volume 4 (North-Holland, Amsterdam) 2111-2245.

[19] Newschaffer, C.J., L. Penberthy, C.E. Desch, et al (1996), The effect of age and comorbidity in the treatment of elderly women with nonmetastatic breast cancer, Archives of Internal Medicine 156, 85-90.

[20] Powell, J.L (1987), Semiparametric Estimation of Bivariate Latent Variable Models, Discussion Paper No. 8704, University of Wisconsin Department of Economics.

[21] Robins, J.M., and A. Rotnitzky (1992), Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers, in: H. Jewell, K.Dietz, and V. Farewell eds., AIDS Epidemiology-Methodological Issues, (Boston) 297-331.

[22] Robins, J.M. and A. Rotnitzky (1995), Semiparametric Efficiency in Multivariate Regression Models with Missing Data, Journal of the American Statistical Association 90, 122-129.

[23] Robins, J.M., A. Rotnitzky and L.P. Zhao (1995), Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data, Journal of the American Statistical Association 90, 106-121.

[24] Robinson, P.M. (1988), Root-N Semiparametric Regression, Econometrica 56, 931-954.

[25] Rosenbaum, P.R. (1987), Model-Based Direct Adjustment, Journal of the American Statistical Association 82, 387-394.

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

[27] Ware J., K.K. Snow, M. Kosinski (2000), Health Survey; Manual and Interpretation Guide, Lincoln, RI: Quality Metric, Incorporated.

[28] White, H., A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, Econometrica 48, 817-838.

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

[30] Wooldridge, J.M. (1995), Selection corrections for panel data models under conditional mean independence assumptions, Journal of Econometrics 68, 115-132.

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

[32] Wooldridge, J.M (2000), Introductory Econometrics: A Modern Approach. Cincinnati, OH: South-Western.

[33] Wooldridge, J.M (2001), Asymptotic Properties of weighted M-estimator for standard stratified samples, Econometric Therory 17, 451-470.

[34] Wooldridge, J.M (2002a), Econometric Analysis of cross section and Panel Data, Cambridge, MA:MIT Press.

[35] Wooldridge, J.M (2002b), Inverse Probability Weighted M-Estimators for sample selection, attrition, and stratification, mimeo, Michigan State University.