MODELING APPROACHES FOR COST AND COST-EFFECTIVENESS ESTIMATION USING OBSERVATIONAL DATA

Jiaqi Li

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Nandita Mitra, Associate Professor of Biostatistics

Graduate Group Chairperson

_____

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Scarlett Bellamy, Associate Professor of Biostatistics

Dylan Small, Professor of Statistics

Elizabeth Handorf, Assistant Research Professor at Fox Chase Cancer Center

Justin Bekelman, Associate Professor of Radiation Oncology

MODELING APPROACHES FOR COST AND COST-EFFECTIVENESS ESTIMATION USING OBSERVATIONAL DATA

ACKNOWLEDGEMENT

My deepest gratitude is to my advisor, Dr. Nandita Mitra. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Nandita has always been there since day one of my graduate studies. The guidance I received from her made my Ph.D. purposeful and fulfilling, and I can not imagine having a better advisor and mentor than her. Nandita taught me how to become a researcher, to question thoughts and express ideas and I could not finish this dissertation without her patience and support.

I would like to thank the chair of my dissertation committee Dr. Scarlett Bellamy, for her insightful comments and constructive criticisms at different stages of my research. I would also like to thank my committee members, Dr. Dylan Small and Dr. Justin Bekelman, who have been there since my master thesis. They have given guidance and had patience over the years, taking their time to carefully read and evaluate my research work. My dissertation member, co-author, and "academic sister" Dr. Elizabeth Handorf has helped me in collaboration projects, dissertation research and given me valuable career advice. Finally, my sincere thanks goes to all the faculty, students, staff and alumni of the Biostatistics Department. I am extremely grateful to several faculty members, including Dr. Mary Putt, Dr. Hongzhe Li, Dr. Benjamin French, Dr. Sharon Xie, as well as Edward Kennedy for their engagement and practical advice throughout my Ph.D career.

Lastly, I would like to thank my husband Shiliang Cui, my parents, my family and my best friends Miao Wang and Ayla Jiang, for the unconditional love and support during the past four years. They all kept me going, and this journal of academic exploration would not be possible with their continuous love and encouragement.

ABSTRACT

MODELING APPROACHES FOR COST AND COST-EFFECTIVENESS ESTIMATION USING OBSERVATIONAL DATA

Jiaqi Li

Nandita Mitra

The estimation of treatment effects on medical costs and cost effectiveness measures is complicated by the need to account for non-independent censoring, skewness and the effects of confounders. In this dissertation, we develop several cost and cost-effectiveness tools that account for these issues. Since medical costs are often collected from observational claims data, we investigate propensity score methods such as covariate adjustment, stratification, inverse probability weighting and doubly robust weighting. We also propose several doubly robust estimators for common cost effectiveness measures. Lastly, we explore the role of big data tools and machine learning algorithms in cost estimation. We show how these modern techniques can be applied to big data manipulation, cost prediction and dimension reduction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Proper medical cost and cost-effectiveness estimation is critical for health economics evaluation and decision-making. We are often most interested in the effect of a new treatment on cost and cost-effectiveness compared to the existing treatment. The gold standard for estimating the treatment effect is a randomized controlled trial. However, it is often not feasible or ethical to carry out trials just for the purpose of collecting medical cost data. Hence, we need to obtain cost and cost-effectiveness measures from observational sources.

1.1. Background

The development of methods for the analysis of cost and cost-effectiveness data has been of interest to many statisticians and health economists. In this section, we review several popular methods that handle the unique features of cost data, including informative censoring, heteroscadasticity, skewness and zero costs. We also provide a review of common cost effectiveness measures. Since observational data play an important role in cost and cost-effectiveness estimation, we review several common propensity score models that handle data from observational sources. Lastly, we give some background on big data and machine learning tools that could be applied to cost estimation.

1.1.1. Background about cost estimations

Medical costs often have very different distributions depending on such the disease and treatment setting. To handle distributional skewness and structural zeros, economists and statisticians have developed two broadly categorized methods:, single equation and multiple equation models. Single equation methods include ordinary least squares regression, generalized linear regression, parametric models (e.g. Weibull, Gamma) with different transformations (e.g. log, Box-Cox) and different variance functions. Some (Lumley et al., 2002) argue that with a large enough sample size ($n > 500$), linear regression and t-tests are appropriate for analysis of highly skewed outcomes, including costs. Mihaylova et al. (2011) carried out a study to review some popular methods for analyzing cost data, and recommended using a simple method such as linear regression that assumes a normal distribution for large samples. Others (O'Hagan and Stevens, 2003) disagree

and advise against ignoring skewness of cost data because variance estimates based on normal approximations may be biased.

Historically, the natural logarithm of costs in ordinary least squares regression (OLS) or generalized linear model (GLM) with log link have been used. However, Manning and Mullahy (2001) found OLS estimators can be biased under heteroscadasticity. Even though GLM estimators are consistent, they can yield imprecise estimates if the log-scale error is heavy-tailed. Manning, Basu, and Mullahy (2005) evaluated OLS, OLS for log cost, standard gamma model and exponential with a log link, and the Weibull model and found that a generalized gamma distribution was the most robust. Several GLM based estimators have also been proposed. For example, Basu and Rathouz (2005) proposed GLM using box cox transformation and parametric models for the variance as a function of the mean. Others have suggested using median regression since the median is less sensitive to skewness and outliers. Bang and Tsiatis (2002); Ying, Jung, and Wei (1995) extended median regression to incorporate simple weights to handle censored cost data. Dodd et al. (2006) compared normal and bootstrapped multiple linear regression, median regression, gamma model with the log link and OLS of log costs. They found that GLM with log link and gamma variance provided the best fit. Lastly, Basu, Manning, and Mullahy (2004) conducted a similar comparison and applied popular medical cost models to different cost data structures and arrived at the same conclusion - GLM with log link and gamma variance is the most robust model.

In addition, multiple equation models have focused on different components of costs such as zero and non-zero costs, and inpatient and outpatient billings (Duan et al., 1983; Leung and Yu, 1996). The rationale behind these multiple equation models are that costs accrued at different times of a patient's history follow different distributions and thus should be modeled differently.

One of the biggest issues present in most cost data is informative censoring due to the lack of a common rate of cost accrual over time among patients. To handle non-ignorable censoring, the popular approaches are either weighting-based (Bang and Tsiatis, 2000; Lin et al., 1997) or survival model based. The latter is less popular; Etzioni et al. (1999) demonstrated that standard survival techniques yield biased estimates. Lin et al. (1997) proposed a non-parametric approach that splits the time period into small intervals and weights mean costs from each interval by survival probabilities estimated from the Kaplan-Meier curve. Lin's method is only consistent if the partitions are chosen so that censored observation and interval boundaries coincide. To overcome this issue,

(Bang and Tsiatis, 2000) proposed two popular methods: simple weighted and partitioned estimators, to estimate mean medical cost. The simple weighted method averages subjects with complete cost information weighted by the probability of not being censored. The The partitioned estimator builds on the same weighting idea but makes use of cost history information and is hence more efficient. Many (Raikou and McGuire, 2004; Zhao, Cheng, and Bang, 2011; Zhao et al., 2007) have studied the properties of these two popular methods. Baser et al. (2004); Lin (2000, 2003) subsequently extended these methods to regression of censored cost. Recent work (Basu and Manning, 2010; Tian and Huang, 2007) has focused on two part models to accommodate significant zeroes, end of life cost and skewness properties of medical cost data.

1.1.2. Background : cost effectiveness estimation

Cost effectiveness (CE) analysis is often used to evaluate the merits of a new health-care intervention (treatment, $Z = 1$) compared to an existing one (control, $Z = 0$). CE measures integrate estimates of costs and effectiveness in a single statistic derived from two components: Δ_E and Δ_C where Δ_E =Effectiveness$_{Z=1}$−Effectiveness$_{Z=0}$ and Δ_C =Cost$_{Z=1}$−Cost$_{Z=0}$.

One common approach to combine the cost and effect outcomes to form Incremental Cost- Effectiveness Ratio (ICER). However, a major limitation of the ICER is its discontinuity when the denominator Δ_E approaches zero. Another issue is that ICER often has an unstable interpretation; when the Δ_E is positive and Δ_C is negative, ICER has a different interpretation than when Δ_C is positive and Δ_E is negative. In addition, estimating the variance of ICER is problematic due to the acknowledged statistical problems associated with ratio statistics. Non-parametric bootstrapping, Fieller's theorem and Bayesian approaches (Heitjan, Moskowitz, and Whang, 1999; Polsky et al., 1997; Willan and O'Brien, 1996) can be applied to estimate the variance of ICER.

Recently, health economists have advocated the use of the Net Monetary Benefit (NMB): NMB $(\lambda) = \lambda\Delta_E - \Delta_C$. NMB is a linear combination of Δ_C and Δ_E and measures the excess benefit given a fixed level of λ. λ is defined as willingness to pay (WTP), which is the maximal monetary value decision-makers are willing to pay for a unit of Δ_E. Typically, λ measures the dollar amount one is willing to pay for one year of additional life. The NMB does not suffer from the singularity problem that the ICER does. Moreover, the interpretation of NMB is straightforward: a positive NMB means a new treatment is more cost-effectives and a negative NMB means a new treat-

ment is less cost-effective compared to the control. It is also easy to estimate its variance as $var(\text{NMB}(\lambda)) = \lambda^2 var(\Delta_E) + var(\Delta_C) - 2\lambda cov(\Delta_E, \Delta_C)$.

A CE acceptability curve builds on the idea underlying the NMB and displays the probability that the treatment is cost-effective (NMB> 0) compared with the control for a range of λ values. To plot the CE acceptability curve, we use bootstrapping to estimate $Pr(\lambda\Delta_E - \Delta_C > 0)$. In practice, we simply count the proportion of bootstrapped samples that yields $\lambda\Delta_E - \Delta_C > 0$ for a range of λ values.

Similar to cost analysis, CE estimation needs to account for informative censoring in cost data. Willan et al. (2002) proposed NMB and ICER estimation methods accounting for censoring in cost data in the setting of randomized trials. A similar approach was taken to extend to developing a linear regression method that accommodates censored outcomes (Willan, Lin, and Manca, 2005).

1.1.3. Background: propensity score methods

Heath care cost information are often collected from observational claims data, thus one must be careful when dealing with potential confounding. The propensity score (PS), first introduced by Rosenbaum and Rubin (1983) is commonly employed to adjusted for confounding in observational studies (Austin, 2011). Propensity scores are often used in covariate adjustment, matching, stratification, and weighting (Lunceford and Davidian, 2004; Rosenbaum, 1987). Covariate adjustment on PS is easy to use but assumes that the relationship between the propensity score and the outcome has been correctly modeled (Austin and Mamdani, 2006), which could be a challenge in cost estimation since there is no one-size-fits-all cost model. Rosenbaum (1987) first introduced inverse probability of treatment weighting (IPTW) while Lunceford and Davidian (2004) presents two other types of weights. Robins, Rotnitzky, and Zhao (1994) provided the theory behind a broader class of weighted estimators. Matching is commonly used in practice where we match subjects in treatment and control groups according to their estimated propensity scores. Stratification based on the quintiles of the PS eliminates approximately 90% of bias due to measured confounders (Cochran, 1968; Rubin and Rosenbaum, 1984). Many have compared the relative performance of these various methods (Austin, Grootendorst, and Anderson, 2007; Austin and Mamdani, 2006; Lunceford and Davidian, 2004). As Rubin (2004) notes, covariate adjustment using PS and IPTW are more sensitive to whether the PS has been accurately estimated.

Recently, a new PS method, doubly robust (DR) estimation based on Robins, Rotnitzky, and Zhao (1994) has become popular. It has the smallest large sample variance among the class of weighted estimators. DR estimation combines outcome regression (regression model) with weighting by PS (PS model) such that it is robust to misspecification of one (but not both) of these models (Bang and Robins, 2005; Tsiatis and Davidian, 2007). The doubly robust property is appealing but can result in biased estimates if both the outcome model and PS model are misspecified (Funk et al., 2011).

1.1.4. Background : big data and machine learning methods

With the recent availability of big data, the role of machine learning in economics has become increasingly important (Varian, 2014). Traditionally cost data have been be stored and manipulated on spreadsheets or by a Structured Query Language (SQL). However, these tools are inadequate for massive data which require special programing paradigms. . Some popular big data storage and manipulation algorithms include Hadoop File Distribution System (Lam, 2010; Shvachko et al., 2010; Venner, 2009) and MapReduce (Dean and Ghemawat, 2008).

Cost prediction has been of great interest to health economists (Folland, Goodman, Stano, et al., 2007). Traditionally, parametric models have been used for cost prediction. In recent studies, machine learning non-parametric algorithms has been shown to have better predictive abilities (Kim, An, and Kang, 2004; Sushmita et al., 2015). Bertsimas et al. (2008) compared the performance of classification trees, clustering, and traditional models for cost bucket estimation, and found that data-mining methods provide more accurate predictions. Popular machine learning prediction algorithms include classification and regression trees, random forests, support vector machines, boosting and Bayesian additive regression trees.

In addition, variable selection can be challenging in cost estimation models. We often have many potential predictors that may need to be narrowed down for model building. Traditionally, researchers use stepwise regression and model complexity measures such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to select important variables. With the rise of big data, we see more and more large datasets with numerous potential predictors where modern dimension reduction methods may serve as important tools in estimating cost. Popular dimension reduction tools such as principle components analysis and Lasso combine the strength of statistical

modeling with machine learning.

1.2. Novel developments

In this dissertation, we develop statistical methods for cost and cost-effectiveness estimation from observational data. The dissertation consist of three parts. In Chapter 2, we investigate propensity score (PS) based methods such as covariate adjustment, stratification and inverse probability weighting taking into account informative censoring of the cost outcome. We compare these more commonly used methods to doubly robust weighting. We then use a machine learning approach called Super-Learner (SL) to 1) choose among conventional regression models to estimate mean models in the DR approach and 2) choose among various covariate specifications for PS estimation. Our simulation studies show that when the PS model is correctly specified, weighting and DR perform well. When the PS model is misspecified, the combined approach of DR with Super Learner can still provide unbiased estimates. SL is especially useful when the underlying cost distribution comes from a mixture of different distributions or when the true PS model is unknown. We apply these approaches to a cost analysis of two bladder cancer treatments, cystectomy versus bladder preservation therapy, using SEER-Medicare data.

In Chapter 3, we propose using separate doubly robust (DR) methods based on propensity scores for estimating CE with and without incorporating cost history and show they are unbiased. We then use cross validation to choose among popular cost models to estimate regression parameters in the DR approach and to choose among various parametric and non-parametric propensity score models. Our simulation studies demonstrate that the proposed DR models perform well even under misspecification of either the PS model or the outcome model. We apply these approaches to a cost-effectiveness analysis of two competing lung cancer surveillance procedures, CT versus chest X-ray, using SEER-Medicare data.

In Chapter 4, we review and explore the use of big data and machine learning techniques in health care cost estimation. Specifically, we look at three areas: big data manipulation, cost prediction and variable dimension reduction. Massive health care cost data calls for the use of big data storage and manipulation algorithms like Hadoop and MapReduce. Traditionally, the focus of cost prediction has been on how to come up with the best parametric model to predict cost. With the rise of modern machine learning techniques, we can use non-parametric prediction models such as

classification and regression trees, random forest, Bayesian adaptive regression trees, and support vector machines. Moreover, popular dimension reduction tools such as LASSO and principle components analysis combine the strengths of statistical modeling with machine learning, and allow us to identify important covariates affecting one's health care cost and to build parsimonious cost models. We demonstrate the use of these state-of-the-art big data and machine learning models using a cohort of lung cancer patients derived from SEER-Medicare.

CHAPTER 2

PROPENSITY SCORE AND DOUBLY ROBUST METHODS FOR ESTIMATING THE EFFECT OF TREATMENT ON CENSORED COST

2.1. Introduction

Proper medical cost estimation is imperative to health economics evaluation and decision-making. Policy makers are often most interested in the average effect treatment effect (ATE) on total costs. Since medical costs are often collected from claims data which are susceptible to confounding, appropriate estimation of the ATE from observational data demands attention. These methods must also account for other complicating features of cost data including informative censoring and skewness.

The primary focus of earlier studies of cost estimation has been on methods for dealing with their distributional skewness. Historically, researchers have used natural logarithm transformed costs in ordinary least square regression (OLS) or used generalized linear models (GLM) with a log link. However, Manning and Mullahy Manning and Mullahy (2001) showed that OLS estimators can be biased under heteroscadasticity and GLM estimators can yield imprecise estimates if the log-scale error is heavy-tailed. Others have suggested using median regression since the median is less sensitive to skewness and outliers (Manning, Basu, and Mullahy, 2005). Several studies (Basu, Manning, and Mullahy, 2004; Basu and Rathouz, 2005; Dodd et al., 2006) have evaluated additional approaches such as OLS, OLS for log cost, standard gamma, standard GLM, generalized gamma, median regression, exponential models with log link, and the weibull model. Dodd et al. (2006) found the generalized gamma model to be the most robust cost model. Recent works (Basu and Manning, 2010; Tian and Huang, 2007) have focused on two part models and Bayesian approaches to accommodate structural zeros and end of life costs.

An important feature of medical costs is censoring, which often occurs if the study terminates after a fixed follow-up period. Even though survival time is non-informatively censored due to end-of-study censoring, cost is not. Censoring in cost is informative since the rate of cost accrual over time may vary greatly among patients. To address this issue, Lin et al. (1997) introduced two estimators

of mean cost by partitioning study period into subintervals and assuming censoring occurs only at the boundaries of these subintervals. Bang and Tsiatis (2000) improved on Lin et al.'s work and proposed two popular methods: the simple weighted method and the partitioned method, to estimate mean medical cost under informative censoring. The simple weighted method averages subjects with complete cost information weighted by the inverse of the probability of not being censored. The partitioned estimator builds on the same weighting idea but also makes use of cost history information and is therefore more efficient. Properties of these methods have been widely studied (Raikou and McGuire, 2004; Zhao, Cheng, and Bang, 2011; Zhao et al., 2007) . Baser et al. (2004); Lin (2000, 2003) have since extended these methods to linear regression and general linear models to incorporate the effect of covariates. Several studies (Bang and Tsiatis, 2002; Ying, Jung, and Wei, 1995) have also applied these techniques to median regression to handle censored cost data.

Heath care cost information is often collected from observational sources, such as Medicare, necessitating the need to adjust for potential confounders. The propensity score (PS), first introduced by Rosenbaum and Rubin (1983) is commonly employed to adjust for confounding in observational studies (Austin, 2011). Propensity scores are often used in covariate adjustment, matching, stratification and weighting (Lunceford and Davidian, 2004; Rosenbaum, 1987). Covariate adjustment of the PS is easily implemented but is sensitive to the assumption that the relationship between the propensity score and the outcome has been correctly modeled (Austin and Mamdani, 2006). Stratification based on PS is also often used as it greatly simplifies implementation over standard methods;Rubin and Rosenbaum (1984) demonstrated that stratification based on the quintiles of the PS eliminates approximately 90% of bias due to measured confounders. More recently, inverse probability of treatment weighting (IPTW) (Rosenbaum, 1987) has become the method of choice. The normalized version of IPTW has been proposed (Busso, DiNardo, and McCrary, 2014; Hirano, Imbens, and Ridder, 2003) which belongs to a broader class of weighted estimators described by Robins, Rotnitzky, and Zhao (1994). Several studies (Austin, Grootendorst, and Anderson, 2007; Lunceford and Davidian, 2004) compared the relative performance of these methods. Covariate adjustment using PS and IPTW has been shown to be more sensitive to whether the PS has been accurately estimated (Austin and Mamdani, 2006; Rubin, 2004).

In this study, we investigate doubly robust (DR) estimation of cost and compare it to more conven-

tional propensity score based approaches. DR estimation combines outcome regression (regression model) with weighting by PS (PS model) such that it is robust to misspecification of one (but not both) of these models (Bang and Robins, 2005; Tsiatis and Davidian, 2007). Lunceford and Davidian (2004) demonstrated that the DR estimator performs better than stratification and IPTW. The doubly robust property is appealing but can still lead to biased estimates if both the regression model and the PS model are misspecified (Funk et al., 2011). When using the DR method, the biggest challenge is to accurately model cost in the regression model. Given the heterogeneous nature of cost distributions and the many possible choices of cost models described above, we propose using an ensemble machine learning approach that relies on V-fold cross validation called Super Learner (SL) (Laan, Polley, and Hubbard, 2007). Using SL, we can incorporate various potential cost models and obtain asymptotically optimal prediction. Moreover, although logistic regression is the most commonly used method for estimating the PS; we can use SL to obtain PS estimates from other potential non-parametric PS models or PS models with different functional forms.

The goal of this study is to develop appropriate PS methods for estimating skewed and censored cost data. In the current literature, Basu, Polsky, and Manning (2011) have discussed several methods for estimating the ATE on health care costs. Anstrom and Tsiatis (2001) have proposed on normalized IPTW for censored cost. We extend this literature by considering PS methods on censored cost. We begin by reviewing some of the existing cost estimation methods and then examine PS covariate adjustment, stratification and weighted approaches. We follow by discuss DR and the application of SL in cost estimation. We provide results from simulation studies that compare the performance of these estimators, and we also highlight the effect of PS mis-specification on treatment effect estimation and demonstrate the merits of SL. Finally, we apply these PS approaches to a cost analysis of two competing bladder cancer treatments, cystectomy versus bladder preservation therapy, using costs derived from SEER-Medicare data.

2.2. Cost estimation - existing methods

Cost estimation has been a great interest in the health economics literature. In this section we give some brief background on existing methods. We are interested in estimation cost up to time L. We define $Y_i(u)$ to be the known accumulated cost up to time u and Y_i is the total cost that subject i

accrues up to L. Let t_i and C_i denote an individual's survival time and censoring time in the duration of interest respectively. Hence the random variable t is bounded by L. L can be considered as a large number such as 100 if we are interested in life time cost. The observables are given by:

$$T_i = \min(t_i, C_i), \text{ time to event or censoring}$$

$$\delta_i = I(t_i \leq C_i), \text{ complete case indicator}$$

$$Y_i = Y_i(t_i), \text{ total cost observed only if } \delta_i = 1$$

We only observe Y_i for the uncensored subjects. For censored subjects, their cost is still accruing hence their total cost Y_i is unknown. in standard survival analysis we say censoring is non-informative if $t \perp\!\!\!\perp C$. In total cost estimation Y is not non-informatively censored since $Y(t) \perp\!\!\!\perp Y(C)$ does not hold. In practice, a patient with high cost at the time of censoring, $Y(C)$, is also likely to have high cost at the time of event $Y(t)$ as that patients may likely have higher cost accrual rate. Hence, censoring of cost is not non-informative and standard survival techniques do not apply. Now, let $K(u) = Pr(C \geq u)$ be the probability of not being censored at time u. $K(u)$ can be estimated from either parametric or non-parametric models. For instance, we can assume a parametric survival model such as an exponential or weibull and estimate $K(u)$ based on maximal likelihood methods. Another approach is to use the Kaplan-Meier estimates $\hat{K}(u)$, based on the data $(\boldsymbol{T}, 1 - \boldsymbol{\delta})$.

Economists and policy makers are often most interested in $E(Y)$. We describe two popular existing methods to estimate $E(Y)$ assuming individual cost history data are not recorded, i.e. only cost at event or censoring time $Y_i(T_i)$ is observed while $Y_i(u), u < T_i$ is unobserved. To estimate mean total cost $E(Y)$, Lin et al. (1997) proposed to partition the study period $(0, L)$ into K subintervals and then "sum up" the cost contribution from subjects who died in each interval. Their method assumes that censoring only occurs at the boundaries of the subintervals. To overcome this limitation, Bang and Tsiatis (2000) propose using cost information from uncensored subjects and then weighting each complete cost observation by the inverse of the probability of not being censored, which is evaluated at the time of the subject's death:

$$\widehat{E(Y)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i Y_i}{\hat{K}(T_i)}$$

This weighted estimator is unbiased as $E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{\hat{K}(T_i)}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\delta_i Y_i}{\hat{K}(T_i)}\middle| T_i\right]\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\hat{K}(T_i)}[E(I(C_i \geq T_i)|T_i]\right] = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = E(Y)$. This estimator is also shown to be consistent regardless of the censoring pattern (Bang and Tsiatis, 2000). Intuitively, a subject that is observed to die at T_i represents $\frac{1}{K(T_i)}$ subjects who would have been observed if there were no censoring.

Lin (2000) also applied the same weighting technique to model the linear relationship between total cost and other covariates $\mathbf{X}$ as $Y = \beta'\mathbf{X}$, when total cost is subjected to informative censoring. If there were no censoring, the least square normal equation can be simply written as$\sum_{i=1}^{n}(Y_i - \beta'\mathbf{X}_i)\mathbf{X}_i = 0$. However, to account for censoring, Lin applied the same weighting idea and modified the above equation as follows:

$$\sum_{i=1}^{n}\frac{\delta_i}{K(T_i)}(Y_i - \beta'\mathbf{X}_i)\mathbf{X}_i = 0$$

This weighting method can also be applied to other regression models such as GLM or median regression as discussed by Lin (2003) and Bang and Tsiatis (2002).

2.3. Propensity score approaches

Cost information is often collected from observational databases which are subjected to confounding, here we develop propensity score approach to modeling censored cost data. Let Z be an indicator of the treatment exposure: $Z = 1$ if treated, $Z = 0$ if control. We adopt the counterfactual framework described by Rubin (1974) and define $Y_i^{(0)}$ to be the total cost of subject i if he were in the control group. Similarly, $Y_i^{(1)}$ is the total cost if the patient had received treatment. Also, let $t_i^{(0)}$ and $t_i^{(1)}$ denote the survival time if the patient were in the control and treatment group respectively.

Although we are most interested in total cost Y, we want to consider both Y and survival time t as Y is dependent on t. We extend the usual assumption of strong ignorability to include both time and total cost as follows

$$(Y^{(0)}, Y^{(1)}, t^{(0)}, t^{(1)}) \perp\!\!\!\perp Z|\boldsymbol{X} \tag{2.1}$$

We also modify the assumption of non-informative censoring to state:

$$C \perp\!\!\!\perp (Y^{(0)}, Y^{(1)}, t^{(0)}, t^{(1)}, \boldsymbol{X})|(, Z) \tag{2.2}$$

In other words, we assume censoring time to be independent of potential failure time and cost outcomes as well of other confounders conditional on covariates and treatment assignment. This assumption is valid for end-of-study and other administrative censoring commonly seen in cost studies; and was first formally introduced by Anstrom and Tsiatis Anstrom and Tsiatis, 2001.

Moreover, let μ be the average causal treatment effect on cost adjusted for covariates $\mathbf{X}$. We use μ_1 and μ_0 to represent $E(Y^{(1)})$ and $E(Y^{(0)})$ respectively. Therefore μ can be defined as:

$$\mu = \mu_1 - \mu_0 = E(Y^{(1)}) - E(Y^{(0)}) \tag{2.3}$$

Further, $K_z(u) = P(C \geq u|Z = z)$ and must be estimated separately for the treatment and control groups since they may have different survival trajectories. For simplicity, we use $\hat{K}(u)$ to denote the treatment-specific estimated probability of being uncensored at time u, $\hat{K}_z(u)$.

Our goal is to estimate μ from observational data utilizing propensity score methods. We extend popular propensity score approaches to handle censored cost data. We also provide general step-by-step guidelines for the proposed methods. First, we need to estimate propensity scores $e(\boldsymbol{X}) = Pr(Z = 1|\boldsymbol{X})$. It is routine to estimate propensity scores from $(\boldsymbol{Z}, \mathbf{X})$ using a logistic regression model:

$$e(\boldsymbol{X}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\boldsymbol{X}\boldsymbol{\beta})} \tag{2.4}$$

For simplicity, we write $e_i = e(\boldsymbol{X}_i, \boldsymbol{\beta})$ and $e_\beta = \boldsymbol{\partial} e_i / \boldsymbol{\partial \beta}$. Moreover, $\boldsymbol{\beta}$ can be estimated using the maximum likelihood method by solving:

$$\sum_{i=1}^{n} \psi(Z_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Z_i - e_i}{e_i(1 - e_i)} e_\beta = \mathbf{0} \tag{2.5}$$

Estimated propensity scores $\hat{e}_i$ can be predicted from the logistic regression model in Equation 2.4.

2.3.1. Covariate Adjustment

In the covariate adjustment approach, the outcome variables $\mathbf{Y}$ is regressed on $\mathbf{Z}$ along with the estimated propensity score $\hat{\boldsymbol{e}}$, and any additional covariates (subset of $\mathbf{X}$). Using an extension of the OLS model described by Lin (2000), we impose the simple weights $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$ to account for censoring in costs. The choice of regression model depends on the nature of the outcome $\boldsymbol{Y}$. Here

we present three popular options:

Normal model The simplest method is a standard linear regression, which assumes that the total cost $\boldsymbol{Y}$ follows a normal distribution, something unlikely to happen in practice. We regress $\mathbf{Y}$ on $\mathbf{Z}$ and $\hat{\boldsymbol{e}}$ weighted by $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$:

$$E(Y_i|Z_i, \boldsymbol{X}_i) = \beta_0 + \beta_1 Z_i + \beta_2 \hat{e}_i \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \tag{2.6}$$

Hence,

$$\hat{\mu}_{ca1} = \hat{\beta}_1 \tag{2.7}$$

Lognormal model This is similar to the linear regression model, except the outcome is transformed using the natural logarithm. This is a popular approach in health economics, as cost is transformed to reduce its skewness. The main shortcoming of this approach is that the analysis does not result in a model for μ in the original scale. Re-transformation to the original scale of interest is problematic (Manning and Mullahy, 2001) especially in the presence of heteroscedasticity. Nevertheless, log transformation of the response variable followed by OLS is still common. Assuming log-scale errors that are normally distributed with mean zero and common variance σ^2, we regress $\log(\boldsymbol{Y})$ on $\mathbf{Z}$ and $\hat{\boldsymbol{e}}$ weighted by $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$.

$$E(\log(Y_i)|Z_i, \boldsymbol{X}_i) = (\beta_0 + \beta_1 Z_i + \beta_2 \hat{e}_i) \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \tag{2.8}$$

Hence,

$$\hat{\mu}_{ca2} = \sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 \hat{e}_i + \hat{\sigma}^2/2) - \exp(\hat{\beta}_0 + \hat{\beta}_2 \hat{e}_i + \hat{\sigma}^2/2) \tag{2.9}$$

Gamma model The gamma distribution has a raw-scale variance function that is proportional to the square of the raw-scale mean function (Equation 2.10), an attribute common to many health applications. To implement this, we regress $\boldsymbol{Y}$ on $\mathbf{Z}$ and $\hat{\boldsymbol{e}}$ in a GLM model weighted by $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$, and specify the variance family to be gamma.

$$E(Y_i|Z_i, \boldsymbol{X}_i) = \exp(\beta_0 + \beta_1 Z_i + \beta_2 \hat{e}_i) \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \tag{2.10}$$

$$\text{and } Var(Y_i|Z_i, \boldsymbol{X}_i) \propto [E(Y_i|Z_i, \boldsymbol{X}_i)]^2 \tag{2.11}$$

Hence,

$$\hat{\mu}_{ca3} = \sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 \hat{e}_i) - \exp(\hat{\beta}_0 + \hat{\beta}_2 \hat{e}_i) \tag{2.12}$$

The variance of $\hat{\mu}$ from covariance adjustment methods can be obtained in several ways. Analytically, the estimated variance of $\hat{\mu}_{ca1}$ equals the variance of $\hat{\beta}_1$ estimated from Equation 2.6. The variances of $\hat{\mu}_{ca2}$ and $\hat{\mu}_{ca3}$ can be derived using the delta method on Equation 2.8 and Equation 2.10. We can also use non-parametric bootstrapping to estimate the variances of $\hat{\mu}_{ca1}$, $\hat{\mu}_{ca2}$ and $\hat{\mu}_{ca3}$.

2.3.2. Stratification

In stratification, subjects are first ranked and stratified into S mutually exclusive subsets based on $\hat{e}_i$. If balance between treatment groups is achieved within each stratum, we can estimate μ by a weighted sum of the difference of sample means of Y_i across strata. Simple weights are imposed to account for informative censoring:

$$\hat{\mu}_s = \sum_{s=1}^{S} \sum_{i=1}^{n} \frac{Y_i Z_i I(\hat{e}_i \in \hat{Q}_s)}{n_{1s}} \times \frac{\delta_i}{\hat{K}_{s1}(T_i)} - \frac{Y_i (1 - Z_i) I(\hat{e}_i \in \hat{Q}_s)}{n_{0s}} \times \frac{\delta_i}{\hat{K}_{s0}(T_i)} \tag{2.13}$$

where Q_s is the sth sample quantile of $\hat{e}$, n_{zs} is the total number of subjects with $Z_i = z$. Here, $\hat{K}_{s0}(T_i)$ denotes the estimated probability of uncensoring for treated subjects in stratum s and $\hat{K}_{s1}(T_i)$ the estimated probability of uncensoring for control subjects in stratum s. Within each stratum, subjects have roughly similar values of the propensity scores. Loosely speaking, we treat S strata as S different independent groups. Therefore, $\hat{K}(T_i)$ needs to be estimated separately for subjects in stratum s and treatment group z.

Notice that δ_i may be correlated with Z_i since subjects on treatment may live longer; hence we are less likely to observe their complete cost information and δ_i is more likely to be zero. However, consistency of $\hat{\mu}$ is still valid. Consistency follows from the fact that $E(\delta_i / \hat{G}(T_i)) = 1$, $Var\left(\frac{\delta_i}{\hat{G}(T_i)}\right)$ is bounded, total cost is bounded (see Appendix 1 of Bang and Tsiatis (2000) for details) and the unbiasedness property of stratification method (Lunceford and Davidian, 2004).

Lunceford and Davidian (2004) recommended approximating the empirical variance by treating $\hat{\mu}$ as the average of S independent, within-stratum, treatment effect estimates. If we further assume

independence of δ_i and Z_i, we have

$$\widehat{Var(\hat{\mu}_s)} = \frac{1}{S^2}\sum_{s=1}^{S}\frac{s_{1j}^2}{n_{1s}} + \frac{s_{0j}^2}{n_{0s}}$$

where s_{1j}^2 and s_{0j}^2 are the sample variance of Y_i for treated and control subjects in stratum s weighted by $\delta_i/\hat{K}(T_i)$. In real life settings, it is unlikely that δ_i is independent of Z_i. Hence, the formula above only serves as a "quick and dirty" variance estimate. In this case, it is preferably to obtain the variance of $\hat{\mu}_s$ via bootstrapping (Jiang and Zhou, 2004).

2.3.3. Weighted approaches

Weighted estimators were first introduced by citetHorvitz1952 and were extended to propensity scores by Rosenbaum (1987). There are many different weight choices; the most popular being the inverse probability of treatment weights (IPTW). IPTW are defined as $w_i = \frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i}$, so that a subject's weight is equal to the inverse of the probability of receiving the treatment the subject was actually given. Again, simple weights $\frac{\delta_i}{\hat{K}(T_i)}$ are applied to account for informative censoring.

$$\hat{\mu}_{iptw1} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{\hat{e_i}} \times \frac{\delta_i}{\hat{K}(T_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}_i} \times \frac{\delta_i}{\hat{K}(T_i)} \tag{2.14}$$

Another popular weight choice is the normalized version of IPTW (Busso, DiNardo, and McCrary, 2014; Hirano, Imbens, and Ridder, 2003), which follows from $E\left(\frac{Z}{e}\right) = E\left(\frac{E(Z|\boldsymbol{X})}{e}\right) = 1$, $E\left(\frac{1-Z}{1-e}\right) = 1$ and the estimating equations $\sum_{i=1}^{n}\frac{Z_i}{\hat{e}_i}\frac{\delta_i}{\hat{K}(T_i)}(Y_i - \mu_1) = 0$, $\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{e}_i}\frac{\delta_i}{\hat{K}(T_i)}(Y_i - \mu_0) = 0$.

$$\hat{\mu}_{iptw2} = \left(\sum_{i=1}^{n}\frac{Z_i}{\hat{e}_i}\frac{\delta_i}{\hat{K}(T_i)}\right)^{-1}\sum_{i=1}^{n}\frac{Z_iY_i}{\hat{e}_i} \times \frac{\delta_i}{\hat{K}(T_i)} - \left(\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{e}_i}\frac{\delta_i}{\hat{K}(T_i)}\right)^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i} \times \frac{\delta_i}{\hat{K}(T_i)} \tag{2.15}$$

As above δ_i may be correlated with Z_i but the consistency of $\hat{\mu}$ is still valid. Consistency of $\hat{\mu}_{iptw1}$ and $\hat{\mu}_{iptw2}$ can also be demonstrated using M estimation.

The variance of $\hat{\mu}_{iptw1}$ and $\hat{\mu}_{iptw2}$ can be obtained in several ways. One option is to use non-parametric bootstrapping. In addition, Anstrom and Tsiatis (2001) derived the analytic form for the variance of $\hat{\mu}_{iptw2}$ when $K(T_i)$ is estimated using the KM method. Similar methods can be used to derive the analytic variance of $\hat{\mu}_{iptw1}$. If $K(T_i)$ is estimated using parametric models, we can use

M-estimation to derive $var(\hat{\mu})$ in Equation 2.14 and 2.15. Here we give the sketch of the derivation when survival time t_i follows an exponential distribution $\exp(\lambda)$:

λ can be estimated using the maximal likelihood $L(\lambda) = \prod_{i=1}^{n}[\lambda \exp(-\lambda T_i)]^{\delta_i}[\exp(-\lambda T_i)]^{1-\delta_i}$. And thus $\hat{\lambda}_{mle} = \frac{\sum_{i=1}^{n}\delta_i}{\sum_{i=1}^{n}T_i}$. Together with Equation 2.5 and Equation 2.14 we have the following estimating equations:

$$\Psi = \sum_{i=1}^{n}\begin{pmatrix}\psi_1\\ \psi_2\\ \psi_3\end{pmatrix} = \sum_{i=1}^{n}\begin{pmatrix}\left(\frac{Z_iY_i}{e_i} - \frac{(1-Z_i)Y_i}{1-e_i}\right)\left(\frac{\delta_i}{K(T_i)}\right) - \mu\\ \frac{Z_i-e_i}{e_i(1-e_i)}e_\beta\\ \delta_i - \lambda T_i\end{pmatrix} = \mathbf{0} \tag{2.16}$$

Using the general framework described by Stefanski and Boos (2002), $var(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1}B(\boldsymbol{\theta})[A(\boldsymbol{\theta})^{-1}]^T$ where $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}, \lambda)^T$. Hence $Var(\mu)$ is the top left corner entry of $var(\boldsymbol{\theta})$.

$$A(\boldsymbol{\theta}) = E\left[-\frac{\partial}{\partial\boldsymbol{\theta}}\Psi\right] = \begin{pmatrix}1 & H & F\\ 0 & e_{\beta\beta} & 0\\ 0 & 0 & E[T_i]\end{pmatrix}$$

where $H = E\left[\frac{\delta}{K(t)}\left(\frac{ZY}{e^2} + \frac{(1-Z)Y}{(1-e)^2}\right)e_\beta\right]$, $F = E\left[-\left(\frac{\delta T}{K(t)}\right)\left(\frac{ZY}{e} - \frac{(1-Z)Y}{1-e}\right)\right]$ and $e_{\beta\beta} = E\left[\frac{e_\beta e_\beta^T}{e(1-e)}\right]$.

$$B(\boldsymbol{\theta}) = E\left[\Psi\Psi^T\right] = \begin{pmatrix}\Sigma^* & H & G_1\\ H & e_{\beta\beta} & G_2\\ G_1 & G_2 & G_3\end{pmatrix}$$

where $\Sigma^* = E\left[\left(\frac{ZY}{e} - \frac{(1-Z)Y}{1-e}\right)\left(\frac{\delta}{K(T)}\right)^2\right] - \mu^2$, $H = E\left[\left(\frac{Y_1}{e} + \frac{Y_0}{1-e}\right)\frac{\delta}{K(T)}e_\beta\right]$,
$G_1 = E\left[\left(\left(\frac{Z_iY_i}{e_i} - \frac{(1-Z_i)Y_i}{1-e_i}\right)\left(\frac{\delta_i}{K(T_i)}\right) - \mu\right)(\delta_i - \lambda T_i)\right]$, $G_2 = E\left[\left(\frac{Z_i-e_i}{e_i(1-e_i)}e_\beta\right)(\delta_i - \lambda t_i)\right]$ and $G_3 = E\left[(\delta_i - \lambda T_i)^2\right]$. The components of all of the above expressions can be estimated from the observed data.

2.4. Doubly Robust Estimation

Doubly Robust (DR) estimation incorporates outcome regression (regression model) and weighting by PS (PS model), and it is robust to misspecification of one (but not both) of these models. There are many forms of DR estimators; here we follow the general procedure described by Robins, Rotnitzky, and Zhao (1994). DR estimator has the smallest large sample variance among the class

of weighted estimators and is locally semi parametric efficient. First, we estimate the regression model for the treated group ($Y \sim \mathbf{X}$ for $Z = 1$) and obtain predicted values for the entire sample: $\hat{m}_1(\mathbf{X}_i)$. We then do the same for the control subjects and obtain predicted values for the entire sample: $\hat{m}_0(\mathbf{X}_i)$. In other words, $m_0(\mathbf{X}_i)$ and $m_1(\mathbf{X}_i)$ are the postulated models for the true regressions $E(Y|Z = 0, \mathbf{X})$ and $E(Y|Z = 1, \mathbf{X})$. Note that simple weights $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$ are applied to the regression models to account for informative censoring. The DR estimator of $\hat{\mu}$ is given by:

$$\hat{\mu}_{dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i \delta_i}{\hat{e}_i \hat{K}(T_i)} - \frac{(Z_i - \hat{e}_i)m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i \hat{K}(T_i)}\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - Z_i)Y_i\delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)} + \frac{(Z_i - \hat{e}_i)m_0(\mathbf{X}_i)\delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)}\right] \quad (2.17)$$

Similar to section 2.2, the regression models $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$ can be modeled in several ways:

Normal model:

$$E(Y_i|Z_i = z, X_i) = \mathbf{X}_i\beta \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \quad (2.18)$$

Lognormal model:

$$E(\log(Y_i)|Z_i = z, X_i) = \mathbf{X}_i\beta \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \quad (2.19)$$

Gamma model:

$$E(Y_i|Z_i = z, X_i) = \exp(\mathbf{X}_i\beta) \text{ weighted by } \frac{\delta_i}{\hat{K}(T_i)} \quad (2.20)$$

The doubly robust estimates are consistent if the propensity score model or the regression model $m_1(\mathbf{X}) = E(Y|Z = 1, \mathbf{X})$ and $m_0(\mathbf{X}) = E(Y|Z = 0, \mathbf{X})$ are correctly specified. To see this, consider $\hat{\mu}_{1,dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i \delta_i}{\hat{e}_i \hat{K}(T_i)} - \frac{(Z_i - \hat{e}_i)m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i \hat{K}(T_i)}\right]$. By the Law of Large Numbers, $\hat{\mu}_{1,dr}$ estimates:

$$\begin{aligned}
& E\left[\frac{ZY\delta}{eK(T)} - \frac{(Z - e)m_1(\mathbf{X})\delta}{eK(T)}\right] \\
= \; & E\left[\frac{ZY^{(1)}\delta}{eK(T)} - \frac{(Z - e)m_1(\mathbf{X})\delta}{eK(T)}\right] \\
= \; & E\left[\frac{\delta}{K(T)}Y^{(1)} + \frac{(Z - e)}{e}\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)\right] \\
= \; & E\left[Y^{(1)}\right] + E\left[\left(\frac{Z}{e} - 1\right)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)\right] \\
= \; & \mu_1 + E\left[\left(\frac{Z}{e} - 1\right)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)\right]
\end{aligned}$$

Hence for $\hat{\mu}_{1,dr}$ to be unbiased, we need the second term $S = E\left[(\frac{Z}{e} - 1)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)\right]$ to be zero. This condition is satisfied when the propensity score model is correctly specified: $E(Z|Y^{(1)}, \mathbf{X}) = E(Z|\mathbf{X}) = e(\mathbf{X}, \beta) = e$ so $S = E\left[E\left[(\frac{Z}{e} - 1)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)|Y^{(1)}, \boldsymbol{X}\right]\right] = E\left[\left(\frac{E(Z|Y^{(1)}, \boldsymbol{X})}{e} - 1\right)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)\right] = 0$. When the regression model $m_1(\boldsymbol{X})$ is correctly specified, $m_1(\boldsymbol{X}) = E(Y|Z = 1, \mathbf{X}) = E(Y^{(1)}|Z = 1, \mathbf{X}) = E(Y^{(1)}|Z, \mathbf{X})$ so $S = E\left[E\left[(\frac{Z}{e} - 1)\left(\frac{\delta}{K(T)}Y^{(1)} - m_1(\boldsymbol{X})\right)|Z, \boldsymbol{X}\right]\right] = E\left[(\frac{Z}{e} - 1)\left(E(\frac{\delta}{K(T)}Y^{(1)}|Z, \boldsymbol{X}) - m_1(\boldsymbol{X})\right)\right] = E\left[(\frac{Z}{e} - 1)\left(E(Y^{(1)}|Z, \boldsymbol{X}) - m_1(\boldsymbol{X})\right)\right] = 0$. Hence, the DR estimator is unbiased if either the propensity score model or the regression model is correctly specified. The doubly robust procedure has benefits over standard estimation but can result in biased estimates if both the regression model and PS model are misspecified (Funk et al., 2011).

2.5. Super-Learning

The Super-learner algorithm (Laan, Polley, and Hubbard, 2007) is an ensemble machine learning approach based on V-fold cross validation. It allows one to specify several candidate prediction models and use them to produce an asymptotically optimal combination. Specifically, data are split into blocks and then each of the candidate algorithms are fitted on the training set and outcomes are predicted using the validation set. The loss function is calculated within each validation set, and averaging across validation sets provides the estimated cross validated risk score for each method The SL algorithm finds the optimal weighted combination of all the methods. Laan, Polley, and Hubbard (2007) proved asymptotic efficiency of the SL algorithm. Further, it is guaranteed to perform at least as well as the best estimators from the candidate models. This machine learning algorithm is available as an R package called `Super Learner` (https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf) and as a SAS macro (Brooks, 2012).

In DR estimation, our primary concern is whether the cost regression models $m_1(\boldsymbol{X})$ and $m_0(\boldsymbol{X})$ are correctly specified. Given the heterogeneous nature of costs, there is no one-size-fits-all regression model. In machine learning literature, it is common to combine predictions from multiple models or multiple parametric and non-parametric predictive algorithms. Hence, one intuitive solution to accommodate the complex features of cost distribution is to employ SL to obtain the optimal prediction from common cost models.

Super-learner methods can also be applied when we are uncertain about model specification in the

propensity model. Untill now, we have assumed the propensity score model to be correctly specified; but this is unlikely to be true in practice. If the correct subset and functional forms of covariates are unknown, we can include all combinations of potential subsets, interactions and quadratic forms of covariates and use SL to find the optimal estimates. Recent studies have proposed to use tree-based methods (Setoguchi et al., 2009), random forests (Lee, Lessler, and Stuart, 2010) and neural networks for estimating the PS. These can be included as candidate PS models, allowing SL to obtain optimal PS estimates from a wide variety of candidate algorithms (Gruber et al., 2015).

2.6. Simulation studies

Using simulation studies, we evaluate the performances of all methods discussed in Section 2, 3 and 4 under various settings, including different survival models, cost models, and censoring distributions. We report the bias, the coverage probability of the resulting 95% confidence interval and the mean square error ratio (MSER) which is the ratio of MSE of each approach with reference to MSE of DR with SL in regression models.

We based choices of our simulation parameters on data from our bladder cancer study (Section 7). We simulated three covariates $\mathbf{X} = \{X_1, X_2, X_3\}$. Since most covariates in our empirical example were categorical, we simulated X_1 and X_2 as binary with success probabilities of 0.5 and 0.25 respectively. X_3 followed a normal distribution with standard deviation 1 and mean 0. Using these covariates, we then defined treatment choice Z using a logit index model where $D \sim$ Bernoulli(p) and

$$\text{logit}(p) = -0.8X_1 - 1.6X_2 + 0.4X_3 \tag{2.21}$$

The coefficients were fixed so that approximately 30% of the population received treatment, to mirror our bladder cancer data. The sample sizes were set to be 1000 and 5000, typical sizes for observational studies.

We drew failure times from weibull and exponential distributions where $f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}$. For weibull failure times, we set $k = 2.5$ and $\lambda = 3.2 + 2Z + 1.2X_1 + 1.4X_2 - 0.6X_3$. For exponential failure times, $k = 1$ and $\lambda = \exp(-Z - 0.8X_1 - 1.2X_2 - 0.6X_3)$. Censoring times were independently simulated from uniform distribution $U(0, 20)$ and $U(0, 12)$ for light and moderate censoring. The probability of censoring was approximately 20% for light censoring and 35% for moderate cen-

soring, respectively. The latter scenario was similar to our bladder cancer example. Observed time was defined as the lesser of survival time and censoring time.

As medical costs are often complex and can come from very different distributions, we generated total medical costs from normal, lognormal and gamma distributions according to the parametrization shown below. The mixed distribution was a weighted average of the normal, lognormal and gamma cases.

$$\begin{aligned}
\text{Normal}: &\quad Y(\text{Normal}) \sim 5.8 + \text{Normal}(0, 0.4) + Z + 0.4X_1 + 0.8X_2 + X_3 \\
\text{Lognormal}: &\quad Y(\text{Lognormal}) \sim \exp(\text{Normal}(0, 0.2) + 1.6Z + 1.2X_1 + 0.8X_2 + 0.2X_3) \\
\text{Gamma}: &\quad Y(\text{Gamma}) \sim \text{Gamma}(\text{shape} = 2.5, \text{scale} = \exp(Z + 0.6X_1 + 0.4X_2 + 0.2X_3)) \\
\text{Mixed}: &\quad Y(\text{Mixed}) \sim \{Y(\text{Normal}) + Y(\text{Lognormal}) + Y(\text{Gamma})\}/3
\end{aligned}$$

Propensity scores were estimated using a logistic regression model assuming correct model specification according to Equation 2.21. We then applied PS covariates adjustment with normal, lognormal and gamma models, stratification, IPTW, normalized IPTW, DR with normal, lognormal and gamma regression models and DR using SL for regression models to estimate μ. Jiang and Zhou (2004) showed that using bootstrap methods to estimate CI of mean cost work well. Bang and Tsiatis (2000) also showed that bootstrap estimates of variance for mean cost are consistent with the analytically derived asymptotic variance estimates. In our analysis, there are several sources of variation for $\hat{\mu}$. For example, when using the DR estimator, we have variation from the PS model, KM model, regression models and the final DR estimation model. This greatly complicates analytic variance estimation but can be easily dealt with by using non-parametric bootstrapping. We used a bootstrap estimate with bias-corrected and accelerated (BCa) correction (Efron, 1987) to construct 95% CI confidence intervals of $\hat{\mu}$. Lastly, we included the naive regression method where total cost is regressed on the main effects of covariates in a linear model to recognize the consequences of analyses that do not properly account for confounding, skewness and censoring.

We simulated each scenario 500 times and summarize results by the empirical percentage bias (%bias), coverage probability of the 95% confidence interval (Coverage) and MSE ratio based on BCa standard errors (MSER). Note that for subjects with large observation time, if the estimated

probability of censored $\hat{K}(t_i)$ was zero, $\min_i \hat{K}(t_i)$ in the specific treatment or treatment-stratum group was used instead to avoid the issue of the denominator of $\frac{\delta_i}{\hat{K}(t_i)}$ being zero. Thus, all empirical estimations of μ were under-estimations. The extent of under-estimation depends on the censoring proportion and method used.

2.6.1. Simulation results

Results of the simulation with various censoring and cost settings and sample size of 1000 appear in Table 2.1. The naive estimator ignoring censoring and confounders is biased under all settings. As anticipated, the PS covariate adjustment performs well when the correct model is specified, but exhibits bias when mis-specified. For example, when cost follows gamma distribution, covariate adjustment with gamma model yields 0.35% bias while the lognormal model had 18.74% bias under light censoring. If cost comes from a mixture of normal, lognormal and the gamma distributions, covariate adjustment methods perform poorly since the true relationship between outcome and PS is unknown. Of the covariate adjustment models, the gamma model is the most robust, with the smallest biases for misspecifed cost distributions, a finding consistent with Dodd et al. (2006) and Basu, Manning, and Mullahy (2004). The PS stratification estimator has large biases and worst MSE among all PS methods. Note that stratification is most susceptible to under-estimation of μ. Since we need to calculate stratum and treatment specific $\hat{K}(u)$, $\hat{K}(u)$ is more likely to be zero for observations with large observation time T.

IPTW estimators yield bias ranging from -0.38% to -6.58%. The normalized IPTW estimator has smaller bias than the typical IPTW, consistent with findings from Lunceford and Davidian (2004). Estimates from DR methods had very small bias, even when the regression model is mis-specified. Since the PS is correctly modeled, DR estimators should be unbiased due to their doubly robust property as demonstrated here. Correct regression model specification in DR has very small effect on bias and coverage since PS model is already correct. Nevertheless, using SL for the regression model results in small bias and MSE among all DR models. Simulations with a sample size of 5000 (data not shown) produce similar results in terms of bias and coverage, but have smaller MSE. As expected, the larger sample size increases overall estimation efficiency.

	normal			lognormal			gamma			mixed		
light censoring	%bias	Coverage	MSER	%bias	Coverage	MSER	%bias	Coverage	MSER	%bias	Coverage	MSER
naive regression	-40.56	0.09	125.14	-34.32	0.04	42.32	-30.23	0.04	4.15	-34.31	0.05	29.36
covariates adjustmt: normal	0.11	0.94	2.34	-12.81	0.52	12.65	-7.86	0.84	0.83	-14.15	0.41	5.79
covariates adjustmt: lognormal	1.82	0.36	3.66	-1.48	0.95	2.80	18.74	0.74	2.43	-15.67	0.17	6.43
covariates adjustmt: gamma	0.35	0.95	2.36	2.62	0.94	1.19	0.35	0.96	0.62	-11.45	0.49	3.90
stratification	-1.43	0.84	10.33	-8.99	0.85	5.84	-3.30	0.90	1.15	-4.73	0.89	3.23
IPTW	-0.81	0.99	49.38	-2.72	0.90	8.18	-0.74	0.92	1.74	-0.38	0.95	6.80
IPTW: normalized	0.13	0.94	39.94	-0.03	0.90	7.26	-0.09	0.90	1.61	-0.08	0.93	5.75
DR: normal	0.02	0.97	1.00	0.83	0.92	1.85	-0.36	0.91	1.00	0.44	0.95	1.34
DR: lognormal	-0.02	0.96	1.08	0.36	0.94	1.01	-0.42	0.92	1.04	0.03	0.97	1.03
DR: gamma	-0.02	0.96	1.07	0.91	0.94	1.00	-0.36	0.92	1.00	0.04	0.94	1.00
DR: SL in regression model	0.01	0.97	1 (ref)	0.96	0.96	1 (ref)	0.07	0.94	1 (ref)	0.01	0.93	1 (ref)
moderate censoring												
naive regression	-67.59	0.04	188.13	-25.61	0.03	35.21	-44.57	0.02	5.59	-47.69	0.08	32.93
covariates adjustmt: normal	0.56	0.92	2.92	-12.18	0.69	7.10	-8.12	0.87	0.76	-14.69	0.48	4.14
covariates adjustmt: lognormal	2.33	0.52	4.14	-1.49	0.97	4.47	33.36	0.59	3.76	-13.47	0.51	3.43
covariates adjustmt: gamma	0.78	0.94	2.95	2.55	0.93	1.07	0.52	0.96	0.59	-11.78	0.60	2.86
stratification	-24.97	0.85	54.56	-2.66	0.92	7.05	-12.30	0.82	1.80	-14.08	0.78	6.05
IPTW	-1.68	0.90	92.04	2.14	0.96	10.49	-6.58	0.90	2.36	-5.51	0.90	8.49
IPTW: normalized	0.73	0.90	86.74	2.10	0.95	10.08	-1.27	0.89	2.24	-1.33	0.88	8.14
DR: normal	-0.70	0.92	1.00	1.93	0.94	1.82	-1.32	0.95	1.01	-1.90	0.94	1.30
DR: lognormal	-0.82	0.94	1.08	2.01	0.95	1.00	-2.17	0.95	1.04	-1.86	0.94	1.03
DR: gamma	-0.79	0.94	1.08	2.01	0.95	0.98	-1.15	0.96	1.01	-1.72	0.94	1.00
DR: SL in regression model	-0.73	0.96	1 (ref)	1.81	0.97	1 (ref)	-1.12	0.97	1 (ref)	-1.41	0.95	1 (ref)

Table 2.1: %Bias, coverage and relative efficiency for estimated treatment effect on cost

2.6.2. Misspecified PS

Next, we explore the case of PS misspecification when the correct model is unknown. We use the same simulation procedure as above changing Equation 2.21 to

$$\text{logit}(p) = -2 - 0.2X_1 - 0.4X_2 - 0.2X_3 + 1.4X_1X_2 - 1.4X_1X_3 + 1.2X_3^2 \tag{2.22}$$

In the simulated data, we estimated PS according to the correct model in Equation 2.22, and also a misspecified PS model with only main effects of X_1, X_2 and X_3. Finally, we used SL to estimate PS using all possible combinations of the second order polynomials of $\mathbf{X}$ and the two way interactions among them. Table 2.2 shows the results for weibull survival time, light censoring, sample size of 1000, gamma and mixed cost models.

	Correct PS			Misspecified PS			SL PS		
gamma model	%bias	Coverage	MSER	%bias	Coverage	MSER	%bias	Coverage	MSER
naive regression	-9.52	0.67	5.81						
covariates adjustmt: normal	6.66	0.92	5.18	7.50	0.88	7.80	5.01	0.92	3.85
covariates adjustmt: lognormal	-21.98	0.22	1.35	-22.67	0.12	2.04	-22.19	0.72	1.26
covariates adjustmt: gamma	0.30	0.94	2.07	4.13	0.94	3.12	0.30	0.96	1.53
stratification	2.85	0.91	10.54	32.58	0.77	20.65	1.77	0.94	5.94
IPTW	0.76	0.96	14.25	46.71	0.28	21.48	0.62	0.93	8.95
IPTW: normalized	-0.36	0.94	1.83	8.88	0.90	2.76	0.21	0.92	1.58
DR: normal	0.61	0.94	1.46	0.75	0.93	1.31	0.30	0.94	1.07
DR: lognormal	-0.71	0.94	1.45	8.91	0.89	1.05	0.38	0.96	0.93
DR: gamma	0.29	0.93	1.53	0.19	0.93	0.94	0.09	0.95	0.90
DR: SL in regression model	0.48	0.92	1 (ref)	3.10	0.94	1 (ref)	0.02	0.90	1 (ref)
mixed									
naive regression	-48.82	0.29	6.70						
covariates adjustmt: normal	44.45	0.54	1.74	45.15	0.54	2.27	42.34	0.60	2.21
covariates adjustmt: lognormal	-16.81	0.85	2.54	-26.51	0.85	5.09	-19.18	0.83	4.21
covariates adjustmt: gamma	23.84	0.79	1.46	15.84	0.79	1.84	21.56	0.84	1.71
stratification	12.64	0.91	5.49	15.47	0.77	16.84	9.25	0.95	3.57
IPTW	8.57	0.97	9.47	10.57	0.97	23.13	4.23	0.92	6.00
IPTW: normalized	-1.00	0.94	1.10	-2.00	0.94	1.88	-0.80	0.92	1.05
DR: normal	-1.81	0.96	1.74	-8.07	0.96	1.03	-1.21	0.95	1.06
DR: lognormal	-2.00	0.97	3.85	-6.77	0.98	1.86	-1.24	0.96	1.07
DR: gamma	-2.07	0.97	1.82	-1.93	0.95	0.97	-0.34	0.95	1.05
DR: SL in regression model	-0.18	0.98	1 (ref)	-5.72	0.97	1 (ref)	-0.02	0.97	1 (ref)

Table 2.2: %Bias, coverage and relative efficiency for estimated treatment effect on cost under different PS estimation methods

When the PS model is mis-specified, estimates from PS covariate adjustment are biased (4.13% to 45.15%). Estimates from IPTW methods are also highly biased (-2.00% to 46.71%) when PS model is mis-specified, in line with Rubin (2004). When the regression models in DR are correctly established, DR estimators have very small bias. However, when both the regression model and the PS model are wrong, as anticipated we see some bias (0.75% to 8.91%). Overall, PS mis-specification affects all of the estimators discussed, especially those that are sensitive to PS. The

only method that is robust to PS misspecification is DR, provided the regression model is correctly established.

When SL is used to estimate PS, we see significant improvement of performance across all estimators. In most cases, using SL in PS estimation yields less bias and better coverage than when the correct PS model is used. Hence, we recommend using SL when the correct PS model is unknown. When true cost comes from a mixture of normal, lognormal and gamma distributions, SL in DR can provide the best regression model estimates. In real life settings, it is highly likely that cost comes from a mixture of different distributions and the correct PS model is unknown. In this case, using SL with DR and PS estimation provides added flexibility which improves estimates substantially.

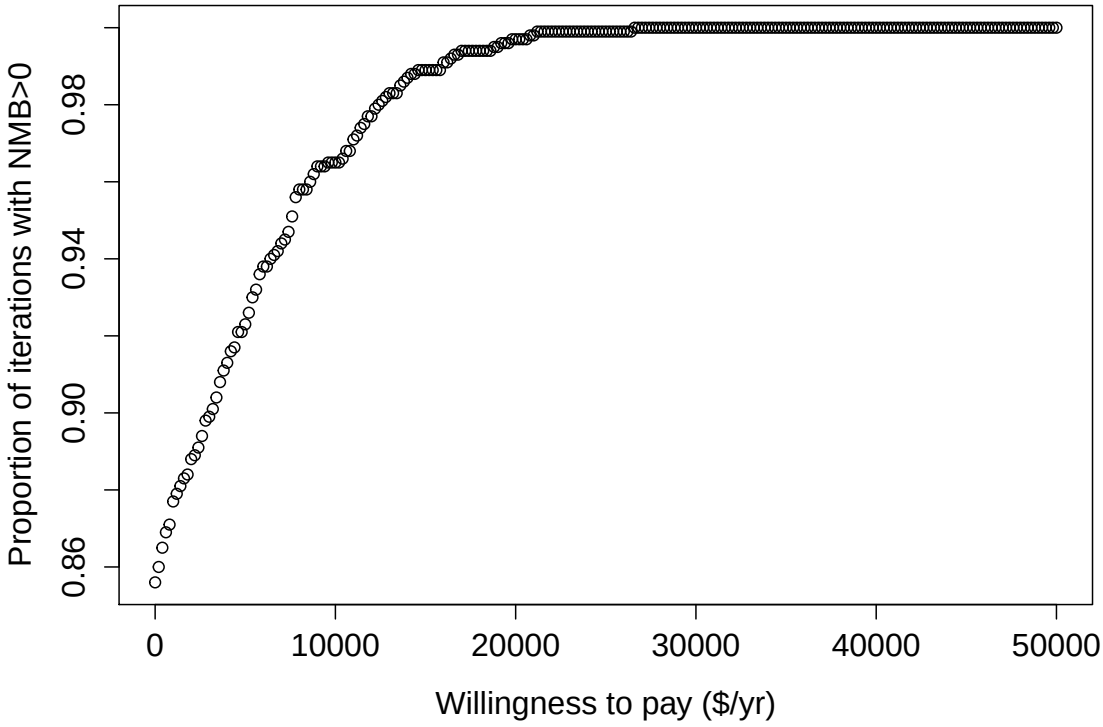
2.7. Costs of Bladder Cancer Therapies

Bladder cancer affects more than 70,000 people annually in the United States and accounts for almost 5% of the total cancer-related costs to Medicare. The guideline recommended treatment for bladder cancer is radical cystectomy (RC) which involves surgical removal of the bladder. Bladder preservation therapy (BPT) is a less aggressive, non-surgical alternative that involves radiation and chemotherapy. Recent studies have shown that BPT may improve quality of life over RC (Efstathiou et al., 2012). We have applied our method to compare the life-time cost of RC and BPT using a cohort of patients derived from SEER-Medicare registry.

We included stage II/III bladder cancer patients diagnosed between 1995 and 2005. See Bekelman et al. (2013) for a detailed description of inclusion/exclusion criterion. 32% of the study cohort were censored at the end of the study. Payment data were extracted from Carrier Claims file, the Outpatient file, and the Medicare Provider Analysis and Review Record. We adjusted all costs to year 2000 dollars using the Medicare Economics Index (Centers for Medicare and Medicaid, 2010). The final cohort sample size was 1860; 422 had BPT and 1438 had RC. The mean uncensored costs were $68,800 for BPT patients and $83,040 for RC patients. Total treatment cost were highly right skewed Figure 2.1 with a maximum observed cost of $511,200. The average observation time was 3.93 years.

In this study, both treatment assignment and total cost may have been affected by covariates such as stage, grade, race, marital status, comorbidities, median income at the census tract level and

Figure 2.1: Density of uncensored total costs in bladder cancer cohort



community size. Hence, we estimated PS using a logistic regression model that was adjusted for all of these potential confounders. We then estimated the difference in total cost between BPT and RC using the approaches described above including: PS covariates adjustment with normal, lognormal and gamma models, stratification, IPTWs, DR with normal, lognormal and gamma regression models and DR using SL in regression model. Naive linear regression ignoring censoring and non-random treatment assignment was used as a reference. Approximate confidence intervals for the treatment effect on cost were constructed using non-parametric bootstrapping with BCa correction.

From Table 2.3, BPT was estimated to be $7,412 cheaper than RC using naive regression. Difference in cost estimated from various propensity score methods ranged from -$10,661 to -$20,937, differed significantly from the naive regression method. Failure to account for censoring and the effect of confounders could lead to biased estimates. Furthermore, covariate adjustment, stratification and weighting methods could be sensitive to the choice of PS model estimation. Unsurprisingly,

	Regular PS Model		SL PS Model	
	Estimates	95% CI	Estimates	95% CI
naive regression	-7,412	(-13,545, -1,279)	-	-
covariates adjustment normal	-12,423	(-22,235, -3,047)	-11,448	(-23,237, -2,689)
covariates adjustment lognormal	-13,877	(-25,729, -3,323)	-13,033	(-26,674, -3,092)
covariates adjustment gamma	-12,482	(-22,171, -2,400)	-11,599	(-24,943, -3,579)
stratification	-17,678	(-28,542, -9,685)	-15,416	(-26,876, -787)
IPTW	-20,937	(-34,244, -7,171)	-22,473	(-30,633, -8,446)
IPTW: normalized weights	-10,661	(-21,073, -1,078)	-11,951	(-21,469, -607)
DR: normal	-12,163	(-23,285, -764)	-12,312	(-23,458, -104)
DR: lognormal	-14,117	(-25,070, -3,444)	-14,086	(-24,449, -3,333)
DR: gamma	-12,144	(-22,920, -172)	-12,179	(-23,745, -235)
DR: SL mean model	-14,163	(-24,216, -3,941)	-14,086	(-26,876, -787)

Table 2.3: Estimated mean cost difference for bladder preserving therapy and radical cyccwetomy

we saw large variation in treatment effect estimates from these models. DR models yielded more consistent treatment effect estimates; BPT was estimated to be -$12,144 to -$14,117 cheaper than RC. Using SL in regression model in DR gave slight different estimations (-$14,163). SL in regression model in DR is likely to be the closest to the true cost estimate as evidenced from the simulation study. Lastly, all CIs did not cross zero, indicating that BPT was significantly less costly than RC.

Next we applied SL in propensity score model to obtain the estimated PS. We specify several potential propensity score models with different covariates functional forms: the basic logistic model where all covariates were included, also a model including all two way interactions between covariates, adding square terms of all covariates and a backwards stepwise selection algorithm with cut-off p-value of 0.1. SL was used to find the optimal combination of predications from these candidate models. We then use this estimated PS to find the differences in cost between BPT and RC.

From Table 2.3, the SL PS models provided similar estimates from the regular PS models. One possible explanation is all covariates were categorical, hence there was little variation in PS due to limited covariate patterns. Interaction and quadratic terms might not have a huge impact on PS estimation for the same reason. SL PS model would be more useful when we have little understanding of the true PS model. Nevertheless, SL PS showed that the estimates of cost differences were between -$11,448 and -$22,473, and 95% CIs strongly suggest the differences in cost between BPT and RC were significant.

All of the approaches discussed above demonstrate that BPT substantially decreases the total medical cost compared to the standard treatment RC. However, we observed significant variations in the ATE estimations and large range in the CIs. From our simulation studies, we believe DR with SL in both regression model and PS model provides the best estimate. Hence, our findings indicate that BPT was $14,086 ($787, $26,876) cheaper than RC.

2.8. Discussion

In this study, we explored propensity score based approaches for estimating the treatment effect on censored costs in an observational study. We extended covariate adjustment, stratification, weighting and doubly robust methods to handle censored medical cost. We also utilized a machine learning algorithm, Super Learner, to better estimate PS and the regression models in DR. Our simulation studies showed that when PS is correctly modeled, stratification and weighting yield unbiased estimates. Covariate adjustment is sensitive to the choice of outcome model, while DR is more robust to misspecification. When the correct PS model is unknown, misspecification could result in biased estimates of the treatment effect even when using DR methods. SL mitigates this bias by producing optimal regression models and PS estimates. In addition, one may consider tree-based methods, random forests and neural networks. These methods can be easily incorporated into SL to obtain optimal PS estimation from both fully parametric and non parametric models.

We note that in this study, we only used total cost data and ignored cost history data which may be available from claims data. Bang and Tsiatis (2000) have proposed partitioned estimators making use of cost history data which they showed to be more efficient than the simple weighted approach we employed. It is unclear what the effect of partitioned estimators would have on PS-based estimation and is worthy of future work.

We have shown that the variance of the IPTW estimator can be obtained analytically. However, multi-parameter or non-parametric survival models add substantial complexity to analyzing variance estimates due to the complex interaction between censoring and propensity scores.

As in any observational study, unobserved or hidden bias may be of concern. We suggest that in addition to a propensity score based analysis of censored cost data, one should conduct a carefully planned sensitivity analysis to assess the effect of an unmeasured confounder on the treatment

effect (Handorf et al., 2013).

2.9. Acknowledgment

Dr. Bekelman was supported by K07-CA163616. Dr. Handorf was supported by NIH grant P30-CA06927. We used the linked SEER-Medicare database and acknowledge the efforts of the Applied Research Program; National Cancer Institute; Office of Research, Development and Information; Centers for Medicare and Medicaid Services; Information Management Services; and SEER program tumor registries in the creation of the SEER-Medicare database.

CHAPTER 3

DOUBLY ROBUST METHODS FOR COST-EFFECTIVENESS ESTIMATION FROM OBSERVATIONAL DATA

3.1. Introduction

Policy makers are often interested in the cost-effectiveness (CE) of health care interventions in their decision making. Many countries, including the United Kingdom, Australia and Canada require CE evidence before a drug is granted reimbursement status (Clement et al., 2009). However, proper CE analysis is complicated by the need to account for features of cost data, including informative censoring and skewness. In addition, medical costs are often collected from claims data, which are susceptible to confounding. In this paper, we aim to estimate CE measures from observational data accounting for the unique features of cost. Common CE measures include the incremental cost effectiveness ratio (ICER), net monetary benefit (NMB) and CE acceptability curves. These measures require one to estimate cost and effectiveness (e.g. survival time, quality adjusted survival) separately (Gomes et al., 2012; Willan and Briggs, 2006; Zhao and Tian, 2001).

Historically, cost estimation methods have focused on two unique features of these data: distributional skewness and informative censoring. To account for the skewness and heteroscadasticity often present in cost data, researchers have used ordinary least square regression (OLS), OLS for log cost, generalized linear models (GLM), generalized gamma models, median regression and the Weibull model. Several studies (Basu, Manning, and Mullahy, 2004; Basu and Rathouz, 2005; Dodd et al., 2006) have evaluated the performance of these various cost models. Dodd et al. (2006) found the generalized gamma model to be the most robust; nevertheless, there is no one-size-fits-all model and Mihaylova, Briggs, and Hagan (2011) concluded that the choice of cost model should depend on the specific structure of the cost data at hand. Another important feature of cost data is excess zeros, especially when analyzing monthly cost data. Two-part models (Duan et al., 1983; Leung and Yu, 1996) have been proposed to accommodate these structural zeros.

Censoring, which occurs when a study terminates before all patients reach their end-points, is also a common attribute of cost data. Although survival time is usually non-informatively censored (e.g.

end-of-study censoring), total cost is informatively censored as patients may have drastically different rates of cost accrual. In practice, a patient with higher costs at the time of censoring is likely to have higher costs at the time of event as well. Lin et al., 1997 and Bang and Tsiatis, 2000 proposed weighted estimators to handle such informative censoring. Several studies have investigated the properties of these methods (Raikou and McGuire, 2004; Young, 2005; Zhao, Cheng, and Bang, 2011; Zhao et al., 2007), and some (Bang and Tsiatis, 2002; Baser et al., 2004; Lin, 2000, 2003) have extended the weighting method to linear regression, generalized linear models and median regression to model the relationship between cost and covariates.

As noted above, effectiveness (typically survival time) is often subject to non-informative censoring. Most CE studies use the area under the Kaplan-Meier (KM) estimate of the survival function to approximate mean survival time (Bang, 2005; Willan, Lin, and Manca, 2005; Zhao and Tian, 2001). However, Willan and Briggs (2006) suggested using the weighting technique often used for cost estimation to estimate effectiveness instead. This is a less intuitive but more flexible approach; to our knowledge there is no study comparing the performances of these two approaches.

Current methodological guidance for CE estimation is primarily in the setting of randomized controlled trials (Glick et al., 2014; Gomes et al., 2012; Willan and Briggs, 2006; Willan, Lin, and Manca, 2005; Zhao and Tian, 2001). In practice, most CE analyses rely on data from observational studies (Kreif, Grieve, and Sadique, 2013), thus necessitating the need to develop CE estimation models for observational data. Previous studies (Anstrom and Tsiatis, 2001 and Li et al., 2015) have investigated propensity score based models for observational cost data only. Goldfeld (2014); Indurkhya, Mitra, and Schrag (2006); Mitra and Indurkhya (2005) proposed propensity score adjustment for estimating the NMB from claims data.

The goal of this study is to develop doubly robust (DR) methods based on propensity scores to estimate CE measures from observational data. Here, we build on existing DR methods for cost estimation (Li et al., 2015) and propose several DR models for time and cost estimation, respectively. DR estimation combines outcome regression with weighting by propensity scores and is robust to misspecification of one (but not both) of these two components (Bang and Robins, 2005; Tsiatis and Davidian, 2007). Lunceford and Davidian (2004) demonstrated that the DR estimator performs better than other propensity score based methods such as stratification and weighting. Given the heterogeneous nature of cost distributions, the many possible choices of cost models de-

scribed above, as well as the challenge of accurately estimating propensity scores, we propose an ensemble machine learning approach based on cross validation called Super Learner (SL) (Laan, Polley, and Hubbard, 2007) to best estimate the outcome and the propensity score components in DR.

We begin with an introduction to common CE measures. We follow by reviewing and comparing two effectiveness estimation methods and then propose a DR effectiveness estimation model. We then introduce two DR models, the simple weighted and the partitioned model, for cost estimation. We present results from extensive simulation studies that compare the performance of the proposed DR estimators. Finally, we apply these DR models to a CE analysis of two lung cancer surveillance procedures, CT scans versus chest X-ray, using SEER-Medicare data.

3.2. Common CE measures

CE analysis is often used to evaluate the merits of a new health-care intervention (treatment, $Z = 1$) compared to an existing one (control, $Z = 0$). CE measures integrate estimates of costs and effectiveness in a single statistic derived from two components: Δ_E and Δ_C where Δ_E =Effectiveness$_{Z=1}-$Effectiveness$_{Z=0}$ and Δ_C =Cost$_{Z=1}-$Cost$_{Z=0}$. The duration of interest can be considered to be $(0, \tau)$ so that cost and effectiveness measures are bounded by τ and cost refers to the total cost from time 0 to τ. In this section, we briefly introduce three of the most commonly used CE measures.

3.2.1. ICER

ICER is an intuitive statistic that is defined as ICER $= \frac{\Delta_C}{\Delta_E}$. A major limitation of the ICER is its discontinuity when the denominator Δ_E approaches zero. In addition, estimating the variance of ICER is problematic due to the acknowledged statistical problems associated with ratio statistics. Non-parametric bootstrapping, Fieller's theorem and Bayesian approaches (Heitjan, Moskowitz, and Whang, 1999; Polsky et al., 1997; Willan and O'Brien, 1996) can be applied to estimate the variance of ICER.

3.2.2. CE acceptability curve

An important concept in CE analysis is called willingness to pay (WTP, denoted by λ), which is the maximal monetary value decision-makers are willing to pay for a unit of Δ_E. Typically, λ measures the dollar amount one is willing to pay for one year of additional life. A CE acceptability curve displays the probability that the treatment is cost-effective compared with the control for a range of λ values. To plot the CE acceptability curve, we use bootstrapping to estimate $Pr(\lambda\Delta_E - \Delta_C > 0)$. In practice, we simply count the proportion of bootstrapped samples that yields $\lambda\Delta_E - \Delta_C > 0$ for a range of λ values.

3.2.3. NMB

Recently, health economists have advocated the use of the Net Monetary Benefit (NMB): NMB $(\lambda) = \lambda\Delta_E - \Delta_C$. NMB is a linear combination of Δ_C and Δ_E; it measures the excess benefit given a fixed level of λ. The NMB does not suffer from the singularity problem that the ICER does and it is straight forward to estimate its variance as $var(\text{NMB}(\lambda)) = \lambda^2 var(\Delta_E) + var(\Delta_C) - 2\lambda cov(\Delta_E, \Delta_C)$.

3.3. CE estimation

3.3.1. Δ_E estimation

In CE studies, effectiveness usually refers to survival time or quality adjusted life years. From here onwards, we will simply use survival time to represent effectiveness. Specifically, we are interested in estimating the *mean* survival time difference Δ_E between two treatment groups in the duration of interest $(0, \tau)$ in the presence of censoring.

Review and comparison of current techniques

Consider a randomized controlled trial where t_i and C_i represent the survival and censoring time for subject i respectively, $T_i = \min(T_i, C_i, \tau)$ and censoring indicator $\delta_i = I(T_i \leq C_i) + I(T_i > C_i) * I(C_i \geq \tau)$. We start by reviewing two popular estimation techniques, the area under the survival curve and the inverse probability weighting to estimate mean time difference $\Delta_E = E(T|Z = 1 - T|Z = 0)$ in the duration of interest $(0, \tau)$. Here, our goal is to accurately estimate mean survival time under censoring.

In the first approach, we integrate the area under the survival curve such as the Kaplan-Meier curve. Let $S(t)$ be the survival function, by definition:

$$\Delta_E = \int_0^\tau S_{Z=1}(t)dt - \int_0^\tau S_{Z=0}(t)dt \tag{3.1}$$

In practice, we integrate the area under the estimated survival curve: $\widehat{\Delta}_E = \sum_{i=1}^{\tau} \hat{S}_{Z=1}(t_i) * (t_{i+1} - t_i) - \sum_{j=1}^{\tau} \hat{S}_{Z=0}(t_j) * (t_{j+1} - t_j)$. The area under the survival curve approach is easy to use and hence very popular in CE studies. It does not require the non-informative censoring assumption and therefore works for dependent censoring too. However, this approach does not accommodate adjustment of confounders. Common regression based survival models such as the Cox proportional hazard model focus on hazard ratio estimation, making mean survival time estimation difficult. In addition, to our knowledge, there is no DR method available for survival models.

An alternative way to handle censoring in mean survival time estimation is to utilize inverse probability weighting (Willan and Briggs, 2006). This weighting technique is often used for cost estimation (Bang and Tsiatis, 2000) and is an application of the general representation theorem for missing data (Robins and Rotnitzky, 1992; Robins, Rotnitzky, and Zhao, 1994). Here we apply the same concept to estimate mean survival times as follows:

$$\Delta_E = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{T_i \delta_i Z_i}{K_{Z=1}(T_i)} - \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{T_j \delta_j (1 - Z_i)}{K_{Z=0}(T_j)} \tag{3.2}$$

where $K_{Z=z}(u) = P(C \geq u | Z = z)$ and n_z is the total number of subjects in treatment group Z. This estimator is simply a weighted average of observed survival times T_i for patients who are not censored. The weight is given by the inverse of the probability of not being censored at the time of death for those who died prior to τ and the inverse of the probability of not being censored at τ for those who survived to τ. We can easily show that $\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{T_i \delta_i Z}{K_{Z=1}(T_i)}$ is an unbiased estimator of $E(T|Z=1)$ as follows: $E\left[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{T_i \delta_i Z}{K_{Z=1}(T_i)}\right] = E\left[\frac{1}{n_1} \sum_i E\left[\frac{T_i \delta_i}{K_{Z=1}(T_i)} \middle| T_i, Z=1\right]\right] = E\left[\frac{1}{n_1} \sum_i \left[\frac{T_i | Z=1}{K_{Z=1}(T_i)} E(I(C_i \geq T_i | T_i, Z=1)\right]\right] = E\left[\frac{1}{n_1} \sum_i T_i | Z=1\right] = E[T|Z=1]$. In practice, we use Kaplan Meier to estimate $K(u)$ based on the data $(T_i, 1-\delta_i)$ so $\hat{\Delta}_E = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{T_i \delta_i Z_i}{\widehat{K}_{Z=1}(T_i)} - \frac{1}{n_0} \sum_{j=1}^{n_0} \frac{T_j \delta_j (1-Z_i)}{\widehat{K}_{Z=0}(T_j)}$.

Hence we have shown that the inverse probability weighting technique provides an unbiased estimate of mean survival time and thus Δ_E. However, this approach assumes censoring to be

non-informative. In other words, censoring in time is independent of other covariates. This assumption is considered to be valid for most observational studies (Anstrom and Tsiatis, 2001; Goldfeld, 2014; Raikou and McGuire, 2004), especially in large population based registries where censoring is administrative due to end of study. Nevertheless, if censoring is dependent on covariates $\mathbf{X}$ in the case of induced dropout or censoring due to non-compliance, we can modify Equation 3.2 by using $P(\delta_i = 0|\mathbf{X}_i, Z)$ instead of $K_z(T_i)$ (Cain and Cole, 2009; Cole and Hernán, 2008). Specifically, we first divide $(0, \tau)$ into discrete time intervals. At each time point k, we estimate $P(\delta_i^k = 0|\mathbf{X}_i, Z, \delta_i^{k-1} = 0)$ using a logistic regression model $\delta_i^k \sim \mathbf{X}_i$ for all subjects with $Z = z$ alive at time k. We can then use $P(\delta_i = 0|\mathbf{X}_i, Z) = \prod_{k<T_i} P(\delta_i^k = 0|\mathbf{X}_i, Z, \delta_i^{k-1} = 0)$ instead of $K(T_i)$ in Equation 3.2. This extension can easily incorporate time varying covariates $\mathbf{X}$ as well.

Although both techniques are unbiased, the weighting technique has more advantages. Specifically, it allows for covariate adjustment, accommodates informative censoring, as we will discuss next, is a natural fit for DR estimation. In section 3.4.1, we carry out simulation studies to compare the performance of these two approaches.

DR method for Δ_E estimation

As CE studies are often based on observational data, we use the conventional counter-factual notation modeling a causal framework. Let $t_i^{(0)}$ and $t_i^{(1)}$ denote the survival time if the patient were in the control and treatment group respectively. Let $\mathbf{X}_i$ be a vector of measured confounders we wish to adjust for. Lastly, propensity scores are denoted by $e = e(\mathbf{X})$. We assume strong ignorability and non-informative censoring (see Li et al. (2015) for further explanation of the assumptions).

We propose the following DR estimator for Δ_E that uses the concept of inverse probability weighting:

$$\widehat{\Delta}_E = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i T_i \delta_i}{\hat{e}_i \hat{K}(T_i)} - \frac{(Z_i - \hat{e}_i) m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i \hat{K}(T_i)}\right] - \left[\frac{(1 - Z_i) T_i \delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)} + \frac{(Z_i - \hat{e}_i) m_0(\mathbf{X}_i)\delta_i}{(1 - \hat{e}_i)\hat{K}(T_i)}\right] \tag{3.3}$$

For simplicity, we use $\hat{K}(u)$ to denote the treatment-specific estimated probability of being uncensored at u, $\hat{K}_z(u)$. Moreover, $m_0(\mathbf{X}_i)$ and $m_1(\mathbf{X}_i)$ are the postulated models for the true regressions $E(T|Z = 0, \mathbf{X})$ and $E(T|Z = 1, \mathbf{X})$. The outcomes models $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$ can be specified in various ways including:

- Normal model: $E(T_i|Z_i = z, X_i) = \mathbf{X}_i\beta$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$
- Lognormal model:$E(\log(T_i)|Z_i = z, X_i) = \mathbf{X}_i\beta$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$
- Gamma model:$E(T_i|Z_i = z, X_i) = \exp(\mathbf{X}_i\beta)$ weighted by $\frac{\delta_i}{\hat{K}(T_i)}$

The proposed doubly robust estimator is consistent if the propensity score model e or the outcome models $m_1(\mathbf{X}) = E(T|Z = 1, \mathbf{X})$ and $m_0(\mathbf{X}) = E(T|Z = 0, \mathbf{X})$ are correctly specified. Notice that weights $\frac{\boldsymbol{\delta}}{\hat{K}(\mathbf{T})}$ are applied to both PS weighting and outcome models to account for censoring. The variance of $\hat{\Delta}_E$ can be estimated using large sample theory to get the sandwich variance estimator (Li et al., 2015) or non-parametric bootstrapping.

Funk et al., 2011 noted that the doubly robust property can lead to biased estimates if both the outcome and the propensity score model are misspecified. In order to best estimate PS and outcome models in DR estimation, we employ a machine learning algorithm, Super Learner (SL) (Laan, Polley, and Hubbard, 2007). SL is an ensemble learning approach based on cross validation that allows us to specify several candidate prediction models and use them to produce an asymptotically optimal combination. Since survival time distributions can be very different for different diseases, we can utilize SL to combine prediction from several possible survival time models to estimate $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$. Recent work suggest using non-parametric machine learning models such as Classification and Regression Trees (CART), random forests and neural networks (Lee, Lessler, and Stuart, 2010; Westreich, Lessler, and Funk, 2010) for propensity score estimation. Thus we can use SL to estimate PS from models of different functional forms as well as the aforementioned non-parametric PS estimation algorithms.

3.3.2. Δ_C estimation

Let $Y_i(u)$ be the known accumulated cost up to time u and Y_i be the total cost that subject i accrues up to τ. Hence total cost $Y_i = Y_i(t_i)$ is not observed if $\delta_i = 0$, when a subject is censored before τ. In other words, we only observe Y_i for uncensored subjects. For censored subjects, their cost will continue to accrue hence their total cost Y_i is unknown. In this section, we introduce two DR estimators, the simple weighted and the partitioned based on Bang and Tsiatis (2000). The simple weighted estimator is appropriate when cost history information $Y(u), u < T_i$ is not available and we only know total cost $Y_i(T_i)$. The partitioned estimator is appropriate when we have access

to cost history information, for example, periodic insurance claims or monthly Medicare payment information.

DR for Δ_C estimation - simple weighted

When cost history data are not available, we propose the simple weighted DR estimator for cost estimation. This estimator is very similar to the Δ_E estimator discussed in section 3.3.1.

$$\widehat{\Delta}_C = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i \delta_i}{\hat{e}_i \hat{K}(T_i)} - \frac{(Z_i - \hat{e}_i) m_1(\mathbf{X}_i)\delta_i}{\hat{e}_i \hat{K}(T_i)}\right] - \left[\frac{(1-Z_i) Y_i \delta_i}{(1-\hat{e}_i) \hat{K}(T_i)} + \frac{(Z_i - \hat{e}_i) m_0(\mathbf{X}_i)\delta_i}{(1-\hat{e}_i) \hat{K}(T_i)}\right] \tag{3.4}$$

Thompson and Nixon (2005) suggested that conclusions from CE analyses are sensitive to choice of cost distribution. Hence, different cost models such as the normal, lognormal and gamma can be used as outcome models for m_0 and m_1. We then apply SL to incorporate all possible outcome models such as those mentioned in section 3.3.1. A proof of the DR property of the simple weighted estimator can be found in Li et al. (2015).

DR methods for Δ_C estimation - partitioned

When cost history information is available, we propose a partitioned DR estimator. This estimator is based on the partitioned total cost estimation method (Bang and Tsiatis, 2000). Specifically, the duration of interest $(0, \tau)$ is partitioned into L sub-intervals $(t_j, t_{j+1}]$, $j = 1, 2, \ldots, L$, $0 = t_1 < t_2 < \cdots < t_{L+1} = \tau$. Intuitively, we estimate the cost difference within each interval and "sum up" contributions from all L intervals.

$$\widehat{\Delta}_C = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}\left[\frac{Z_i Y_{ij} \delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)} - \frac{(Z_i - \hat{e}_i) m_1^j(\mathbf{X}_i)\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)}\right] - \left[\frac{(1-Z_i) Y_{ij} \delta_i^j}{(1-\hat{e}_i) \hat{K}_j(T_i^j)} - \frac{(Z_i - \hat{e}_i) m_0^j(\mathbf{X}_i)\delta_i^j}{(1-\hat{e}_i) \hat{K}_j(T_i^j)}\right] \tag{3.5}$$

where $Y_{ij} = Y_i(t_j) - Y_i(t_{j-1})$ is subject *i*'s cost accrued in the interval $(t_{j-1}, t_j]$. $T_i^j = \min(T_i^{t_j}, C_i)$ and $\delta_i^j = I(\min(T_i, t_j) \leq C_i)$, the censoring indicator for subject i at time t_j. $\hat{K}_j(T_i^j)$ are the KM survival estimates of $K_j(T_i^j)$ based on the data $(\min(T_i^{t_j}, C_i), 1 - \delta_i^j)$. $m_0^j(\mathbf{X}_i)$ and $m_1^j(\mathbf{X}_i)$ are the postulated models for the true regressions $E(Y_{ij}|Z_i = 0, \mathbf{X}_i)$ and $E(Y_{ij}|Z_i = 1, \mathbf{X}_i)$.

Similarly we use SL to estimate the postulated outcome models and PS model. For outcome models, we estimate within each interval. For example, the normal estimating equation for the pos-

tulated outcome models is $\sum_{i=1}^{n} \frac{\delta_i^j}{\hat{K}_j(T_i^j)}(Y_{ij} - \beta_k' X_i)X_i = 0$ and thus $\hat{\beta}_k = \left\{\sum_{i=1}^{n} \frac{\delta_i^j}{\hat{K}_j(T_i^j)} X_i' X_i\right\}^{-1} \sum_{i=1}^{n} \frac{\delta_i^j}{\hat{K}_j(T_i^j)} Y_{ij} X_i$. See Appendix A for a proof of the doubly robust property. Variance of $\widehat{\Delta}_C$ can be estimated using large sample theory to obtain the sandwich variance estimator or via non-parametric bootstrapping.

In monthly claims data, it is common to observe subjects with zero costs in specific months. Thus in addition to the outcome models mentioned in section 3.3.1, we propose the two-part model first introduced by Duan et al. (1983) to account for structural zeros:

- Part 1: a logit/probit model to model the probability of zero cost: $I(Y_{ij} = 0) \sim X_i\beta$.
- Part 2: a normal/lognormal/gamma model for positive costs: $f(Y_{ij}|Y_{ij} > 0) \sim X_i\beta$.

3.3.3. CE estimation

After we have estimated Δ_E and Δ_C according to the methods presented above, we can combine them to estimate common cost effectiveness measures:

$$\widehat{ICER} = \frac{\hat{\Delta}_C}{\hat{\Delta}_E} \tag{3.6}$$

and

$$\widehat{NMB}(\lambda) = \lambda\hat{\Delta}_E - \hat{\Delta}_C \tag{3.7}$$

For the CE acceptability curve, we can use bootstrapping to count the proportion of iterations with $\lambda\hat{\Delta}_E - \hat{\Delta}_C > 0$ for a range of λ values. The variance of NMB can be estimated directly as mentioned in section 3.2, although bootstrapping (Briggs, Wonderling, and Mooney, 1997) is preferred due to the complexity of the DR models.

3.4. Simulation studies

Using simulation studies, we first compare the two effectiveness estimation methods: inverse probability weighting vs. area under the survival curve. Then, we evaluate the performance of the DR models for Δ_E, Δ_C and NMB discussed in section 3.3 under various settings. We chose to focus on NMB over other CE measures because of its attractive statistical properties and popularity in modern CE analyses. We report the percentage bias (bias), the coverage probability (cvrg) of the

resulting 95% confidence interval and the empirical standard error (SE).

3.4.1. Δ_E estimation: area under the survival curve vs. inverse probability weighting

In section 3.3.1, we reviewed these two popular techniques for mean time estimation. To our knowledge, there are no studies in the literature comparing their empirical performance. In this study, exponential (mean=6), weibull (shape=2, scale=6) and uniform ([0,10]) survival times were simulated. Two levels of censorship were introduced with censoring time being uniformly distributed on [0,20] and [0,12.5], corresponding to 20% and 30% censoring, respectively. The average 10-year mean survival time $E(T), \tau = 10$ serves as the parameter of interest. Sample size n, was chosen to be 100 and 1000. 500 simulations were conducted and confidence intervals were constructed using non-parametric bootstrapping with the BCa correction. The simulation study settings were chosen according to those used by Bang and Tsiatis (2000).

Table 3.1: $E(T)$ simulation results: inverse probability weighting vs. area under the survival curve

			Inverse probability weighting			Area under the survival curve		
n	Censoring	Survival time	bias	SE	cvrg	bias	SE	cvrg
100	Light	Uniform	-0.006	0.309	0.952	-1.243	0.311	0.950
		Weibull	-0.297	0.280	0.944	-3.360	0.313	0.896
		Exponential	-0.327	0.371	0.964	-1.230	0.371	0.960
	Heavy	Uniform	-0.673	0.349	0.940	-1.337	0.331	0.942
		Weibull	-0.009	0.336	0.940	-1.758	0.326	0.926
		Exponential	-0.218	0.438	0.962	-0.891	0.395	0.958
1000	Light	Uniform	0.002	0.098	0.952	-0.116	0.098	0.948
		Weibull	-0.038	0.089	0.958	-0.408	0.091	0.954
		Exponential	0.037	0.118	0.942	-0.052	0.118	0.938
	Heavy	Uniform	-0.013	0.105	0.954	-0.114	0.105	0.950
		Weibull	0.163	0.097	0.946	-0.063	0.098	0.942
		Exponential	-0.046	0.125	0.926	-0.138	0.125	0.918

As expected, mean survival time $E(T)$ estimation using inverse probability weighting and area under the survival curve both yielded very small bias (Table 1) and comparably coverage. Note that for subjects with large observation time, if the estimated probability of censoring $\hat{K}(T_i)$ was zero, then $\min \hat{K}(T_i)$ was used instead to avoid the denominator being zero. Similarly, for the area under the survival curve, the estimated probability of the largest censored observation was underestimated. Thus, we see some downward bias for the empirical estimation of mean survival time. Results from Table 1 show that inverse probability weighting has comparable, if not slightly better empirical performance compared to the area under the survival curve method, especially with smaller sample sizes.

3.4.2. CE estimation

We developed simulation studies to represent CE analyses from observational data. We first simulated three covariates $\mathbf{X} = (X_1, X_2, X_3)$, where X_1 was binary with success probability 0.5. X_2 and X_3 were normally distributed with means of 2 and 1, respectively and common standard error of 1. Using these covariates, treatment choice Z was defined using a logit index model with $D \sim$ Bernoulli(p) and logit$(p) = 0.5 + X_1 + 0.25X_2 + 0.5X_3 - 0.5X_1X_2 - 0.25X_2X_3 - 0.5X_3^2$. The sample size was set to be 1000. The mean average five year Δ_E, Δ_C and NMB were chosen as parameters of interest.

We drew failure times from Weibull and exponential distributions. For exponential failure times, we set the mean to be $1/\exp(0.25 - Z + 0.5X_1 - 0.25X_2 - 0.25X_3)$. For Weibull, the shape parameter was 2 and the scale parameter was $\exp(-0.15 + Z - 0.5X_1 + 0.25X_2 + 0.25X_3)$. Censoring times were independently drawn from $U(0, 10)$ and $U(0, 6)$ for light and heavy censoring. The rate of censoring was approximately 20% for light censoring and 40% for heavy censoring.

For each subject, costs were generated for each month until the end of the study period τ. The total costs were generated from three different components: an initial diagnostic cost, an ongoing monthly cost and a cost accrued at the time of death if the subject's death was observed before τ. Four different cost distributions: normal, gamma, mixed and excess zeros, were generated as demonstrated in Table 2. For the excess zero cost model, proportion p_0 (logit$(p0) = -Z - 2X_1 - .5X_2 + .5X_3$) of the patients were assumed to experience zero cost each month.

Table 3.2: Simulation set up: four cost distributions

	Initial cost	Ongoing cost	Dying cost
Normal	N($\mu = 10 + 10Z + 5X_1 + 5X_2 + 5X_3$,5)	N($.1\mu$,.5)	N(0.2μ,.1)
Gamma	Gamma($\alpha = 2.5$, $\beta = \exp(-Z - .5X_1 - .2X_2 - .2X_3)$)	Gamma($\alpha, .1\beta$)	Gamma($\alpha, .2\beta$)
Mixed	a 50/50 mixture of normal and gamma		
Excess zero	N($\mu, 5$)	$P(p_0) = 0$, $P(1 - p_0) = N(.1\mu, .5)$	N(0.2μ,.1)

We first estimated Δ_E and Δ_C with a commonly used in practice approach that we refer to as a "conventional" approach, in which Δ_C was estimated from a linear regression and Δ_E was derived using area under the survival curve. We then estimated Δ_C and Δ_E using inverse probability of treatment weighting (IPTW) based on propensity scores with censoring correction $\frac{\delta}{K(T)}$ (Li et al.,

2015). Propensity score (PS) models were either correctly specified or mis-specified. In the correctly specified case, covariates and their correct functional forms were included in the logistic regression model; in the misspecified case, only the main effects X_1, X_2, X_3 were included. Then, our proposed DR model for Δ_E and our two DR models, simple weighted and partitioned for Δ_C were applied. Propensity scores were estimated utilizing the Super Learner algorithm. Candidate PS models included a logistic regression model, a logistic regression model with all interactions and two non-parametric algorithms: generalized additive model (GAM) and k nearest network (KNN). Similarly the outcome parameters in all DR models were estimated using the Super Learner algorithm in conjunction with linear regression, GLM, generalized gamma model and the two parts model. We estimated NMB with a standard willingness to pay of $\lambda = 50,000$/yr.

Results from Table 3 show that the conventional area under the survival curve method yields biased estimates because it failed to account for confounders. IPTW method was unbiased; but produced biased estimates (-13.2 % to 15.0%) when the propensity score model was mis-specified. In addition, even when the propensity score model was correct, IPTW had large standard errors and inflated coverage (0.96-0.99). The proposed DR method worked well and had very small bias (-0.23% to -0.05 %), small standard error and good coverage. DR model was also robust to propensity score model misspecification since propensity scores were estimated using several parametric and non-parametric algorithms.

For Δ_C estimation, results show that conventional linear regression produced biased estimates under all scenarios. IPTW yielded unbiased estimates with large standard errors and failed when the propensity score model was mis-specified. Both simple and partitioned DR had negligible biases ranging from (-0.036% to 0.349%). The two also had similar standard errors and their coverage probabilities were around 95%. Note that even for the excessive zero cost structure, the two DR methods had similar performance. This result again confirms the doubly robust property of the proposed estimators since the propensity scores were estimated correctly using Super Learner.

For NMB estimation, the conventional method produced very biased estimates (-278% to 55%). Estimates from IPTW exhibited the same problem with large standard errors, inflated coverage and biased results when the propensity score model was mis-specified. Our proposed DR methods had superior performance compared to all other approaches.

Table 3.3: Simulation results: cost, effectiveness and NMB estimation

Effectiveness Δ_E

		Conventional			IPTW			DR		
PS	Time dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	-21.061	0.09	0.028	-0.319	0.578	0.99	-0.065	0.082	0.952
	Exp	-21.574	0.104	0.198	-1.024	0.332	0.964	-0.209	0.107	0.948
Mis	Weibull	-20.901	0.097	0.026	-13.192	0.1	0.372	-0.065	0.082	0.952
	Exp	-21.09	0.112	0.304	-15.026	0.124	0.634	-0.209	0.107	0.948

Cost Δ_C

			Conventional			IPTW			DR - simple			DR - partitioned		
PS	Time dis	Cost dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	normal	-38.134	4.545	0	0.404	44.871	0.99	0.268	4.021	0.938	-0.131	4.025	0.95
		mixed	-34.142	3.779	0	-1.845	18.147	0.964	-0.121	3.433	0.958	0.064	3.428	0.95
		gamma	-29.046	3.386	0	0.912	33.183	0.988	-0.354	3.132	0.956	0.349	3.207	0.958
		zero	-32.713	4.234	0	-1.702	25.871	0.968	-0.036	3.809	0.95	-0.092	3.803	0.938
Mis		normal	-38.134	4.545	0	-15.192	5.771	0.204	0.268	4.021	0.938	-0.131	4.025	0.95
		mixed	-34.142	3.779	0	-14.071	4.623	0.198	-0.121	3.433	0.958	0.064	3.428	0.95
		gamma	-29.046	3.386	0	-13.145	3.966	0.25	-0.354	3.132	0.956	0.349	3.207	0.958
		zero	-32.713	4.234	0	-13.93	4.988	0.182	-0.036	3.809	0.95	-0.092	3.803	0.938

NMB

			Δ_E, Δ_C :Conventional			Δ_E, Δ_C :IPTW			Δ_E:DR,Δ_C:simple DR			Δ_E:DR,Δ_C:partitioned DR		
PS	Time dis	Cost dis	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg	bias	SE	cvrg
Correct	Weibull	normal	-119.442	3.037	0.000	3.269	16.402	0.988	1.282	1.559	0.950	-1.007	1.550	0.946
		mixed	-278.966	2.749	0.000	-2.363	3.279	0.970	-2.893	1.328	0.952	0.603	1.431	0.954
		gamma	54.670	2.906	0.624	1.718	6.026	0.988	4.718	1.693	0.952	-1.671	1.659	0.962
		zero	-107.682	3.300	0.008	-8.007	7.858	0.964	-0.232	2.123	0.948	-0.639	2.109	0.964
Mis		normal	-119.442	3.037	0.000	-22.820	1.980	0.432	1.282	1.559	0.950	-1.007	1.550	0.946
		mixed	-278.966	2.749	0.000	-32.122	1.469	0.804	-2.893	1.328	0.952	0.603	1.431	0.954
		gamma	54.670	2.906	0.624	-11.882	1.693	0.900	4.718	1.693	0.952	-1.671	1.659	0.962
		zero	-107.682	3.300	0.008	-18.326	2.401	0.804	-0.232	2.123	0.948	-0.639	2.109	0.964

3.5. Lung cancer surveillance data

Lung cancer is responsible for the largest number of cancer-related deaths worldwide (Siegel, Naishadham, and Jemal, 2012). In addition, the overall economic burden of lung cancer on society is large and growing (Goodwin and Shepherd, 1998). Patients that undergo curative resection for lung cancer are at risk of developing a recurrence or a new primary lung cancer in the future. Therefore, imaging surveillance has become standard of care after lung surgery. Currently, the two most common approaches to surveillance are use of chest X-ray or chest CT. The optimal surveillance strategy is unknown; there are no randomized trials that have directly compared the effect of imaging strategy (CT vs. X-ray) on overall survival following lung cancer resection. Here we apply our CE estimation approach to compare the three year cost-effectiveness of chest X-ray versus CT using a cohort of patients derived from the SEER-Medicare registry.

We included stage I-IIIA non-small cell lung cancer patients diagnosed between 2007 and 2009 and treated with curative intent surgery. See Ciunci et al. (2015) for a detailed description of inclusion/exclusion criterion. 59.1 percent of the study cohort were censored at the end of the study. Payment data were extracted from Medicare claims from the inpatient MEDPAR, outpatient SAF and non-institutional Carrier files covering 2007 through 2010. For each patient, we calculated total spending as the sum of payments made to the provider by Medicare, the patient, and other payers. Payments were calculated in consecutive 30-day periods starting 181 days after the surgery index date and lasting until patient death or censoring (December 31, 2010). We did not adjust for inflation due to the short time span covered in this study (2007 to 2010). The final cohort sample size was 3389; 1058 of whom had chest X-ray and 2331 had CT surveillance. Three year total cost was highly right skewed, with a maximum observed cost of $722,100. The average observation duration was 22 months.

In this study, both the choice of surveillance strategy and three year cost may have been influenced by covariates such as age, sex, median income, marital status, Charlson score, histology, chemotherapy and radiation. We first estimated CE measures, including ICER and NMB, using the “conventional” method, where Δ_C is estimated from a linear model and Δ_E is derived using area under the survival curve. We then estimated Δ_C and Δ_E using IPTW based on propensity scores. Lastly, DR models proposed in section 3.3.1 and section 3.3.2 were applied. Propensity

scores were estimated utilizing the Super Learner algorithm. Candidate PS algorithms included a logistic regression model, a logistic regression model with all interactions, GAM and KNN. Similarly, the regression parameters in the DR model were estimated using the Super Learner algorithm in conjunction with linear regression, GLM and generalized gamma models. Approximate confidence intervals for ICER and NMB (WTP=$50,000/yr) were constructed using non-parametric bootstrapping with BCa correction.

Table 3.4: CE analysis of CT vs. X-ray (reference) for lung cancer surveillance

			NMB (WTP=50,000)	
Method	$\Delta_E(mths)$	$\Delta_C(\$)$	estimate	95% CI
Conventional	3.14	410	156,482	5,939, 19,359
IPTW	3.12	-2,539	158,390	7,330, 23,455
DR	3.65	-3,512	185,990	11,490, 27,472

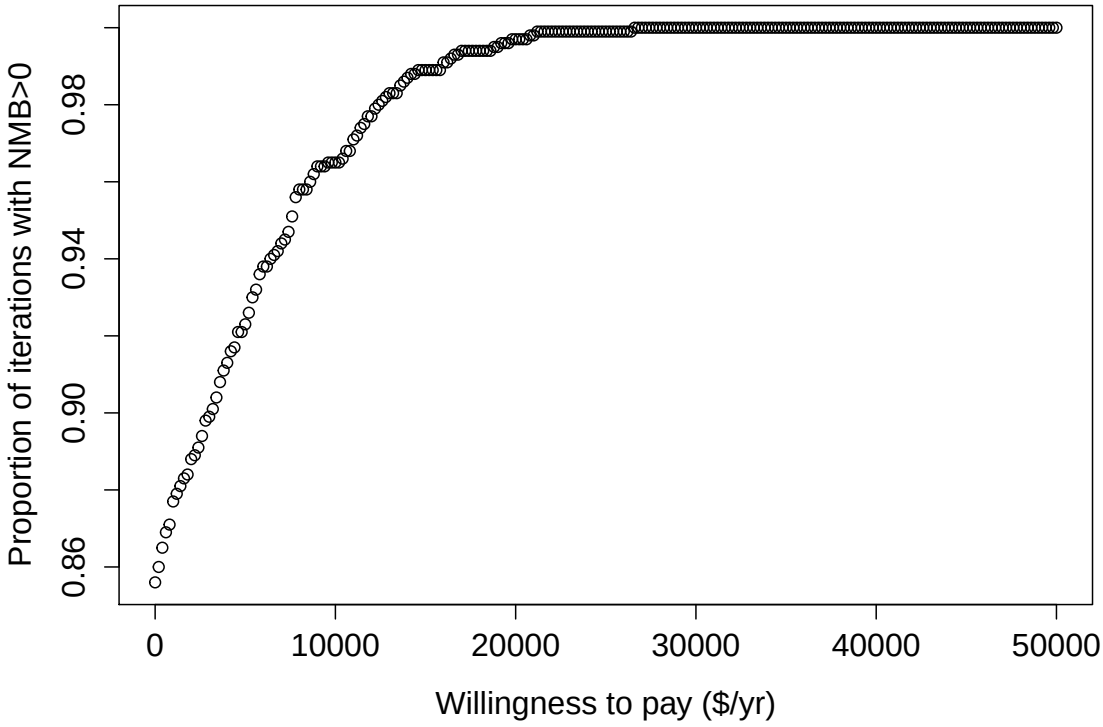
From Table 4, patients on CT were estimated to live on average 3.12 to 3.65 months longer than patients on X-ray. This result is consistent with Ciunci et al. (2015) that demonstrated CT is associated with lower hazard of death. Δ_C were vastly different between the "conventional" method and IPTW or DR. Although all three approaches suggest that CT is significantly more cost-effective than X-ray, the DR approach produced a much higher NMB estimate, indicating that CT is notably more cost-effective. In addition, DR yielded tighter 95% confidence interval than IPTW.

In Figure 3.1, we plotted the CE acceptability curve from bootstrapped samples under a wide range of WTP values. This figure provides a visual demonstration of when CT becomes significantly more cost-effective compared to X-ray. We see that around $\lambda = \$8,000$/yr, over 95% of the bootstrap iterations yield positive NMB. In other words, CT was significantly more cost-effective compared to X-ray with a WTP of more than $\$8,000$/yr.

3.6. Summary

In policy making and health services evaluation where an emphasis is placed on estimating not only the effectiveness but the cost-effectiveness of interventions, it is imperative to estimate CE measures accurately and robustly. We propose DR models based on propensity scores to estimate the ICER and the NMB from censored observational data. These models draw on the strengths of propensity score weighting and outcome regression fitting utilizing machine learning algorithms. Thus, we have demonstrated the merit of both causal inference models and modern machine learn-

Figure 3.1: CE acceptability curve of CT vs. X-ray



ing approaches in CE analysis. We note that the partitioned DR Δ_C estimator, although theoretically more efficient than the simple weighted one, is more computationally intensive. Hence, we suggest using the simple weighted estimator. With smaller sample sizes, the partitioned DR may perform better, but further investigation is needed.

As in any observational study, unobserved or hidden bias may be of concern. Hence, in addition to DR based CE analysis, we suggest conducting sensitivity analyses to assess the effect of unmeasured confounders on the treatment effect (Handorf et al., 2013).

Acknowledgment

We used the linked SEER-Medicare database and acknowledge the efforts of the Applied Research Program; National Cancer Institute; Office of Research, Development and Information; Centers for

Medicare and Medicaid Services; Information Management Services; and SEER program tumor registries in the creation of the SEER-Medicare database.

CHAPTER 4

MODERN STATISTICAL AND MACHINE LEARNING APPROACHES FOR HEALTH CARE COST ESTIMATION FROM BIG DATA

4.1. Introduction

With the rise of big data, the role of machine learning in economics has gathered attention (Varian, 2014). Economists, especially econometricians have employed these modern techniques for model building, prediction and model selection to various economics research questions (Ahmed et al., 2010; Ghose, Ipeirotis, and Sundararajan, 2007; Scott and Varian, 2013). However, machine learning approaches has not been commonly used to answer health economics questions. In this paper, we review big data and machine learning techniques and their potential application to health care cost estimation, with an emphasis on providing statistical insights underlying each of these state-of-the-art approaches.

In the era of big data, health care cost related data have grown exponentially. Traditionally cost data could be stored and manipulated on spreadsheets or using a Structured Query Language (SQL). However, these tools are inadequate for massive data which require special programing paradigms. Some popular big data storage and manipulation algorithms include Hadoop File Distribution System (Lam, 2010; Shvachko et al., 2010; Venner, 2009) and MapReduce (Dean and Ghemawat, 2008).

Cost prediction, which is of great interest in health economic evaluations (Folland, Goodman, Stano, et al., 2007), typically presents challenges because of the skewness and heterogeneity inherent in cost data. Historically, researchers used parametric models such as ordinary least squares regression, generalized linear regression, and other parametric models (e.g. Weibull, Gamma) with different transformations (e.g. log, Box-Cox) and different variance functions(Ash et al., 2001; Manning, 1998; Montez-Rath et al., 2006). For cost prediction, previous studies have evaluated the performances of various cost prediction models (Basu and Rathouz, 2005; Dodd et al., 2006). In recent studies, machine learning non-parametric algorithms have been shown to have better predictive ability (Bertsimas et al., 2008; Kim, An, and Kang, 2004; Sushmita et al., 2015). In this paper,

we review some popular machine learning prediction algorithms such as classification and regression trees, random forest, supporting vector machines, boosting and Bayesian additive regression trees.

In addition, variable selection can be challenging in cost estimation models. We often have many potential predictors that may need to be narrowed down for model building. Traditionally, researchers use stepwise regression and model complexity measures such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to select important variables. With the rise of big data, we see more and more large datasets with numerous potential predictors where modern dimension reduction methods may serve as important tools in estimating cost. Popular dimension reduction tools such as principle components analysis and Lasso combine the strength of statistical modeling with machine learning.

We demonstrate the application of these cutting edge big data and machine learning approaches using a cohort of lung cancer patients derived from SEER-Medicare. Lung cancer causes the largest number of cancer-related deaths worldwide (Siegel, Naishadham, and Jemal, 2012) and is the second most frequently diagnosed cancer in the United States (Disease Control and Prevention. 2014). The overall economic burden of lung cancer and the associated cost of treatment on society is large and growing (Goodwin and Shepherd, 1998). In the United States, per-patient total health care costs were estimated to be US$34,191 to $47,941 for lung cancer patients following diagnosis for different populations (Baker et al., 1991; Fireman et al., 1997; Hillner et al., 1998; Kutikova et al., 2005). In this study, we use a group of patients derived from the SEER-Medicare registry to demonstrate the merits of big data and machine learning algorithms. Specifically, we included all non-small cell lung cancer patients diagnosed between 1998 and 2009. Payment data were extracted from Medicare claims from the inpatient MEDPAR, outpatient SAF and non-institutional Carrier files covering 1998 through 2010. For each patient, we calculated total spending as the sum of payments made to the provider by Medicare, the patient, and other payers. Payments were calculated in consecutive 30-day periods starting 181 days after the surgery index date and lasting until patient death or December 31, 2010. We adjusted for inflation according to consumer price index (Economic Analysis, 2015).

The goal of this paper is to elucidate big data and machine learning tools that can be used for cost estimation and prediction. We provide statistical insight underlying behind these “black box”

algorithms. Section 2 introduces big data storage and manipulation tools and section 3 reviews several popular machine learning prediction algorithms. In section 4 we discuss dimension reduction and variable selection. Lastly, Section 5 touches on some new topics and unanswered questions. Throughout the paper, we use the lung cancer cost example to demonstrate these tools.

4.2. Big data storage and manipulation

With the rise of computing capabilities, there has been an explosive growth of health care cost data and therefore a move from data that fits in a spreadsheet to data that lives on multi-server databases (Einav and Levin, 2013). Examples of large health cost care data includes national claims databases, patient level bills for large medical centers and insurance claims data for insurance and reinsurance companies. Thus, there is a pressing need to store and handle such large volumes of data. In this section, we introduce some popular big data storage and manipulation algorithms such as MapReduce and Hadoop.

Big data storage often requires specialized hardware infrastructure and storage mechanisms. In order to achieve consistency and availability in storing large volumes of data, these storage mechanisms need to be more primitive and less flexible than relational databases such as those that can be accessed using a Structured Query Language (SQL) (Chen et al., 2014). The most popular storage mechanism is the Hadoop Distributed File System (HDFS) from Apache Hadoop. HDFS is an open source, distributed database processing platform designed to store big data across several thousands nodes (Lam, 2010; Shvachko et al., 2010; Venner, 2009) and was inspired by Google's File System (Shafer, Rixner, and Cox, 2010). To see how HDFS works, image that we need to store a file that contains all insurance claims from all 50 states in the past 24 months; the claims from different months might be stored on different servers. All servers together constitute the entire claims data file. As the number of servers increases, server failures will be inevitable. Thus, a major component of HDFS has to do with dealing with such failures. HDFS can divide and replicate data into multiple pieces (the default is three) to be stored on different servers. This redundancy means higher availability when a server fails. In reality, it is often cheaper to rent data storage clusters from cloud computing providers such as Google, IBM, Amazon, rather than to build and maintain a data storage system.

Next, we need a programming paradigm to handle analytics across hundreds or thousands of

servers. MapReduce (Dean and Ghemawat, 2008), originally proposed by Google, is one of the most popular programming models that handles big data analytics. MapReduce Performs two different tasks, the Map task and Reduce task. In the Map stage, the query is "mapped" to the servers and is then applied in parallel to the different components of the data. The partial calculations are then combined or aggregated ("Reduced") to create the summary statistics of interest (Chu et al., 2007; Dean and Ghemawat, 2010). Figure 4.1 demonstrates the "divide and conquer" idea behind the MapReduce algorithm. For example, if we are interested in finding out the maximum insurance claims from each state, then mappers would work in parallel and summarize the maximum claims by state from each month. After processing, the reducer receives the summary statistics from all mappers and then calculates the maximum claims by state. One key feature of MapReduce is fault-tolerant; a master server oversees the entire procedures and each slave server periodically reporting its status to the master server. If a node fails, the master server reassigns that piece of the job to other available servers. Slave servers work in parallel and thus save computing time. In addition, most of the computing takes place on servers with data on local servers so that network traffic is reduced.

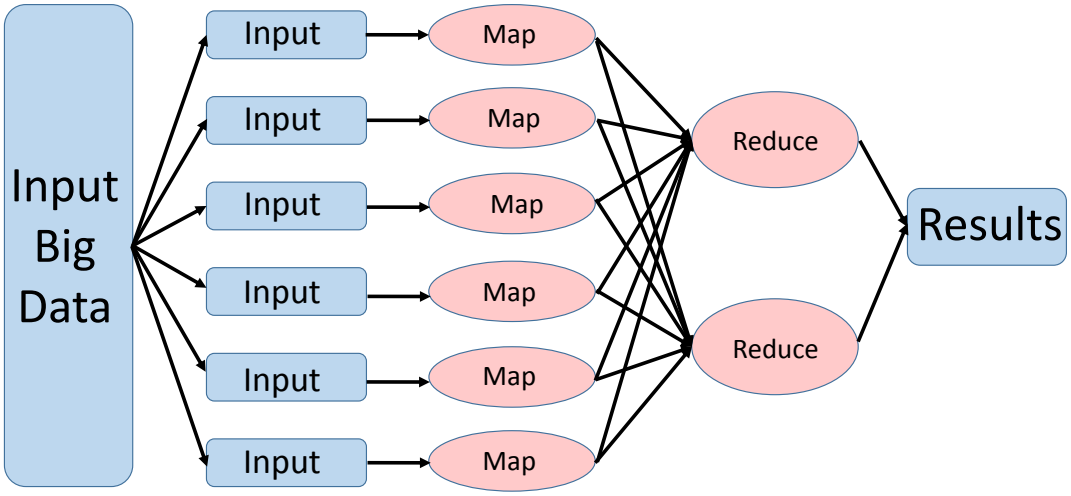


Figure 4.1: MapReduce paradigm

Lastly, we can also connect data stored in Hadoop using Excel 2013 (Hortonworks, 2016). Hortonworks, part of Apache Hadoop, provides an option to access big data stored in their Hadoop

platform using Excel 2013. One can use Power View feature of Excel 2013 to easily summarizes and visualize large data which is otherwise not accessible by Excel (Mohammed, Far, and Naugler, 2014).

4.3. Cost prediction

Health care cost prediction has been of interest to health economists and policy makers. While economists and statisticians are generally looking to draw inference; machine learning specialists are more concerned with developing algorithms with high predicting power. Historically, researchers used parametric models such as ordinary least square regression, log normal and gamma models (Ash et al., 2001; Diehr et al., 1999; Manning, 1998; Montez-Rath et al., 2006) with or without transformation to model health care cost and then make predictions. However, machine learning based non-parametric algorithms has been shown to have better predicting abilities (Bertsimas et al., 2008; Sushmita et al., 2015). In this section, we review some popular machine learning algorithms and demonstrate how they can be used in cost prediction.

Assume we have a training sample of n observations; the outcome of interest is $Y = \{Y_1 \dots Y_n\}$ and Y can be continuous or discrete. The p predictor variables are $X_1, \dots, X_p$. Our goal is to build a model for predicting Y from X and later use this model to predict Y from new X values called the testing sample.

4.3.1. Tree based models

Tree based models including Classification and Regression Trees (CART), pruned CART and random forest are some of the most widely used method in machine learning (Michie, Spiegelhalter, and Taylor, 1994). All tree based models are based on stratifying or segmenting the predictor space X into several simple regions (James et al., 2013).

CART. The idea of CART is intuitive: we divide the predictor space R into L disctinct and non-overlapping partitions and each terminal node in the tree represents partition R_l (Breiman et al., 1984). For each observation that falls into that terminal node and thus the associated partition, we make the same prediction which is the mean of the Y values for the training observations in R_l for continuous Y and the majority of Y categories for categorical Y (Loh, 2011). At each step, we grow two more new branch further down on the tree and thus make more partitions. This process

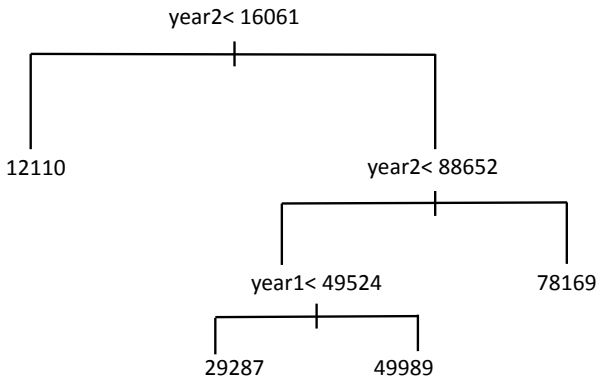
is repeated recursively until a stopping criterion such as each terminal nodes has fewer than some pre-determined minimum of observations is reach. Here we provide the idea behind CART:

1. We take a top down approach and start from the root node
2. For each X_i, find the cut off point s_i such that the resulting tree yields the minimal risk score. Consider all X_is, choose the X_i and its associated cut off s_i that has the overall small risk score.
3. If a stopping criterion is reached, then stop. Otherwise, repeat step 2.

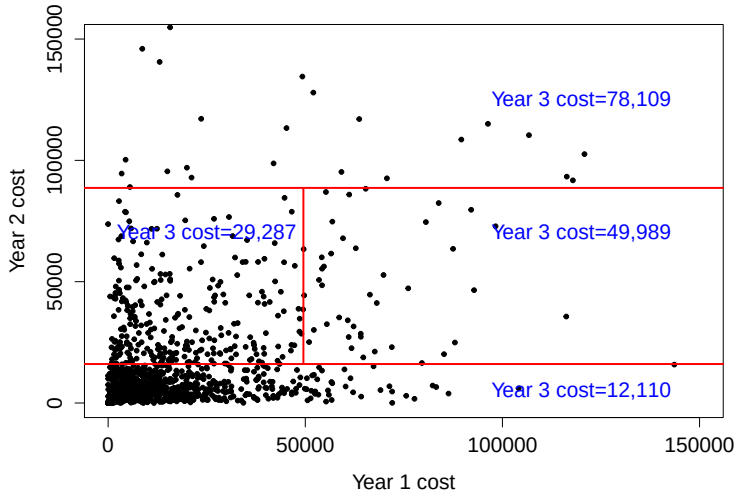
Other tree based algorithms such as C4.5 follow a similar idea (Quinlan, 1996). For categorical Y (classification tree), C4.5 uses entropy to calculate its risk function $R(X, Y)$, where CART uses the Gini index. For continuous Y (regression tree), root mean square error is used as the risk function for both. Since we can grow a very large tree with excellent in sample prediction by changing the stopping criterion, one obvious disadvantage of these tree based models is over-fitting. Many algorithms have been proposed to overcome the issue of over fitting such as pruning (Quinlan, 2014), where we first grow a large tree and then apply cost complexity pruning to the large tree. A common restriction is the number of terminal nodes T, thus changing its risk function to $R(X, Y) + \alpha T$, α can be determined using cross-validation to provide the best out of sample prediction. CART is available as R packages `rpart` and `tree`.

Classification and regression trees have nice graphical representations. Here we present an over-simplification of the regression and classification trees with only two X variables. For the lung cancer patients who were diagnostic in 2007, we want to use their year 1 and year 2 cost to predict their year 3 cost. Figure 4.2a shows regression tree results obtained from using the `rpart` package in R. At each terminal node, the estimated year 3 cost is simple the average of all year 3 cost that falls into the terminal code in the training set. Figure 4.2b depicts the corresponding plot where the predictor space formed by year 1 and year 2 costs are partitioned into four different regions according to the regression tree.

Random Forest. Random Forest builds on the idea of CART. Instead of constructing a single classification or regression tree, Random Forest grows many de-correlated trees to correct for over-fitting (Breiman, 2001; Liaw and Wiener, 2002). We then combine all the trees and form a



(a) Graphic representation of CART



(b) Regression tree of year 3 cost

"forest". Thus, random forest is a representation of bagging (bootstrap aggregation), where we build an ensemble model by combining many different, often weaker, models. The Random Forest idea can be summarized as:

1. A number $q \approx \sqrt{p}$ is specified.

2. Sample n cases from the original training set $\{X, Y\}$ with replacement, $\{X', Y'\}$

3. Grow a tree according the new training set $\{X', Y'\}$ to the largest extent possible. At each node, q variables of X are selected at random to grow the next brunch. A fresh sample of q variables of X is taken at each node but q is held constant during the forest growing

4. Repeat step 2-3 to get a forest of many trees.

For a new observation, run all trees and use majority vote or averaging to get the final prediction. The idea of random forest is to decrease correlation between any two trees in the forest while maintain the strength of each individual tree. Since we only consider a small subset of the predictors at each node, the trees in the forest are less likely to be correlated. In addition, we can calculate the information such as decrease of accuracy in predictions for out of bag samples at each node and average over all nodes in all tress we find the average information for each variable X_j. Therefore, we can rank all variables according to their average information to obtain the importance of all X (Criminisi, Shotton, and Konukoglu, 2012). One disadvantage of random forest is the lack of simple and intuitive summaries of relationships in the data. Unlike CART, we cannot visualize the "forest". Random forest is available as R package `randomForest`.

4.3.2. Support Vector Machines (SVM)

Support vector classifiers are based on the idea that we can use a class of hyperplanes to "separate" outcome Y based on predictors X (Suykens and Vandewalle, 1999). For a p dimensional space, hyperplane is a $p-1$ dimensional subspace. In our earlier example, the predictor space $X = (X_1, X_2)$ is two dimensional where X_1 is the year 1 cost and X_2 is year 2 cost. In this case, a hyperplane is simply a line defined by $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. If the predictor space can be completely separated, then we want to maximize the margin M, the distance between the hyperplane the points that are perpendicularly closest (support vectors) to the hyperplane. If we wish to predict whether patients are considered high cost with year 3 cost $Y \geq 100,000$, we essentially want to draw a hyperplane that separates the predictor space X into two regions corresponding to year 3 high cost and low cost subjects. In practice, a single separating hyperplane usually does not exist. In Figure 4.2, no single line can separate year 3 low and high cost patients. For a p dimensional space, the SVM classifiers are hydroplanes defined as $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0$. In order for these hyperplanes to identify binary Ys, they should satisfy $y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) > 0$ for all i.

Machine learning researchers extend the concept of separating hyperplane to develop a hyperplane that "almost" separate different classes using a "soft margin" (Cristianini and Shawe-Taylor, 2000). In other words, we want to maximize margin M while allowing misclassification. Thus, support

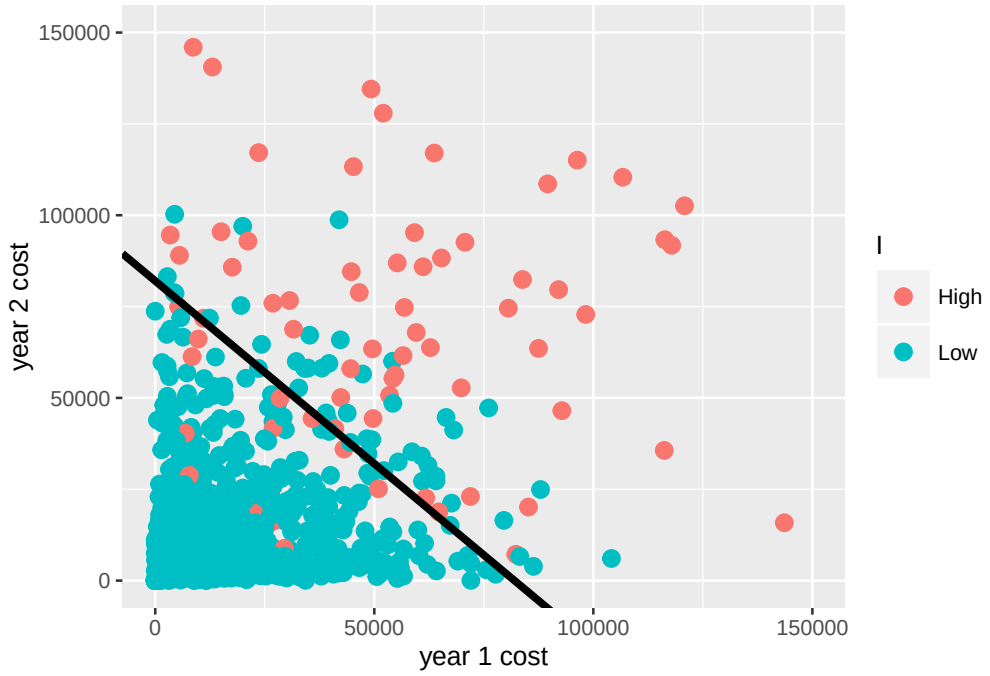


Figure 4.2: Separating year 3 high and low cost patients

machine classifiers are simply the solution to maximize M conditional on:

$$\sum_{j=1}^{p} \beta_i^2 = 1$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) > M(1 - \epsilon_i)$$
$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

where ϵ_i are slack variables that allow individual points to be on the wrong side of the hyperplane and C is a non negative tuning parameter. Loosely speaking, the first two equations defined the supporting vectors and the separating margins the last equation allows for observations to violate the margin. The tuning parameter C controls for the amount of violation allowed. In practice, C is generally chosen by cross validation (Hsu, Chang, Lin, et al., 2003).

Moreover, SVM is an extension of support vector classifiers. Instead of linear boundaries like in the one in Figure 4.2, SVM deals with non-linear class boundaries by incorporating higher order terms of X. For example, we can include all second order terms of X and our predictor space would have dimension $2p$: (X_i, X_i^2). And we can modify the equations for supporting vector classifiers, thus

SVM is defined as the solution to maximize M conditional on:

$$\sum_{j=1}^{p} \beta_{j1}^2 + \beta_{j2}^2 = 1$$
$$y_i(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2) > M(1 - \epsilon_i)$$
$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Similarly, SVMs can work with continuous outcome or categorical outcome with more than two categories. One big advantage of SVM is flexibility in the choice of the forms of the threshold separating hyperplanes (Auria and Moro, 2008). In stead of incorporating higher order spaces of X, one can also use any generalized kernel function. Similar to random forest, SVM is a "black box" while we cannot visualize the relationships in the data. SVM is available as R package `e1071`.

4.3.3. Boosting

The intuition beyond boosting is simple: we combine weak learners that usually have low variance and do not over-fit to produce a strong learner (Schapire, 1990, 1999). Thus we draw strength from many weak learners that are good at different parts of the predictor space. The key of boosting lay in two areas: how to weight output from different weak learners and how to force weak learners to learn about different parts of the predictor space. To answer these questions, all boosting algorithms builds on the idea of weighting misclassified observations in such a way that they get properly classified in future iterations (Schapire, 2003).

Similarly, in statistical modeling we often face the issue that some data points are more important than the others as they affect our model and thus predictions more. Boosting takes consideration of this by targeting these important observations. All boosting algorithms follow the same general framework:

1. start by assigning an equal distribution $D_1(i) = 1/n$ to all data points
2. at iteration l, train base learner with distribution D_l
3. get predictions from base learner. Update the weights D_{l+1} by how incorrectly it was predicted

4. train next base learner with distribution $D_l + 1$

5. Report stem 2-4, the final is a linear combination of the predictions of the different learners weighted by their strength.

In the example of trees, we grow small trees with no pruning and then improve based on that. Boosting is another example of "slow learning" where we fit a tree using the current residuals so that we improve fitting in areas where it does not perform well (James et al., 2013). In other words, these algorithms learns and improves from previous trees errors and finally make a weighted sum of all the trees.

Popular boosting algorithms include Gradient Boosting (Friedman, 2001) and AdaBoost (Freund and Schapire, 1997). Boosting is a very popular choice on the statistical and machine learning modeling competition site Kaggle. Although boosting often has stellar prediction performances, it is less intuitive and even more of a "black box". Adaboost is available as R package `Adabag` and gradient boosting as `gbm`.

4.3.4. Bayesian additive regression trees (BART)

BART builds on the idea of Bayesian CART where we place CART within a Bayesian framework by specifying a prior on tree space (Chipman, George, and McCulloch, 1998, 2002; Denison, Mallick, and Smith, 1998). The intuition behind BART is to form a probability distribution over the space of possible trees explored using Markov Chain Monte Carlo methods. If we let the true model to be $y(x) = f(x) + \epsilon, \epsilon \sim N(0, \sigma^2)$, then a CART model says $y(x) = g(x; T, M) + \epsilon$ where T denotes the tree structure and $M_T = (\mu_1 \ldots, \mu_b)$ represents parameters at the b terminal nodes. Then in a Bayesian framework we have $P(T, M, \sigma^2) = P(M|T, \sigma^2)P(T|\sigma^2)P(\sigma^2)$ where the three components model the tree structure itself, the terminal node parameters given the tree structure and the error variance which is independent of the tree structure and terminal node parameters respectively.

Compare to Bayesian CART, BART models the outcome Y by adding many regression trees, thus the name additive regression trees. Thus our regression tree model becomes
$y(x) = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon$ where $g(x; T_j, M_j)$ represents one of m regression trees. And in Bayesian framework, we have $P((T_{1\ldots m}, M_{1\ldots m}), \sigma^2) = \sum_{j=1}^{m} P(M_j|T_j, \sigma^2)P(T_j|\sigma^2)P(\sigma^2)$. Here

we allow different forms for $\mu_i(x)$ such as constant (as seen in CART), linear and Gaussian process Denison, Mallick, and Smith, 1998; Gramacy and Lee, 2012; Wu, Tjelmeland, and West, 2007. Furthurmore, a Gibbs sampler is used to generate draws from the posterior distribution of $P((T_{1...m}, M_{1...m}), \sigma^2|y)$ according to the following:

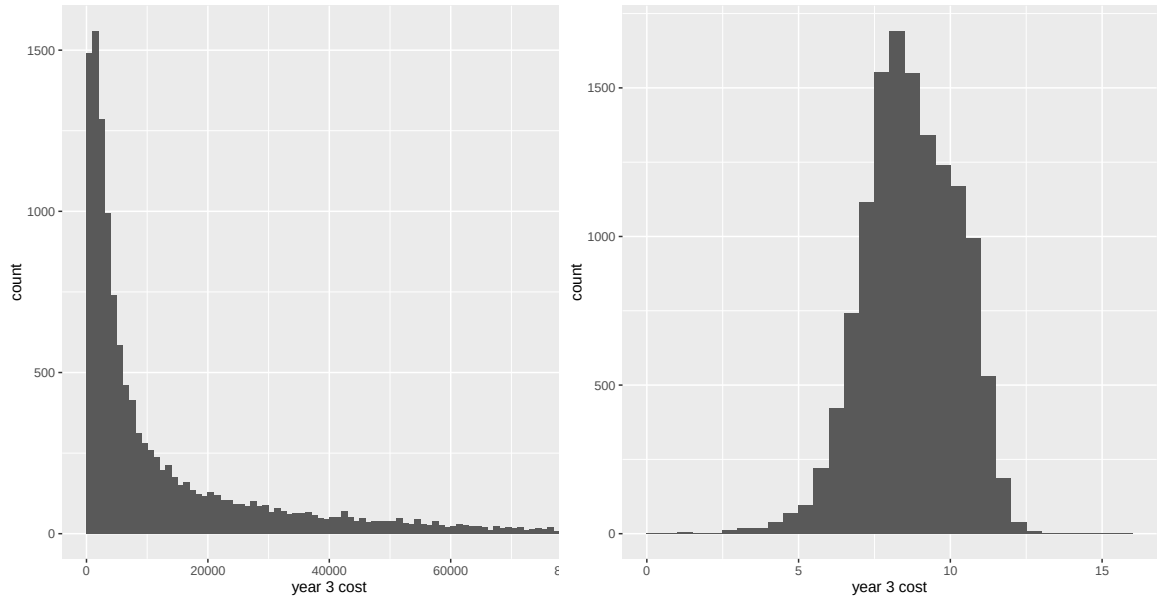
1. draw $T_i|R_i, \sigma$ (Metropolis-Hastings step)
2. draw $M_i|T_i, R_i, \sigma^2$ (Gibbs step)
3. repeat step 1-2, and draw $\sigma|T_{1...m}, M_{1...m}$ (Gibbs step)

Specifically, drawing of $T_i|R_i, \sigma$ (step 1) involves introducing a small perturbations that grows a terminal node, pruning two nodes or change a splitting rule. Therefore, we can get multiple tree realizations of one tree by introducing perturbations at each iteration; we average over the posterior to form predictions. BART has shown to have excellent prediction performance. However it is relatively slow due to its MCMC nature and somewhat less intuitive to a layperson. BART is available as R package `BayesTree` and `bartMachine`.

4.3.5. Lung cancer cost prediction

In this section, we demonstrate how the machine learning algorithms introduced can be used in cost prediction and compare their performances to traditional parametric cost prediction models. Our data set includes 13,063 patients with at least three years of complete cost information 181 days after the surgery index data. We divide our data into learning (2/3) and testing (1/3) cohorts. The learning cohort is used to build and calibrate our prediction models while the testing set is used to evaluate the performance of the various models.

We are interested in predicting patients' third year cost from their baseline demographic, medical and past cost information. Baseline demographic variables include age at diagnosis, gender, race, marital status, year of diagnosis, median income of primary residence at zip code level and reporting sources. In addition, we include medical and treatment related variables including Charlson commodity score, stage, histology, primary surgery site. Lastly, we include year 1 and 2 cost, the maximum monthly cost and the number of months with above average costs in the previous two years.



(a) Histogram of year 3 cost

(b) Histogram of log year 3 cost

Figure 4.3a shows that outcome of interest, year 3 cost is highly right skewed with a maximal observed cost of $443,278. Log transformation of year 3 cost Figure 4.3b results in a somewhat normally distributed shape. To make meaningful comparison, we incorporate baseline methods including linear regression with log transformed year 3 cost and the generalized gamma model with log link and gamma distribution of the error terms. All 56 predictor variables are used for these baseline methods.

We measure the performance of prediction models with two error measurements : mean square error (MSE) and mean absolute error (MAE). The latter, although often used in published cost prediction studies (Bertsimas et al., 2008; Diehr et al., 1999; Moran et al., 2007), is not as a strong indicator for model predictability as MSE. Prediction results from the test cohort is summarized in Table 4.1 where mean and se represents the mean and standard errors of the predictions.

	Mean	SE	MSE	MAE
Linear regression - logged	8801	27757	30181	13102
Generalized gamma model	16380	26287	27664	14557
CART	15876	15606	25305	14577
Random forest	16518	11345	23046	14275
Svm	10073	10545	24447	12590
Gradient Boosting	15734	12264	23033	13902
BART	16433	12804	24262	14271

Table 4.1: Predicted year 3 cost and model performance indices

Results from Table 4.1 shows that linear regression with logged year 3 cost provides poor prediction while generalized gamma model has better. This is not surprising given the right skewed distribution of year 3 cost; and is consistent with past studies Dodd et al., 2006; Moran et al., 2007. All machine learning algorithms show improved prediction and yield 15% to 33% reduction in MSE compared to linear regression. Among the five popular machine learning models, random forest and gradient boosting machines have the best prediction. Such result is due to the ensemble nature of both models where we combine strength of many weak learners. Overall, machine learning algorithms provide accurate prediction and could be powerful tools for prediction of health care cost.

4.4. Dimension reduction / variable selection

4.4.1. Principle component analysis (PCA)

The goal of PCA is to extract important information from correlated predictors and express this information as a set of new orthogonal variables called principle components (Abdi and Williams, 2010). These principle components can be calculated by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix (Jolliffe, 2002). These principle components are the most important information from X so we can compress the size of X by keeping only the essential information. In practice, we often construct the first M principle components $Z_1, \ldots, Z_M$ and then use them as predictors in regression model. M is often much smaller than p and all Z_i are uncorrected with each other. In practice, M is chosen according by cross-validation. One disadvantage of PCA is that the principle components are hard to interpret. Thus, PCA is mostly used making predictive models. PCA is available as R package `pls`.

We use our lung cancer cost model which have 88 predictors to see how PCA works in action. Applying PCA and using cross validation, M is chosen to be 11. 97.84% of information about the predictors is captured with 11 principle components. We these principle components to a regression model on the training set and evaluate its test set performance we get MSE of 26487. This test set MSE is slightly larger than what we obtain from other machine learning algorithms, but still shows an improvement over traditional cost models.

4.4.2. Lasso

Least Absolute Shrinkage and Selection Operator (Lasso), organically proposed by Tibshirani (1996), is a modern regression technique that works well for variable selection. In a linear regression model, the estimated coefficients β minimizes residual sums of squares (RSS): $\hat{\beta} = \arg\min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$. However, if p is large, lasso can force some of coefficients estimates to be zero and thus performs variables selection Efron et al., 2004. The lasso coefficients is defined as:

$$\hat{\beta_\lambda} = \arg\min \sum_{i=1}^{n} \left(y_i(\beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The tuning parameter λ controls the strength of the penalty. When $\lambda = 0$, lasso will give the same estimate with linear regression. When $\lambda > 0$, we essentially fit a linear model while shrinking some of the coefficients to zero. Such shrunken is caused by the nature of the ℓ_1 penalty; as λ increases, more penalty is imposed ans thus more shrinkage is employed. In practice, we often choose λ using cross-validation. Lasso can also work when $n < p$, the sample size is smaller than the number of predictors. Lasso is available as R package `glmnet` and `lars`.

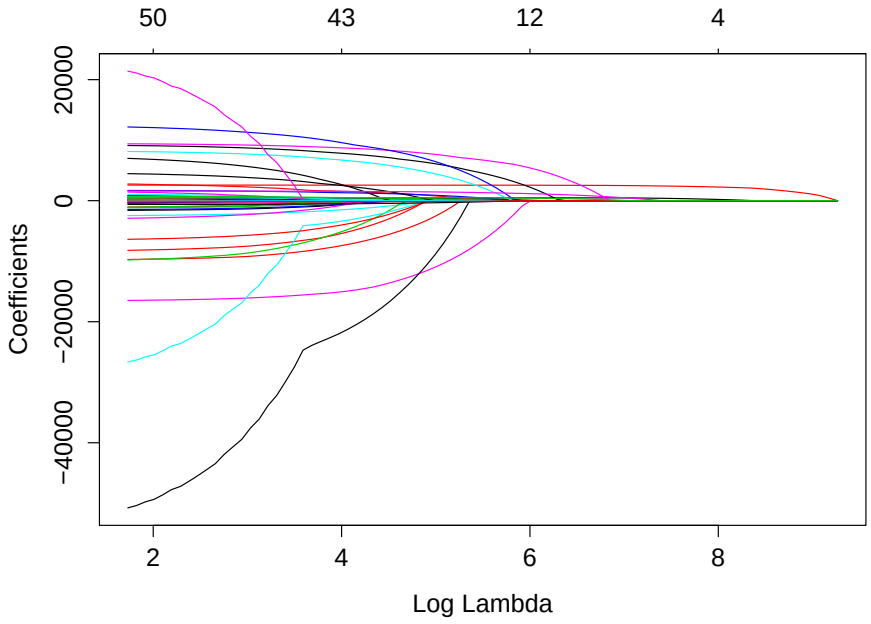


Figure 4.3: Lasso plot: λ versus coefficients of variables

In our lung cancer cost data, we applied Lasso to select important variables. In Figure 4.3, each

curve corresponds to a variable and the figure shows the paths of its coefficients against the $\log(\lambda)$. As λ increase, we allow for more and more predictors in our model. Using cross validation, we find our optimum λ value to be 233, corresponding to including 22 variables in our regression model.

4.5. Discussion

To summarize, modern big data and machine learning tools are highly relevant to health care cost estimation. In this paper, we see some applications in health care cost modeling, prediction, variable selection. The obvious advantage of machine learning algorithms over traditional statistical methods is efficiency: we see better predictive powers and efficiently reduced models. On the other hand, many of machine learning algorithms' lack of interpretable models or its "black box" property is a double edged sword: it is highly automatic but at the same time makes statistical and economics inference difficult. Therefore, depending on the research question, we should be cautious with these modern techniques.

Big data and machine learning tools have many other applications such as causal inference in cost. Propensity score based methods are often used in causal inference where the propensity scores are traditionally estimated using a logistic regression mode. Recent work suggest using tree based and neural networks (Lee, Lessler, and Stuart, 2010; Westreich, Lessler, and Funk, 2010) for propensity score estimation. Li et al. (2015) has looked at cost and cost effectiveness estimation combining traditional propensity score models and machine learning tools.

A developing area in machine learning now is how to handle missing data in predictors. While most existing models such as the ones we mentioned in this paper already have missing data algorithm built in, new methods, especially Bayesian based methods have been proposed (Marlin, 2008). In practice, we often see missing data in predictors when modeling health care cost, thus these methods are more appealing over traditional simple or multiple imputation. Another area of statistical learning, unsupervised learning, can be applied in many real life health care cost scenarios. Unlike supervised learning methods where we do know the outcome, unsupervised learning can be used to find clusters of similar objects (Hastie, Tibshirani, and Friedman, 2009). For example, insurance company could unsupervised learning to group patients into high/low cost groups when they are first enrolled.

CHAPTER 5

DISCUSSION

5.1. Conclusion

Policy making and health services evaluations rely on appropriate and accurate estimation of the cost and cost effectiveness of interventions. In this dissertation, we developed cost and cost effectiveness evaluation models from observational sources. In Chapter 2, we proposed several propensity score based models including covariate adjustment, stratification, weighting and doubly robust estimation to estimate the treatment effect on cost. These methods allow an investigator to compare the causal effect of two treatments on cost. We discussed the issue of informative censoring for cost data, and showed the proposed methods utilizing inverse probability weighting idea are unbiased. We also discussed the variance estimation of the proposed methods. Using simulations, we showed that the doubly robust method gives unbiased results compared to other models. Finally, we demonstrated the use of our proposed methods in a study comparing the costs of two treatments for Stage II/III bladder cancer using an observational cohort derived from SEER-Medicare.

In Chapter 3, we build on the cost estimation idea from Chapter 2 and proposed doubly robust modeling strategies for cost-effectiveness. Specifically, we argue inverse probability weighting should be used to estimate effectiveness instead of the traditional area under the survival cure method. We then propose doubly robust estimators for effectiveness estimation based on inverse probability weighting. We also proposed two estimators for cost estimation, with and without incorporating cost history data. Our simulation studies demonstrate that the proposed DR models perform well even under misspecification of either the propensity score model or the outcome model. We apply these approaches to a cost-effectiveness analysis of two competing lung cancer surveillance procedures, CT versus chest X-ray, using SEER-Medicare data.

In Chapter 4, we review and explore the use of big data and machine learning techniques in cost estimation, especially in big data manipulation, cost prediction and variable dimension reduction. We reviewed popular algorithms including Hadoop, MapReduce, classification and regression trees, random forest, Bayesian additive regression trees, supported vector machines, principle compo-

nents analysis and LASSO. Throughout the chapter, we also demonstrate these state of the art big data and machine learning models using a cohort of lung cancer patients derived from SEER-Medicare.

5.2. Future Directions

This dissertation motivates several further areas of research. In both Chapters 2 and 3, we recommended using the non-parametric bootstrapping to estimate the variance of the doubly robust cost and cost-effectiveness estimators due to model complexity. However, it would be desirable to derive the variance formula and compare the variance estimated from bootstrapping and analytic derivation.

Another future area of research is more in-depth understanding of the simple weighted and partitioned doubly robust cost estimator. In Chapter 3, we showed that the two are both unbiased and had similar bias under large sample sizes. However, we suspect with smaller sample size, the portioned estimator may perform better, but further investigation is needed. In addition, the partitioned estimator may have better empirical performance if the cost partitions, for example monthly cost, follows drastically different distributions.

APPENDIX A

DR PROPERTY OF PARTITIONED Δ_C

From 3.3.2, consider $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}\left[\frac{Z_i Y_{ij}\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)} - \frac{(Z_i-\hat{e}_i)m_1^j(\mathbf{X}_i)\delta_i^j}{\hat{e}_i \hat{K}_j(T_i^j)}\right]$. By the Law of Large Numbers, $\widehat{\mu}_1$ estimates:

$$E\left[\sum_{j=1}^{L}\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right] = \sum_{j=1}^{L}E\left[\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right]$$

Take an arbitrary interval j, $\hat{\mu}_{1,j}$ estimates:

$$\begin{aligned}
& E\left[\frac{ZY^j\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right] \\
= \;& E\left[\frac{ZY^{(1),j}\delta^j}{eK_j(T^j)} - \frac{(Z-e)m_1^j(\mathbf{X})\delta}{eK_j(T^j)}\right] \\
= \;& E\left[\frac{\delta^j}{K_j(T^j)}Y^{(1),j} + \frac{(Z-e)}{e}\left(\frac{\delta^j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)\right] \\
= \;& E\left[Y^{(1),j}\right] + E\left[\left(\frac{Z}{e}-1\right)\left(\frac{\delta^j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)\right] \\
= \;& \mu_{1,j} + E\left[\left(\frac{Z}{e}-1\right)\left(\frac{\delta^j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)\right]
\end{aligned}$$

Hence for $\hat{\mu}_{1,j}$ to be unbiased, we need the second term $S = E\left[(\frac{Z}{e}-1)\left(\frac{\delta^j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)\right]$ to be zero. This condition is satisfied when the propensity score model is correctly specified: $E(Z|Y^{(1)},\mathbf{X}) = E(Z|\mathbf{X}) = e(\mathbf{X},\beta) = e$ so

$$\begin{aligned}
S \;&=\; E\left[E\left[\left(\frac{Z}{e}-1\right)\left(\frac{\delta^j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)|Y^{(1)},\boldsymbol{X}\right]\right] \\
&=\; E\left[\left(\frac{E(Z|Y^{(1)},\boldsymbol{X})}{e}-1\right)\left(\frac{\delta}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)\right] \\
&=\; 0
\end{aligned}$$

When the outcome model $m_1^j(\boldsymbol{X})$ is correctly specified, $m_1^j(\boldsymbol{X}) = E(Y^j|Z=1,\mathbf{X}) = E(Y^{(1),j}|Z=1,\mathbf{X}) = E(Y^{(1),j}|Z,\mathbf{X})$ so

$$\begin{aligned} S &= E\left[E\left[\left(\frac{Z}{e}-1\right)\left(\frac{\delta_j}{K_j(T^j)}Y^{(1),j} - m_1^j(\boldsymbol{X})\right)|Z,\boldsymbol{X}\right]\right] \\ &= E\left[\left(\frac{Z}{e}-1\right)\left(E(\frac{\delta_j}{K_j(T^j)}Y^{(1),j}|Z,\boldsymbol{X}) - m_1^j(\boldsymbol{X})\right)\right] \\ &= E\left[\left(\frac{Z}{e}-1\right)\left(E(Y^{(1),j}|Z,\boldsymbol{X}) - m_1^j(\boldsymbol{X})\right)\right] \\ &= 0 \end{aligned}$$

Hence $\mu_{1,j}$ is unbiased if either the propensity score model e or the outcome model m_1^j is correctly specified. Adding all the L intervals together, the DR property holds for Δ_E as long as either the propensity score model e or the outcome models m_0^j and m_1^j are correct.

BIBLIOGRAPHY

Abdi, H and Williams, LJ (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, 433–459.

Ahmed, NK, Atiya, AF, Gayar, NE, and El-Shishiny, H (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews* 29.5-6, 594–621.

Anstrom, KJ and Tsiatis, AA (2001). Utilizing Propensity Scores to Estimate Causal Treatment Effects with Censored Time-Lagged Data. *Biometrics* 57.4, 1207–1218.

Ash, AS, Zhao, Y, Ellis, RP, and Kramer, MS (2001). Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. *Health services research* 36.6 Pt 2, 194.

Auria, L and Moro, RA (2008). Support vector machines (SVM) as a technique for solvency analysis.

Austin, PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46.3, 399–424.

Austin, PC, Grootendorst, P, and Anderson, GM (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 26.4, 734–753.

Austin, PC and Mamdani, MM (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 25.12, 2084–2106.

Baker, MS, Kessler, LG, Urban, N, and Smucker, RC (1991). Estimating the treatment costs of breast and lung cancer. *Medical care*, 40–49.

Bang, H (2005). Medical cost analysis: application to colorectal cancer data from the SEER Medicare database. *Contemporary clinical trials* 26.5, 586–597.

Bang, H and Robins, JM (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61.4, 962–973.

Bang, H and Tsiatis, A (2000). Estimating medical costs with censored data. *Biometrika* 87.2, 329–343.

Bang, H and Tsiatis, A (2002). Median Regression with Censored Cost Data. *Biometrics* 58.3, 643–649.

Baser, O, Gardiner, JC, Bradley, CJ, and Given, CW (2004). Estimation from Censored Medical Cost Data. *Biometrical Journal* 46.3, 351–363.

Basu, A and Manning, WG (2010). Estimating lifetime or episode-of-illness costs under censoring. *Health Economics* 19, 1010–1028.

Basu, A, Manning, WG, and Mullahy, J (2004). Comparing alternative models: log vs Cox proportional hazard? *Health Economics* 13.8, 749–766.

Basu, A, Polsky, D, and Manning, WG (2011). Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology* 11.1-2, 1–26.

Basu, A and Rathouz, PJ (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 6.1, 93–109.

Bekelman, JE, Handorf, EA, Guzzo, T, Pollack, CE, Christodouleas, J, Resnick, MJ, Swisher-McClure, S, Vaughn, D, Ten Have, T, Polsky, D, et al. (2013). Radical cystectomy versus bladder-preserving therapy for muscle-invasive urothelial carcinoma: examining confounding and misclassification biasin cancer observational comparative effectiveness research. *Value in Health* 16.4, 610–618.

Bertsimas, D, Bjarnadóttir, MV, Kane, MA, Kryder, JC, Pandey, R, Vempala, S, and Wang, G (2008). Algorithmic prediction of health-care costs. *Operations Research* 56.6, 1382–1392.

Breiman, L (2001). Random forests. *Machine learning* 45.1, 5–32.

Breiman, L, Friedman, J, Stone, CJ, and Olshen, RA (1984). *Classification and regression trees*. CRC press.

Briggs, AH, Wonderling, DE, and Mooney, CZ (1997). Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health economics* 6.4, 327–340.

Brooks, JC (2012). *Super Learner and Targeted Maximum Likelihood Estimation for Longitudinal Data Structures with Applications to Atrial Fibrillation*. PhD thesis. University of California, Berkeley.

Busso, M, DiNardo, J, and McCrary, J (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* 96.5, 885–897.

Cain, LE and Cole, SR (2009). Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Statistics in medicine* 28.12, 1725–1738.

Centers for Medicare and Medicaid (2010). *Medicare economic index*. URL: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareProgramRatesStats/Downloads/mktbskt-summary.pdf (visited on 01/01/2015).

Chen, M, Mao, S, Zhang, Y, and Leung, VC (2014). Big data storage. In: *Big Data*. Springer, 33–49.

Chipman, HA, George, EI, and McCulloch, RE (1998). Bayesian CART model search. *Journal of the American Statistical Association* 93.443, 935–948.

Chipman, HA, George, EI, and McCulloch, RE (2002). Bayesian treed models. *Machine Learning* 48.1-3, 299–320.

Chu, C, Kim, SK, Lin, YA, Yu, Y, Bradski, G, Ng, AY, and Olukotun, K (2007). Map-reduce for machine learning on multicore. *Advances in neural information processing systems* 19, 281.

Ciunci, CN, Mitra, N, Yang, J, Epstein, A, Langer, C, DeMichele, A, and Vachani, A (2015). Patterns and effectiveness of surveillance after curative intent surgery in older stage I-IIIA non-small cell lung cancer patients. In preparation.

Clement, FM, Harris, A, Li, JJ, Yong, K, Lee, KM, and Manns, BJ (2009). Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. *Jama* 302.13, 1437–1443.

Cochran, WG (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295–313.

Cole, SR and Hernán, MA (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology* 168.6, 656–664.

Criminisi, A, Shotton, J, and Konukoglu, E (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision* 7.2–3, 81–227.

Cristianini, N and Shawe-Taylor, J (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dean, J and Ghemawat, S (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM* 51.1, 107–113.

Dean, J and Ghemawat, S (2010). MapReduce: a flexible data processing tool. *Communications of the ACM* 53.1, 72–77.

Denison, DG, Mallick, BK, and Smith, AF (1998). A bayesian CART algorithm. *Biometrika* 85.2, 363–377.

Diehr, P, Yanez, D, Ash, A, Hornbrook, M, and Lin, D (1999). Methods for analyzing health care utilization and costs. *Annual review of public health* 20.1, 125–144.

Disease Control, C for and Prevention. (2014). *CDC WONDER On-line Database*.

Dodd, S, Bassi, A, Mrcp, KB, and Williamson, P (2006). A comparison of multivariable regression models to analyses cost data. *Journal of Evaluation in Clinical Practice* 12.1, 76–86.

Duan, N, Manning, WG, Morris, CN, and Newhouse, JP (1983). Comparison of for Alternative Care Models for the Demand Medical. *Journal of Business & Economic Statistics* 1.2, 115–126.

Economic Analysis, B of (2015). *National Income Product Accounts Tables, Section 1 Domestic Product and Income, Table 1.1.4 Price Indexes for Gross Domestic Products*.

Efron, B (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association* 82.397, 171–185.

Efron, B, Hastie, T, Johnstone, I, Tibshirani, R, et al. (2004). Least angle regression. *The Annals of statistics* 32.2, 407–499.

Efstathiou, JA, Spiegel, DY, Shipley, WU, Heney, NM, Kaufman, DS, Niemierko, A, Coen, JJ, Skowronski, RY, Paly, JJ, McGovern, FJ, et al. (2012). Long-term outcomes of selective bladder preservation by combined-modality therapy for invasive bladder cancer: the MGH experience. *European Urology* 61.4, 705–711.

Einav, L and Levin, JD (2013). *The data revolution and economic analysis*. Tech. rep. National Bureau of Economic Research.

Etzioni, RD, Feuer, EJ, Sullivan, SD, Lin, D, Hu, C, and Ramsey, SD (1999). On the use of survival analysis techniques to estimate medical care costs. *Journal of health economics* 18.3, 365–380.

Fireman, BH, Quesenberry, CP, Somkin, CP, Jacobson, AS, et al. (1997). Cost of care for cancer in a health maintenance organization. *Health care financing review* 18.4, 51.

Folland, S, Goodman, AC, Stano, M, et al. (2007). *The economics of health and health care*. Vol. 6. Pearson Prentice Hall New Jersey.

Freund, Y and Schapire, RE (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55.1, 119–139.

Friedman, JH (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Funk, MJ, Westreich, D, Wiesen, C, Stürmer, T, Brookhart, MA, and Davidian, M (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology* 173.7, 761–767.

Ghose, A, Ipeirotis, PG, and Sundararajan, A (2007). "Opinion mining using econometrics: A case study on reputation systems". In: *annual meeting-association for computational linguistics*. Vol. 45. 1, 416.

Glick, HA, Doshi, JA, Sonnad, SS, and Polsky, D (2014). *Economic evaluation in clinical trials*. Oxford University Press.

Goldfeld, K (2014). Twice-weighted multiple interval estimation of a marginal structural model to analyze cost-effectiveness. *Statistics in medicine* 33.7, 1222–1241.

Gomes, M, Ng, ESW, Grieve, R, Nixon, R, Carpenter, J, and Thompson, SG (2012). Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Medical Decision Making* 32.2, 350–361.

Goodwin, PJ and Shepherd, FA (1998). Economic issues in lung cancer: a review. *Journal of clinical oncology* 16.12, 3900–3912.

Gramacy, RB and Lee, HK (2012). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*.

Gruber, S, Logan, RW, Jarrín, I, Monge, S, and Hernán, MA (2015). Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statistics in Medicine* 34.1, 106–117.

Handorf, Ea, Bekelman, JE, Heitjan, DF, and Mitra, N (2013). Evaluating costs with unmeasured confounding: A sensitivity analysis for the treatment effect. *Annals of Applied Statistics* 7.4, 2062–2080.

Hastie, T, Tibshirani, R, and Friedman, J (2009). *Unsupervised learning*. Springer.

Heitjan, DF, Moskowitz, AJ, and Whang, W (1999). Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health economics* 8.3, 191–201.

Hillner, BE, McDonald, MK, Desch, CE, Smith, TJ, Penberthy, LT, Maddox, P, and Retchin, SM (1998). Costs of care associated with non-small-cell lung cancer in a commercially insured cohort. *Journal of clinical oncology* 16.4, 1420–1424.

Hirano, BYK, Imbens, GW, and Ridder, G (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71.4, 1161–1189.

Hortonworks (2016). *Hortonworks and Microsoft: We do Hadoop in the Cloud*. URL: `http://hortonworks.com/partner/microsoft/` (visited on 02/15/2016).

Hsu, CW, Chang, CC, Lin, CJ, et al. (2003). A practical guide to support vector classification.

Indurkhya, A, Mitra, N, and Schrag, D (2006). Using propensity scores to estimate the cost–effectiveness of medical therapies. *Statistics in medicine* 25.9, 1561–1576.

James, G, Witten, D, Hastie, T, and Tibshirani, R (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jiang, H and Zhou, XH (2004). Bootstrap confidence intervals for medical costs with censored observations. *Statistics in Medicine* 23.21, 3365–3376.

Jolliffe, I (2002). *Principal component analysis*. Wiley Online Library.

Kim, GH, An, SH, and Kang, KI (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and environment* 39.10, 1235–1242.

Kreif, N, Grieve, R, and Sadique, MZ (2013). Statistical Methods For Cost-Effectiveness Analyses That Use Observational Data: A Critical Appraisal Tool And Review Of Current Practice. *Health economics* 22.4, 486–500.

Kutikova, L, Bowman, L, Chang, S, Long, SR, Obasaju, C, and Crown, WH (2005). The economic burden of lung cancer and the associated costs of treatment failure in the United States. *Lung Cancer* 50.2, 143–154.

Laan, MJ Van der, Polley, EC, and Hubbard, AE (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* 6.1.

Lam, C (2010). *Hadoop in action*. Manning Publications Co.

Lee, BK, Lessler, J, and Stuart, EA (2010). Improving propensity score weighting using machine learning. *Statistics in medicine* 29.3, 337–346.

Leung, SF and Yu, S (1996). On the choice between sample selection and two-part models. *Journal of econometrics* 72.1, 197–229.

Li, J, Handorf, E, Bekelman, J, and Mitra, N (2015). Propensity score and doubly robust methods for estimating the effect of treatment on censored cost. *Statistics in medicine*.

Liaw, A and Wiener, M (2002). Classification and regression by randomForest. *R news* 2.3, 18–22.

Lin, DY, Feuer, EJ, Etzioni, R, and Wax, Y (1997). Estimating Medical Costs from Incomplete Follow-Up Data. *Biometrics* 53.2, 419–434.

Lin, D (2000). Linear regression analysis of censored medical costs. *Biostatistics* 1.1, 35–47.

Lin, D (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine* 22.7, 1181–1200.

Loh, WY (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, 14–23.

Lumley, T, Diehr, P, Emerson, S, and Chen, L (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health* 23.1, 151–169.

Lunceford, JK and Davidian, M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23.19, 2937–2960.

Manning, WG (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of health economics* 17.3, 283–295.

Manning, WG, Basu, A, and Mullahy, J (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 24.3, 465–488.

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

Marlin, BM (2008). *Missing data problems in machine learning*. PhD thesis. University of Toronto.

Michie, D, Spiegelhalter, DJ, and Taylor, CC (1994). Machine learning, neural and statistical classification.

Mihaylova, B, Briggs, A, and Hagan, AO (2011). REVIEW OF STATISTICAL METHODS FOR ANALYSING HEALTHCARE RESOURCES AND COSTS. *Health Economics* 916.August 2010, 897–916.

Mihaylova, B, Briggs, A, O'Hagan, A, and Thompson, SG (2011). Review of statistical methods for analysing healthcare resources and costs. *Health economics* 20.8, 897–916.

Mitra, N and Indurkhya, A (2005). A propensity score approach to estimating the cost–effectiveness of medical therapies from observational data. *Health economics* 14.8, 805–815.

Mohammed, EA, Far, BH, and Naugler, C (2014). Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. *BioData mining* 7.1, 1.

Montez-Rath, M, Christiansen, CL, Ettner, SL, Loveland, S, and Rosen, AK (2006). Performance of statistical models to predict mental health and substance abuse cost. *BMC medical research methodology* 6.1, 53.

Moran, JL, Solomon, PJ, Peisach, AR, and Martin, J (2007). New models for old questions: generalized linear models for cost prediction. *Journal of evaluation in clinical practice* 13.3, 381–389.

O'Hagan, A and Stevens, JW (2003). Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health economics* 12.1, 33–49.

Polsky, D, Glick, HA, Willke, R, and Schulman, K (1997). Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health economics* 6.3, 243–252.

Quinlan, JR (1996). Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research*, 77–90.

Quinlan, JR (2014). *C4. 5: programs for machine learning*. Elsevier.

Raikou, M and McGuire, A (2004). Estimating medical care costs under conditions of censoring. *Journal of Health Economics* 23.3, 443–470.

Robins, JM and Rotnitzky, A (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In: *AIDS Epidemiology*. Springer, 297–331.

Robins, JM, Rotnitzky, A, and Zhao, LP (1994). Estimation of Regression Coefficients When Some of Regression Coefficients Estimation Regressors Are Not Always Observed. *Journal of the American Statistical Association* 89.427, 846–866.

Rosenbaum, PR and Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.

Rosenbaum, PR (1987). Model-Based Direct Adjustment. *Journal of the American Statistical Association* 82.398, 387–394.

Rubin, D (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.

Rubin, DB (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* 13.April, 855–857. ISSN: 10538569. DOI: 10.1002/pds.968.

Rubin, DB and Rosenbaum, PR (1984). Reducing Bias in Observational Studies Using Score on the Propensity Subolassification. *The American Economic Review* 79.387, 516–524.

Schapire, RE (1990). The strength of weak learnability. *Machine learning* 5.2, 197–227.

Schapire, RE (1999). "A brief introduction to boosting". In: *Ijcai*. Vol. 99, 1401–1406.

Schapire, RE (2003). The boosting approach to machine learning: An overview. In: *Nonlinear estimation and classification*. Springer, 149–171.

Scott, SL and Varian, HR (2013). *Bayesian variable selection for nowcasting economic time series*. Tech. rep. National Bureau of Economic Research.

Setoguchi, S, Schneeweiss, S, Brookhard, M, Glynn, R, and Cook, F (2009). NIH Public Access. *Pharmacoepidemiology and Drug Safety* 27.6, 417–428. ISSN: 1937-1209. DOI: 10.1055/s-0029-1237430.. arXiv:NIHMS150003.

Shafer, J, Rixner, S, and Cox, AL (2010). "The hadoop distributed filesystem: Balancing portability and performance". In: *Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on*. IEEE, 122–133.

Shvachko, K, Kuang, H, Radia, S, and Chansler, R (2010). "The hadoop distributed file system". In: *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 1–10.

Siegel, R, Naishadham, D, and Jemal, A (2012). Cancer statistics, 2012. *CA: a cancer journal for clinicians* 62.1, 10–29.

Stefanski, L and Boos, D (2002). The Calculus of M-Estimation. *The American Statistician* 56, 29–38.

Sushmita, S, Newman, S, Marquardt, J, Ram, P, Prasad, V, Cock, MD, and Teredesai, A (2015). "Population cost prediction on public healthcare datasets". In: *Proceedings of the 5th International Conference on Digital Health 2015*. ACM, 87–94.

Suykens, JA and Vandewalle, J (1999). Least squares support vector machine classifiers. *Neural processing letters* 9.3, 293–300.

Thompson, SG and Nixon, RM (2005). How sensitive are cost-effectiveness analyses to choice of parametric distributions? *Medical Decision Making* 25.4, 416–423.

Tian, L and Huang, J (2007). A two-part model for censored medical cost data. *Statistics in Medicine* 26, 4273–4292. DOI: 10.1002/sim.

Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tsiatis, AA and Davidian, M (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science: a review journal of the Institute of Mathematical Statistics* 22.4, 569.

Varian, HR (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives* 28.2, 3–27.

Venner, J (2009). *Pro Hadoop*. Apress.

Westreich, D, Lessler, J, and Funk, MJ (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology* 63.8, 826–833.

Willan, AR and Briggs, AH (2006). *Statistical analysis of cost-effectiveness data*. Vol. 37. John Wiley & Sons.

Willan, AR, Lin, D, and Manca, A (2005). Regression methods for cost-effectiveness analysis with censored data. *Statistics in medicine* 24.1, 131–145.

Willan, AR and O'Brien, BJ (1996). Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. *Health economics* 5.4, 297–305.

Willan, A, Lin, D, Cook, R, and Chen, E (2002). Using inverse-weighting in cost-effectiveness analysis with censored data. *Statistical methods in medical research* 11.6, 539–551.

Wu, Y, Tjelmeland, H, and West, M (2007). Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics* 16.1, 44–66.

Ying, Z, Jung, S, and Wei, L (1995). Survival analysis with median regression models. *Journal of the American Statistical Association* 90.429, 178–184.

Young, TA (2005). Estimating mean total costs in the presence of censoring. *Pharmacoeconomics* 23.12, 1229–1242.

Zhao, H, Cheng, Y, and Bang, H (2011). Some insight on censored cost estimators. *Statistics in Medicine* 30.19, 2381–2388.

Zhao, H and Tian, L (2001). On Estimating Medical Cost and Incremental Cost-Effectiveness Ratios with Censored Data. *Biometrics* 57, 1002–1008.

Zhao, H, Bang, H, Wang, H, and Pfeifer, PE (2007). On the equivalence of some medical cost estimators with censored data. *Statistics in Medicine* 26, 4520–4530.