LIKELIHOOD-BASED INSTRUMENTAL VARIABLE ANALYSIS IN THE PRESENCE OF AN UNOBSERVED LATENT CONFOUNDER

BY ANJUN CAO

A Dissertation submitted to the

Graduate School – New Brunswick

School of Public Health – Piscataway

University of Medicine and Dentistry of New Jersey

in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biostatistics

written under the direction of

Dirk F. Moore

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2011

ABSTRACT OF THE DISSERTATION

Likelihood-Based Instrumental Variable Analysis In the Presence of an Unobserved Latent Confounder

BY ANJUN CAO

Dissertation Director:

Dirk F. Moore

Instrumental variable analysis (IVA) is used to control unobserved confounders and estimate average causal effects in observational studies. Classical IVA involves a two-stage procedure with two ordinary linear models. The first stage relates the treatment or intervention to the instrument, and the second relates the outcome to the expected treatment predicted by the first stage. The average causal effect can be estimated using the difference in outcomes between the strata of the instrumental variable. D.B. Rubin in a series of papers (summarized in Angrist, Imbens, and Rubin, 1996) re-framed IVA in terms of a causal model which can be applied to binary outcome variables when the instrumental variable and treatment status are also binary. However, the average causal effect is typically expressed as a difference. When causal effects expressed as rate ratios or odds ratios are desired in nonlinear models, it is problematic to obtain the unbiased estimators for these parameters. We propose a two-stage likelihood-based IVA model. In both stages, the estimates of parameters of interest are obtained using maximum likelihood functions. In the first stage, patient compliance with the instrumental variable is estimated. Treatment effect is then imputed in the second stage with the adjustment of compliance. Essentially, the likelihood function is formulated using the joint

distribution of outcome and instrumental variables by integrating out the treatment and unknown confounder, assuming the distribution of the confounder is known, and the associations between the confounder and treatment, and confounder and outcome are also known or can be estimated. This likelihood function is maximized to obtain an estimator of the coefficient of the treatment variable. The variance of this maximum likelihood estimation (MLE) of treatment effect can be estimated using average Fisher's information matrix.

We illustrate this two-stage likelihood-based IVA model using data from a study of primary androgen deprivation therapy (PADT) in men with localized prostate cancer (Lu-Yao, Albertsen, Moore, et al. 2008). We also examine the optimal minimum sample size needed for each health service area in order to reduce the misclassifications, and obtain unbiased estimates of the average causal effect.

Acknowledgements

I would like to give my greatest appreciation to my advisor, Professor Dirk F. Moore. Without his encouragement and his wise instruction in hundreds of hours during the past five years, I would be unable to present this dissertation to my committee members. His professional mentoring will affect my attitude in my career as well as in life eternally.

I would also like to express my gratitude to my committee members. I give my sincere thanks to Professor Weichung Joe Shih, who directed me to this promising topic at the beginning of my research. I give my sincere thanks to Professor Yong Lin, who read my work and corrected my mistakes in detail. I give my sincere thanks to Professor Grace Lu-Yao for her valuable data and information on prostate cancer, and allowing me to join her research team.

Table of Contents

List of Tables

List of Figures

List of Acronyms

2SLS	Two-Stage Least Squares
ACE	Average Causal Effect
CDF	Cumulative Distribution Function
CI	Confidence Interval
CM	Conservative Management (for prostate cancer)
COR	Causal Odds Ratio
CRR	Causal Relative Risk
DAG	Directed Acyclic Graph
GLM	Generalized Linear Model
GMM	Generalized Method of Moments
IV	Instrumental Variable
IVA	Instrumental Variable Analysis
LATE	Local Average Treatment Effect
MLE	Maximum Likelihood Estimation
OR	Odds Ratio
PADT	Primary Androgen Deprivation Therapy (for prostate cancer)
PDF	Probability Density Function
PSA	Prostate Specific Antigen

RR	Relative Risk
SEER	Surveillance, Epidemiology, and End Results
SUTVA	Stable Unit Treatment Value Assumption

Chapter 1

Introduction

Evaluating the causal effect of a new treatment compared to a current treatment or placebo is the mainstay of pharmaceutical statisticians. Identifying the causality effect between disease and exposure is the ultimate goal for epidemiologists. Both tasks rely on well designed research studies and proper statistical analysis methods. Associations discovered between response and treatment or disease and exposure may not always be causal. Often the associations are caused by factors other than the true cause-effect relationship. Non-causal factors include chance, bias or confounding. Chance associations can be evaluated using p-values and confidence intervals. Bias is a systematic error introduced during the study conduct. General types of bias include recall bias, selection bias, and interviewer bias. Sicker subjects may recall more details of the exposure. Investigators may apply non-comparable criteria when enrolling study participants into different treatment groups. Interviewers may focus on particular questions for subjects treated with active drug, and collect biased information between different treatment groups. Confounding is another common phenomenon that interferes with the treatment-outcome relationship. A confounder is a factor that is correlated with treatment or exposure, and can independently affect the magnitude of response or development of disease. Failure to control confounders results in under- or over-estimates of the true treatment-outcome relationship.

Observed confounders can be controlled in the study design or in the data analysis by statistical adjustment. This is particularly important in observational

studies. By contrast, intervention studies such as randomized clinical trials are designed to control both observed and unobserved confounders. With sufficient sample size, all potential confounders, whether observed or not, are supposed to be evenly distributed among the treatment groups by randomization. Therefore, results from randomized clinical trials are treated as a "gold standard" when they are compared to the results from other studies with different designs.

One of the disadvantages of clinical trials is the cost. According to a report from a business intelligence firm (Cutting Edge Information, in 2006) running phase 3 trials in pharmaceutical companies can cost more than $26,000 per patient on average, and in phase 3 clinical trials, companies typically recruit several hundreds to several thousands of patients. The cost is huge, but the randomized clinical trial is still the most favorable design in pharmaceutical companies because it controls for unknown confounders, something which cannot ordinarily be achieved by other designs.

Another disadvantage of randomized clinical trials involves ethics. When the exposure is harmful, it will not be ethical to randomize any participants to that group. For example, in studies of smoking and lung cancer, investigators will never randomize non-smokers to the smoking group.

Other limitations of clinical trials include difficulty in recruiting patients, particularly elderly or seriously sick patients. For example, hemophilia is a rare congenital bleeding disorder that affects about 18,000 people in the United States (National Heart Lung and Blood Institute, 2008). It is highly possible that a sponsor is not able to enroll enough patients in a phase 3 trial to test a new treatment. If the clinical trials take years to complete, maintaining compliance and preventing drop-outs also become challenging tasks, as well. Compliance with the treatment in the

elderly is particularly difficult. In addition, trial participants are also hard to follow-up if they move to other areas, or withdraw consent.

1.1 Methods for controlling observed confounders in an observational study

When clinical trials are not feasible, statisticians will use observational studies to examine the associations between the health outcome and treatment or exposure. Although observational studies cost less, and are easier to conduct when compared to interventional studies, control of confounders becomes one of the outstanding issues. There are several statistical methods for controlling observed confounders, including stratification analysis, use of regression to adjust for confounders, and propensity score analysis.

Stratification is done by evaluating associations between treatment and effect or exposure and disease separately among the levels of the confounders. Stratification is often used when the confounding variables are categorical, such as sex, race or cigarette smoking status. For example, when the association between alcohol consumption and cardiovascular disease is studied, cigarette smoking can be a strong confounder because alcohol consumers are more likely to be cigarette smokers, too. Therefore, the alcohol exposure status is correlated with smoking status. There are more cigarettes smokers in the exposure group than non-exposure group. Furthermore, cigarette smoking alone can affect the outcome of cardiovascular disease, so the outcome of cardiovascular disease could be a mixed effect from both alcohol intake and cigarette smoking. With separate analyses for smokers and non-smokers, we assure that the outcome is independent of the confounder of cigarette smoking.

Regression analysis is the most frequently used statistical method to control observed confounders simultaneously. Potential confounders are named as covariates in the regression model. In an ordinary linear regression model, the outcome is placed on the left side of the equation and treated as a dependent variable. Treatment or exposure status is treated as a fixed effect and fitted on the right side of the equation along with a set of covariates such as age, sex, race, or body mass index. With the adjustment for these covariates, the estimated association between outcome and treatment is consistent with the true treatment effect.

Propensity score analysis is another statistical method to control observed confounders in observational studies. A propensity score is defined as the probability of assignment to treatment, conditional on observed covariates which are potential confounders,

$$e(\underset{\sim}{X}) = pr(T = 1 | \underset{\sim}{X})$$

where $\underset{\sim}{X}$ is a vector of covariates, and T is the assignment to treatment, 1 or 0.

For large size samples, Rosenbaum (1983) has presented a large-sample theory of propensity score analysis, and in particular presented this theorem:

> **Theorem:** Treatment assignment and the observed covariates are conditionally independent given the propensity score, that is
>
> $$\underset{\sim}{X} \perp T | e(\underset{\sim}{X})$$

This theorem states that, given the propensity score $e(\underset{\sim}{X})$, the treatment assignment is random and independent on any covariates of $\underset{\sim}{X}$. With the adjustment on propensity score, any association discovered between treatment and outcome is independent on those observed confounders.

A propensity score is usually expressed as a function of a vector of covariates. For example, it can be modelled as a logistic regression with a vector of covariates as independent predictors,

$$\log\left(\frac{e(\underset{\sim}{X})}{1-e(\underset{\sim}{X})}\right)=\underset{\sim}{\beta}^{T}\cdot\underset{\sim}{X}$$

The propensity score summarizes the multi-dimensional covariates with a uni-dimensional score. With this single dimensional propensity score, statisticians are able to conduct matched sampling conveniently. Sub-groups of population with similar propensity score can be easily identified. The sample mean difference of matched treatment groups with the same propensity score provides an unbiased estimate for the true treatment effect.

In practice, propensity scores are first calculated for every subject based on the observed confounders. Subjects with nearest propensity scores are then matched between treatment group and control group. Statistical analyses are applied on the selected groups. Dehejia and Wahba (2002) applied propensity score matching methods to the data from National Supported Work experiment. The National Support Work experiment (LaLonde, 1986) was a randomized trial to evaluate the effect of a nine months to one year's training program on trainee earnings. The treated group received on-the-job training, and the control group did not. Dehejia and Wahba (2002) created matched control groups from databases of the Population Survey of Income Dynamics and the Current Population Survey based on individual's propensity score. Propensity scores were estimated from a logistic regression with independent predictors of age, number of school years, race, marriage status, previous annual earnings, and employment history. When propensity scores were matched in both treated and control groups, these potential confounders were also comparable for both

groups. Any observed differences in the trainee earnings between the treated group and selected control groups were then independent of these confounders. Therefore, the estimated difference in earnings was an unbiased estimator for the average effect of the training program. The result from one of the propensity score matching methods showed an average raise of $1473 and $1616 per person year for the treated group when it is compared to the two control groups. These numbers were very close to the $1672 raise from the randomized experiment.

1.2 Methods for controlling unobserved confounders in an observational study

In reality, not all the confounders are observable or measurable. We examined the ordinary linear regression model:

$$Y_i = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \varepsilon_i \tag{1.2.1}$$

In equation (1.2.1), Y_i is the response from subject i. T_i is the treatment received by subject i. U_i is an observed confounder. If the observed confounder is controlled, β_1 is the true treatment effect.

In the situation where the confounder is unobserved or unknown, the term $\beta_2 \cdot U_i$ gets omitted from equation (1.2.1).

$$Y_i = \beta_0^* + \beta_1^* \cdot T_i + \varepsilon_i^* \tag{1.2.2}$$

Equation 1.2.2 is not equivalent to the true model equation (1.2.1).

How can we control the unobserved confounders in observational studies? One of the ideal solutions will be to find something very similar to randomization.

Instrumental variable analysis (IVA) has been used by economists and epidemiologists for decades, and it is close to the solution we are looking for. IVA

controls the unobserved confounders by introducing a third variable called the instrumental variable (IV). A valid instrumental variable is correlated with the treatment or exposure status only, and independent of all potential observed or unobserved covariates, that is, $Z \perp \underset{\sim}{X}$ and $Z \perp \underset{\sim}{U}$, where Z represents an instrumental variable. It is important that the instrumental variable itself does not cause variation in outcome response. It affects the outcome indirectly through the unevenly distributed treatment or exposure status among the strata of instrumental variable. The diagram $Z \rightarrow T \rightarrow Y$ illustrates a path of an instrumental variable Z causing a outcome Y. T is a treatment or exposure variable which facilitates the effect of Y from Z. There is no direct path from Z to Y.

Examples of IVA can be traced back to 1854. There was an epidemic of cholera in London. British physician John Snow observed that there were higher death rates from cholera among the residents who received their drinking water from the Lambeth Company or the Southwark and Vauxhall Company as opposed to from other households with different water supply companies. Further investigation discovered that the two companies drew water from the Thames River at a point polluted with main sewage discharge. The contaminated drinking water was the source of the outbreak of cholera. In this story, water companies served as an instrumental variable. Water companies themselves were not able to cause the disease, but they were highly correlated to the exposure, and indirectly affected the death rates. In addition, baseline characteristics such as occupation, health and monetary conditions were comparable between the two groups of people who received water from different water companies.

Geographic location is another widely used instrumental variable because it is often likely to correlate with certain treatments or exposures. In 2008, investigators

from the Cancer Institute of New Jersey published their results on the primary androgen deprivation therapy (PADT) among men with localized prostate cancer (Lu-Yao, et al., 2008). In this population-based cohort study, IVA was utilized. Investigators noticed that the PADT usage rates were highly differentiated among health service areas within the U.S (Shahinian, Vahakn B, Kuo, Yong-fang, Freeman, Jean L, et al., 2005). The variation was not from the medical consideration, but from the preference of local health service practice. This finding indicated that the health service areas could serve as a valid instrumental variable. In the statistical analysis, these health service areas were then categorized into two classes, high PADT usage areas and low PADT usage areas. Patients' survivals were compared between these two types of areas to evaluate the effectiveness of PADT. The direct comparison between patients with PADT and conservative management (CM) was believed to be inappropriate because some unobserved confounders, particularly prostate specific antigen (PSA) level, could bias the results.

Randomized treatment assignment in clinical trials is actually a perfect instrumental variable when the sample size is sufficient large. Patients are randomized to treatment group or control group with an equal probability. Confounders associated with patients' characteristics are hence randomized into treatment group or control group with equal probability. If all patients fully comply with the randomized assignment, the sample difference in outcome between treatment group and control group is truly an unbiased estimator of treatment effect. In some cases, a few patients take a treatment other than the one they are assigned to mistakenly. Outcomes are still compared between the randomized treatment groups rather than as treated groups. This is so called the intention-to-treat method which is equivalent to the instrumental variable analysis. In clinical trials, because patients are closely monitored, the

inconsistency rate between assigned treatment and actual treatment is very low. Investigators should not adjust the results with this non-compliance rate. This intention-to-treat analysis provides conservative estimates, and additionally, penalizes careless monitoring of patient compliance during the trial. In observational studies, the compliance to the instrumental variable is much lower than the compliance to the randomization codes in clinical trials, so the inconsistency rate needs to be adjusted.

In Chapter 2, we review current IVA methodologies and make comparisons between them. In Chapter 3, we discuss drawbacks of these IVA methodologies, and propose a two-stage likelihood-based IVA model. We apply this two-stage likelihood-based IVA to generalized linear models. In Chapter 4, the two-stage likelihood-based IVA model is extended to a survival data analysis. In Chapter 5, optimal minimum sample size is explored when the instrumental variable is not binary in nature. Some instruments are continuous variables. When the instrumental variable in categorical form is desired, the continuous data need to be converted to categorical data. During this procedure, sample size of the subunits can become one of technical detail. In Chapter 6, we discuss future research possibilities including IVA in cluster data analysis.

We use study of PADT on localized prostate cancer (Lu-Yao, et al., 2008) as an example to develop the two-stage likelihood-based IVA throughout this dissertation. All statistical analyses are performed using R, version 2.10.1.

Chapter 2

Review of IVA Methodologies

We begin this chapter with the classical IVA method, two-stage least squares. This method was described as early as in 1954 by Durbin. We consider the rationale of the method, and from there, we present more IVA models currently used in linear and nonlinear statistical analyses.

2.1 Two-stage Least Squares (2SLS)

When a linear model includes all important predictors, the coefficients of the predictors consistently measure the causal-effect relationship between the outcome and predictors.

$$Y_i = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \varepsilon_i, \quad \text{for subject } i = 1, 2, ..., N \tag{2.1.1}$$

In model (2.1.1), assuming ε_i is identical independently distributed with mean 0, β_1 quantifies the causal effect from T to Y, and similarly, β_2 quantifies the causal effect from U to Y. If U is a confounder of T, this is true only when both predictors are included in the model and no other important confounders are omitted from the model. Plots of residuals ε_i versus all predictors can be helpful for diagnosing the appropriateness of the model. If the plots show the following,

$$\text{cov}(T_i, \varepsilon_i) = 0 \quad \text{and} \quad \text{cov}(U_i, \varepsilon_i) = 0 \tag{2.1.2}$$

we can interpret β_1 as the treatment effect on outcome Y, and $\hat{\beta}_1$, the least square estimator of β_1, is an unbiased estimator of treatment effect. If any of the important predictors, particularly confounders, are omitted from model (2.1.1),

$$Y_i = \beta_0^* + \beta_1^* \cdot T_i + \varepsilon_i^*, \qquad \text{for subject } i = 1, 2, ..., N \tag{2.1.3}$$

plots of estimated residuals $\hat{\varepsilon}_i^*$ from model (2.1.3) versus T_i will most likely show a deviation from independence, that is,

$$\operatorname{cov}\left(T_i, \varepsilon_i^*\right) \neq 0 \tag{2.1.4}$$

It can be shown that $\hat{\beta}_1^*$, which is the estimator of β_1^* in model (2.1.3) is a biased estimator of β_1 in model (2.1.1). Let $\underset{\sim}{T}$ and $\underset{\sim}{U}$ be vectors of size N, and $\bar{T}$ be mean of $\underset{\sim}{T}$. By least squares:

$$\begin{aligned} E\left(\hat{\beta}_1^* \mid \underset{\sim}{T}, \underset{\sim}{U}\right) &= E\left[\frac{\sum_{i=1}^{N}\left\{Y_i \cdot \left(T_i - \bar{T}\right)\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2}\right] = \frac{\sum_{i=1}^{N}\left\{\left(T_i - \bar{T}\right) \cdot E\left(Y_i\right)\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2} \\ &= \frac{\sum_{i=1}^{N}\left\{\left(T_i - \bar{T}\right) \cdot E\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right)\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2} = \beta_1 + \beta_2 \cdot \frac{\sum_{i=1}^{N}\left\{\left(T_i - \bar{T}\right) \cdot U_i\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2} \end{aligned} \tag{2.1.5}$$

The bias is $\beta_2 \cdot \dfrac{\sum_{i=1}^{N}\left\{\left(T_i - \bar{T}\right) \cdot U_i\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2}$. When U is not a confounder of T, that is, T and U are independent, the bias is zero.

In order to obtain an unbiased estimator of the treatment effect, economists, epidemiologists, and statisticians have paid most attention to a classical IVA of two-stage least squares model. With the involvement of an instrumental variable (IV), the two-stage least squares model includes two ordinary linear regression models. In the

first stage, an instrumental variable is used to predict the treatment allocation. This predicted treatment assignment is then used in the second stage as one of the independent variables to forecast the outcome. The actual treatment status should not be fitted as one of the predictors in the second stage. Instead, it is used in the first stage as the dependent variable to estimate the probabilities of the treatment received based on the values of instrumental variable.

$$\text{First stage:} \qquad T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i \tag{2.1.6}$$

$$\text{Second stage:} \quad Y_i = \beta_0 + \beta_1 \cdot \hat{T}_i + \varepsilon_i \tag{2.1.7}$$

In equation 2.1.6, Z is the instrumental variable and is used to predict the treatment status T. In equation (2.1.7), outcome Y is fitted with the predicted treatment status $\hat{T}$ from the first stage. The coefficient β_1 reflects the treatment effect, and therefore is the parameter of interest.

As in the ordinary regression model, assumptions made for the two-stage least squares model are:

$$\text{cov}(Z_i, \varepsilon_i) = 0 \quad \text{and} \quad \text{cov}(Z_i, \upsilon_i) = 0$$

In addition, the instrumental variable Z must be correlated to the treatment status T:

$$\text{cov}(Z_i, T_i) \neq 0 \quad \text{that is} \quad \alpha_1 \neq 0$$

The coefficient β_1 from the second stage of the least squares is estimated by the ratio of the estimated covariance between Z and Y to the estimated covariance between Z and T (Durbin, 1954) and is called an IV estimand (Angrist, et al., 1996):

$$\hat{\beta}_{IV} = \frac{\widehat{\text{cov}}(Z, Y)}{\widehat{\text{cov}}(Z, T)}$$

It can be proved that $\hat{\beta}_{IV}$ is an unbiased estimator of treatment effect β_1. From regression (2.1.6), α_1 is estimated by:

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}}{\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2} \tag{2.1.8}$$

$$\begin{aligned} E\left(\hat{\beta}_{IV} \mid \underset{\sim}{Z}, \underset{\sim}{T}\right) &= E\left[\frac{\sum_{i=1}^{N}\left\{Y_i \cdot \left(Z_i - \bar{Z}\right)\right\}}{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}}\right] = \frac{\sum_{i=1}^{N}\left\{\left(Z_i - \bar{Z}\right) \cdot E\left(Y_i\right)\right\}}{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}} \\ &= \frac{\sum_{i=1}^{N}\left\{\left(Z_i - \bar{Z}\right) \cdot E\left(\beta_0 + \beta_1 \cdot \hat{T}_i\right)\right\}}{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}} = \beta_1 \cdot \frac{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}}{\sum_{i=1}^{N}\left\{T_i \cdot \left(Z_i - \bar{Z}\right)\right\}} \\ &= \beta_1 \end{aligned}$$

$$E\left(\hat{\beta}_{IV}\right) = E\left[E\left(\hat{\beta}_{IV} \mid \underset{\sim}{Z}, \underset{\sim}{T}\right)\right] = \beta_1 \tag{2.1.9}$$

Imbens and Angrist proved that $\hat{\beta}_{IV}$ has an asymptotic normal distribution (Durbin, 1954; Imbens and Angrist, 1994).

$$\hat{\beta}_{IV} \sim AN\left(\beta_1, \frac{\operatorname{var}\left(\varepsilon_i\right) \cdot \sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2}{\left[\sum_{i=1}^{N}\left\{\left(Z_i - \bar{Z}\right) \cdot \left(T_i - \bar{T}\right)\right\}\right]^2}\right). \tag{2.1.10}$$

Comparing the variance of $\hat{\beta}_{IV}$ to the variance of $\hat{\beta}_1$, we see that an instrumental variable causes a certain loss of efficiency (Durbin, 1954)

$$\begin{aligned} \operatorname{var}\left(\hat{\beta}_1 \mid \underset{\sim}{T}, \underset{\sim}{U}\right) &= \operatorname{var}\left[\frac{\sum_{i=1}^{N}\left\{Y_i \cdot \left(T_i - \bar{T}\right)\right\}}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2}\right] = \frac{\sum_{i=1}^{N}\left\{\left(T_i - \bar{T}\right)^2 \cdot \operatorname{var}\left(Y_i\right)\right\}}{\left\{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2\right\}^2} \\ &= \frac{\operatorname{var}\left(\varepsilon_i\right)}{\sum_{i=1}^{N}\left(T_i - \bar{T}\right)^2} \end{aligned} \tag{2.1.11}$$

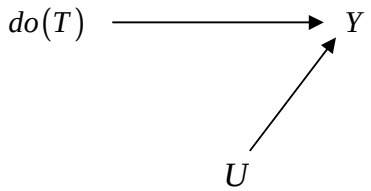
$$\frac{\operatorname{var}\left(\hat{\beta}_1\right)}{\operatorname{var}\left(\hat{\beta}_{IV}\right)} = \frac{\operatorname{var}\left(\varepsilon_i\right)}{\sum_{i=1}^{N}\left(T_i-\bar{T}\right)^2} \cdot \frac{\left[\sum_{i=1}^{N}\left\{\left(Z_i-\bar{Z}\right)\cdot\left(T_i-\bar{T}\right)\right\}\right]^2}{\operatorname{var}\left(\varepsilon_i\right)\cdot\sum_{i=1}^{N}\left(Z_i-\bar{Z}\right)^2}$$
$$= \frac{\left[\sum_{i=1}^{N}\left\{\left(Z_i-\bar{Z}\right)\cdot\left(T_i-\bar{T}\right)\right\}\right]^2}{\sum_{i=1}^{N}\left(Z_i-\bar{Z}\right)^2\cdot\sum_{i=1}^{N}\left(T_i-\bar{T}\right)^2} \le 1 \qquad (2.1.12)$$

A valid instrumental variable is one that causes variation in treatment status across the levels of the instrumental variable, and is uncorrelated with the unobserved confounders. An IV estimand measures the correlation between the instrumental variable and the outcome, which is then adjusted for the correlation between the instrumental variable and treatment status. In other words, the correlation between treatment and outcome is assessed indirectly by comparing both variables to a common reference variable.

2.2 Pearl's causal effect and IV core conditions

Pearl (2009) denoted the causal effect of T on Y as $P\left(Y \mid do\left(T\right)\right)$. In Pearl's notation, $P\left(Y \mid do\left(T\right)\right)$ is different from the conditional distribution of $P\left(Y \mid T\right)$. $do\left(T\right)$ stands for an intervention of T to induce the outcome of Y. The intervention of T is randomly performed, and theoretically is independent of any unobserved confounders. In the diagram 2.2.1, there is no arrow from U to $do\left(T\right)$.

Diagram 2.2.1



The average causal effect (ACE) can be expressed as the difference in expectations under different interventions of T (Didelez, and Sheehan, 2007):

$$ACE(t_1, t_2) = E(Y \mid do(T = t_1)) - E(Y \mid do(T = t_2)) \quad (2.2.1)$$

When the intervention is treatment with binary values, for example, t_1 is an active drug and t_2 is a placebo, the average causal effect of the active drug on the outcome is

$$E(Y \mid do(T = 1)) - E(Y \mid do(T = 0)) \quad (2.2.2)$$

When the intervention T is a continuous variable, an ordinary linear regression model $E(Y \mid do(T = t)) = \beta_0 + \beta_1 \cdot t$ is used to examine the causal effect of T on Y. The average causal effect is evaluated by β_1.

The definition of $do(T)$ is similar to Rosenbaum and Rubin's "causal effect" (1983) in terms of counterfactuals. For subject i, the response would be r_{1i} if he/she had received treatment 1, and r_{0i} if he/she had received treatment 0. The causal effect would be $(r_{1i} - r_{0i})$. The notation of $do(T)$ has the advantage of capturing both the counterfactual concept and randomized intervention.

Estimates of $P(Y \mid do(T))$ are not always available. Instead, the conditional probability $P(Y \mid T)$ from an observational study is often used to estimate the causal effect. Although $P(Y \mid T)$ is also a function of T, T is possibly correlated with unobserved confounders. With the assistance of an instrumental variable, unobserved confounders are adjustable and the average causal effect is identifiable.

Motivated by Pearl's causality, Didelez and Sheehan (2007) defined three core conditions for an instrumental variable in IVA. The notation and terminology were adapted from Greenland (2000) and Dawid (2003).

Condition 1: $Z \perp U$: Z must be independent of confounding between T and Y.

Condition 2: $Z \not\perp T$: Z must not be independent of T.

Condition 3: $Y \perp Z \,|\, (T,U)$: Conditionally on T and U, Z must be independent of Y.

The joint distribution of the 4 variables is:

$$P(Y,T,U,Z) = P(Y \,|\, T,U,Z) \cdot P(T \,|\, U,Z) \cdot P(U \,|\, Z) \cdot P(Z) \tag{2.2.3}$$

Because $Y \perp Z \,|\, (T,U)$ and $U \perp Z$, this may be expressed more compactly as

$$P(Y,T,U,Z) = P(Y \,|\, T,U) \cdot P(T \,|\, U,Z) \cdot P(U) \cdot P(Z) \tag{2.2.4}$$

Pearl (2009) used a directed acyclic graph (DAG) to illustrate the joint probability function.

Diagram 2.2.2

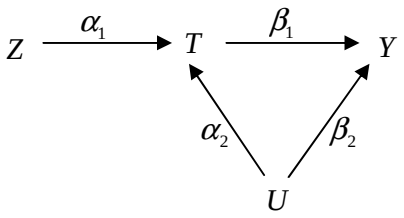


Diagram 2.2.2 presents the causal relationships among the four variables. Y is dependent on T and U, while T is conditional on Z and U. Z and U are completely independent.

Using the probability functions, we derive the expectations given in the 2SLS:

$$E(T \mid U,Z) = E\left(\alpha_0 + \alpha_1 \cdot Z + \alpha_2 \cdot U + \upsilon\right)$$
$$E\left(Y \mid U,T\right) = E\left(\beta_0 + \beta_1 \cdot T + \beta_2 \cdot U + \varepsilon\right) \tag{2.2.5}$$

Because $Y \perp Z \mid (T,U)$,

$$E\left(Y \mid U,T,Z\right) = E\left(Y \mid U,T\right)$$
$$= E\left(\beta_0 + \beta_1 \cdot T + \beta_2 \cdot U + \varepsilon\right) = \beta_0 + \beta_1 \cdot T + \beta_2 \cdot U \tag{2.2.6}$$

$$E\left(Y \mid Z\right) = E_U\left[E_{T|Z,U}\left[E\left(Y \mid U,T\right)\right]\right]$$
$$= E_U\left[E_{T|Z,U}\left(\beta_0 + \beta_1 \cdot T + \beta_2 \cdot U\right)\right]$$
$$= E_U\left[\beta_0 + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z + \alpha_2 \cdot U\right) + \beta_2 \cdot U\right] \tag{2.2.7}$$
$$= \beta_0 + \beta_1 \cdot \alpha_0 + \beta_1 \cdot \alpha_1 \cdot Z + \left(\beta_1 \cdot \alpha_2 + \beta_2\right) \cdot E\left(U\right)$$
$$= \beta_0^* + \beta_1 \cdot \alpha_1 \cdot Z \qquad \text{because} \quad Z \perp U$$

Therefore $\beta_{IV} = \beta_1 \cdot \alpha_1 / \alpha_1 = \beta_1$. (2.2.8)

2.3 Rubin's causal model

In 1996, Angrist, Imbens, and Rubin brought up a special IVA named Rubin's causal model. This model is well designed for the studies with a binary instrumental variable, binary treatment status, and binary outcome variable. The Rubin's IV estimand is imputed as a ratio of the difference in probability of developing disease between the two strata of instrument to the difference of exposure rates between the two strata of instrument.

For a sample of size N, let the instrument Z be coded with a dummy value of 1 or 0. The sample probabilities of developing disease for each stratum of the instrument are:

$$\overline{Y}_{Z=1} = \frac{\sum_{i=1}^{N} Y_i \cdot Z_i}{\sum_{i=1}^{N} Z_i} \qquad \text{and} \qquad \overline{Y}_{Z=0} = \frac{\sum_{i=1}^{N} Y_i \cdot \left(1 - Z_i\right)}{\sum_{i=1}^{N} \left(1 - Z_i\right)} \quad ,$$

Their difference is:

$$\bar{Y}_{Z=1}-\bar{Y}_{Z=0}=\frac{\sum_{i=1}^{N}Y_i\cdot Z_i}{\sum_{i=1}^{N}Z_i}-\frac{\sum_{i=1}^{N}Y_i\cdot(1-Z_i)}{\sum_{i=1}^{N}(1-Z_i)}=\frac{\left(N-\sum_{i=1}^{N}Z_i\right)\cdot\left(\sum_{i=1}^{N}Y_i\cdot Z_i\right)-\sum_{i=1}^{N}Z_i\cdot\sum_{i=1}^{N}Y_i\cdot(1-Z_i)}{\sum_{i=1}^{N}Z_i\cdot\left(N-\sum_{i=1}^{N}Z_i\right)}$$

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot Z_i-\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}{\frac{1}{N}\sum_{i=1}^{N}Z_i-\frac{1}{N}\sum_{i=1}^{N}Z_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}$$

Let the treatment be exposed ($T=1$) and not exposed ($T=0$). The sample exposure rates from each stratum of the instrument are:

$$\bar{T}_{Z=1}=\frac{\sum_{i=1}^{N}T_i\cdot Z_i}{\sum_{i=1}^{N}Z_i}\qquad\text{and}\qquad\bar{T}_{Z=0}=\frac{\sum_{i=1}^{N}T_i\cdot(1-Z_i)}{\sum_{i=1}^{N}(1-Z_i)}$$

Their difference is:

$$\bar{T}_{Z=1}-\bar{T}_{Z=0}=\frac{\sum_{i=1}^{N}T_i\cdot Z_i}{\sum_{i=1}^{N}Z_i}-\frac{\sum_{i=1}^{N}T_i\cdot(1-Z_i)}{\sum_{i=1}^{N}(1-Z_i)}=\frac{\left(N-\sum_{i=1}^{N}Z_i\right)\cdot\left(\sum_{i=1}^{N}T_i\cdot Z_i\right)-\sum_{i=1}^{N}Z_i\cdot\sum_{i=1}^{N}T_i\cdot(1-Z_i)}{\sum_{i=1}^{N}Z_i\cdot\left(N-\sum_{i=1}^{N}Z_i\right)}$$

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}T_i\cdot Z_i-\frac{1}{N}\sum_{i=1}^{N}T_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}{\frac{1}{N}\sum_{i=1}^{N}Z_i-\frac{1}{N}\sum_{i=1}^{N}Z_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}$$

The ratio of the two differences is called Rubin's IV estimand.

$$\hat{\beta}_{IV,R}=\frac{\bar{Y}_{Z=1}-\bar{Y}_{Z=0}}{\bar{T}_{Z=1}-\bar{T}_{Z=0}}=\frac{\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot Z_i-\frac{1}{N}\sum_{i=1}^{N}Y_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}{\frac{1}{N}\sum_{i=1}^{N}T_i\cdot Z_i-\frac{1}{N}\sum_{i=1}^{N}T_i\cdot\frac{1}{N}\sum_{i=1}^{N}Z_i}=\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum_{i=1}^{N}(T_i-\bar{T})(Z_i-\bar{Z})}=\frac{\widehat{\mathrm{cov}}(Z_i,Y_i)}{\widehat{\mathrm{cov}}(Z_i,T_i)}$$

$\hat{\beta}_{IV,R}$ is also named as Local Average Treatment Effect (LATE). It is a consistent estimator of the average causal effect of T on Y from marginal population if and

only if the following five assumptions are satisfied. A marginal population is defined as those patients who receive the same treatment as they are assigned to.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)

a. If $Z_i = Z_i'$, then $T_i(Z) = T_i(Z')$

b. If $Z_i = Z_i'$ and $T_i = T_i'$, then $Y_i(Z,T) = Y_i(Z',T')$

SUTVA assumes that one unit's outcome is not affected by another unit's treatment assignment. It goes beyond the concept of independence (Wikipedia, 2010). A violation example is given in Wikipedia, The Free Encyclopedia. Joe and Mary live in the same house. They both receive anti-hypertension treatment. Mary cooks for both of them. Mary does not cook salty foods if she does not take the drug, but she does cook salty foods when she takes the drug. Mary's treatment assignment affects both Mary and Joe's diet, and a high salt diet is supposed to increase blood pressure. Therefore, not only is Joe's blood pressure affected by his treatment assignment, but it is also affected by Mary's treatment assignment. The unstable unit treatment value causes a difficulty in identifying the causal effect from the treatment.

Assumption 2: Random Assignment

$\Pr(Z = c) = \Pr(Z = c')$ for any $c \,\&\, c'$

Random assignment assumes that the probability of being assigned to any value of the instrumental variable is equal for all patients. This assumption assures that the values of the instrumental variable are independent on all observed or unobserved confounders.

Assumption 3: Exclusion Restriction

$Y(Z,T) = Y(Z',T)$ for all Z, Z' and for all T

The exclusion restriction assumes that the effect of treatment on a patient's outcome stays the same on any level of the instrumental variable.

Assumption 4: Nonzero Average Causal Effect of Z on T

Assumption 4 states that the probabilities of receiving treatment are determined by the value of the instrumental variable. On average, there is a nonzero linear correlation between the treatment values and instrument values.

Assumption 5: Monotonicity

Under the assumption of monotonicity, the causal effect of Z on T is one-way, not two-way. For example, $Z=1$ can cause $T=1$, or $Z=1$ can cause $T=0$, but $Z=1$ is not allowed to cause $T=1$ in one case, but $T=0$ in another case. In other words, no patients are allowed to intentionally get the opposite treatment to the one they are assigned to.

We may illustrate Rubin's causal model using an example on evaluation of the effect of serving in the military on health outcomes. During the Vietnam War, in the United States, being drafted into military service was determined by randomly assigned lottery numbers. Those with low lottery numbers ($Z=1$) would have served in the military ($T=1$), and those with high lottery numbers ($Z=0$) would not have served in the military ($T=0$). In reality, there were non-compliers who always served in the military or who never served in the military regardless of which lottery numbers they were assigned. The marginal population include all the compliers, i.e., those who received low lottery numbers ($Z=1$), and served in the military ($T=1$), or those who received high lottery numbers ($Z=0$), and did not serve in the military ($T=0$). The lottery numbers are treated as an instrumental variable in this example. Health outcomes from the study population are compared between those who received low lottery numbers and those who received high lottery numbers. Elevated mortality is

found among men with low lottery numbers (Hearst, Newman, and Hulley 1986). After the five assumptions are carefully examined, it is concluded that the elevated mortality is from the marginal population due to their history of military service (Angrist, et al. 1996). Military service during the Vietnam War ($T = 1$) has a negative impact on the mortality.

Assumption 1: SUTVA. One person's health outcomes were not affected by another person's military service status.

Assumption 2: Random Assignment. The lottery numbers ($Z = 1$ or $Z = 0$) were randomly assigned to men who were born between 1950 and 1952 based on their birth dates.

Assumption 3: Exclusion Restriction. The amount of impact of military service on health outcomes was the same for all men regardless of high or low number assignment.

Assumption 4: Nonzero Average Causal Effect of Z on T. The majority of the participants were compliers, that is, men who received low lottery numbers ($Z = 1$) were more likely to serve in the army ($T = 1$), and men who received high lottery numbers ($Z = 0$) were more likely not to serve in the army ($T = 0$).

Assumption 5: Monotonicity. Non-compliers include those who ignored the lottery numbers, and always voluntarily committed to military service ($T = 1$), or who never entered into military service ($T = 0$). Non-compliers who were induced to avoid military service ($T = 0$) by the low lottery numbers ($Z = 1$) are not allowed.

2.4 Comparison of the assumptions in 2SLS, IV core conditions, and Rubin's causal model

Despite different notations and terminologies, assumptions required in the IVA are comparable. Comparisons are made in Table 2.4.1.

Table 2.4.1 Comparison of the assumptions in 2SLS, IV core conditions, and Rubin's causal model

Two-stage Least Squares	**IV Core Conditions**	**Rubin Causal Model**
Y_is are independent observations.		SUTVA
$\text{cov}(Z_i, \varepsilon_i) = 0$, and $\text{cov}(Z_i, \upsilon_i) = 0$	$Z \perp U$	Random Assignment
$\text{cov}(Z_i, Y_i) \neq 0$ if and only if $\text{cov}(Y_i, T_i) \neq 0$	$Y \perp Z \mid (T, U)$	Exclusion Restriction
$\text{cov}(Z_i, T_i) \neq 0$ or $\alpha_1 \neq 0$	$Z \not\perp T$	Nonzero Average Causal Effect of Z on T
either $\text{cov}(Z_i, T_i) < 0$ or $\text{cov}(Z_i, T_i) > 0$, but not both in the study		Monotonicity

2.5 Generalized method of moments (GMM)

Foster (1997) first applied IVA to nonlinear models such as logistic regression by using the Generalized Method of Moments (GMM). Johnston, Gustafson, Levy, et al. (2008) extended the GMM instrumental variable analysis (GMM IVA) to the other generalized linear models, such as Poisson regression. In their opinion, the 2SLS would not produce the consistent parameter estimates in nonlinear models by simply replacing the second stage of ordinary least square with a generalized linear model. IVA could be conducted using GMM. In GMM, a set of estimator-defining equations (Hansen, 1982 and 1985; Hansen and Singleton, 1982) are identified first. These equations include population moments, and are solved simultaneously for solutions of the population moments. The estimated population moments are believed to be a

consistent approach to the true values. A simple example of the estimator-defining equations is from the ordinary least square:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \cdots + \beta_J \cdot X_{iJ} + \varepsilon_i \qquad i = 1, ..., N \tag{2.5.1}$$

i represents the ith observations, and there are a total of J explanatory variables. One way to obtain the solutions of the parameters β_1 to β_J is to minimize the sum of squares of the residuals.

$$\sum_{i=1}^{N} (\varepsilon_i)^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \beta_2 \cdot X_{i2} - \cdots - \beta_J \cdot X_{iJ})^2 \tag{2.5.2}$$

Take the derivative on the right side of equation (2.5.2) with respect to the βs, and set them equal to zero:

$$\begin{aligned}
&\sum_{i=1}^{N} X_{i1} \cdot (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \beta_2 \cdot X_{i2} - \cdots - \beta_J \cdot X_{iJ}) = 0 \\
&\sum_{i=1}^{N} X_{i2} \cdot (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \beta_2 \cdot X_{i2} - \cdots - \beta_J \cdot X_{iJ}) = 0 \\
&\qquad\qquad \vdots \\
&\sum_{i=1}^{N} X_{iJ} \cdot (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \beta_2 \cdot X_{i2} - \cdots - \beta_J \cdot X_{iJ}) = 0
\end{aligned} \tag{2.5.3}$$

Equations (2.5.3) are actually the estimator-defining equations. By solving these equations simultaneously, we are able to obtain consistent estimators for all the βs.

Another way to view these equations is that all the explanatory variables are assumed to be uncorrelated with residuals.

$$\text{cov}(\underset{\sim}{X}_i, Y_i - \underset{\sim}{\beta}^T \cdot \underset{\sim}{X}_i) = 0 \tag{2.5.4}$$

where $\underset{\sim}{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iJ} \end{pmatrix}$ and $\underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_J \end{pmatrix}$

Both equations (2.5.2) and (2.5.4) result in the same set of equations in (2.5.3).

In generalized linear models, there are three components. The random component indicates the distribution of independent outcomes, Y_i s. The systematic components specifies a linear regression on a set of explanatory variables, X_{ij} s. Finally, the link is a function links the mean of the random component to the systematic components as in equation (2.5.5).

$$g\left(E\left(Y_i\right)\right)=\beta_0+\beta_1\cdot X_{i1}+\beta_2\cdot X_{i2}+\cdots+\beta_J\cdot X_{iJ} \tag{2.5.5}$$

Equation (2.5.5) can be rewritten as (2.5.6):

$$E\left(Y_i\right)=f\left(\underset{\sim}{X}_i;\underset{\sim}{\beta}\right) \tag{2.5.6}$$

where $f\left(X;\beta\right)$ is a linear or nonlinear function.

$$Y_i=E\left(Y_i\right)+\varepsilon_i=f\left(\underset{\sim}{X}_i;\underset{\sim}{\beta}\right)+\varepsilon_i \tag{2.5.7}$$

The estimator-defining equations in GMM for the nonlinear regression are expressed as:

$$\sum_{i=1}^{N}X_{ij}\cdot\varepsilon_i=\sum_{i=1}^{N}X_{ij}\cdot\left[Y_i-f\left(\underset{\sim}{X}_i;\underset{\sim}{\beta}\right)\right]=0 \qquad j=1,...,J \tag{2.5.8}$$

Again, equations in (2.5.8) assume that the explanatory variables and residuals do not co-vary.

In the case that there are unobserved confounders, the equations in (2.5.8) do not hold. A set of instrumental variables are then introduced. They replace the corresponding explanatory variables to form the estimator-defining equations. This is called GMM IVA. As defined, the instrumental variables are independent of the residuals.

$$\sum_{i=1}^{N}Z_{ik}\cdot\varepsilon_i=\sum_{i=1}^{N}Z_{ik}\cdot\left[Y_i-f\left(\underset{\sim}{X}_i;\underset{\sim}{\beta}\right)\right]=0 \qquad k=1,...,K \tag{2.5.9}$$

k represents the kth instrumental variable. Explanatory variables Xs that are independent to the residuals are counted as their own instruments (Foster, 1997).

2.6 Nonlinear Wald type methods

After Pearl's causal effect using the $do(.)$ operator (2009) was established, Didelez, Meng, and Sheehan (2010) further presented nonlinear Wald type methods in IVA based on the three IV core conditions (Didelez, and Sheehan, 2007) and an additional assumption of no interaction terms in the models. In the case of a binary instrumental variable Z, let

$$\alpha_1 = E(T_i \mid Z_i = 1) - E(T_i \mid Z_i = 0). \tag{2.6.1}$$

α_1 in equation (2.6.1) is essentially from the model (2.6.2),

$$E(T_i \mid Z_i = z, U_i = u) = \alpha_1 \cdot z + h_1(u) \tag{2.6.2}$$

where $h_1(u)$ in equation (2.6.2) is a function of the unobserved confounder U.

For a log-linear model,

$$\log E(Y_i \mid T_i = t, U_i = u) = \log E(Y_i \mid do(T_i = t), U_i = u) = \beta_1 \cdot t + h_2(u) \tag{2.6.3}$$

Again, $h_2(u)$ in equation (2.6.3) is a function of the unobserved confounder U, but it is different from $h_1(u)$. Since T is correlated with U, and function $h_2(u)$ is unknown, the instrumental variable Z replaces U and T.

$$\log E(Y_i \mid Z_i = z) = \gamma_1 \cdot z + h_3(u) \tag{2.6.4}$$

With the assumption of Z independent of U, it is not necessary to collect the actual values of $h_3(u)$, which is another function of the unobserved confounder U. Omitting $h_3(u)$ from model (2.6.4) does not change the value of γ_1. It is called

collapsibility for γ_1 over U (Greenland, Robins, and Pearl, 1999). In generalized linear models, collapsibility occurs for an identity link function or a log link function, but not for a logit link function (Gail, Wieand and Piantadosi, 1984; Gail, 1986). Equation (2.6.2) has a identity link function, so omitting $h_1(u)$ from model (2.6.2) does not change the value of α_1. β_1 can then be imputed as a ratio of the coefficient from the log-linear regression model of Y on Z to the coefficient of the linear regression model of T on Z. The Wald relative risk (WaldRR) is just an exponential of β_1. This method is called the two-stage quasi maximum likelihood (Mullahy, 1997).

$$\beta_1 = \frac{\log E(Y \mid Z=1) - \log E(Y \mid Z=0)}{E(T \mid Z=1) - E(T \mid Z=0)} \tag{2.6.5}$$

$$RR(Y \mid Z) = \exp\left[\log E(Y \mid Z=1) - \log E(Y \mid Z=0)\right]$$

$$WaldRR = RR(Y \mid Z)^{1/\alpha_1} = \exp(\beta_1) \tag{2.6.6}$$

In a logistic regression model,

$$\text{logit}\left\{E(Y_i \mid T_i = t, U_i = u)\right\} = \text{logit}\left\{E(Y_i \mid do(T_i = t), U_i = u)\right\} = \beta_1 \cdot t + h_2(u) \tag{2.6.7}$$

Let Z replacing T and U,

$$\text{logit}\left\{E(Y_i \mid Z_i = z)\right\} = \gamma_1 \cdot z + h_3(u), \tag{2.6.8}$$

However, for logistic regression, γ_1 is not collapsible over U. Omitting $h_3(u)$ changes the value γ_1. $\hat{\beta}_1$ obtained as a ratio of estimated coefficient from the logistic regression model of Y on Z to the estimated coefficient of the linear regression model of T on Z is a biased estimator of the true causal odds ratio (COR). The Wald type odds ratio is given in (2.6.9).

$$OR(Y \mid Z) = \exp\left[\text{logit}\left\{E(Y \mid Z=1)\right\} - \text{logit}\left\{E(Y \mid Z=0)\right\}\right]$$

$$WaldOR = OR(Y \mid Z)^{1/\alpha_1} \tag{2.6.9}$$

Wald's estimator is originally brought up in the case of fitting straight lines with two variables, both of them having uncorrelated errors (Wald, 1940). When Wald's estimators, such as WaldRR and WaldOR, are used in IVA as IV estimators, they are called Wald type estimators by the authors. In the logistic regression model, WaldOR is approximately equal to the COR when sample size is large. As another measurement of the causal effect, the true causal relative risk (CRR) can be calculated by integrating out U in the logistic regression model (Didelez, et al., 2010).

$$CRR = \frac{\int \left[1+\exp\left(-\beta_0-\beta_1-\beta_2 \cdot U\right)\right]^{-1} f(U)\, dU}{\int \left[1+\exp\left(-\beta_0-\beta_2 \cdot U\right)\right]^{-1} f(U)\, dU} \tag{2.6.10}$$

where $f(U)$ is a density function of U.

Chapter 3

IVA in Generalized Linear Models (GLM)

As stated in Chapter 2, in linear regression models, the IVA can be implemented using 2SLS, and Rubin's causal model. Rubin's causal model is a variation of the 2SLS when the instrumental variable and treatment variable are both binary. Both 2SLS and Rubin's causal model estimate the difference in treatment effects. In nonlinear regression models, so far, there are GMM and Wald type methods which can be applied in IVA. They are designed to provide unbiased estimators of the multiplicative treatment effect such as rate ratio and odds ratio.

In this chapter, we discuss issues in GMM IVA. We use principal stratification to illustrate the problem of estimating nonlinear causal effects in IVA. We propose a two-stage likelihood-based IVA model to estimate the nonlinear causal effect assuming the distribution of the unobserved confounder is known.

3.1 GMM

Let's consider a simple example of GMM. If we only have one explanatory variable that is the treatment T, the generalized linear model is:

$$g\left(E\left(Y_i\right)\right) = \beta_0^* + \beta_1^* \cdot T_i \tag{3.1.1}$$

$$E\left(Y_i\right) = f\left(\beta_0^* + \beta_1^* \cdot T_i\right)$$

$$Y_i = f\left(\beta_0^* + \beta_1^* \cdot T_i\right) + \varepsilon_i^*$$

By assuming that the treatment T does not co-vary with ε^*, we obtain the estimator-defining equation:

$$\sum_{i=1}^{N} T_i \cdot \varepsilon_i^* = \sum_{i=1}^{N} T_i \cdot \left[Y_i - f\left(\beta_0^* + \beta_1^* \cdot T_i\right) \right] = 0 \tag{3.1.2}$$

Ordinary least squares is a special case of GMM when $f\left(\beta_0 + \beta_1 \cdot T\right) = \beta_0 + \beta_1 \cdot T$ (Foster, 1997).

If we have an unobserved confounder that is omitted from the model (3.1.1), the true model should be:

$$g\left(E\left(Y_i\right)\right) = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i \tag{3.1.3}$$

$$E\left(Y_i\right) = f\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right)$$

$$Y_i - f\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) = \varepsilon_i$$

The estimator defining equation becomes:

$$\sum_{i=1}^{N} T_i \cdot \varepsilon_i = \sum_{i=1}^{N} T_i \cdot \left[Y_i - f\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) \right] = 0 \tag{3.1.4}$$

Since equation (3.1.4) holds true, equation (3.1.2) does not hold any more. In GMM IVE (Johnston, et al. 2008), it results in a new estimator defining equation that involves an instrumental variable Z.

$$\sum_{i=1}^{N} Z_i \cdot \varepsilon_i^* = \sum_{i=1}^{N} Z_i \cdot \left[Y_i - f\left(\beta_0^* + \beta_1^* \cdot T_i\right) \right] = 0 \tag{3.1.5}$$

Equation (3.1.5) is not always true when $f\left(\beta_0^* + \beta_1^* \cdot T_i\right) \neq f\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right)$

$$\begin{aligned} & E\left(Z_i \cdot \varepsilon_i^*\right) = E\left\{ Z_i \cdot \left[Y_i - f\left(\beta_0^* + \beta_1^* \cdot T_i\right) \right] \right\} \\ & \neq E\left\{ Z_i \cdot \left[Y_i - f\left(\beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) \right] \right\} = E\left(Z_i \cdot \varepsilon_i\right) = 0 \end{aligned} \tag{3.1.6}$$

When the unobserved confounder has a linear relationship with the expected outcome, equation (3.1.3) becomes:

$$Y_i = f\left(\beta_0^* + \beta_1^* \cdot T_i\right) + \beta_2^* \cdot U_i + \varepsilon_i \tag{3.1.7}$$

The estimator-defining equation in GMM IVE does hold because $\text{cov}\left(Z_i \cdot \varepsilon_i\right)$=0 and $\text{cov}\left(Z_i \cdot U_i\right)$=0.

$$\begin{aligned} & E\left(Z_i \cdot \varepsilon_i^*\right) = E\left\{Z_i \cdot \left[Y_i - f\left(\beta_0^* + \beta_1^* \cdot T_i\right)\right]\right\} \\ & = E\left\{Z_i \cdot \left[Y_i - f\left(\beta_0^* + \beta_1^* \cdot T_i\right) - \beta_2^* \cdot U_i + \beta_2^* \cdot U_i\right]\right\} \\ & = E\left\{Z_i \cdot \left[\varepsilon_i + \beta_2^* \cdot U_i\right]\right\} = 0 \end{aligned} \tag{3.1.8}$$

However, model (3.1.7) is not a standard generalized linear model. The association between the treatment T and confounder U is not easy to define and interpret.

3.2 Principal stratification

The study of PADT among men with localized prostate cancer is used as an example in our presentation. Using the algorithm and notation from Zhang (2004), we partition study patients into four categories.

- Compliers: patients who lived in PADT high usage areas and received PADT, or patients who lived in PADT low usage areas and received conservative management (CM).
- Always-takers: patients who received PADT regardless of where they lived.
- Never-takers: patients who received CM regardless of where they lived.
- Defiers: patients who intentionally receive CM as residents of PADT high usage areas, or patients who intentionally receive PADT as residents of PADT low usage areas.

A subtle distinction is that the patient alone is not a “complier”, “always taker”, or “never-taker”. Rather, it is really the patient and doctor together, a combinational unit, that is a “complier”, “always taker”, or “never-taker”. Let $\phi_n, \phi_a, \phi_c,$ and ϕ_d denote the population proportions of never-takers, always-takers, compliers, and defiers

respectively. Let $g_{CZ}(y)$ be the distribution of outcome Y_i for patients of category C $(C = n, a, c, d)$ and $Z = 0,1$. Let $f_{ZT}(y)$ be the distribution of observed outcome Y_i for patients of $Z = 0,1$ and $T = 0,1$. Under the exclusion restriction assumption, the distribution of outcome Y_i of never-takers or always-takers does not vary with the values of $Z = 0,1$. That is, $g_{n0}(y) = g_{n1}(y) = g_n(y)$ and $g_{a0}(y) = g_{a1}(y) = g_a(y)$. From observed outcome Y_i, they are identified as $g_n(y) = f_{10}(y)$ and $g_a(y) = f_{01}(y)$. Under the assumption of monotonicity, the population proportion of defiers is 0. Since $\phi_n = \Pr(T_i = 0 \mid Z_i = 1), \phi_a = \Pr(T_i = 1 \mid Z_i = 0)$, so $\phi_c = 1 - \phi_n - \phi_a$.

For patients from low PADT usage areas who receive CM, the distribution of Y_i is a mixture distribution from compliers and never-takers. Similarly, for patients from high PADT usage areas and receiving PADT, the distribution of Y_i is a mixture distribution from compliers and always-takers.

$$f_{00}(y) = \frac{\phi_c}{\phi_c + \phi_n} g_{c0}(y) + \frac{\phi_n}{\phi_c + \phi_n} g_n(y) \tag{3.2.1}$$

$$f_{11}(y) = \frac{\phi_c}{\phi_c + \phi_a} g_{c1}(y) + \frac{\phi_a}{\phi_c + \phi_a} g_a(y) \tag{3.2.2}$$

$$f_{10}(y) = g_n(y) \tag{3.2.3}$$

$$f_{01}(y) = g_a(y) \tag{3.2.4}$$

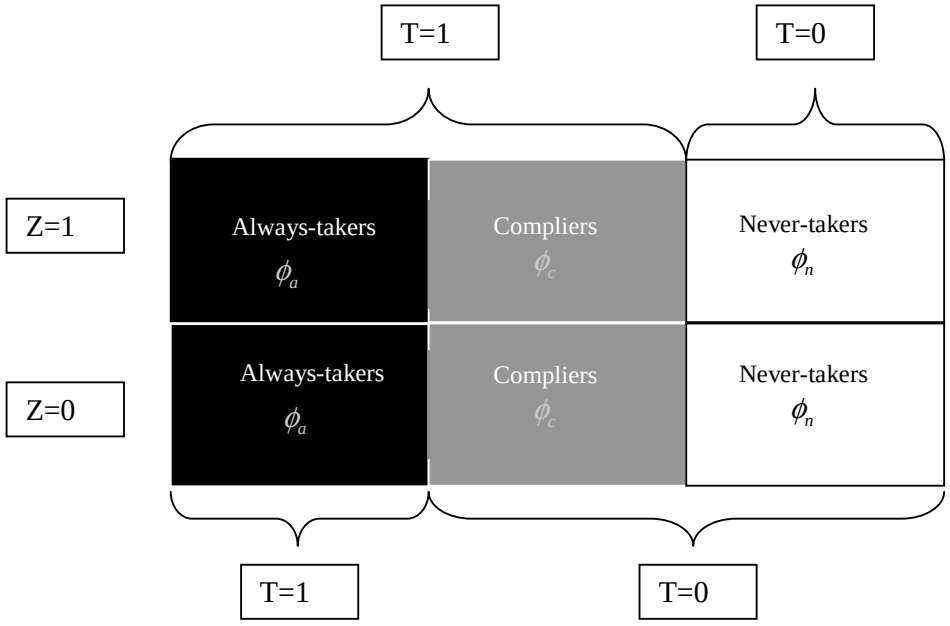
Solve equations to get $g_{c0}(y)$ and $g_{c1}(y)$:

$$g_{c0}(y) = \frac{\phi_n + \phi_c}{\phi_c} f_{00}(y) - \frac{\phi_n}{\phi_c} f_{10}(y) \tag{3.2.5}$$

$$g_{c1}(y) = \frac{\phi_a + \phi_c}{\phi_c} f_{11}(y) - \frac{\phi_a}{\phi_c} f_{01}(y) \tag{3.2.6}$$

Diagram 3.2.1 shows always-takers in black, never-takers in white, and compliers in grey. If outcomes of patients from $T=1$ and $T=0$ are compared, it implies that always-takers plus compliers in the high usage areas and never-takers plus compliers in the low usage areas are compared. These two groups are not comparable. Instead, if outcomes of patients from $Z=1$ and $Z=0$ are compared, outcomes from always-takers and never-takers are cancelled out for the high usage areas and low usage areas. The comparison is actually conducted on the compliers in different treatment groups. Therefore, the results represent the unbiased treatment effect.

Diagram 3.2.1



The expected outcome means from the high usage areas and low usage areas are:

$$E\left(Y_{Z=1}\right)=\phi_a\cdot E\left(Y_a\right)+\phi_c\cdot E\left(Y_{c,Z=1}\right)+\phi_n\cdot E\left(Y_n\right)$$
$$=\phi_a\cdot h\left(\beta_0+\beta_1\cdot T+\beta_2\cdot U_a\right)+\phi_c\cdot h\left(\beta_0+\beta_1\cdot T+\beta_2\cdot U_c\right)+\phi_n\cdot h\left(\beta_0+\beta_1+\beta_2\cdot U_n\right) \quad (3.2.7)$$

$$E\left(Y_{Z=0}\right)=\phi_a\cdot E\left(Y_a\right)+\phi_c\cdot E\left(Y_{c,Z=0}\right)+\phi_n\cdot E\left(Y_n\right)$$
$$=\phi_a\cdot h\left(\beta_0+\beta_1\cdot T+\beta_2\cdot U_a\right)+\phi_c\cdot h\left(\beta_0+\beta_1+\beta_2\cdot U_c\right)+\phi_n\cdot h\left(\beta_0+\beta_1+\beta_2\cdot U_n\right) \quad (3.2.8)$$

The expected difference in outcome means from the two types of health service areas is:

$$\begin{aligned} E\left(Y_{Z=1}-Y_{Z=0}\right) &= \phi_c \cdot \left[E\left(Y_{c,Z=1}\right)-E\left(Y_{c,Z=0}\right)\right] \\ &= \phi_c \cdot \left[h\left(\beta_0+\beta_1 \cdot T+\beta_2 \cdot U_c\right)-h\left(\beta_0+\beta_1+\beta_2 \cdot U_c\right)\right] \end{aligned} \tag{3.2.9}$$

Equation (3.2.9) shows that the sample mean difference between the two types of health service areas can be used to estimate the actual treatment effect in differences, especially, when $h\left(\beta_0+\beta_1 \cdot T+\beta_2 \cdot U\right)$ has an identity link, the sample mean difference is an unbiased estimator of β_1. If $h\left(\beta_0+\beta_1 \cdot T+\beta_2 \cdot U\right)$ is a nonlinear model, to calculate β_1 is not usually possible without knowing the distribution of U.

3.3 Likelihood function in IVA with linear models

As we have presented in section 2.2, in 2SLS, the expectation of the outcome variable given the value of instrumental variable is:

$$E\left(Y_i \mid Z_i\right)=\beta_0^*+\beta_1 \cdot \alpha_1 \cdot Z_i \tag{3.3.1}$$

We are able to show the same result as equation (3.3.1) using the likelihood function with the assumption of normality from all Y_i, T, and U. Let $Y_i \mid T,U$ has a normal density function with mean of $\beta_0+\beta_1 \cdot T+\beta_2 \cdot U$ and variance of σ_Y^2, $T_i \mid U,Z_i$ has a normal density function with mean of $\alpha_0+\alpha_1 \cdot Z_i+\alpha_2 \cdot U$ and variance of σ_T^2, and U has a normal density function with mean of μ_U and variance of σ_U^2.

$$L(\beta_0,\beta_1|\alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z}) = \prod_{i=1}^{N} f(Y_i|Z_i) = \prod_{i=1}^{N}\iint_{T\,U} f(Y_i,T,U|Z_i)\cdot dU\cdot dT$$

$$= \prod_{i=1}^{N}\iint_{T\,U} f(Y_i|T,U)\cdot f(T|U,Z_i)\cdot f(U)\cdot dU\cdot dT$$

$$= \prod_{i=1}^{N}\int_{U} \frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left[-\frac{(U-\mu_U)^2}{2\sigma_U^2}\right]\cdot$$

$$\int_{T} \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left[-\frac{(Y_i-\beta_0-\beta_1\cdot T-\beta_2\cdot U)^2}{2\sigma_Y^2}\right]\cdot\frac{1}{\sqrt{2\pi\sigma_T^2}}\exp\left[-\frac{(T-\alpha_0-\alpha_1\cdot Z_i-\alpha_2\cdot U)^2}{2\sigma_T^2}\right]\cdot dT\cdot dU$$

$$= \prod_{i=1}^{N}\int_{U} \frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left[-\frac{(U-\mu_U)^2}{2\sigma_U^2}\right]\cdot$$

$$\int_{T} \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left[-\frac{(Y_i-\beta_0-\beta_2\cdot U-\beta_1\cdot T)^2}{2\sigma_Y^2}\right]\cdot\frac{1}{\sqrt{2\pi\sigma_T^2}}\exp\left[-\frac{(\beta_1\cdot T-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U)^2}{2\beta_1^2\cdot\sigma_T^2}\right]\cdot dT\cdot dU$$

(3.3.2)

Using the substitution rule, let $X = \beta_1\cdot T$, then $dX = \beta_1\cdot dT$.

$$L(\beta_0,\beta_1|\alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z}) = \prod_{i=1}^{N}\int_{U} \frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left[-\frac{(U-\mu_U)^2}{2\sigma_U^2}\right]\cdot$$

$$\int_{X} \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left[-\frac{(Y_i-\beta_0-\beta_2\cdot U-X)^2}{2\sigma_Y^2}\right]\cdot\frac{1}{\sqrt{2\pi\cdot\beta_1^2\sigma_T^2}}\exp\left[-\frac{(X-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U)^2}{2\beta_1^2\cdot\sigma_T^2}\right]\cdot dX\cdot dU$$

(3.3.3)

We apply convolution integrals to the normal distribution functions (Vinga, and Almeida, 2004), i.e., if $G_1(X)$ and $G_2(X)$ are normal distributions of $N(a,A)$ and $N(b,B)$, the convolution $G_1\cdot G_2$ is defined as:

$$G_1\cdot G_2(W) = \int G_1(X)\cdot G_2(W-X)\cdot dX = G(W;a+b,A+B) \qquad (3.3.4)$$

where $G(W;a+b,A+B)$ is a density function of normal distribution W with mean of $a+b$ and variance of $A+B$

Let $W_i = Y_i-\beta_0-\beta_2\cdot U$,

$a_i = \beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U$, and $b_i = 0$,

$A=\beta_1^2\cdot\sigma_T^2$, and $B=\sigma_Y^2$.

Using the convolution formula, we see that $G(W_i)$ has a density function of normal distribution with mean of $\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U$ and variance of $\beta_1^2\cdot\sigma_T^2+\sigma_Y^2$.

$$\int_X \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left[-\frac{(Y_i-\beta_0-\beta_2\cdot U-X)^2}{2\sigma_Y^2}\right]\cdot\frac{1}{\sqrt{2\pi\cdot\beta_1^2\sigma_T^2}}\exp\left[-\frac{(X-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U)^2}{2\beta_1^2\cdot\sigma_T^2}\right]\cdot dX$$
$$=\int_X \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left[-\frac{(W_i-X)^2}{2\sigma_Y^2}\right]\cdot\frac{1}{\sqrt{2\pi\cdot\beta_1^2\sigma_T^2}}\exp\left[-\frac{(X-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U)^2}{2\beta_1^2\cdot\sigma_T^2}\right]\cdot dX$$
$$=\frac{1}{\sqrt{2\pi\cdot(\sigma_Y^2+\beta_1^2\sigma_T^2)}}\exp\left[-\frac{(W_i-a_i-b_i)^2}{2(\sigma_Y^2+\beta_1^2\sigma_T^2)}\right]$$
$$=\frac{1}{\sqrt{2\pi\cdot(\sigma_Y^2+\beta_1^2\sigma_T^2)}}\exp\left[-\frac{(Y_i-\beta_0-\beta_2\cdot U-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-\beta_1\cdot\alpha_2\cdot U)^2}{2(\sigma_Y^2+\beta_1^2\sigma_T^2)}\right]$$
$$=\frac{1}{\sqrt{2\pi\cdot(\sigma_Y^2+\beta_1^2\sigma_T^2)}}\exp\left[-\frac{(Y_i-\beta_0-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-(\beta_2+\beta_1\cdot\alpha_2)\cdot U)^2}{2(\sigma_Y^2+\beta_1^2\sigma_T^2)}\right] \tag{3.3.5}$$

Returning to the likelihood function (3.3.2):

$$L(\beta_0,\beta_1|\alpha_0,\alpha_1,\alpha_2,,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z})=\prod_{i=1}^{N}f(Y_i|Z_i)=\prod_{i=1}^{N}\iint_{T\,U}f(Y_i,T,U|Z_i)\cdot dU\cdot dT$$
$$=\prod_{i=1}^{N}\iint_{U}f(U)\cdot\int_T f(Y_i|T,U)\cdot f(T|U,Z_i)\cdot dT\cdot dU$$
$$=\prod_{i=1}^{N}\iint_{U}\frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left[\frac{(U-\mu_U)^2}{2\sigma_U^2}\right]\cdot\frac{1}{\sqrt{2\pi(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2)}}\exp\left[-\frac{(Y_i-\beta_0-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-(\beta_2+\beta_1\cdot\alpha_2)\cdot U)^2}{2(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2)}\right]\cdot dU$$
$$=\prod_{i=1}^{N}\iint_{U}\frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left[-\frac{\left[(\beta_2+\beta_1\cdot\alpha_2)U-(\beta_2+\beta_1\cdot\alpha_2)\mu_U\right]^2}{2\sigma_U^2\cdot(\beta_2+\beta_1\cdot\alpha_2)^2}\right]\cdot$$
$$\frac{1}{\sqrt{2\pi(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2)}}\exp\left[-\frac{(Y_i-\beta_0-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-(\beta_2+\beta_1\cdot\alpha_2)\cdot U)^2}{2(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2)}\right]\cdot dU \tag{3.3.6}$$

Again, using the substitution rule, let $X=(\beta_2+\beta_1\cdot\alpha_2)\cdot U$, then

$dX=(\beta_2+\beta_1\cdot\alpha_2)\cdot dU$.

$$
\begin{aligned}
&L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z}) = \prod_{i=1}^{N} f(Y_i \mid Z_i) \\
&= \prod_{i=1}^{N} \int_X \frac{1}{\sqrt{2\pi\sigma_U^2 \cdot (\beta_2+\beta_1\cdot\alpha_2)^2}} \exp\left[\frac{\left[X-(\beta_2+\beta_1\cdot\alpha_2)\mu_U\right]^2}{2\sigma_U^2\cdot(\beta_2+\beta_1\cdot\alpha_2)^2} \right] \cdot \\
&\frac{1}{\sqrt{2\pi\left(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2\right)}} \exp\left[\frac{(Y_i-\beta_0-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-X)^2}{2\left(\sigma_Y^2+\beta_1^2\cdot\sigma_T^2\right)} \right] \cdot dX
\end{aligned}
\tag{3.3.7}
$$

Using the same technique of convolution integrals for normal distribution functions,

let $W_i = Y_i - \beta_0 - \beta_1 \cdot \alpha_0 - \beta_1 \cdot \alpha_1 \cdot Z_i$,

$a = (\beta_2 + \beta_1 \cdot \alpha_2) \cdot \mu_U$, and $b = 0$,

$A = \sigma_U^2 \cdot (\beta_2 + \beta_1 \cdot \alpha_2)^2$, and $B = \sigma_Y^2 + \beta_1^2 \cdot \sigma_T^2$

By convolution, $G(W_i)$ has a normal density function with mean of $(\beta_2 + \beta_1 \cdot \alpha_2) \cdot \mu_U$ and variance of $\sigma_U^2 \cdot (\beta_2 + \beta_1 \cdot \alpha_2)^2 + \beta_1^2 \cdot \sigma_T^2 + \sigma_Y^2$.

$$
\begin{aligned}
&L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z}) = \prod_{i=1}^{N} f(Y_i \mid Z_i) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left[\sigma_U^2 \cdot (\beta_2+\beta_1\cdot\alpha_2)^2 + \beta_1^2\cdot\sigma_T^2 + \sigma_Y^2\right]}} \exp\left[-\frac{\left[Y_i-\beta_0-\beta_1\cdot\alpha_0-\beta_1\cdot\alpha_1\cdot Z_i-(\beta_2+\beta_1\cdot\alpha_2)\cdot\mu_U\right]^2}{2\left[\sigma_U^2\cdot(\beta_2+\beta_1\cdot\alpha_2)^2+\beta_1^2\cdot\sigma_T^2+\sigma_Y^2\right]} \right] \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left[\sigma_U^2 \cdot (\beta_2+\beta_1\cdot\alpha_2)^2 + \beta_1^2\cdot\sigma_T^2 + \sigma_Y^2\right]}} \exp\left[-\frac{\left[Y_i-\beta_0^*-\beta_1\cdot\alpha_1\cdot Z_i\right]^2}{2\left[\sigma_U^2\cdot(\beta_2+\beta_1\cdot\alpha_2)^2+\beta_1^2\cdot\sigma_T^2+\sigma_Y^2\right]} \right]
\end{aligned}
\tag{3.3.8}
$$

The likelihood function of (3.3.8) shows that Y given Z has a normal density function with mean of $\beta_0^* + \beta_1 \cdot \alpha_1 \cdot Z_i$. This result matches equation (3.3.1).

If the second stage in the 2SLS is a nonlinear equation, the conditional expectation of Y given Z is:

$$
E(Y \mid Z) = E_U \left\{ E_{T|Z,U} \left[h(\beta_0 + \beta_1 \cdot T + \beta_2 \cdot U) \right] \right\} \tag{3.3.9}
$$

Because the expectation of a function is not a function of expectation (Kelejian, 1971), the confounder of U is unable to be absorbed as a constant of $E(U)$. We need to find the distribution of the unobserved confounder and numerically integrate it over the probability measurement.

3.4 Likelihood function in IVA with nonlinear models

Inspired by the idea of true causal relative risk in equation (2.6.10) (Didelez, et al., 2010), we propose a two-stage likelihood-based IVA to estimate the multiplicative treatment effect. In general:

Stage 1:

$$L(\alpha_0,\alpha_1 \mid \alpha_2;\underset{\sim}{T},\underset{\sim}{Z}) = \prod_{i=1}^{N} f(T_i \mid Z_i) = \prod_{i=1}^{N} \int_U f(T_i \mid Z_i, U)\cdot dF_U(U) \tag{3.4.1}$$

Stage 2:

$$\begin{aligned} &L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z}) = \prod_{i=1}^{N} f(Y_i \mid Z_i) \\ &= \prod_{i=1}^{N} \int_U \int_{T|U} f(Y_i \mid U,T)\cdot dF_{T|U}(T,Z_i)\cdot dF_U(U) \end{aligned} \tag{3.4.2}$$

Assumptions: α_2, β_2 and $F_U(U)$ are known.

U is distributed with a probability density function (pdf) of $f_U(U)$ and a cumulative distribution function (cdf) of $F_U(U)$. Stage 2 is a conditional likelihood function given α_0 and α_1 obtained from stage 1. If T and U are continuous variables, the two-stage likelihood-based IVA is presented as (3.4.3) and (3.4.4). The example has been given in section 3.3 where both T and U are normally distributed.

Stage 1:

$$L(\alpha_0,\alpha_1\mid\alpha_2;\underset{\sim}{T},\underset{\sim}{Z})=\prod_{i=1}^{N}f(T_i\mid Z_i)=\prod_{i=1}^{N}\int_U f(T_i\mid Z_i,U)\cdot f(U)\cdot dU \tag{3.4.3}$$

Stage 2:

$$\begin{aligned}&L(\beta_0,\beta_1\mid\alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z})=\prod_{i=1}^{N}f(Y_i\mid Z_i)\\&=\prod_{i=1}^{N}\int_U\int_{T\mid U,Z_i}f(Y_i\mid U,T)\cdot f(T\mid U,Z_i)\cdot f(U)\cdot dT\cdot dU\end{aligned} \tag{3.4.4}$$

If T and U are discrete variables, the two-stage likelihood-based IVA model is expressed using probability mass functions. Our next example is based on the equations (3.4.5) and (3.4.6), where both T and U follow a Bernoulli distribution.

Stage 1:

$$L(\alpha_0,\alpha_1\mid\alpha_2;\underset{\sim}{T},\underset{\sim}{Z})=\prod_{i=1}^{N}P(T_i\mid Z_i)=\prod_{i=1}^{N}\sum_U P(T_i\mid Z_i,U)\cdot P(U) \tag{3.4.5}$$

Stage 2:

$$\begin{aligned}&L(\beta_0,\beta_1\mid\alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z})=\prod_{i=1}^{N}P(Y_i\mid Z_i)\\&=\prod_{i=1}^{N}\sum_U\sum_{T\mid U,Z_i}P(Y_i\mid U,T)\cdot P(T\mid U,Z_i)\cdot P(U)\end{aligned} \tag{3.4.6}$$

We let Y, Z, and U be all binary variables scored as 0 or 1. Logistic regression models are used in both stages.

$$\log\frac{E(T_i\mid Z_i,U_i)}{1-E(T_i\mid Z_i,U_i)}=\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U_i \tag{3.4.7}$$

$$\log\frac{E(Y_i\mid T_i,U_i)}{1-E(Y_i\mid T_i,U_i)}=\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i \tag{3.4.8}$$

Both equations (3.4.7) and (3.4.8) are non-collapsible for α_1 or β_1 over U (Greenland, Robins, and Pearl, 1999), that is, if the unobserved confounder U is omitted in the models, $\alpha_1^*\neq\alpha_1$, and $\beta_1^*\neq\beta_1$.

$$\log\frac{E\left(T_i \mid Z_i, U_i\right)}{1-E\left(T_i \mid Z_i, U_i\right)} = \alpha_0^* + \alpha_1^* \cdot Z_i \tag{3.4.9}$$

$$\log\frac{E\left(Y_i \mid T_i, U_i\right)}{1-E\left(Y_i \mid T_i, U_i\right)} = \beta_0^* + \beta_1^* \cdot T_i \tag{3.4.10}$$

We use likelihood functions in equations (3.4.5) and (3.4.6) to solve α_1 and β_1.

$$\begin{aligned}
&L\left(\alpha_0, \alpha_1 \mid \alpha_2; \underset{\sim}{T}, \underset{\sim}{Z}\right) = \prod_{i=1}^{N} P\left(T_i \mid Z_i\right) = \prod_{i=1}^{N} \sum_{U} P\left(T_i \mid Z_i, U\right) \cdot P\left(U\right) \\
&= \prod_{i=1}^{N} \sum_{U} \left[\left\{ \frac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{T_i} \cdot \left\{ \frac{1}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{(1-T_i)} \cdot \left(\mu_U\right)^U \cdot \left(1-\mu_U\right)^{(1-U)} \right] \\
&= \prod_{i=1}^{N} \left[\left\{ \frac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2\right)}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2\right)} \right\}^{T_i} \cdot \left\{ \frac{1}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2\right)} \right\}^{(1-T_i)} \cdot \left(\mu_U\right) + \right. \\
&\qquad \left. \left\{ \frac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i\right)}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i\right)} \right\}^{T_i} \cdot \left\{ \frac{1}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i\right)} \right\}^{(1-T_i)} \cdot \left(1-\mu_U\right) \right]
\end{aligned} \tag{3.4.11}$$

$$\begin{aligned}
&l\left(\alpha_0, \alpha_1 \mid \alpha_2; \underset{\sim}{T}, \underset{\sim}{Z}\right) = \log L\left(\alpha_0, \alpha_1 \mid \alpha_2; \underset{\sim}{T}, \underset{\sim}{Z}\right) = \sum_{i=1}^{N} \log\left\{ \sum_{U} P\left(T_i \mid Z_i, U\right) \cdot P\left(U\right) \right\} \\
&= \sum_{i=1}^{N} \log\left[\sum_{U} \left\{ \frac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{T_i} \cdot \left\{ \frac{1}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{(1-T_i)} \cdot \left(\mu_U\right)^U \cdot \left(1-\mu_U\right)^{(1-U)} \right]
\end{aligned} \tag{3.4.12}$$

Estimators of α_0 and α_1 are obtained by maximizing the log-likelihood function in (3.4.12).

Let:

$$M_i = \sum_{U} \left\{ \frac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{T} \cdot \left\{ \frac{1}{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)} \right\}^{(1-T)} \cdot \left(\mu_U\right)^U \cdot \left(1-\mu_U\right)^{(1-U)}$$

$$\frac{dl}{d\alpha_0} = \sum_{i=1}^{N} \frac{\sum_{U} \left(\mu_U\right)^U \cdot \left(1-\mu_U\right)^{(1-U)} \cdot (-1)^{(1-T_i)} \cdot \dfrac{\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)}{\left\{1+\exp\left(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U\right)\right\}^2}}{M_i} \tag{3.4.13}$$

$$\frac{dl}{d\alpha_1}=\sum_{i=1}^{N}\frac{\sum_{U}(\mu_U)^{U}\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-T_i)}\cdot Z_i\cdot\dfrac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{\{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)\}^2}}{M_i} \quad (3.4.14)$$

$\hat{\alpha}_0$ and $\hat{\alpha}_1$ can be numerically solved using equation (3.4.13) and (3.4.14). When the sample size is large, the variances of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ can be ignored, and they are treated as constants.

$$\begin{aligned}
&L(\beta_0,\beta_1\,|\,\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=\prod_{i=1}^{N}P(Y_i\,|\,Z_i)\\
&=\prod_{i=1}^{N}\sum_{U}\sum_{T|Z_i,U}P(Y_i\,|\,T,U,Z_i)\cdot P(T\,|\,U,Z_i)\cdot P(U)\\
&=\prod_{i=1}^{N}\sum_{U}\sum_{T|Z_i,U}P(Y_i\,|\,T,U)\cdot P(T\,|\,U,Z_i)\cdot P(U)\\
&=\prod_{i=1}^{N}\sum_{U}\sum_{T|Z_i,U}\left[\left\{\frac{\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{(1-Y_i)}\cdot\right.\\
&\quad\left.\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{T}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^{U}\cdot(1-\mu_U)^{(1-U)}\right]\\
&=\prod_{i=1}^{N}\left[\left\{\frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1+\beta_2)}\right\}^{(1-Y_i)}\cdot\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2)}\right\}\cdot\mu_U+\right.\\
&\quad\left\{\frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_2)}\right\}^{(1-Y_i)}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2)}\right\}\cdot\mu_U+\\
&\quad\left\{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1)}\right\}^{(1-Y_i)}\cdot\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i)}\right\}\cdot(1-\mu_U)+\\
&\quad\left.\left\{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0)}\right\}^{(1-Y_i)}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i)}\right\}\cdot(1-\mu_U)\right]
\end{aligned} \quad (3.4.15)$$

$$l(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=\log L(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})$$
$$=\sum_{i=1}^{N}\log\left\{\sum_U\sum_{T|Z_i,U}P(Y_i\mid T,U)\cdot P(T\mid U,Z_i)\cdot P(U)\right\}$$
$$=\sum_{i=1}^{N}\log\left[\sum_U\sum_{T|Z_i,U}\left\{\frac{\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{(1-Y_i)}\cdot\right.$$
$$\left.\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{T}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\right] \quad (3.4.16)$$

Let:

$$L_i=\sum_U\sum_{T|Z_i,U}\left[\left\{\frac{\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{(1-Y_i)}\cdot\right.$$
$$\left.\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{T}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\right] \quad (3.4.17)$$

$$\xi=\frac{\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}$$

$$\zeta_i=\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}$$

$$\theta=\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)$$

$$\frac{dl}{d\xi}=\sum_{i=1}^{N}\frac{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}}{L_i}$$

$$\frac{d\xi}{d\theta}=\frac{1}{(1+\theta)^2}$$

$$\frac{d\theta}{d\beta_0}=\theta \quad \text{and} \quad \frac{d\theta}{d\beta_1}=\theta\cdot T$$

$$\frac{dl}{d\beta_0} = \frac{dl}{d\xi} \cdot \frac{d\xi}{d\theta} \cdot \frac{d\theta}{d\beta_0} =$$

$$\sum_{i=1}^{N} \frac{\sum_{U} \sum_{T|Z_i,U} (\zeta_i)^T \cdot (1-\zeta_i)^{(1-T)} \cdot (\mu_U)^U \cdot (1-\mu_U)^{(1-U)} \cdot (-1)^{(1-Y_i)} \cdot \frac{\theta}{(1+\theta)^2}}{L_i} \tag{3.4.18}$$

$$\frac{dl}{d\beta_1} = \frac{dl}{d\xi} \cdot \frac{d\xi}{d\theta} \cdot \frac{d\theta}{d\beta_1} =$$

$$\sum_{i=1}^{N} \frac{\sum_{U} \sum_{T|Z_i,U} (\zeta_i)^T \cdot (1-\zeta_i)^{(1-T)} \cdot (\mu_U)^U \cdot (1-\mu_U)^{(1-U)} \cdot (-1)^{(1-Y_i)} \cdot \frac{\theta}{(1+\theta)^2} \cdot T}{L_i} \tag{3.4.19}$$

Set equations (3.4.18) and (3.4.19) to 0, and solve them for $\hat{\beta}_0$ and $\hat{\beta}_1$. The maximum likelihood estimator of β_1 is a consistent estimator of the treatment effect. The closed form for $\hat{\beta}_1$ is difficult to obtain because the derivatives of equation (3.4.19) involve a natural logarithm of the summation. $\hat{\beta}_1$ is obtained numerically.

In summary, α_2 quantifies the association between unobserved confounder and treatment status. β_2 quantifies the association between unobserved confounder and outcome. The maximum likelihood estimators of α_0 and α_1 from model (3.4.1) are calculated with a pre-defined value of α_2. $\hat{\beta}_{0IV}$ and $\hat{\beta}_{1IV}$, which are the estimators of β_0 and β_1 in model (3.4.2) using two-stage likelihood-based IVA, are then calculated with the pre-defined value of β_2.

The average Fisher's information matrix is used to estimate the sample variance-covariance matrix for $\hat{\beta}_{0IV}$ and $\hat{\beta}_{1IV}$.

$$I_{A11}(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\frac{dl}{d\beta_0}\right)^2=$$

$$\frac{1}{N}\cdot\sum_{i=1}^{N}\left\{\frac{\sum_{U}\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}}{L_i}\right\}^2$$

$$I_{A22}(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\frac{dl}{d\beta_1}\right)^2=$$

$$\frac{1}{N}\cdot\sum_{i=1}^{N}\left\{\frac{\sum_{U}\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T}{L_i}\right\}^2$$

$$I_{A12}(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=I_{A21}(\beta_0,\beta_1\mid\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})$$

$$=\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\frac{dl}{d\beta_0}\cdot\frac{dl}{d\beta_1}\right)$$

$$=\frac{1}{N}\cdot\sum_{i=1}^{N}\left[\left\{\frac{\sum_{U}\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}}{L_i}\right\}\cdot\right. \tag{3.4.20}$$

$$\left.\left\{\frac{\sum_{U}\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T}{L_i}\right\}\right]$$

The expected Fisher's information matrix can be obtained from following equations.

$$I_{11}(\beta_0,\beta_1\,|\,\beta_2,\alpha_0,\alpha_1,\alpha_2)=E\left[\frac{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\dfrac{\theta}{(1+\theta)^2}}{L_i}\right]^2$$

$$=\sum_Z\sum_{Y|Z}\left[\left\{\frac{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\dfrac{\theta}{(1+\theta)^2}}{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}}\right\}^2\cdot P(Y\,|\,Z)\cdot P(Z)\right]$$

$$=\sum_Z\sum_{Y|Z}\left[\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\dfrac{\theta}{(1+\theta)^2}\right\}^2}{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}}\cdot(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\right]$$

$$=\sum_Z\left\{(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\cdot\left[\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\dfrac{\theta}{(1+\theta)^2}\right\}^2}{\sum_U\sum_{T|Z,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\xi\right\}}+\right.\right.$$

$$\left.\left.\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\dfrac{\theta}{(1+\theta)^2}\right\}^2}{\sum_U\sum_{T|Z,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(1-\xi)\right\}}\right]\right\}$$

$$I_{22}(\beta_0,\beta_1\,|\,\beta_2,\alpha_0,\alpha_1,\alpha_2)=E\left[\frac{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\dfrac{\theta}{(1+\theta)^2}\cdot T}{L_i}\right]^2$$

$$=\sum_Z\sum_{Y|Z}\left[\left\{\frac{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\dfrac{\theta}{(1+\theta)^2}\cdot T}{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}}\right\}^2\cdot P(Y\,|\,Z)\cdot P(Z)\right]$$

$$=\sum_Z\sum_{Y|Z}\left[\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\dfrac{\theta}{(1+\theta)^2}\cdot T\right\}^2}{\sum_U\sum_{T|Z_i,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}\right\}}\cdot(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\right]$$

$$=\sum_Z\left\{(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\cdot\left[\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\dfrac{\theta}{(1+\theta)^2}\cdot T\right\}^2}{\sum_U\sum_{T|Z,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\xi\right\}}+\right.\right.$$

$$\left.\left.\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\dfrac{\theta}{(1+\theta)^2}\cdot T\right\}^2}{\sum_U\sum_{T|Z,U}\left[(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(1-\xi)\right]}\right]\right\}$$

$$
\begin{aligned}
&I_{12}(\beta_0,\beta_1|\beta_2,\alpha_0,\alpha_1,\alpha_2)=I_{21}(\beta_0,\beta_1|\beta_2,\alpha_0,\alpha_1,\alpha_2)\\
&=E\left[\left\{\frac{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}}{L_i}\right\}\cdot\right.\\
&\left.\left\{\frac{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y_i)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T}{L_i}\right\}\right]\\
&=\sum_Z\sum_{Y|Z}\left[\left\{\frac{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\frac{\theta}{(1+\theta)^2}}{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}}\right\}\cdot\right.\\
&\left.\left\{\frac{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T}{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}}\right\}\cdot P(Y|Z)\cdot P(Z)\right]\\
&=\sum_Z\sum_{Y|Z}\left[\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\frac{\theta}{(1+\theta)^2}\right\}}{\sum_U\sum_{T|Z,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}\right\}}\cdot\right.\\
&\left.\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(-1)^{(1-Y)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T\right\}\cdot(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\right]\\
&=\sum_Z\left\{(\mu_Z)^Z\cdot(1-\mu_Z)^{(1-Z)}\cdot\right.\\
&\left[\frac{\left\{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\frac{\theta}{(1+\theta)^2}\right\}\cdot\left\{\sum_U\sum_{T|Z_i,U}(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T\right\}}{\sum_U\sum_{T|Z_i,U}\left\{(\zeta_i)^T\cdot(1-\zeta_i)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\xi\right\}}\right.\\
&\left.\left.+\frac{\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\frac{\theta}{(1+\theta)^2}\right\}\cdot\left\{\sum_U\sum_{T|Z,U}(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T\right\}}{\sum_U\sum_{T|Z,U}\left\{(\zeta)^T\cdot(1-\zeta)^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot(1-\xi)\right\}}\right]\right\}
\end{aligned}
\tag{3.4.21}
$$

If the values of confounder U are observed, IVA is not needed. The expected Fisher's information matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained with the logistic regression model in (3.4.22).

$$\log\frac{E(Y_i\mid T_i,U_i)}{1-E(Y_i\mid T_i,U_i)}=\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i \tag{3.4.22}$$

$$\log\frac{E(T_i\mid U_i)}{1-E(T_i\mid U_i)}=\alpha_0^{*}+\alpha_1^{*}\cdot U_i \tag{3.4.23}$$

Let $\xi_i=\dfrac{\exp(\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i)}{1+\exp(\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i)}$ and $\theta_i=\exp(\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i)$

$$L(\beta_0,\beta_1,\beta_2\mid \underset{\sim}{Y},\underset{\sim}{T},\underset{\sim}{U})=\prod_{i=1}^{N}P(Y_i\mid T_i,U_i)=\prod_{i=1}^{N}\left\{\xi_i^{Y_i}\cdot(1-\xi_i)^{(1-Y_i)}\right\}$$

$$l(\beta_0,\beta_1,\beta_2\mid \underset{\sim}{Y},\underset{\sim}{T},\underset{\sim}{U})=\log L(\beta_0,\beta_1,\beta_2\mid \underset{\sim}{Y},\underset{\sim}{T},\underset{\sim}{U})=\sum_{i=1}^{N}\left\{Y_i\cdot\log(\xi_i)+(1-Y_i)\cdot\log(1-\xi_i)\right\}$$

$$\frac{dl}{d\beta_0}=\sum_{i=1}^{N}\frac{Y_i-\xi_i}{\xi_i\cdot(1-\xi_i)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_0}$$

$$\frac{dl}{d\beta_1}=\sum_{i=1}^{N}\frac{Y_i-\xi_i}{\xi_i\cdot(1-\xi_i)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_1}$$

$$I_{11}(\beta_0,\beta_1,\beta_2)=E\left[\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_0}\right]^2=E\left[\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right]^2$$

$$=\sum_{U}\sum_{T|U}\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right\}^2\cdot P(Y|T,U)\cdot P(T|U)\cdot P(U)\right]$$

$$=\sum_{U}\sum_{T|U}\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right\}^2\cdot\xi^{Y}\cdot(1-\xi)^{(1-Y)}\cdot\right.$$

$$\left.\left(\frac{\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}{1+\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}\right)^{T}\cdot\left(\frac{1}{1+\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}\right)^{(1-T)}\cdot(\mu_U)^{U}\cdot(1-\mu_U)^{(1-U)}\right]$$

$$=\sum_{U}\sum_{T|U}\left\{\frac{\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}{1+\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}\right\}^{T}\cdot\left\{\frac{1}{1+\exp(\alpha_0^{*}+\alpha_1^{*}\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^{U}\cdot(1-\mu_U)^{(1-U)}\cdot\left[\frac{\left\{\frac{\theta}{(1+\theta)^2}\right\}^2}{\xi}+\frac{\left\{\frac{\theta}{(1+\theta)^2}\right\}^2}{1-\xi}\right]$$

$$I_{22}(\beta_0,\beta_1,\beta_2)=E\left[\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_1}\right]^2=E\left[\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T_i\right]^2$$

$$=\sum_U\sum_{T|U}\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T\right\}^2\cdot P(Y|T_i,U_i)\cdot P(T|U)\cdot P(U)\right]$$

$$=\sum_U\sum_{T|U}\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T_i\right\}^2\cdot\xi^Y\cdot(1-\xi)^{(1-Y)}\cdot\right.$$

$$\left.\left\{\frac{\exp(\alpha_0^*+\alpha_1^*\cdot U)}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^T\cdot\left\{\frac{1}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\right]$$

$$=\sum_U\sum_{T|U}\left\{\frac{\exp(\alpha_0^*+\alpha_1^*\cdot U)}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^T\cdot\left\{\frac{1}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\left[\frac{\left\{\frac{\theta}{(1+\theta)^2}\cdot T\right\}^2}{\xi}+\frac{\left\{\frac{\theta}{(1+\theta)^2}\cdot T\right\}^2}{1-\xi}\right]$$

$$I_{12}(\beta_0,\beta_1,\beta_2)=I_{21}(\beta_0,\beta_1,\beta_2)=E\left[\left\{\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_0}\right\}\cdot\left\{\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{d\xi}{d\theta}\cdot\frac{d\theta}{d\beta_1}\right\}\right]$$

$$=E\left[\left\{\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right\}\cdot\left\{\frac{Y_i-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T_i\right\}\right]$$

$$=\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right\}\cdot\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\cdot T_i\right\}\cdot P(Y|T,U)\cdot P(T|U)\cdot P(U)\right]$$

$$=\sum_U\sum_{T|U}\sum_{Y|T,U}\left[\left\{\frac{Y-\xi}{\xi\cdot(1-\xi)}\cdot\frac{\theta}{(1+\theta)^2}\right\}^2\cdot T_i\cdot(\xi)^Y\cdot(1-\xi)^{(1-Y)}\cdot\right.$$

$$\left.\left\{\frac{\exp(\alpha_0^*+\alpha_1^*\cdot U)}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^T\cdot\left\{\frac{1}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\right]$$

$$=\sum_U\sum_{T|U}\left\{\frac{\exp(\alpha_0^*+\alpha_1^*\cdot U)}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^T\cdot\left\{\frac{1}{1+\exp(\alpha_0^*+\alpha_1^*\cdot U)}\right\}^{(1-T)}\cdot(\mu_U)^U\cdot(1-\mu_U)^{(1-U)}\cdot\left[\frac{\left\{\frac{\theta}{(1+\theta)^2}\right\}^2\cdot T}{\xi}+\frac{\left\{\frac{\theta}{(1+\theta)^2}\right\}^2\cdot T}{1-\xi}\right]$$

(3.4.24)

3.5 Efficiency loss by using IVA

The efficiency loss is estimated by the ratio of the variances of estimators. Let β_1 represent the treatment effect from ordinary linear regression without IVA, and β_{IV} represent the treatment effect from two-stage linear models of IVA. In section 2.1, we showed the ratio of the two variances of corresponding treatment effect estimators.

$$\frac{\mathrm{var}\left(\hat{\beta}_1\right)}{\mathrm{var}\left(\hat{\beta}_{IV}\right)} = \frac{\left[\sum_{i=1}^{N}\left\{\left(Z_i-\bar{Z}\right)\cdot\left(T_i-\bar{T}\right)\right\}\right]^2}{\sum_{i=1}^{N}\left(Z_i-\bar{Z}\right)^2\cdot\sum_{i=1}^{N}\left(T_i-\bar{T}\right)^2} \le 1 \qquad (3.5.1)$$

The ratio of the variances is determined by the correlation of instrumental variable and treatment status. When the correlation between instrumental variable and treatment status is 1, there is no efficiency loss. The example can be found in well monitored clinical trials without non-compliers. Patients are all treated with the assigned drugs following the randomization codes. In observational studies, patients' compliance is usually much lower than 100%. Therefore, the correlation between T and Z is lower than 1. The confidence interval of the treatment effect estimator from the 2SLS is wider than the one from the regular linear regression model without IVA. When the compliance is very poor, there will be no power to detect any treatment effect using IVA 2SLS.

In IVA with two-stage likelihood-based model, the formula for the efficiency evaluation is much more complicated than the one in (3.5.1). In general, we use notation in (3.5.2) to stand for the expected Fisher's information matrix of β_{0IV} and β_{1IV} in the two-stage likelihood-based IVA. The expected Fisher's information matrix in (3.4.21) is one of the examples.

$$I\left(\beta_{0IV},\beta_{1IV}\right)=\begin{bmatrix} I_{11}\left(\beta_{0IV},\beta_{1IV}\right) & I_{12}\left(\beta_{0IV},\beta_{1IV}\right) \\ I_{21}\left(\beta_{0IV},\beta_{1IV}\right) & I_{22}\left(\beta_{0IV},\beta_{1IV}\right) \end{bmatrix} \tag{3.5.2}$$

In the same way, we use notation in (3.5.3) to stand for the expected Fisher's information matrix of β_0 and β_1 without IVA. The expected Fisher's information matrix in (3.4.24) is one of the cases expanded in detail.

$$I\left(\beta_0,\beta_1\right)=\begin{bmatrix} I_{11}\left(\beta_0,\beta_1\right) & I_{12}\left(\beta_0,\beta_1\right) \\ I_{21}\left(\beta_0,\beta_1\right) & I_{22}\left(\beta_0,\beta_1\right) \end{bmatrix} \tag{3.5.3}$$

The determinant of the ratio of the two matrixes (3.5.2) and (3.5.3) measures the cost of efficiency by using the two-stage likelihood-based IVA.

$$\det\left[I^{-1}\left(\beta_0,\beta_1\right)\cdot I\left(\beta_{0IV},\beta_{1IV}\right)\right] \tag{3.5.4}$$

Particularly, the ratio of the variances of treatment effect estimators from non-IVA and IVA is given by

$$\frac{\left[I^{-1}\left(\beta_0,\beta_1\right)\right]_{22}}{\left[I^{-1}\left(\beta_{0IV},\beta_{1IV}\right)\right]_{22}} \tag{3.5.5}$$

To sum up, for IVA with nonlinear regression models, we develop a two-stage likelihood-based model. As in the 2SLS, the first stage is used to adjust for non-compliance. In the second stage, maximum likelihood estimator of the treatment effect is imputed.

$$E\left(T_i\right)=h_1\left(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U_i\right) \tag{3.5.6}$$

$$E\left(Y_i\right)=h_2\left(\beta_0+\beta_1\cdot T_i+\beta_2\cdot U_i\right) \tag{3.5.7}$$

We can express the two-stage likelihood-based IVA in a more general form in terms of the likelihood functions in (3.4.1) and (3.4.2). In the application of the two-stage likelihood-based model, the distribution of the unknown confounder U, the

association between T and U (α_2), and the association between Y and U (β_2) must be assessed in advance.

3.6 Simulation of two-stage likelihood-based IVA model

It is well known that the prostate specific antigen (PSA) screening test affects the usage of PADT. Patients with high PSA values are more likely to receive PADT. In the meantime, high values on PSA test cause higher mortality comparing to normal PSA values. In the study of PADT among men with localized prostate cancer, information on PSA testing is missing. There is no doubt that we miss an important confounder when we try to evaluate the treatment effect of PADT on the mortality. In our simulation, we assume that the distribution of PSA is binary, high or normal, with a mean of 0.2. Y is the outcome variable, e.g., 10-year mortality of the patient being dead or alive. T is the treatment of PADT or CM. U stands for the unobserved confounder of PSA. Finally, Z is the instrumental variable recording of where patients lived in, high PADT usage areas or low usage areas. The true parameter values are given in Table 3.6.1.

Whether a patient lived in a high PADT usage area or a low PADT usage is completely random before the disease is developed. It is reasonable to assume that a patient has an equal chance of living in either kind of health service area. The probability of being treated with PADT in low PADT usage areas and with normal PSA is assumed to be 0.1. The probability of being treated with PADT in low PADT usage areas but with high PSA is assumed to be 0.7. On the other hand, in high PADT usage areas, patients with normal PSA have a probability of 0.3 to be treated with PADT, and patients with high PSA have a probability of 0.9 to be treated with PADT. In addition, patients being treated with PADT with normal PSA levels are assumed to

be suffering the lowest mortality, which is 0.2. Patients being treated with CM with normal PSA had slightly higher mortality, which is 0.3. For patients with high PSA, even though they received PADT, the 10-year mortality can be as high as 0.7. If they received CM, the mortality is even higher at 0.8.

We use the two-stage likelihood-based IVA with two stages of logistic regression models to estimate the true PADT treatment effect. With all the assumptions listed on the left side of Table 3.6.1, It is not difficult to calculate the true values of $\alpha_0, \alpha_1, \alpha_2$, and $\beta_0, \beta_1, \beta_2$. The true α_1 value is log(3.86)=1.35. This number provides the information on compliance of the treatment given the instrumental variable. The true α_2 value is log(21)=3.04. This number provides information on the association between the treatment status and PSA result. The true β_1 which reflects the treatment effect is found to be log(0.58)=-0.54. The odds ratio of $\exp(\beta_1)$ =0.58 on the 10-year mortality of PADT versus CM implies that the PADT lowers mortality. The true β_2 which measures the association between mortality and PSA level is found to be log(9.3)=2.23.

Table 3.6.1 Parameters used in simulation of two-stage logistic regression model

$$P(U=1)=0.2$$
$$P(Z=1)=0.5$$

$$\left.\begin{array}{l} P(T=1\mid Z=0,U=0)=0.1 \\ P(T=1\mid Z=0,U=1)=0.7 \\ P(T=1\mid Z=1,U=0)=0.3 \\ P(T=1\mid Z=1,U=1)=0.9 \end{array}\right\} \Rightarrow \begin{array}{l} \dfrac{odds(T=1\mid Z=1,U)}{odds(T=1\mid Z=0,U)}=3.86 \\ \dfrac{odds(T=1\mid U=1,Z)}{odds(T=1\mid U=0,Z)}=21 \end{array}$$

$$\left.\begin{array}{l} P(Y=1\mid T=0,U=0)=0.3 \\ P(Y=1\mid T=0,U=1)=0.8 \\ P(Y=1\mid T=1,U=0)=0.2 \\ P(Y=1\mid T=1,U=1)=0.7 \end{array}\right\} \Rightarrow \begin{array}{l} \dfrac{odds(Y=1\mid T=1,U)}{odds(Y=1\mid T=0,U)}=0.58 \\ \dfrac{odds(Y=1\mid U=1,T)}{odds(Y=1\mid U=0,T)}=9.3 \end{array}$$

100 sets of data are generated with 30,000 subjects in each data set. Besides the two-stage likelihood-based model, we also fit data to several other models for the purpose of comparison. Means and their empirical standard deviations of the 100 sets of estimators are calculated, and results are presented in Table 3.6.2. The true values of the coefficients are listed in column **a**. For clarification, in this section, we use a subscript of "GMM" indicating the estimators are from the GMM method, subscript of "Wald" indicating the estimators are from Wald method, and a subscript of "IV" indicating the estimators are from the two-stage likelihood-based IVA method.

In column **b**, the estimated coefficients are from the regular logistic regressions with the unobserved confounder omitted in the models. $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are estimated from model (3.6.1). $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from model (3.6.2). Because the unobserved confounder is not adjusted for, the estimated treatment effect goes to the opposite direction of the true value.

$$\log\frac{E(T_i \mid Z_i)}{1-E(T_i \mid Z_i)} = \alpha_0 + \alpha_1 \cdot Z_i \tag{3.6.1}$$

$$\log\frac{E(Y_i \mid T_i)}{1-E(Y_i \mid T_i)} = \beta_0 + \beta_1 \cdot T_i \tag{3.6.2}$$

In column **c**, the regular logistic regression is also used, but the models include the unobserved confounder with values from simulated data. $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are estimated from model (3.6.3). $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from model (3.6.4). Since the unobserved confounder is adjusted in the models, the estimated coefficients are very close to the true values.

$$\log\frac{E(T_i \mid Z_i, U_i)}{1-E(T_i \mid Z_i, U_i)} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_i \tag{3.6.3}$$

$$\log\frac{E\left(Y_i \mid T_i, U_i\right)}{1-E\left(Y_i \mid T_i, U_i\right)} = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i \tag{3.6.4}$$

In column **d**, the MLEs are obtained from the two-stage logistic regression model as presented in section 3.4. Instead of generating simulated values for the unobserved confounder, we numerically integrate out the unobserved confounder over its probability measurement.

Column **e** gives the estimators from GMM IVA (Johnston, et al, 2008). The estimator-defining equation (3.6.5) is also called M-estimation.

$$\sum_{i=1}^{N}\varphi = \sum_{i=1}^{N}\left(Z_i \cdot \varepsilon_i\right) = \sum_{i=1}^{N}\left[Z_i \cdot \left\{Y_i - \frac{\exp\left(\beta_0 + \beta_1 \cdot T_i\right)}{1+\exp\left(\beta_0 + \beta_1 \cdot T_i\right)}\right\}\right] = 0 \tag{3.6.5}$$

$\hat{\beta}_{1GMM}$ is obtained from the solution of (3.6.5) with the estimator of β_0 from model (3.6.2). The variance of $\hat{\beta}_{1GMM}$ is imputed from the sandwich matrix. The empirical estimator of the sandwich matrix is

$$V_N\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) = A_N^{-1}\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) \cdot B_N\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) \cdot A_N^{-1}\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) \tag{3.6.6}$$

where

$$A_N\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) = -\frac{1}{N} \cdot \sum_{i=1}^{N}\frac{d\varphi}{d\hat{\beta}_{1GMM}} = \frac{1}{N} \cdot \sum_{i=1}^{N}\left[Z_i \cdot T_i \cdot \frac{\exp\left(\hat{\beta}_{0GMM} + \hat{\beta}_{1GMM} \cdot T_i\right)}{\left\{1+\exp\left(\hat{\beta}_{0GMM} + \hat{\beta}_{1GMM} \cdot T_i\right)\right\}^2}\right],$$

and

$$B_N\left(\hat{\beta}_{1GMM} \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right) = \frac{1}{N} \cdot \sum_{i=1}^{N}\left[Z_i \cdot \left\{Y_i - \frac{\exp\left(\hat{\beta}_{0GMM} + \hat{\beta}_{1GMM} \cdot T_i\right)}{1+\exp\left(\hat{\beta}_{0GMM} + \hat{\beta}_{1GMM} \cdot T_i\right)}\right\}\right]^2.$$

The variance of $\hat{\beta}_1$ is estimated by $\dfrac{V_N\left(\hat{\beta}_1 \mid \underset{\sim}{Z}, \underset{\sim}{Y}, \underset{\sim}{T}\right)}{N}$.

In column **f**, Wald type estimators are derived by using equations (3.6.7) and (3.6.8).

$$T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i \tag{3.6.7}$$

$$\log \frac{E(Y_i \mid Z_i)}{1 - E(Y_i \mid Z_i)} = \beta_0 + \beta_1 \cdot \hat{T}_i = \beta_0 + \beta_1 \cdot (\alpha_0 + \alpha_1 \cdot Z_i) \tag{3.6.8}$$

Essentially, in Wald type methods, $\log \frac{E(Y_i \mid T_i)}{1 - E(Y_i \mid T_i)}$ is treated as a linear function of Z_i. Same as in 2SLS, we obtained $\hat{\beta}_{1Wald} = \frac{\widehat{\beta_1 \alpha_1}}{\hat{\alpha}_1}$, and $SE\left(\hat{\beta}_{1Wald}\right) = \frac{SE\left(\widehat{\beta_1 \alpha_1}\right)}{\hat{\alpha}_1}$.

Table 3.6.2 Simulation results from the two-stage logistic regression model and its comparative models with a binary distributed confounder

a	b	c	d	e	f
True value	**Estimator from Logistic Regression[1]**	**Estimator from Logistic Regression[2]**	**Estimator from two-stage Logistic Regression**	**Estimator from GMM**	**Estimator from Wald method**
$\alpha_0 = -2.20$	$\hat{\alpha}_0 = -1.27$ SE=0.02	$\hat{\alpha}_0 = -2.20$ SE=0.03	$\hat{\alpha}_{0IV} = -2.15$ SE=0.029		
$\alpha_1 = 1.35$	$\hat{\alpha}_1 = 0.94$ SE=0.02	$\hat{\alpha}_1 = 1.35$ SE=0.03	$\hat{\alpha}_{1IV} = 1.32$ SE=0.036		
$\alpha_2 = 3.04$		$\hat{\alpha}_2 = 3.04$ SE=0.03			
$\beta_0 = -0.85$	$\hat{\beta}_0 = -0.71$ SE=0.01	$\hat{\beta}_0 = -0.85$ SE=0.01	$\hat{\beta}_{0IV} = -0.84$ SE=0.04		
$\beta_1 = -0.54$	$\hat{\beta}_1 = 0.51$ SE=0.03	$\hat{\beta}_1 = -0.54$ SE=0.03	$\hat{\beta}_{1IV} = -0.56$ SE=0.14	$\hat{\beta}_{1GMM} = 0.29$ SE=0.03	$\hat{\beta}_{1wald} = -0.45$ SE=0.11
$\beta_2 = 2.23$		$\hat{\beta}_2 = 2.23$ SE=0.04			

1. Models omit the unobserved confounder.
2. Models include the unobserved confounder with simulated values as fixed effects.

In Table 3.6.2, it is obvious that coefficient estimators from both regular logistic regression models omitting the unobserved confounder (column **b**), and

GMM IVA (column **e**) are biased. The coefficient estimators from the two-stage logistic regression model (column **d**) are close to the true values in column **a**. Estimator of β_1 from Wald method (column **f**) slightly deviates from the true value. This result confirms the non-collapsibility in the logistic regression which is discussed in section 2.6. The coefficient estimators from the regular logistic regression model with simulated values of the unobserved confounder (column **c**) are closest to the true parameters values. They also have small standard errors. Standard errors of $\hat{\alpha}_{0IV}$ and $\hat{\alpha}_{1IV}$ differ little from the standard errors of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ in column **c**, but the standard error of $\hat{\beta}_{1IV}$ is much larger than the standard error of $\hat{\beta}_1$ in column **c**. The standard error of $\hat{\beta}_{1Wald}$ is slightly smaller than that of $\hat{\beta}_{1IV}$. 95% confidence intervals of the estimated coefficients from the two-stage logistic regression model and Wald method both cover the true coefficients values. 95% confidence interval of $\hat{\beta}_{1IV}$ is given in (3.6.9):

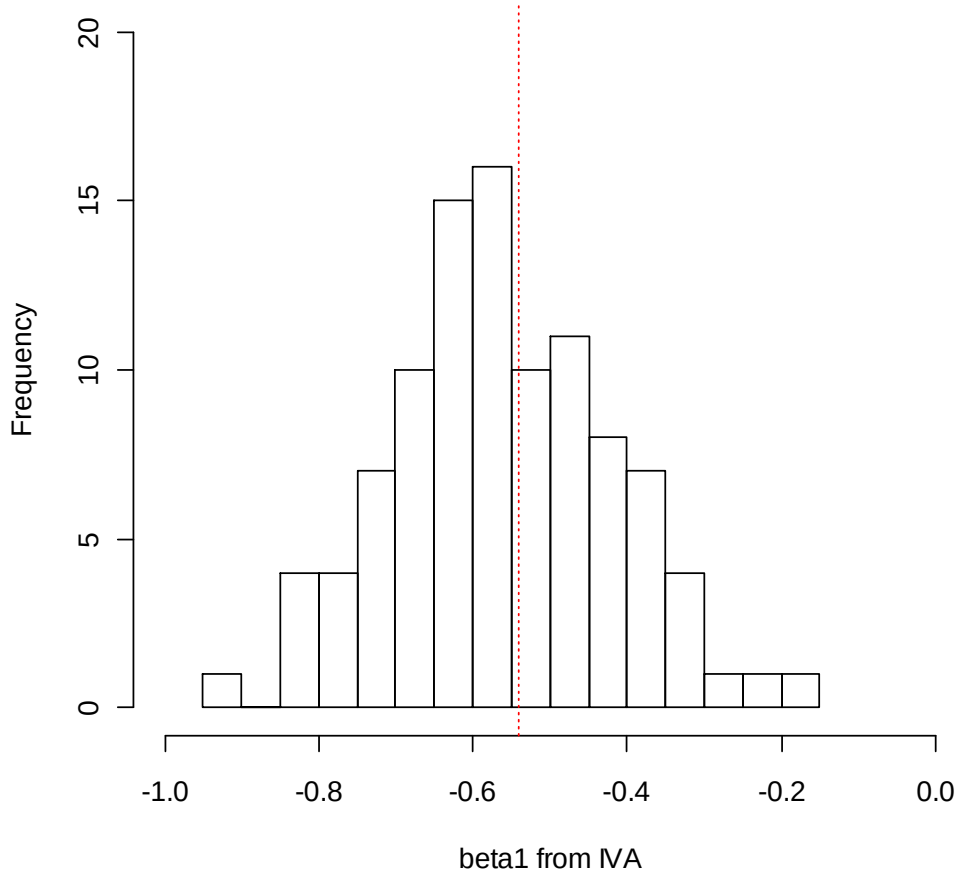
$$95\%CI = \left(-0.56 - 1.96 \cdot 0.14; -0.56 + 1.96 \cdot 0.14\right) = \left(-0.84; -0.28\right) \quad (3.6.9)$$

95% confidence interval of $\hat{\beta}_{1Wald}$ is given in (3.6.10):

$$95\%CI = \left(-0.45 - 1.96 \cdot 0.11; -0.45 + 1.96 \cdot 0.11\right) = \left(-0.66; -0.23\right) \quad (3.6.10)$$

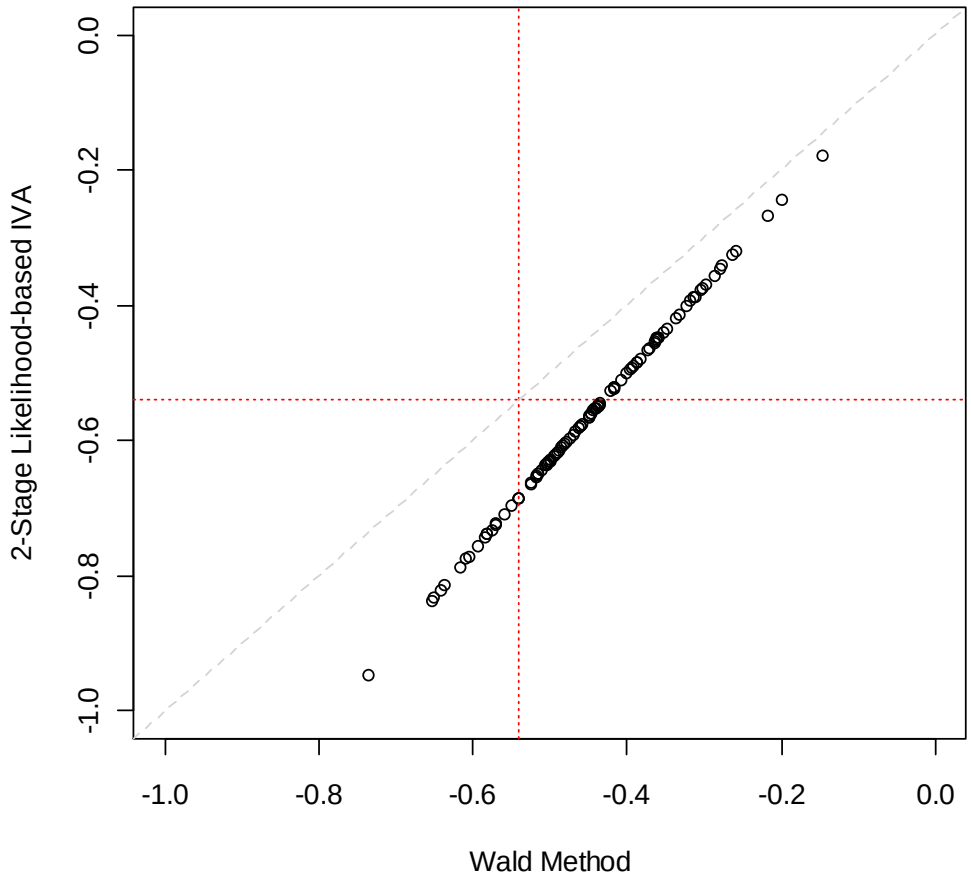
A histogram of the 100 $\hat{\beta}_{1IV}$ is illustrated in Figure 3.6.1. A plot of $\hat{\beta}_{1IV}$ versus $\hat{\beta}_{1Wald}$ (Figure 3.6.2) clearly displays the linear association between the two types of treatment effect estimators.

Figure 3.6.1 Histogram of $\hat{\beta}_{1IV}$ from two-stage likelihood-based IVA – binomial distribution



Note: Dotted line indicates the true value.

Figure 3.6.2 Treatment effect estimators $\hat{\beta}_{1IV}$ vs $\hat{\beta}_{1Wald}$ - binomial distribution



Note: Dotted lines indicate the true value.

The efficiency of the two-stage likelihood-based IVA is further studied. The variances of $\hat{\beta}_{1IV}$ from column **d** are compared to the variances of $\hat{\beta}_1$ from column **c** for three scenarios: PADT superior to CM, PADT equivalent to CM, and PADT inferior to CM. The true parameters values are given in Tables 3.6.3 and 3.6.4. I_{22} from expected Fisher's information matrix, (3.4.21) and (3.4.24), is used to calculate the variances of $\hat{\beta}_{1IV}$ and $\hat{\beta}_1$, respectively.

Table 3.6.3 Parameters in the first stage used in simulation to estimate efficiency loss in IVA

$$P(U=1)=0.2$$
$$P(Z=1)=0.5$$

$$P(T=1|Z=0,U=0)=0.1$$
$$P(T=1|Z=0,U=1)=0.7$$
$$P(T=1|Z=1,U=0)=0.3$$
$$P(T=1|Z=1,U=1)=0.9$$

$$\Downarrow$$

$$\frac{odds(T=1|Z=1,U)}{odds(T=1|Z=0,U)}=3.86$$

$$\frac{odds(T=1|U=1,Z)}{odds(T=1|U=0,Z)}=21$$

$$\Downarrow$$

$$\alpha_0=\log(0.1/0.9)=-2.20$$
$$\alpha_1=\log(3.86)=1.35$$
$$\alpha_2=\log(21)=3.04$$

$$\Rightarrow$$

$$P(T=1|U=0)=0.2$$
$$P(T=1|U=1)=0.8$$

$$\Downarrow$$

$$\frac{odds(T=1|U=1)}{odds(T=1|U=0)}=16$$

$$\Downarrow$$

$$\alpha_0^*=\log(0.2/0.8)=-1.39$$
$$\alpha_1^*=\log(16)=2.77$$

$P(T=1|U=0)$ and $P(T=1|U=1)$ in Table 3.6.3 were obtained by equations (3.6.11) and (3.6.12).

$$P(T=1|U=0)=P(T=1|U=0,Z=0)\cdot P(Z=0)+P(T=1|U=0,Z=1)\cdot P(Z=1)$$
$$=0.1\cdot 0.5+0.3\cdot 0.5=0.2 \qquad (3.6.11)$$

$$P(T=1|U=1)=P(T=1|U=1,Z=0)\cdot P(Z=0)+P(T=1|U=1,Z=1)\cdot P(Z=1)$$
$$=0.7\cdot 0.5+0.9\cdot 0.5=0.8 \qquad (3.6.12)$$

Table 3.6.4 Parameters in the second stage used in simulation to estimate efficiency loss in IVA

(1) PADT superior to conservative management:

$$\begin{aligned} &P(Y=1|T=0,U=0)=0.3\\ &P(Y=1|T=0,U=1)=0.8\\ &P(Y=1|T=1,U=0)=0.2\\ &P(Y=1|T=1,U=1)=0.7 \end{aligned} \Longrightarrow \begin{aligned} &\frac{odds(Y=1|T=1,U)}{odds(Y=1|T=0,U)}=0.58\\ &\frac{odds(Y=1|U=1,T)}{odds(Y=1|U=0,T)}=9.3 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \beta_0 &= \log(0.3/0.7) = -0.85\\ \beta_1 &= \log(0.58) = -0.54\\ \beta_2 &= \log(0.93) = 2.23 \end{aligned}$$

(2) PADT equivalent to conservative management:

$$\begin{aligned} &P(Y=1|T=0,U=0)=0.3\\ &P(Y=1|T=0,U=1)=0.8\\ &P(Y=1|T=1,U=0)=0.3\\ &P(Y=1|T=1,U=1)=0.8 \end{aligned} \Longrightarrow \begin{aligned} &\frac{odds(Y=1|T=1,U)}{odds(Y=1|T=0,U)}=1\\ &\frac{odds(Y=1|U=1,T)}{odds(Y=1|U=0,T)}=9.3 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \beta_0 &= \log(0.3/0.7) = -0.85\\ \beta_1 &= \log(1) = 0\\ \beta_2 &= \log(0.93) = 2.23 \end{aligned}$$

(3) PADT inferior to conservative management:

$$\begin{aligned} &P(Y=1|T=0,U=0)=0.2\\ &P(Y=1|T=0,U=1)=0.7\\ &P(Y=1|T=1,U=0)=0.3\\ &P(Y=1|T=1,U=1)=0.8 \end{aligned} \Longrightarrow \begin{aligned} &\frac{odds(Y=1|T=1,U)}{odds(Y=1|T=0,U)}=1.71\\ &\frac{odds(Y=1|U=1,T)}{odds(Y=1|U=0,T)}=9.3 \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \beta_0 &= \log(0.2/0.8) = -1.39\\ \beta_1 &= \log(1.71) = 0.54\\ \beta_2 &= \log(0.93) = 2.23 \end{aligned}$$

The results in Table 3.6.5 show a loss of approximately 90% of the efficiency with 20% of the marginal population, where the 20% are from equations (3.6.13) and (3.6.14). Standard errors do not vary when the treatment effect changes.

$$P(T=1|Z=1)=P(T=1|Z=1,U=1)\cdot P(U=1)+P(T=1|Z=1,U=0)\cdot P(U=0)$$
$$=0.9\cdot 0.2+0.3\cdot 0.8=0.42 \tag{3.6.13}$$

$$P(T=1|Z=0)=P(T=1|Z=0,U=1)\cdot P(U=1)+P(T=1|Z=0,U=0)\cdot P(U=0)$$
$$=0.7\cdot 0.2+0.1\cdot 0.8=0.22 \tag{3.6.14}$$

Table 3.6.5 Simulation results of efficiency loss in IVA

True value of β_1	-0.54	0	0.54
Expected SE of $\hat{\beta}_1$	0.024	0.024	0.024
Expected SE of $\hat{\beta}_{1IV}$	0.075	0.076	0.077
Ratio of SE $\left[\frac{SE(\hat{\beta}_1)}{SE(\hat{\beta}_{1IV})}\right]$	0.32	0.31	0.31

Next, we replace the binary distribution with a normal distribution for the unobserved confounder in the two-stage likelihood-based IVA model. We assume that the mean of U is 0.34 and the standard deviation is 0.63.

$$L(\beta_0,\beta_1|\beta_2,\alpha_0,\alpha_1,\alpha_2;\underset{\sim}{Y},\underset{\sim}{Z})=\prod_{i=1}^{N}P(Y_i|Z_i)=\prod_{i=1}^{N}\int_U\sum_{T|Z_i,U}f(Y_i,T,Z_i,U)\cdot dU$$
$$=\prod_{i=1}^{N}\int_U\sum_{T|Z_i,U}P(Y_i|T,U)\cdot P(T|U,Z_i)\cdot f(U)$$
$$=\prod_{i=1}^{N}\int_U\sum_{T|Z_i,U}\left\{\frac{\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{Y_i}\cdot\left\{\frac{1}{1+\exp(\beta_0+\beta_1\cdot T+\beta_2\cdot U)}\right\}^{(1-Y_i)}\cdot$$
$$\left\{\frac{\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{T}\cdot\left\{\frac{1}{1+\exp(\alpha_0+\alpha_1\cdot Z_i+\alpha_2\cdot U)}\right\}^{(1-T)}\cdot\frac{1}{\sqrt{2\pi\sigma_U^2}}\exp\left\{\frac{-(U-\mu_U)^2}{2\sigma_U^2}\right\}\cdot dU \tag{3.6.15}$$

100 sample datasets of size 30000 are generated. Gaussian quadrature is used to integrate out the normal distribution.

Table 3.6.6 Simulation results from the two-stage logistic regression model and its comparative models with a normally distributed confounder

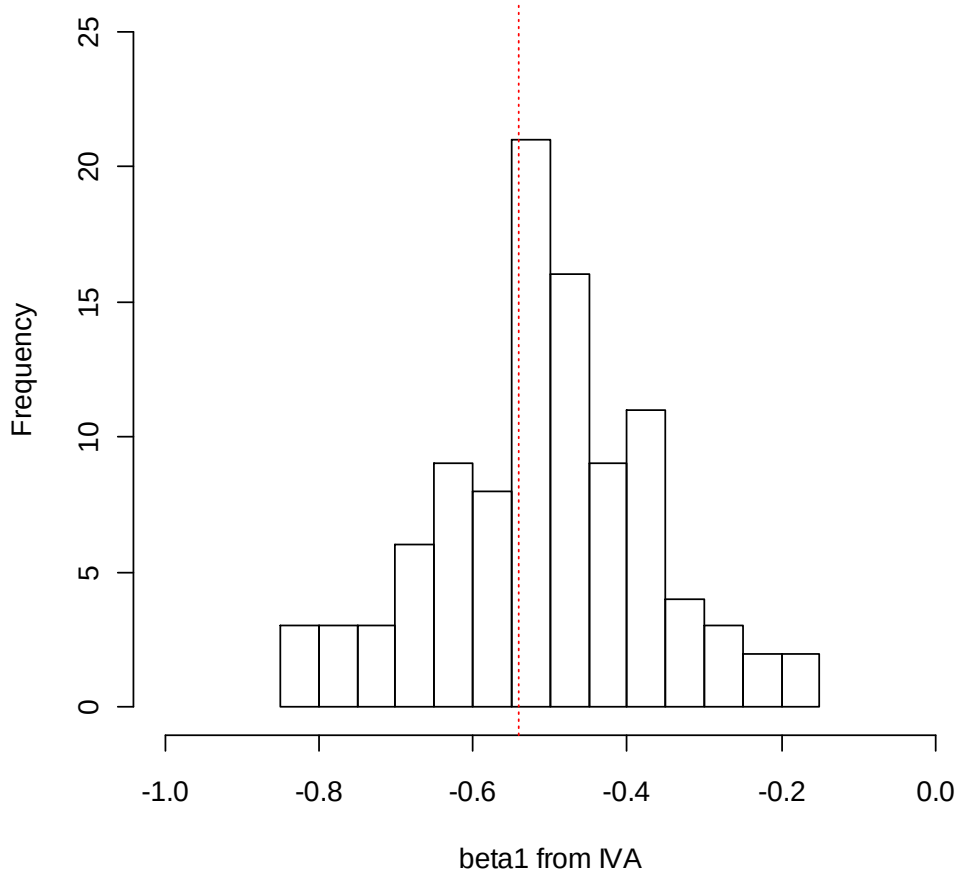
a	b	c	d	e
True value	**Estimator from Logistic Regression[1]**	**Estimator from Logistic Regression[2]**	**Estimator from two-stage Logistic Regression**	**Estimator from Wald method**
$\alpha_0 = -2.20$		$\hat{\alpha}_0 = -2.16$ SE=0.03	$\hat{\alpha}_{0IV} = -2.19$ SE=0.03	
$\alpha_1 = 1.35$		$\hat{\alpha}_1 = 1.30$ SE=0.03	$\hat{\alpha}_{1IV} = 1.33$ SE=0.04	
$\alpha_2 = 3.04$		$\hat{\alpha}_2 = 3.00$ SE=0.04		
$\beta_0 = -0.85$	$\hat{\beta}_0 = -0.61$ SE=0.02	$\hat{\beta}_0 = -0.85$ SE=0.02	$\hat{\beta}_{0IV} = -0.86$ SE=0.06	
$\beta_1 = -0.54$	$\hat{\beta}_1 = 0.85$ SE=0.03	$\hat{\beta}_1 = -0.54$ SE=0.03	$\hat{\beta}_{1IV} = -0.51$ SE=0.14	$\hat{\beta}_{1wald} = -0.43$ SE=0.12
$\beta_2 = 2.23$		$\hat{\beta}_2 = 2.23$ SE=0.03		

1. Models omit the unobserved confounder.
2. Models include the unobserved confounder with simulated values as a covariate.

The results (Table 3.6.6) are consistent with the previous findings. The treatment effect estimator from a regular logistic regression model including treatment only but omitting the unobserved confounder indicates the treatment caused higher mortality than placebo while the true treatment effect is to lower the mortality. In contrast, the two-stage likelihood-based IVA and Wald type method provide closer estimators to the true values although the standard errors are quite large. 95% confidence intervals of the $\hat{\beta}_{1IV}$ and $\hat{\beta}_{1Wald}$ contain the true value of β_1.

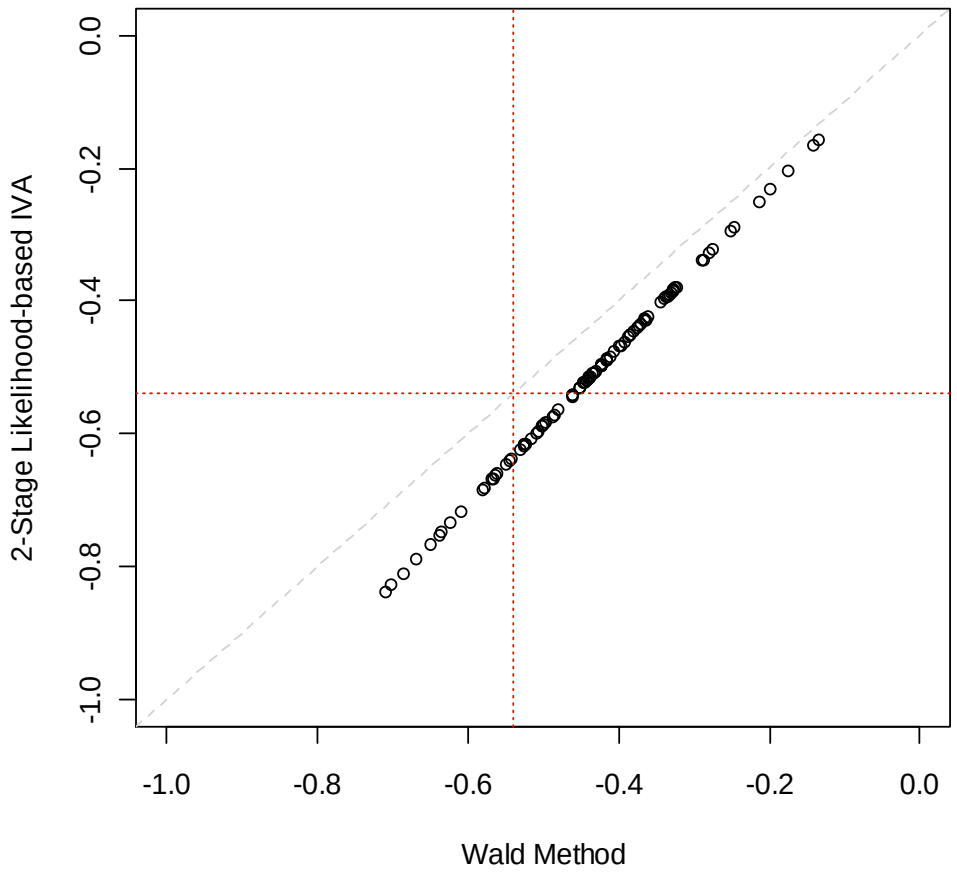
A histogram of $\hat{\beta}_{1IV}$ is graphed in Figure 3.6.3. $\hat{\beta}_{1IV}$ and $\hat{\beta}_{1Wald}$ are highly linearly correlated as shown in Figure 3.6.4.

Figure 3.6.3 Histogram of $\hat{\beta}_{1IV}$ from the two-stage likelihood-based IVA – binomial distribution with a normally distributed confounder



Note: Dotted line indicates the true value.

Figure 3.6.4 Treatment effect estimators $\hat{\beta}_{1IV}$ vs $\hat{\beta}_{1Wald}$ - binomial distribution with a normally distributed confounder



Note: Dotted lines indicate the true value.

Chapter 4

IVA in Survival Regression Model

IVA in survival analysis was mentioned by Dunn, Maracy, and Tomenson (2005). IVA has been used previously to analyze the prostate cancer survival data (Lu-Yao, et al., 2008), Other similar examples of IVA with censored survival data include a comparison of prostate specific survival of hormone therapy to radiotherapy (Zeliadt, Potosky, Penson, et al., 2006), an evaluation on effects of invasive Cardiac Management on acute myocardial infarction survival (Stukel, Fisher, Wennberg, et al., 2007), and an example comparing survival of lung cancer patients treated with chemotherapy as compared to no chemotherapy (Earle, Tsai, Gelber, et al., 2001). All of these papers made use of the SEER-Medicare linked database, and compared survival in high-use to low-use health service areas. Nevertheless, there has been very little methodological research on the use of IVA using survival analysis methodology. It could be very complicated when the estimation of hazard ratio in the Cox Proportional Hazard model involves a partial likelihood function.

4.1 Two-stage likelihood-based model in survival analysis

We extend the IVA two-stage likelihood-based model to survival data analysis. The general form is almost the same as in (3.4.1) and (3.4.2). In stage two, a survival function with censoring data is used.

Stage 1:

$$L\left(\alpha_0,\alpha_1 \mid \alpha_2;\underset{\sim}{T},\underset{\sim}{Z}\right)=\prod_{i=1}^{N} f\left(T_i \mid Z_i\right)=\prod_{i=1}^{N}\int_U f\left(T_i \mid Z_i,U\right)\cdot dF_U\left(U\right) \tag{4.1.1}$$

Stage 2:

$$L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) = \prod_{i=1}^{N} f(Y_i,C_i \mid Z_i)$$
$$= \prod_{i=1}^{N} \int_U \int_{T|U} f(Y_i,C_i \mid U,T)\cdot dF_{T|U}(T,Z_i)\cdot dF_U(U) \tag{4.1.2}$$

Assumptions: α_2, β_2 and $F_U(U)$ are known. In addition, the censoring status is random.

In special cases, stage two can be formulated from a Weibull distribution function, an exponential function, or a proportional hazard function. The likelihood function at stage two may be written in the standard form for right-censored survival data, and summing over U and $T \mid U$, as follows:

$$L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C})$$
$$= \prod_{i=1}^{N} \int_U \int_{T|U} f(T \mid Z_i,U)\cdot f(U)\cdot h(Y_i \mid T,U)^{C_i}\cdot S(Y_i \mid T,U)\cdot dT\cdot dU \tag{4.1.3}$$

where $h(Y_i \mid T,U)$ is the hazard function and $S(Y_i \mid T,U)$ is the survival function. C_i indicates the censoring status. When the survival time is a continuous variable, and T and U are binary variables, the likelihood function is expressed as in (4.1.4).

$$L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) = \prod_{i=1}^{N} f(Y_i,C_i \mid Z_i)$$
$$= \prod_{i=1}^{N} \sum_U \sum_{T|U,Z_i} \left\{ P(T \mid U,Z_i)\cdot P(U)\cdot h(Y_i \mid T,U)^{C_i}\cdot S(Y_i \mid T,U) \right\} \tag{4.1.4}$$

If the survival data follow an exponential distribution, equation (4.1.4) becomes:

$$L(\beta_0,\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) = \prod_{i=1}^{N} f(Y_i,C_i \mid Z_i)$$
$$= \prod_{i=1}^{N} \sum_U \sum_{T|U,Z_i} \left\{ f(Y_i,C_i \mid T,U)\cdot P(T \mid U,Z_i)\cdot P(U) \right\} \tag{4.1.5}$$
$$= \prod_{i=1}^{N} \sum_U \sum_{T|Z_i,U} \left\{ \lambda^{C_i}\cdot \exp(-\lambda\cdot Y_i)\cdot P(T \mid U,Z_i)\cdot P(U) \right\}$$

where hazard $\lambda = \exp(-\beta_0 - \beta_1 \cdot T - \beta_2 \cdot U)$, and it is a constant. If the survival data follow a Weibull distribution, equation (4.1.4) becomes:

$$L(\beta_0, \beta_1 \mid \alpha_0, \alpha_1, \alpha_2, \beta_2; \underset{\sim}{Y}, \underset{\sim}{Z}, \underset{\sim}{C}) = \prod_{i=1}^{N} \sum_{U} \sum_{T|Z_i,U} \left[\left\{ \lambda \cdot \gamma \cdot (\lambda \cdot Y_i)^{(\gamma-1)} \right\}^{C_i} \cdot \exp\left\{ -(\lambda \cdot Y_i)^{\gamma} \right\} \cdot P(T \mid U, Z_i) \cdot P(U) \right] \quad (4.1.6)$$

where $\lambda = \frac{1}{\gamma} \cdot \exp(-\beta_0 - \beta_1 \cdot T - \beta_2 \cdot U)$ is a shape parameter, and γ is a scale parameter.

4.2 Two-stage likelihood-based model in piecewise constant hazard function

The Cox proportional hazard regression model is very popular in survival analysis. We model the first stage as a logistic regression and the second stage as a Cox proportional hazard regression:

$$\log \frac{E(T_i \mid Z_i, U_i)}{1 - E(T_i \mid Z_i, U_i)} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_i \quad (4.2.1)$$

$$h_i(Y_i \mid T_i, U_i) = h_0 \cdot \exp(\beta_1 \cdot T_i + \beta_2 \cdot U_i) \quad (4.2.2)$$

$h(Y)$ is a hazard function at time Y, and h_0 is the baseline hazard function at time Y when covariates of T and U equal 0. Outcome Y stands for time to event or time to censored event. However, because the Cox proportional hazard regression model involves more complicated partial likelihood functions to estimate the parameters, we replace the Cox proportional hazard regression model with a piecewise constant hazard model. Piecewise constant hazard functions have been used in the Surveillance, Epidemiology, and End Results (SEER) data to examine prostate cancer mortality (Goodman, Li, and Tiwari, 2011). The hazard function is:

$$h_{ij}(Y_{ij} \mid T_i, U_i) = h_0(Y_j) \cdot \exp(\beta_1 \cdot T_i + \beta_2 \cdot U_i), \quad (4.2.3)$$

for $\tau_{j-1} < Y_{ij} \le \tau_j, j = 1, 2, ..., J$, and $0 < \tau_1 < \tau_2 < \cdots < \tau_j < \cdots < \tau_J$ and $\tau_0 = 0$

In (4.2.3), time to death Y_i is divided into J intervals. Y_{ij} denotes the time within jth interval for subject i. h_{ij} is constant within the time interval of $\tau_{j-1} < Y_{ij} \le \tau_j$. h_0 is the corresponding baseline hazard within the same interval.

$$h_0\left(Y_j\right) = \lambda_j = \exp\left(\beta_{0j}\right), \quad \tau_{j-1} < Y_j \le \tau_j, j = 1, 2, ..., J \tag{4.2.4}$$

The baseline survival function is given by

$$S_0\left(Y_i\right) = \exp\left\{-\lambda_1 \cdot \left(\tau_1 - \tau_0\right) - \lambda_2 \cdot \left(\tau_2 - \tau_1\right) - ... - \lambda_{J_i^*} \cdot \left(Y - \tau_{J_i^*-1}\right)\right\}, \quad \tau_{J_i^*-1} < Y_i \le \tau_{J_i^*} \tag{4.2.5}$$

where J_i^* is the index for which $\tau_{J_i^*-1} < Y_i \le \tau_{J_i^*}$. Instead of Y_{ij}, Y_i is used in equation (4.2.5) because the survival function is derived from the cumulative hazard function covering J_i^* intervals. For individual i with survival and censoring values Y_i and C_i, and receiving treatment T_i and having confounder U_i, the hazard function is:

$$h\left(Y_{ij}\right) = \begin{cases} \exp\left(\beta_{01} + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) & \tau_0 < Y_{ij} \le \tau_1 \\ \exp\left(\beta_{02} + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) & \tau_1 < Y_{ij} \le \tau_2 \\ \quad \vdots & \\ \exp\left(\beta_{0J_i^*} + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) & \tau_{J_i^*-1} < Y_{ij} \le \tau_{J_i^*} \end{cases} \tag{4.2.6}$$

The survival function is:

$$S(Y_i) = \begin{cases} \exp\left\{-\lambda_1 \cdot (Y_i - \tau_0) \cdot \exp\left(\beta_1 \cdot T_i + \beta_2 \cdot U_i\right)\right\}, & \tau_1 < Y_i \le \tau_0 \\ \exp\left[\left\{-\lambda_1 \cdot (\tau_1 - \tau_0) - \lambda_2 \cdot (Y_i - \tau_1)\right\} \cdot \exp\left(\beta_1 \cdot T_i + \beta_2 \cdot U_i\right)\right], & \tau_2 < Y_i \le \tau_1 \\ \quad \vdots & \\ \exp\left[\left\{-\lambda_1 \cdot (\tau_1 - \tau_0) - \lambda_2 \cdot (\tau_2 - \tau_1) - ... - \lambda_{J_i^*} \cdot \left(Y_i - \tau_{J_i^*-1}\right)\right\} \cdot \exp\left(\beta_1 \cdot T_i + \beta_2 \cdot U_i\right)\right], & \tau_{J_i^*-1} < Y_i \le \tau_{J_i^*} \end{cases} \tag{4.2.7}$$

When the confounder U is not observed, we have to use the two-stage likelihood-based IVA model to estimate the treatment effect β_1. First, we estimate α_0 and α_1 from the likelihood function at stage one with pre-specified value of α_2 and pre-specified distribution of U. This function is the same as (3.4.11) because we use the same logistic regression at stage one as in section 3.4. The likelihood function at stage two with a piecewise constant hazard model is:

$$
\begin{aligned}
&L(\beta_{01},\beta_{02},...,\beta_{0J},\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) = \prod_{i=1}^{N} f(Y_i, C_i \mid Z_i) \\
&= \prod_{i=1}^{N} \sum_{U} \sum_{T|U,Z_i} \left\{ P(T \mid U, Z_i) \cdot P(U) \cdot \left\{ \lambda_{J_i^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \right\}^{C_i} \cdot \right. \\
&\qquad \left. \exp\left[-\left\{ \sum_{j=1}^{J_i^*-1} \lambda_j \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right\} - \lambda_{J_i^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot \left(Y_i - \tau_{J_i^*-1}\right) \right] \right\}
\end{aligned}
\tag{4.2.8}
$$

where $\lambda_j = \exp(\beta_{0j})$ is the baseline hazard during the time interval of $(\tau_{j-1}, \tau_j]$.

The log likelihood function of (4.2.8) is:

$$
\begin{aligned}
&l(\beta_{01},\beta_{02},...,\beta_{0J},\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) \\
&= \log L(\beta_{01},\beta_{02},...,\beta_{0J},\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2,\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C}) \\
&= \sum_{i=1}^{N} \log\left\{ f(Y_i, C_i \mid Z_i) \right\} \\
&= \sum_{i=1}^{N} \log\left[\sum_{U} \sum_{T|U,Z_i} \left\{ \frac{\exp(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U)}{1 + \exp(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U)} \right\}^{T} \cdot \left\{ \frac{1}{1 + \exp(\alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U)} \right\}^{(1-T)} \cdot (\mu_U)^{U} \cdot (1-\mu_U)^{(1-U)} \cdot \right. \\
&\left. \left\{ \lambda_{J_i^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \right\}^{C_i} \cdot \exp\left\{ -\left[\sum_{j=1}^{J_i^*-1} \lambda_j \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right] - \lambda_{J_i^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot \left(Y_i - \tau_{J_i^*-1}\right) \right\} \right]
\end{aligned}
\tag{4.2.9}
$$

It is not easy to find the score functions explicitly for the log likelihood functions as the one in (4.2.9). Given pre-specified values α_2 and β_2, solutions for the MLEs of

the parameters and their variances at both stage one and stage two are handled numerically.

After we obtain the MLEs $\hat{\beta}_{01}, \hat{\beta}_{02}, ..., \hat{\beta}_{0J}$ and $\hat{\beta}_1$, we are able to estimate the survival functions for given T and U. The standard error of the survival probability at any time point Y can be approximated using the delta method (Valenta and Weissfeld, 2002).

$$S(Y) = \exp\left[-\left\{ \sum_{j=1}^{J^*-1} \lambda_j \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right\} - \lambda_{J^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (Y - \tau_{J^*-1}) \right]$$
$$= \exp\left[-\left\{ \sum_{j=1}^{J^*-1} \exp(\beta_{0j} + \beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right\} - \exp(\beta_{0J^*} + \beta_1 \cdot T + \beta_2 \cdot U) \cdot (Y - \tau_{J^*-1}) \right] \quad (4.2.10)$$

where J^* is the index for which $\tau_{J^*-1} < Y \le \tau_{J^*}$.

$$g(Y \mid \underset{\sim}{\beta}) = \log S(Y) = -\left\{ \sum_{j=1}^{J^*-1} \lambda_j \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right\} - \lambda_{J^*} \cdot \exp(\beta_1 \cdot T + \beta_2 \cdot U) \cdot (Y - \tau_{J^*-1})$$
$$= -\left\{ \sum_{j=1}^{J^*-1} \exp(\beta_{0j} + \beta_1 \cdot T + \beta_2 \cdot U) \cdot (\tau_j - \tau_{j-1}) \right\} - \exp(\beta_{0J^*} + \beta_1 \cdot T + \beta_2 \cdot U) \cdot (Y - \tau_{J^*-1}) \quad (4.2.11)$$

$$\operatorname{var}\left(\log \hat{S}(Y)\right) \approx g'\left(Y \mid \underset{\sim}{\hat{\beta}}\right)^T \cdot \operatorname{var}\left(\underset{\sim}{\hat{\beta}}\right) \cdot g\left(Y \mid \underset{\sim}{\hat{\beta}}\right) \quad (4.2.12)$$

In the approximate equation (4.2.12), $\underset{\sim}{\hat{\beta}}$ is a vector of parameter estimators, $\hat{\beta}_{01}, \hat{\beta}_{02}, ..., \hat{\beta}_{0J}, \hat{\beta}_1$. $\operatorname{var}\left(\underset{\sim}{\hat{\beta}}\right)$ is the variance-covariance matrix of $\underset{\sim}{\hat{\beta}}$:

$$\begin{bmatrix} \operatorname{var}(\hat{\beta}_{01}) & \cdots & \operatorname{cov}(\hat{\beta}_{01}, \hat{\beta}_{0J^*}) & \operatorname{cov}(\hat{\beta}_{01}, \hat{\beta}_1) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(\hat{\beta}_{01}, \hat{\beta}_{0J^*}) & \cdots & \operatorname{var}(\hat{\beta}_{0J^*}) & \operatorname{cov}(\hat{\beta}_{0J^*}, \hat{\beta}_1) \\ \operatorname{cov}(\hat{\beta}_{01}, \hat{\beta}_1) & \cdots & \operatorname{cov}(\hat{\beta}_{0J^*}, \hat{\beta}_1) & \operatorname{var}(\hat{\beta}_1) \end{bmatrix}$$

$g'\left(Y \mid \underset{\sim}{\beta}\right)^T$ is the transpose matrix of $g'\left(Y \mid \underset{\sim}{\beta}\right)$, and $g'\left(Y \mid \underset{\sim}{\beta}\right)$ is the first order derivative of $g\left(Y \mid \underset{\sim}{\beta}\right)$.

$$g'\left(Y \mid \underset{\sim}{\beta}\right) = \begin{bmatrix} -(\tau_1 - \tau_0)\cdot\exp(\beta_{01} + \beta_1 \cdot T + \beta_2 \cdot U) \\ \vdots \\ -\left(Y - \tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*} + \beta_1 \cdot T + \beta_2 \cdot U\right) \\ -T\cdot\left\{\sum_{j=1}^{J^*-1}\left(\tau_j - \tau_{j-1}\right)\cdot\exp\left(\beta_{0j} + \beta_1 \cdot T + \beta_2 \cdot U\right)\right\} - T\cdot\left(Y - \tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*} + \beta_1 \cdot T + \beta_2 \cdot U\right) \end{bmatrix}$$

The 95% confidence interval for the survival probability is:

$$\exp\left\{\log \hat{S}(Y) \pm 1.96 \cdot \sqrt{\widehat{\text{var}}\left(\log \hat{S}(Y)\right)}\right\} \tag{4.2.13}$$

4.3 Comparison in survivals between treatment groups

The survival probabilities at time Y between two treatment groups can be compared using rate ratio or rate difference. Variances of the estimated rate ratio or rate difference may also be obtained using delta method.

Let r denote the log of rate ratio, and r' denote the vector of derivatives. We have

$$\log\left(\textit{rate ratio}\right) = r\left(Y \mid \underset{\sim}{\beta}; U\right) = \log S\left(Y \mid T = 1, U\right) - \log S\left(Y \mid T = 0, U\right) \tag{4.3.1}$$

$$\begin{aligned} r\left(Y \mid \underset{\sim}{\beta}\right) = &-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j} + \beta_1 + \beta_2 \cdot U\right)\cdot\left(\tau_j - \tau_{j-1}\right)\right\} - \exp\left(\beta_{0J^*} + \beta_1 + \beta_2 \cdot U\right)\cdot\left(Y - \tau_{J^*-1}\right) \\ &+\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j} + \beta_2 \cdot U\right)\cdot\left(\tau_j - \tau_{j-1}\right)\right\} + \exp\left(\beta_{0J^*} + \beta_2 \cdot U\right)\cdot\left(Y - \tau_{J^*-1}\right) \end{aligned}$$

$$\text{var}\left[\hat{r}(Y \mid \hat{\underset{\sim}{\beta}})\right] \approx r'(Y \mid \hat{\underset{\sim}{\beta}})^T \,\text{cov}(\hat{\underset{\sim}{\beta}}) r'(Y \mid \hat{\underset{\sim}{\beta}}) \tag{4.3.2}$$

The derivatives $r(Y \mid \underset{\sim}{\beta})$ are given by

$$r'\left(Y\,|\,\underset{\sim}{\beta}\right)=\begin{bmatrix} -(\tau_1-\tau_0)\cdot\exp(\beta_{01})\cdot\{\exp(\beta_1+\beta_2\cdot U)-\exp(\beta_2\cdot U)\} \\ \vdots \\ -\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)\cdot\{\exp(\beta_1+\beta_2\cdot U)-\exp(\beta_2\cdot U)\} \\ -\left\{\sum_{j=1}^{J^*-1}\left(\tau_j-\tau_{j-1}\right)\cdot\exp\left(\beta_{0j}+\beta_1+\beta_2\cdot U\right)\right\}-\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}+\beta_1+\beta_2\cdot U\right) \end{bmatrix}$$

The survival rate ratio at time Y is estimated by $\exp\left(\hat{r}(Y\,|\,\hat{\underset{\sim}{\beta}})\right)$. 95% confidence interval of the estimated rate ratio is given by $\exp\left(\hat{r}(Y\,|\,\hat{\underset{\sim}{\beta}})\pm 1.96*\widehat{\text{var}}[\hat{r}(Y\,|\,\hat{\underset{\sim}{\beta}})]\right)$.

If one is interested in the survival rate difference between two treatment groups at time Y as presented by Lu-Yao et al. (Lu-Yao et al. 2008), the difference denoted as d is given in equation (4.3.3).

$$\textit{rate difference}=d(Y\,|\,\underset{\sim}{\beta};U)=S(Y\,|\,T=1,U)-S(Y\,|\,T=0,U) \tag{4.3.3}$$

$$d(Y\,|\,\underset{\sim}{\beta})=\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}+\beta_1+\beta_2\cdot U\right)\cdot\left(\tau_j-\tau_{j-1}\right)\right\}-\exp\left(\beta_{0J^*}+\beta_1+\beta_2\cdot U\right)\cdot\left(Y-\tau_{J^*-1}\right)\right]$$
$$-\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}+\beta_2\cdot U\right)\cdot\left(\tau_j-\tau_{j-1}\right)\right\}-\exp\left(\beta_{0J^*}+\beta_2\cdot U\right)\cdot\left(Y-\tau_{J^*-1}\right)\right]$$

Similar, the variance can be obtained using the delta method.

$$\text{var}\left[\hat{d}(Y\,|\,\hat{\underset{\sim}{\beta}})\right]\approx d'(Y\,|\,\hat{\underset{\sim}{\beta}})^T\,\text{cov}(\hat{\underset{\sim}{\beta}})d'(Y\,|\,\hat{\underset{\sim}{\beta}}) \tag{4.3.4}$$

where $d'\left(Y\,|\,\hat{\underset{\sim}{\beta}}\right)$ is the vector of derivatives of $d(Y\,|\,\hat{\underset{\sim}{\beta}})$.

Let

$$A=\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}+\beta_1+\beta_2\cdot U\right)\cdot\left(\tau_j-\tau_{j-1}\right)\right\}-\exp\left(\beta_{0J^*}+\beta_1+\beta_2\cdot U\right)\cdot\left(Y-\tau_{J^*-1}\right)\right]$$

and

$$B=\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}+\beta_2\cdot U\right)\cdot\left(\tau_j-\tau_{j-1}\right)\right\}-\exp\left(\beta_{0J^*}+\beta_2\cdot U\right)\cdot\left(Y-\tau_{J^*-1}\right)\right]$$

The derivatives $d'(Y\,|\,\hat{\underset{\sim}{\beta}})$ are given by

$$d'\left(Y\,|\,\underset{\sim}{\beta}\right)=\begin{bmatrix} -\left(\tau_1-\tau_0\right)\cdot\exp\left(\beta_{01}\right)\cdot\left\{A\cdot\exp\left(\beta_1+\beta_2\cdot U\right)-B\cdot\exp\left(\beta_2\cdot U\right)\right\} \\ \vdots \\ -\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)\cdot\left\{A\cdot\exp\left(\beta_1+\beta_2\cdot U\right)-B\cdot\exp\left(\beta_2\cdot U\right)\right\} \\ -A\cdot\left\{\sum_{j=1}^{J^*-1}\left(\tau_j-\tau_{j-1}\right)\cdot\exp\left(\beta_{0j}+\beta_1+\beta_2\cdot U\right)\right\}-A\cdot\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}+\beta_1+\beta_2\cdot U\right) \end{bmatrix}$$

4.4 Estimated marginal survivals

The confounder variable U is unknown. We assume that it takes values 0 and 1 with pre-specified probabilities π_0 and $\pi_1=1-\pi_0$. Marginal survival distribution can be estimated using a mixture survival distribution. Hazards from marginal survival distributions of $T=0$ and $T=1$ are not proportional.

$$\begin{aligned} S\left(Y\,|\,\underset{\sim}{\beta};T\right)&=\pi_0\cdot S\left(Y|U=0,T\right)+\pi_1\cdot S\left(Y|U=1,T\right) \\ &=\pi_0\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1\cdot T\right)\right]+ \\ &\pi_1\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1\cdot T+\beta_2\right)\right] \end{aligned} \tag{4.4.1}$$

Let $g\left(Y\,|\,\underset{\sim}{\beta}\right)=S\left(Y\,|\,\underset{\sim}{\beta}\right)$, and

$$A=\pi_0\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1\cdot T\right)\right],$$

$$B=\pi_1\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1\cdot T+\beta_2\right)\right].$$

The corresponding derivatives are given as follows:

$$g'\left(Y|\underset{\sim}{\beta}\right)=\begin{bmatrix} -\left\{A\cdot\exp(\beta_1\cdot T)+B\cdot\exp(\beta_1\cdot T+\beta_2)\right\}\cdot(\tau_1-\tau_0)\cdot\exp(\beta_{01}) \\ \vdots \\ -\left\{A\cdot\exp(\beta_1\cdot T)+B\cdot\exp(\beta_1\cdot T+\beta_2)\right\}\cdot\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right) \\ -T\cdot\left\{A\cdot\exp(\beta_1\cdot T)+B\cdot\exp(\beta_1\cdot T+\beta_2)\right\}\cdot\left\{\sum_{j=1}^{J^*-1}\left(\tau_j-\tau_{j-1}\right)\cdot\exp\left(\beta_{0j}\right)+\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)\right\} \end{bmatrix}$$

Let:

$$a=\pi_0\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1\right)\right]$$

$$b=\pi_1\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_1+\beta_2\right)\right]$$

$$c=\pi_0\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\right]$$

$$d=\pi_1\cdot\exp\left[-\left\{\sum_{j=1}^{J^*-1}\exp\left(\beta_{0j}\right)\cdot\left(\tau_j-\tau_{j-1}\right)+\exp\left(\beta_{0J^*}\right)\cdot\left(Y-\tau_{J^*-1}\right)\right\}\cdot\exp\left(\beta_2\right)\right]$$

The log of survival rate ratio is:

$$r\left(Y|\underset{\sim}{\beta}\right)=\log(a+b)-\log(c+d) \tag{4.4.2}$$

The corresponding derivatives are:

$$r'\left(Y|\underset{\sim}{\beta}\right)=\begin{bmatrix} \dfrac{-\left\{a\cdot\exp(\beta_1)+b\cdot\exp(\beta_1+\beta_2)\right\}\cdot(\tau_1-\tau_0)\cdot\exp(\beta_{01})}{a+b}+\dfrac{\left\{c+d\cdot\exp(\beta_2)\right\}\cdot(\tau_1-\tau_0)\cdot\exp(\beta_{01})}{c+d} \\ \vdots \\ \dfrac{-\left\{a\cdot\exp(\beta_1)+b\cdot\exp(\beta_1+\beta_2)\right\}\cdot\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)}{a+b}+\dfrac{\left\{c+d\cdot\exp(\beta_2)\right\}\cdot\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)}{c+d} \\ \dfrac{-\left\{a\cdot\exp(\beta_1)+b\cdot\exp(\beta_1+\beta_2)\right\}\cdot\left\{\sum_{j=1}^{J^*-1}\left(\tau_j-\tau_{j-1}\right)\cdot\exp\left(\beta_{0j}\right)+\left(Y-\tau_{J^*-1}\right)\cdot\exp\left(\beta_{0J^*}\right)\right\}}{a+b} \end{bmatrix}$$

The estimated survival rate ratio is then obtained from $\exp\left(\hat{r}\left(Y \mid \hat{\underset{\sim}{\beta}}\right)\right)$. The 95% CI of the estimated rate ratio is given by $\exp\left(\hat{r}(Y \mid \hat{\underset{\sim}{\beta}}) \pm 1.96 * \widehat{\text{var}}[\hat{r}(Y \mid \hat{\underset{\sim}{\beta}})]\right)$ where

$$\text{var}\left[\hat{r}(Y \mid \hat{\underset{\sim}{\beta}})\right] \approx r'(Y \mid \hat{\underset{\sim}{\beta}})^T \, \text{cov}(\hat{\underset{\sim}{\beta}}) r'(Y \mid \hat{\underset{\sim}{\beta}})$$

The survival difference is:

$$d(Y \mid \underset{\sim}{\beta}) = a + b - c - d \tag{4.4.3}$$

The corresponding derivatives are:

$$d'\left(Y \mid \underset{\sim}{\beta}\right) = \begin{bmatrix} -\left\{a \cdot \exp(\beta_1) + b \cdot \exp(\beta_1 + \beta_2)\right\} \cdot (\tau_1 - \tau_0) \cdot \exp(\beta_{01}) + \left\{c + b \cdot \exp(\beta_2)\right\} \cdot (\tau_1 - \tau_0) \cdot \exp(\beta_{01}) \\ \vdots \\ -\left\{a \cdot \exp(\beta_1) + b \cdot \exp(\beta_1 + \beta_2)\right\} \cdot \left(Y - \tau_{J^*-1}\right) \cdot \exp\left(\beta_{0J^*}\right) + \left\{c + b \cdot \exp(\beta_2)\right\} \cdot \left(Y - \tau_{J^*-1}\right) \cdot \exp\left(\beta_{0J^*}\right) \\ -\left\{a \cdot \exp(\beta_1) + b \cdot \exp(\beta_1 + \beta_2)\right\} \cdot \left\{\sum_{j=1}^{J^*-1} (\tau_j - \tau_{j-1}) \cdot \exp(\beta_{0j}) + \left(Y - \tau_{J^*-1}\right) \cdot \exp\left(\beta_{0J^*}\right)\right\} \end{bmatrix}$$

4.5 Piecewise constant hazard function and the Poisson distribution

Holford (1980) and Laird and Oliver (1981) discovered that the maximum likelihood estimators from an exponential distribution are identical to the ones from a Poisson distribution. This finding extends directly to a piecewise constant hazard function. The maximum likelihood estimators of the parameters from a piecewise constant hazard function happen to be the same as those from an equivalent Poisson density function.

Let d_{ij} be the event number for subject i within the time interval of $(\tau_{j-1}, \tau_j]$, where the left boundary of the time interval is excluded and the right boundary is included. d_{ij} is equal to the censoring status C_{ij} for subject i during the time interval of $(\tau_{j-1}, \tau_j]$. For instance, if a subject i died at year 7, for the time interval $(0, 6]$,

$d_{ij} = C_{ij} = 0$, but for the time interval $(6,12]$, $d_{ij} = C_{ij} = 1$. Assume the distribution of d_{ij} following a Poisson density function.

$$f\left(d_{ij}\right) = \frac{\mu_{ij}{}^{d_{ij}} \cdot \exp\left(-\mu_{ij}\right)}{d_{ij}!} \tag{4.5.1}$$

where μ_{ij} is the mean of the Poisson distribution for patient i within the time interval of $(\tau_{j-1}, \tau_j]$. In fact, $\mu_{ij} = \lambda_{ij} \cdot Y_{ij}$, where λ_{ij} is the constant hazard for subject i during the time interval of $(\tau_{j-1}, \tau_j]$. In log-linear regression model,

$$\log\left(\mu_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \log\left(Y_{ij}\right)$$

$$\lambda_{ij} = \exp\left(\beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i\right) \tag{4.5.2}$$

where the baseline hazard during the time interval of $(\tau_{j-1}, \tau_j]$ is $\lambda_j = \exp\left(\beta_{0j}\right)$. Equation (4.5.2) is the same as (4.2.6).

Y_{ij} in the equation is the total exposure time in person years for subject i during time interval $(\tau_{j-1}, \tau_j]$. For the previous example, if subject i died at year 7, for the time interval $(0,6]$, $Y_{ij} = 6$, but for the time interval $(6,12]$, $Y_{ij} = 1$. Y_{ij} is also called an offset in log-linear regression model.

In the two-stage likelihood-based IVA model, if we use a Poisson mass function instead of the corresponding piecewise constant hazard model at stage two, will we obtain the same parameter estimators?

$$\log E\left(C_{ij} \mid T_i, U_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \log\left(Y_{ij}\right) \tag{4.5.3}$$

We examine the likelihood function with a Poisson mass function at stage two.

$$L(\beta_{01},\beta_{02},...,\beta_{0J},\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{d}) = \prod_{i=1}^{N}\prod_{j=1}^{J_i^*} P\left(d_{ij} \mid Z_i, Y_{ij}\right)$$

$$= \prod_{i=1}^{N}\sum_{U}\sum_{T|U,Z_i}\left[P(T \mid U,Z_i)\cdot P(U)\cdot \prod_{j=1}^{J_i^*} P\left(d_{ij} \mid T,U,Y_{ij}\right)\right]$$

$$= \prod_{i=1}^{N}\sum_{U}\sum_{T|U,Z_i}\left[P(T \mid U,Z_i)\cdot P(U)\cdot \prod_{j=1}^{J_i^*} \frac{\left\{\lambda_j \cdot \exp\left(\beta_1 \cdot T + \beta_2 \cdot U\right)\cdot Y_{ij}\right\}^{d_{ij}} \cdot \exp\left\{-\lambda_j \cdot \exp\left(\beta_1 \cdot T + \beta_2 \cdot U\right)\cdot Y_{ij}\right\}}{d_{ij}!}\right] \tag{4.5.4}$$

In equation (4.5.4), $\lambda_j = \exp\left(\beta_{0j}\right)$, $d_{ij} = C_{ij}$, and $d_{ij}!$ is always 1 because d_{ij} can be only 0 or 1 for subject i. $d_{ij} = 0$ for $j = 1,2,...J_i^* - 1$, and $d_{iJ^*} = C_i$. Equation (4.5.4) is equivalent to:

$$L(\beta_{01},\beta_{02},...,\beta_{0J},\beta_1 \mid \alpha_0,\alpha_1,\alpha_2,\beta_2;\underset{\sim}{Y},\underset{\sim}{Z},\underset{\sim}{C})$$

$$= \prod_{i=1}^{N}\sum_{U}\sum_{T|U,Z_i}\left[P(T \mid U,Z_i)\cdot P(U)\cdot\left\{\lambda_{J_i^*}\cdot \exp\left(\beta_1 \cdot T + \beta_2 \cdot U\right)\right\}^{C_i}\cdot\left(Y_i - \tau_{J_i^*-1}\right)^{C_i}\cdot \right.$$

$$\left. \exp\left[-\left\{\sum_{j=1}^{J_i^*-1}\lambda_j \cdot \exp\left(\beta_1 \cdot T + \beta_2 \cdot U\right)\cdot\left(\tau_j - \tau_{j-1}\right)\right\} - \lambda_{J_i^*}\cdot \exp\left(\beta_1 \cdot T + \beta_2 \cdot U\right)\cdot\left(Y_i - \tau_{J_i^*-1}\right)\right]\right] \tag{4.5.5}$$

for $\tau_{J_i^*-1} < Y_i \le \tau_{J_i^*}$.

Equation (4.5.5) has an additional term of $\left(Y_i - \tau_{J_i^*-1}\right)^{C_i}$ compared to equation (4.2.8). Since the likelihood function contains summations over T and U, the score function from (4.5.5) is no longer the same as the score function from (4.2.8). Therefore, in our two-stage likelihood-based model, the density function of a piecewise constant hazard model cannot be replaced by a Poisson mass function.

In the following sections, we will use the Poisson distribution approach to the piecewise constant hazard function in the simulations when the two-stage likelihood-based model is not used.

4.6 Simulation using piecewise constant hazard model in IVA

Simulation is conducted using a two-piecewise constant hazard model in the second stage. A procedure described by Walke (2010) is adapted to generate random times from a two-piecewise constant hazard function. Although we can arbitrarily select values for the baseline hazards in the simulation, we looked up previous study and observed survival probability at year 10 for moderately differentiated prostate cancer patients is 0.884. Assuming the survivals follow an exponential distribution with a constant hazard, we obtain the value of the hazard as $-\log(0.884)/10 \approx 0.123$. This value is used as the piece one hazard from patients with $T=1$ and $U=0$. For piece two, we increase the hazard by 1.5 times to be 0.01845. Consequently, with pre-defined hazard ratios between levels of T and U, we derive the baseline hazards being 0.018 for the years of $(0,6]$, and 0.028 for the years of $(6,12]$. $\tau_0 = 0$, $\tau_1 = 6$, and $\tau_2 = 12$. Other parameters are pre-defined in Table 4.6.1.

Table 4.6.1 Parameters used in simulation of piecewise constant hazard model in IVA

$$P(U=1)=0.2$$
$$P(Z=1)=0.5$$

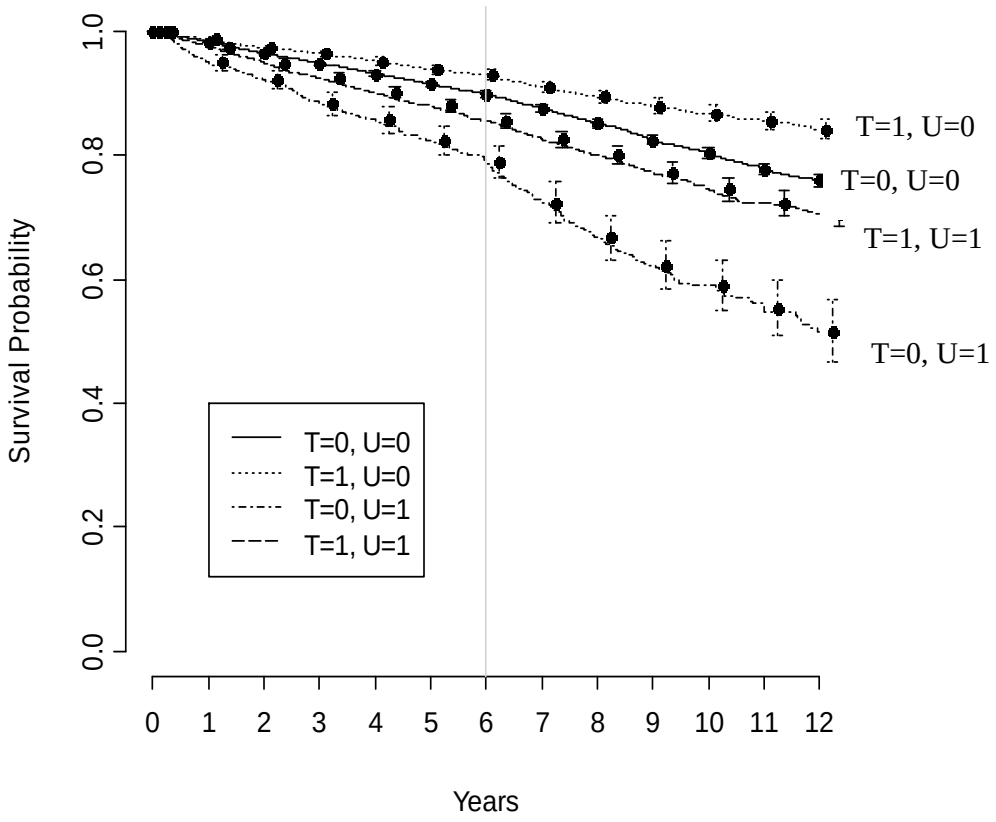
$$\left.\begin{array}{l} P(T=1\mid Z=0,U=0)=0.1 \\ P(T=1\mid Z=0,U=1)=0.7 \\ P(T=1\mid Z=1,U=0)=0.3 \\ P(T=1\mid Z=1,U=1)=0.9 \end{array}\right. \Longrightarrow \begin{array}{l} \dfrac{odds(T=1\mid Z=1,U)}{odds(T=1\mid Z=0,U)}=3.86 \\ \dfrac{odds(T=1\mid U=1,Z)}{odds(T=1\mid U=0,Z)}=21 \end{array}$$

$$hazard(T=1,U=0)=0.0123 \quad \text{for} \quad 0<Y\le 6$$
$$hazard(T=1,U=0)=0.01845 \quad \text{for} \quad 6<Y\le 12$$

$$\frac{hazard(T=0,U=0)}{hazard(T=1,U=0)}=1.5 \qquad \frac{hazard(T=1,U=1)}{hazard(T=1,U=0)}=2 \qquad \frac{hazard(T=0,U=1)}{hazard(T=1,U=0)}=3$$

As in section 3.6, 100 loops are simulated with a sample size of 30,000 patients for each loop. Kaplan-Meier survival curves are plotted in Figure 4.6.1 for one of the 100 datasets. The hazard is higher after year 6 than the hazard before year 6 as shown by the steeper slope from year 6 to year 12 than from year 0 to year 6.

Figure 4.6.1 Kaplan-Meier survival curve – simulated data from a two-piecewise constant hazard model



The models from Table 4.6.2 are fitted to the randomly generated data. In model **a**, data are fitted to a two-piecewise constant hazard model using treatment as a fixed effect but omitting the unobserved confounder in the model. In model **b**, data are fitted to a two-piecewise constant hazard model with both treatment and confounder, which are randomly generated from binomial distributions, as fixed

effects. Both model **c** and model **d** give Wald type estimators for the treatment effect. In model **c**, data are fitted to a two-piecewise constant hazard model with the instrumental variable as the only fixed effect. In the same way, in model **d**, the instrumental variable is also included as the only fixed effect assuming the data follow a Cox proportional hazard model. Coefficient estimators of the instrumental variable are then divided by the difference in treatment proportions between the two classes of instrumental variable. Finally, in model **e**, the two-stage likelihood-based model is examined for its treatment effect estimation.

Table 4.6.2 Comparative models used in simulation of piecewise constant hazard model in IVA

Model a	$\log E\left(C_{ij} \mid T_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \log\left(Y_{ij}\right)$ where $j = 1, 2$
Model b	$\log E\left(C_{ij} \mid T_i, U_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \log\left(Y_{ij}\right)$ where $j = 1, 2$
Model c	$T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i$ $\log E\left(C_{ij} \mid Z_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z_i\right) + \log\left(Y_{ij}\right)$ where $j = 1, 2$
Model d	$T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i$ $\log h_i\left(Y_i \mid Z_i\right) = \log h_0 + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z_i\right) = \beta_0 + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z_i\right)$
Model e	$\log \frac{E\left(T_i \mid Z_i, U_i\right)}{1 - E\left(T_i \mid Z_i, U_i\right)} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_i$ $\log \lambda_{ij}\left(Y_{ij} \mid T_i, U_i\right) = \log \lambda_j + \beta_1 \cdot T_i + \beta_2 \cdot U_i = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i$ where $j = 1, 2$

Estimated coefficient means for models **a-e** and their empirical standard deviations from the 100 samples are given in Table 4.6.3. To distinguish the estimated parameters from different models, in section 4.6 to section 4.8, we use the subscript “Wald” to indicate the estimated parameters are from the Wald method, and subscript “IV” to indicate the estimated parameters are from two-stage likelihood-based IVA method. The estimated baseline hazards and hazard

ratios along with their 95% confidence intervals are obtained from the estimated parameters in Table 4.6.3. They are given in Table 4.6.4. Parameter estimators from models **b** to **e** are all very close to the true values. The treatment effect estimator from model **a** is far away from the true value because the model does not take the unobserved confounder into account. Model **b** has the smallest variance for the treatment effect estimator because it assumes the confounder was observed. Both Wald type estimators and two-stage likelihood-based IVA estimator have much larger variances than the simple log-linear model **b.** The large variances come from the third variable of instrument.

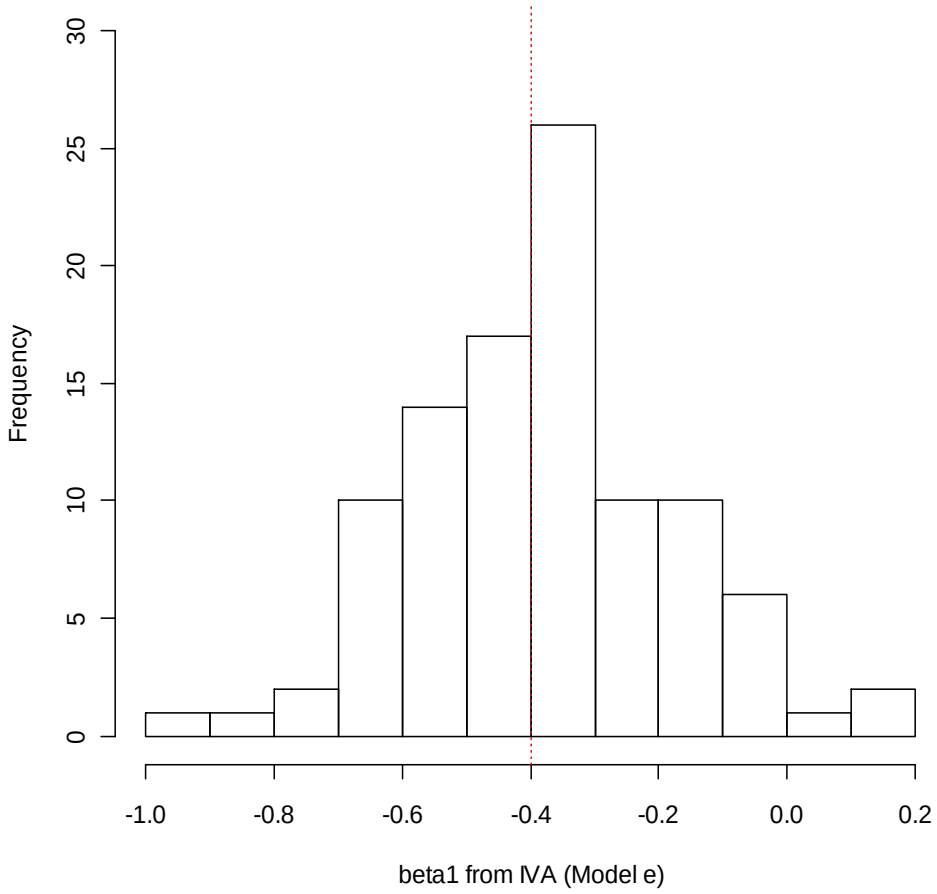
Table 4.6.3 Estimated coefficients from simulation of models in Table 4.6.2

True value	**Model a**	**Model b**	**Model c**	**Model d**	**Model e**
$\alpha_0 = -2.20$					$\hat{\alpha}_{0IV} = -2.20$ SE=0.03
$\alpha_1 = 1.35$					$\hat{\alpha}_{1IV} = 1.35$ SE=0.04
$\alpha_2 = 3.04$					
$\beta_{01} = -3.99$	$\hat{\beta}_{01} = -3.94$ SE=0.02	$\hat{\beta}_{01} = -3.99$ SE=0.02			$\hat{\beta}_{01IV} = -4.00$ SE=0.07
$\beta_{02} = -3.59$	$\hat{\beta}_{02} = -3.54$ SE=0.03	$\hat{\beta}_{02} = -3.59$ SE=0.03			$\hat{\beta}_{02IV} = -3.60$ SE=0.07
$\beta_1 = -0.40$	$\hat{\beta}_1 = -0.06$ SE=0.04	$\hat{\beta}_1 = -0.40$ SE=0.04	$\hat{\beta}_{1Wald} = -0.35$ SE=0.18	$\hat{\beta}_{1Wald} = -0.35$ SE=0.18	$\hat{\beta}_{1IV} = -0.38$ SE=0.20
$\beta_2 = 0.69$		$\hat{\beta}_2 = 0.69$ SE=0.04			

Table 4.6.4 Estimated hazards and hazard ratios from simulation of models in Table 4.6.2

	Baseline Hazard for 0<Years≤6	Baseline Hazard for 6<Years≤12	Treatment Effect (Hazard Ratio)	Unobserved Confounder Effect (Hazard Ratio)
True value	0.018	0.028	0.667	2
Model a (95% CI)	0.019 (0.019, 0.020)	0.029 (0.027, 0.031)	0.940 (0.875, 1.010)	
Model b (95% CI)	0.018 (0.018, 0.019)	0.028 (0.026, 0.030)	0.667 (0.614, 0.725)	2.001 (1.844, 2.172)
Model c (95% CI)			0.701 (0.490, 1.002)	
Model d (95% CI)			0.701 (0.490, 1.003)	
Model e (95% CI)	0.018 (0.016, 0.021)	0.027 (0.024, 0.031)	0.684 (0.460, 1.016)	

Figure 4.6.2 Histogram of $\hat{\beta}_{1IV}$ from the two-stage likelihood-based IVA - two-piecewise constant hazard model

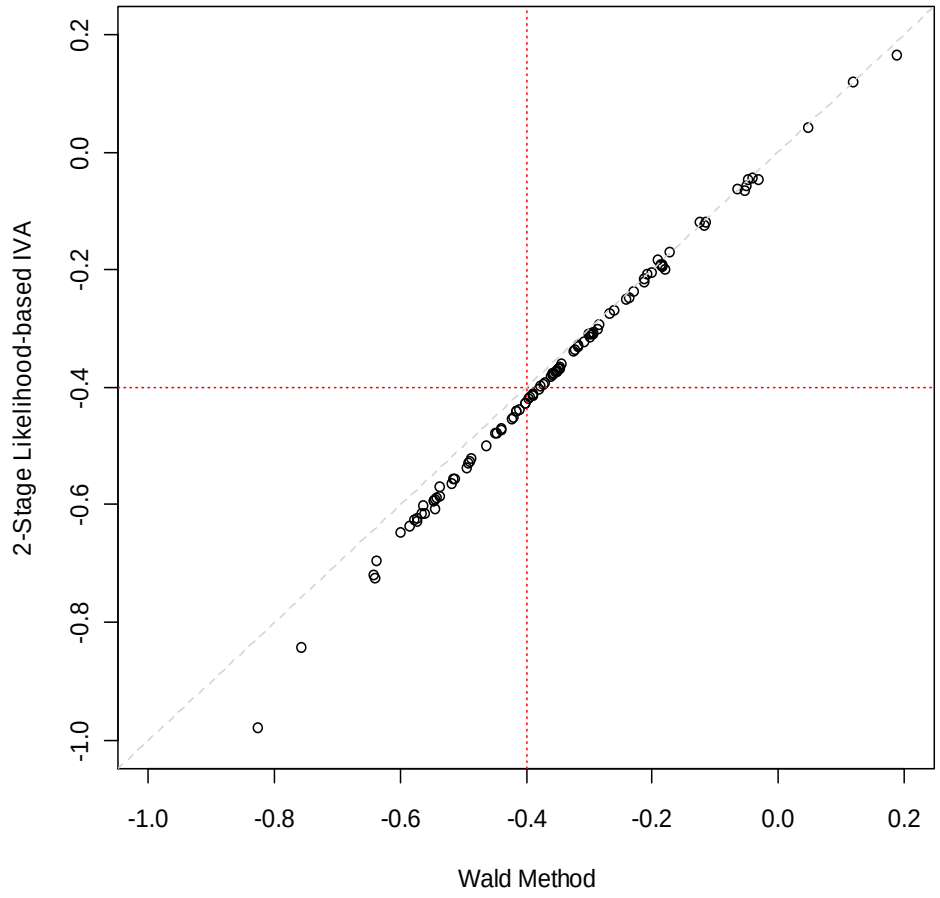


Note: Dotted line indicates the true value.

A histogram of the 100 treatment effect estimators, $\hat{\beta}_{1IV}$, is displayed in Figure 4.6.2. The 95% confidence interval is (-0.78; 0.02). Because of the large variance, there is not enough power to detect treatment effect, particularly when the baseline hazard is as low as 0.018.

Treatment estimators from Wald type method are compared with that from two-stage likelihood-based IVA. Plot (Figure 4.6.3) of $\hat{\beta}_{1Wald}$ from model **c** versus $\hat{\beta}_{1IV}$ from model **e** shows linear correlation between them.

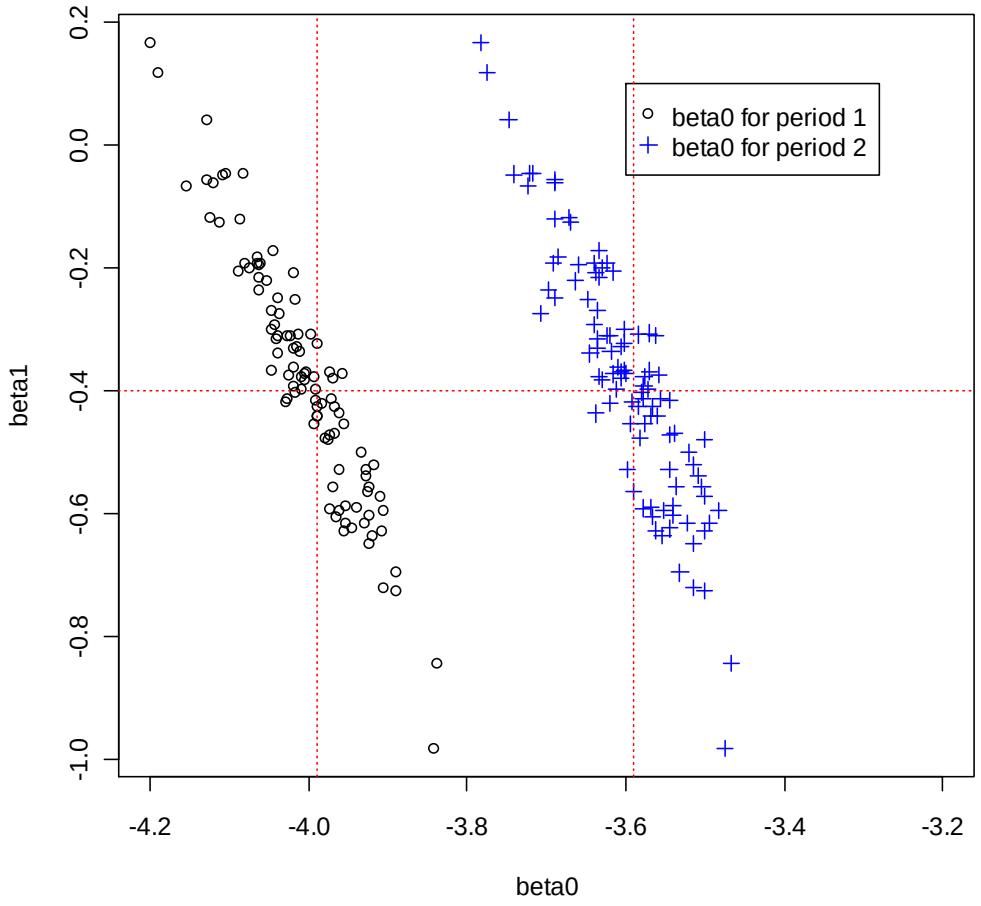
Figure 4.6.3 Treatment effect estimators $\hat{\beta}_{1IV}$ vs $\hat{\beta}_{1Wald}$ - two-piecewise constant hazard model



Note: Dotted lines indicate the true value.

It was also noticed that in two-stage likelihood-based model, the treatment estimators are linearly correlated to the baseline hazard estimators. Figure 4.6.4 illustrates this evidence.

Figure 4.6.4 Estimated hazard ratio vs estimated baseline hazards in two-stage likelihood-based IVA



Note: Dotted lines indicate the true value.

With the estimated means of $\hat{\underset{\sim}{\beta}}_{IV}$ from the two-stage likelihood-based IVA (model **e**), we estimate the marginal survival function using pre-specified probabilities $\pi_0 = 0.8$ and $\pi_1 = 0.2$ for $U = 0$ and $U = 1$ respectively. The 5-year and 10-year survival probabilities are calculated and compared between $T = 0$ and $T = 1$ with regard to the estimated survival rate ratio and survival rate difference. Standard errors of the 5-year and 10-year survival probabilities are calculated using the delta method as described in section 4.4. The delta method is also used to construct the 95% confidence intervals for the estimated survival rate ratio and survival rate difference.

In addition to the delta method, empirical standard deviations of the estimators and empirical 95% confidence intervals of the estimators from the 100 samples are also given in Table 4.6.5 and Table 4.6.6.

In Table 4.6.5, the estimated 5-year survival probabilities for both $T=1$ and $T=0$ are very close to the true values, as does the estimated survival rate ratio between the two groups. The differences between the true values and estimated values are less than 1%. The estimated survival rate difference between the two groups deviates from the true value by about 9%. In Table 4.6.6, we observe almost the same results as in Table 4.6.5 for 10-year survival probabilities. The differences between the estimated values and true values are less than 1% for 10-year survival probabilities in both groups and survival rate ratio of $T=1$ vs $T=0$. The estimated survival rate difference of $T=1$ vs $T=0$ at year 10 is 9% larger than the true value. We assume the estimated survival probabilities, the estimated log of survival rate ratios, and the estimated survival rate differences are all asymptotically normally distributed with the means approximated from the functions of the maximum likelihood estimators. However, the approximation of the means on the survival rate difference is not as accurate as on the survival rate ratio.

In both Table 4.6.5 and Table 4.6.6, the empirical standard deviations and empirical 95% confidence intervals of the estimators are close to the ones from the delta method. The empirical 95% confidence intervals are narrower compared to the ones from the delta methods.

Table 4.6.5 Estimated 5-year survivals using marginal survival function from two-stage likelihood-based IVA on simulated data

	True value	Log of estimator	Estimator	95% CI of the estimator
Survival probability $T=1$	0.929		0.931 SE1=0.0081 SE2=0.0080	CI1=(0.916, 0.947) CI2=(0.916, 0.946)
Survival probability $T=0$	0.896		0.895 SE1=0.0051 SE2=0.0051	CI1=(0.885, 0.905) CI2=(0.886, 0.904)
Survival rate ratio	1.037	0.040 SE1=0.0138 SE2=0.0138	1.040	CI1=(1.013, 1.069) CI2=(1.015, 1.065)
Survival rate difference	0.033		0.036 SE1=0.0126 SE2=0.0126	CI1=(0.011, 0.061) CI2=(0.013, 0.058)

SE1 and CI1 = Standard error of the mean or 95% CI from delta method
SE2 and CI2 = Standard error of the mean or 95% CI from empirical distribution

Table 4.6.6 Estimated 10-year survivals using marginal survival function from two-stage likelihood-based IVA on simulated data

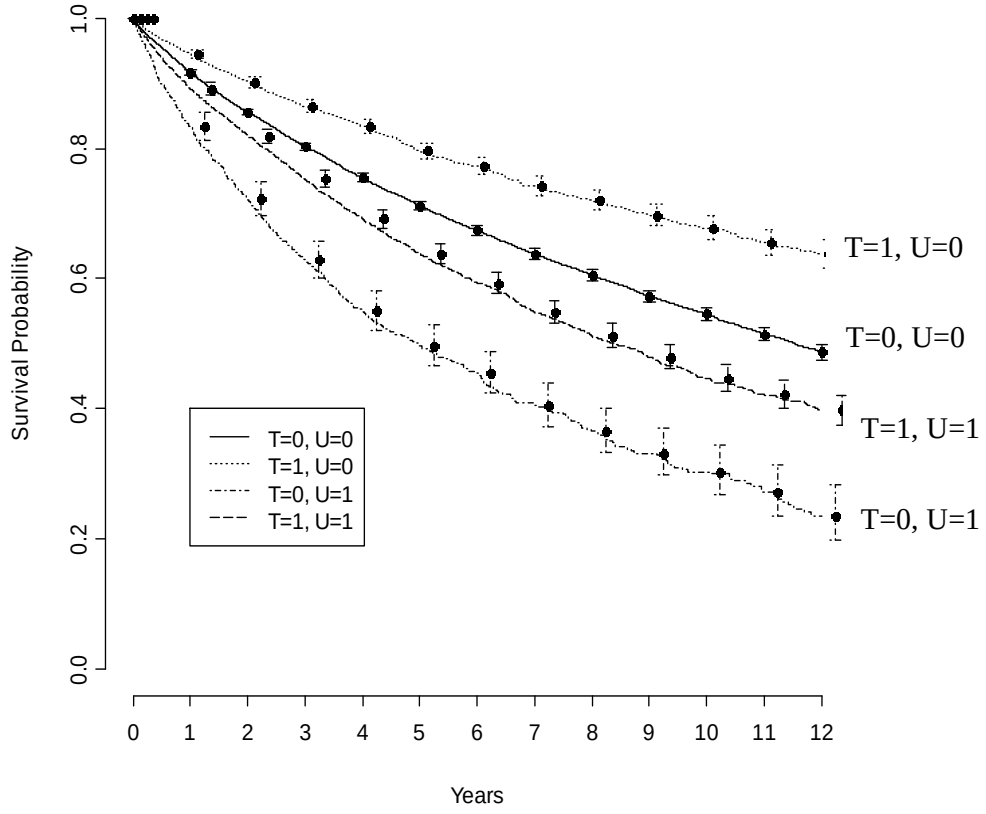
	True Value	Log of estimator	Estimator	95% CI of the estimator
Survival probability $T=1$	0.838		0.844 SE1=0.0171 SE2=0.0169	CI1=(0.810, 0.877) CI2=(0.811, 0.876)
Survival probability $T=0$	0.770		0.768 SE1=0.0104 SE2=0.0103	CI1=(0.748, 0.789) CI2=(0.751, 0.787)
Survival rate ratio	1.089	0.094 SE1=0.0328 SE2=0.0327	1.098	CI1=(1.030, 1.171) CI2=(1.034, 1.162)
Survival rate difference	0.069		0.075 SE1=0.0267 SE2=0.0266	CI1=(0.023, 0.128) CI2=(0.027, 0.122)

SE1 and CI1 = Standard error of the mean or 95% CI from delta method
SE2 and CI2 = Standard error of the mean or 95% CI from empirical distribution

4.7 Simulation using Weibull distribution in IVA

Simulations are also performed for the survival data following a Weibull distribution in the second stage. Instead of pre-defining the baseline hazards, we first randomly generate a set of data as a control group for patients receiving PADT $(T=1)$ and having normal PSA $(U=0)$. The survival time in the control group follows Weibull distribution with 5-year survival rate of 0.799 and 10-year survival rate of 0.668. Then, we specify hazard ratios of 1.5, 2, and 3 to the control group to generate survival times for patients who receive conservative management $(T=0)$ and have normal PSA $(U=0)$, who receive PADT $(T=1)$ and have high PSA $(U=1)$, and who receive conservative management $(T=0)$ and have high PSA $(U=1)$ respectively. The rest of the parameters remain the same as in Table 4.6.1. The Kaplan-Meier survival curves for the four groups are shown in Figure 4.7.1.

Figure 4.7.1 Kaplan-Meier survival curve – simulated data from Weibull distribution



As in Table 4.6.2, the five models from Table 4.7.1 are examined using the simulated data. In models **a** and **b**, a parametric survival model with Weibull distribution is used. Model **a** includes treatment as the only predictor. Model **b** includes both treatment and confounder from simulated data as predictors. It is well known that the parametric survival regression model with an underlying Weibull distribution can also be expressed as a proportional hazard regression model. Due to this, we fit the data with a Cox proportional hazard model in model **c** with both treatment and confounder from simulated data as predictors. In model **d**, the Weibull distributed survival times are regressed on the instrumental variable. Finally, model **e** is our two-stage likelihood-based IVA model which is composed of a log odds

regression model in the first stage and a Weibull survival regression model in the second stage.

Table 4.7.1 Comparative models used in simulation of Weibull regression model in IVA

Model a	$\log(Y_i) = \beta_0^* + \beta_1^* \cdot T_i + \frac{\varepsilon_i}{\gamma}$
Model b	$\log(Y_i) = \beta_0^* + \beta_1^* \cdot T_i + \beta_2^* \cdot U_i + \frac{\varepsilon_i}{\gamma}$
Model c	$\log h_i(Y_i \mid T_i, U_i) = \log h_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i = \beta_0 + \beta_1 \cdot T_i + \beta_2 \cdot U_i$
Model d	$T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i$ $\log(Y_i) = \beta_0^* + \beta_1^* \cdot (\alpha_0 + \alpha_1 \cdot Z_i) + \frac{\varepsilon_i}{\gamma}$
Model e	$\log \frac{E(T_i \mid Z_i, U_i)}{1 - E(T_i \mid Z_i, U_i)} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_i$ $\log(Y_i) = \beta_0^* + \beta_1^* \cdot T_i + \beta_2^* \cdot U_i + \frac{\varepsilon_i}{\gamma}$

In our simulation, we parameterize the coefficients of the equations in Table 4.7.1 as in model **c,** i.e., the parameterization in the Cox regression model. The relationships between β and β^* are:

$$\beta_0 = -\beta_0^* \cdot \gamma, \quad \beta_1 = -\beta_1^* \cdot \gamma, \quad \text{and} \quad \beta_2 = -\beta_2^* \cdot \gamma$$

In addition, with 5-year survival rate of 0.799 and 10-year survival rate of 0.668 in the control group of patients receiving PADT and having normal PSA, we obtain the two parameters values of the Weibull distribution.

$$\lambda = \exp(\beta_0 + \beta_1 \cdot T) = 0.057 \tag{4.7.1}$$

and $\gamma = 0.846$.

For $\beta_1 = \log(1/1.5)$, we solve equation (4.6.1), and obtain $\beta_0 = -2.45$.

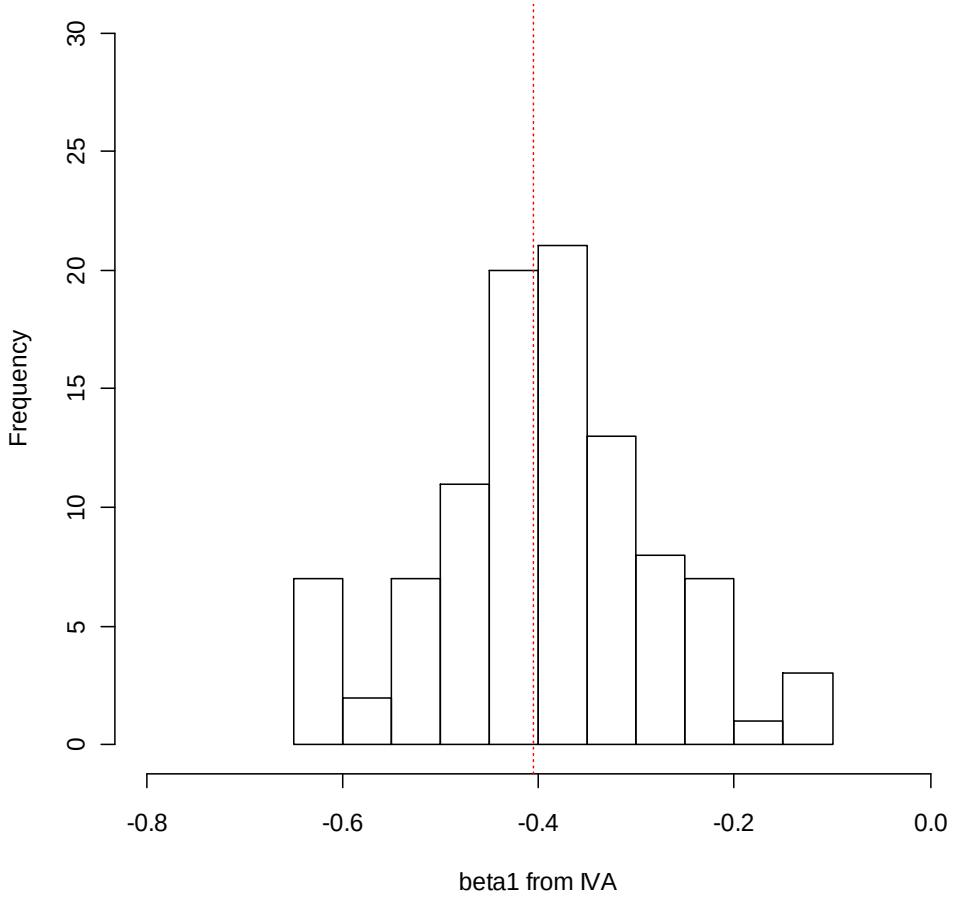
With a sample size of 30,000 in each loop, parameters from models in Table 4.7.1 are estimated 100 times. Their means and standard errors are obtained and listed in Table 4.7.2.

Table 4.7.2 Estimates from simulation of models in Table 4.7.1

True value	**Model a**	**Model b**	**Model c**	**Model d**	**Model e**
$\alpha_0 = -2.20$					$\hat{\alpha}_{0IV} = -2.20$ SE=0.03
$\alpha_1 = 1.35$					$\hat{\alpha}_{1IV} = 1.35$ SE=0.03
$\alpha_2 = 3.04$					
$\gamma = 0.85$	$\hat{\gamma} = 0.84$ SE=0.01	$\hat{\gamma} = 0.84$ SE=0.01			$\hat{\gamma}_{IV} = 0.85$ SE=0.01
$\beta_0 = -2.45$	$\hat{\beta}_0 = -2.39$ SE=0.02	$\hat{\beta}_0 = -2.45$ SE=0.02			$\hat{\beta}_{0IV} = -2.45$ SE=0.04
$\beta_1 = -0.41$	$\hat{\beta}_1 = -0.07$ SE=0.02	$\hat{\beta}_1 = -0.41$ SE=0.03	$\hat{\beta}_1 = -0.41$ SE=0.03	$\hat{\beta}_{1Wald} = -0.37$ SE=0.10	$\hat{\beta}_{1IV} = -0.39$ SE=0.11
$\beta_2 = 0.69$		$\hat{\beta}_2 = 0.70$ SE=0.03	$\hat{\beta}_2 = 0.70$ SE=0.03		

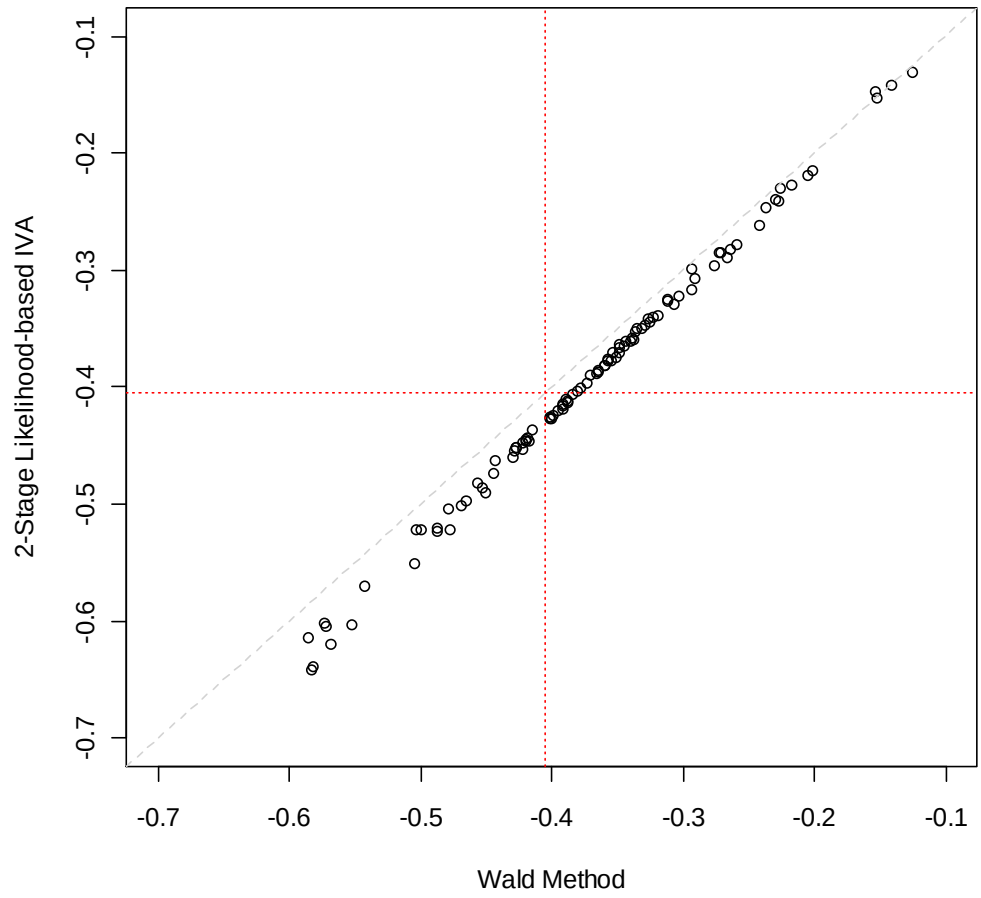
As expected, models **b** and **c** estimate the parameters accurately because the unobserved confounder is included in the model. Models **d** and **e** also give point estimators close to the parameters, but the standard errors are much larger due to the variation from the instrumental variable. Figure 4.7.2 displays the histogram of the 100 $\hat{\beta}_{1IV}$. $\hat{\beta}_{1IV}$ is always linearly correlated with $\hat{\beta}_{1wald}$ as shown in Figure 4.7.3.

Figure 4.7.2 Histogram of $\hat{\beta}_{1IV}$ from two-stage likelihood-based IVA – Weibull distribution



Note: Dotted line indicates the true value.

Figure 4.7.3 Treatment effect estimators $\hat{\beta}_{1IV}$ vs $\hat{\beta}_{1Wald}$ - Weibull distribution



Note: Dotted lines indicate the true value.

Using the estimated parameters from the two-stage likelihood-based IVA model, we plot the hazard functions in Figure 4.7.4 and the survival curves in Figure 4.7.5. Figure 4.7.5 is almost the same as Figure 4.7.1.

Figure 4.7.4 Hazard function based on estimated parameters from two-stage likelihood-based IVA – Weibull distribution

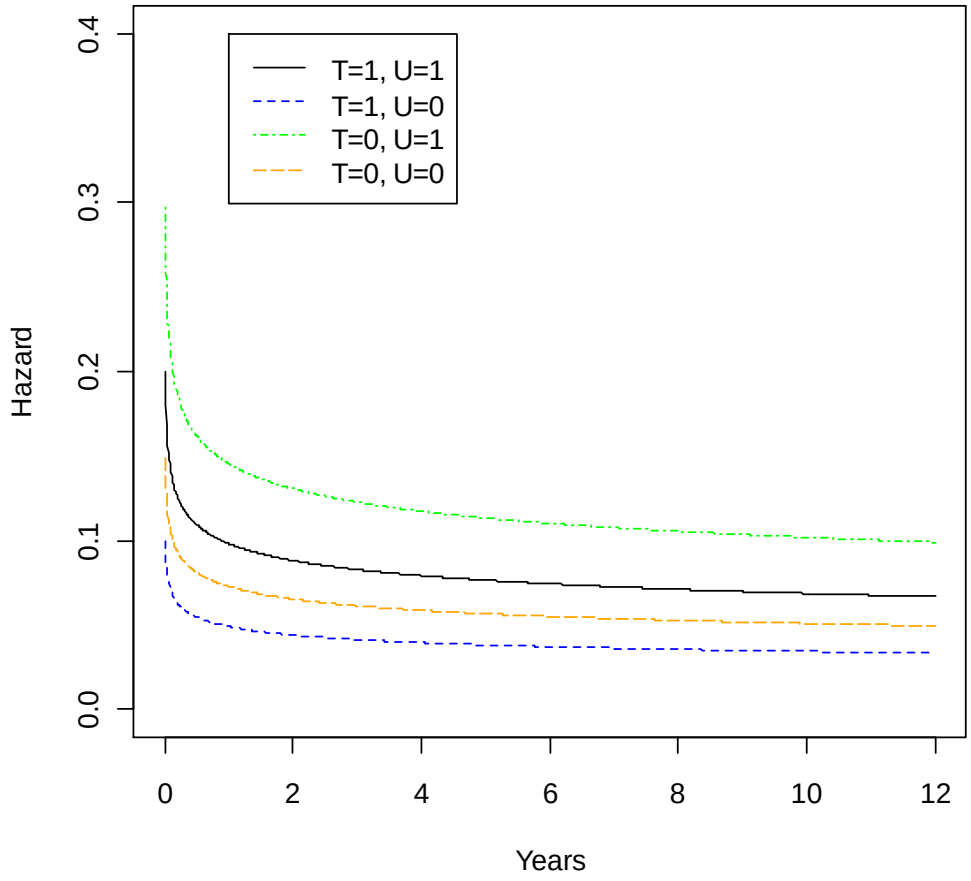
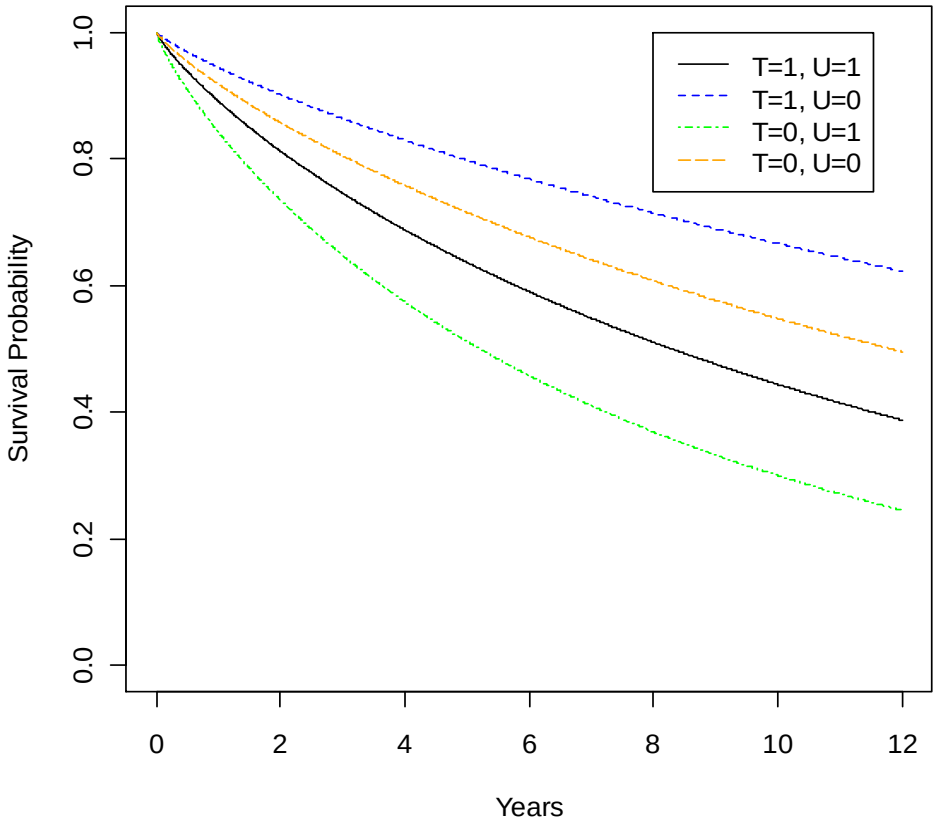


Figure 4.7.5 Survival curve based on estimated parameters from two-stage likelihood-based model – Weibull distribution



4.8 Fit Weibull distributed data with a piecewise constant hazard model

In this section, we fit Weibull distributed data which are generated as described in section 4.7 with a two-piecewise constant hazard function in the two-stage likelihood-based IVA. The mean treatment effect estimator is compared to the true value. The purpose of this simulation is to investigate how well the piecewise constant hazard function is fitted in the proportional hazard model. We fit the data with four models in Table 4.8.1. In model **d**, a two-piecewise constant hazard function

with a cut-off point at year 1 is used in the second stage. The results are given in Table 4.8.2.

Table 4.8.1 Comparative models used in simulation of piecewise constant hazard model in IVA

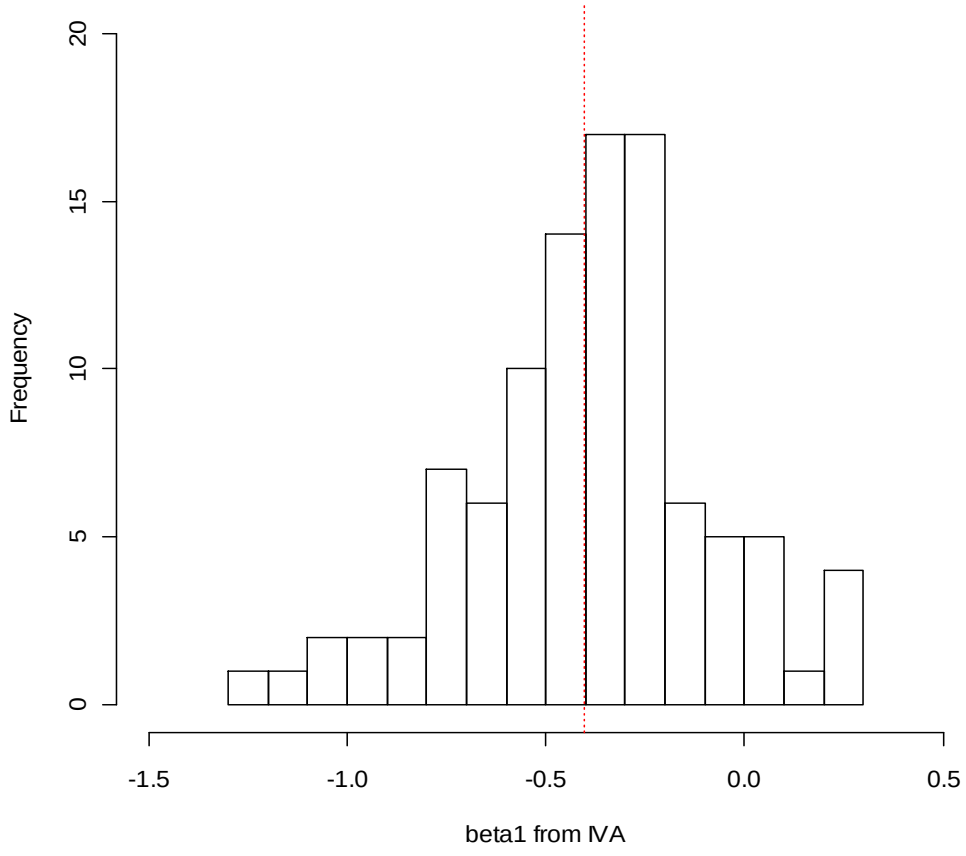
Model a	$\log E\left(C_{ij} \mid T_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \log\left(Y_{ij}\right)$ where $j = 1, 2$
Model b	$\log E\left(C_{ij} \mid T_i, U_i, Y_{ij}\right) = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i + \log\left(Y_{ij}\right)$ where $j = 1, 2$
Model c	$T_i = \alpha_0 + \alpha_1 \cdot Z_i + \upsilon_i$ $\log h_i\left(Y_i \mid Z_i\right) = \log h_0 + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z_i\right) = \beta_0 + \beta_1 \cdot \left(\alpha_0 + \alpha_1 \cdot Z_i\right)$
Model d	$\log \dfrac{E\left(T_i \mid Z_i, U_i\right)}{1 - E\left(T_i \mid Z_i, U_i\right)} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_i$ $\log \lambda_{ij}\left(Y_{ij} \mid T_i, U_i\right) = \log \lambda_j + \beta_1 \cdot T_i + \beta_2 \cdot U_i = \beta_{0j} + \beta_1 \cdot T_i + \beta_2 \cdot U_i$ where $j = 1, 2$

Table 4.8.2 Estimated coefficients from simulation of models in Table 4.8.1 – fit Weibull curve with two-piecewise constant hazard curve

True value	**Model a**	**Model b**	**Model c**	**Model d**
$\alpha_0 = -2.20$		$\hat{\alpha}_0 = -2.20$ SE=0.03		$\hat{\alpha}_{0IV} = -2.20$ SE=0.03
$\alpha_1 = 1.35$		$\hat{\alpha}_1 = 1.35$ SE=0.03		$\hat{\alpha}_{1IV} = 1.35$ SE=0.03
$\alpha_2 = 3.04$		$\hat{\alpha}_2 = 3.04$ SE=0.04		
	$\hat{\beta}_{01} = -2.39$ SE=0.02	$\hat{\beta}_{01} = -2.45$ SE=0.02		$\hat{\beta}_{01IV} = -2.46$ SE=0.09
	$\hat{\beta}_{02} = -0.38$ SE=0.02	$\hat{\beta}_{02} = -0.37$ SE=0.02		$\hat{\beta}_{02IV} = -2.82$ SE=0.09
$\beta_1 = -0.41$	$\hat{\beta}_1 = -0.07$ SE=0.02	$\hat{\beta}_1 = -0.41$ SE=0.03	$\hat{\beta}_{1Wald} = -0.37$ SE=0.10	$\hat{\beta}_{1IV} = -0.39$ SE=0.30
$\beta_2 = 0.69$		$\hat{\beta}_2 = 0.71$ SE=0.03		

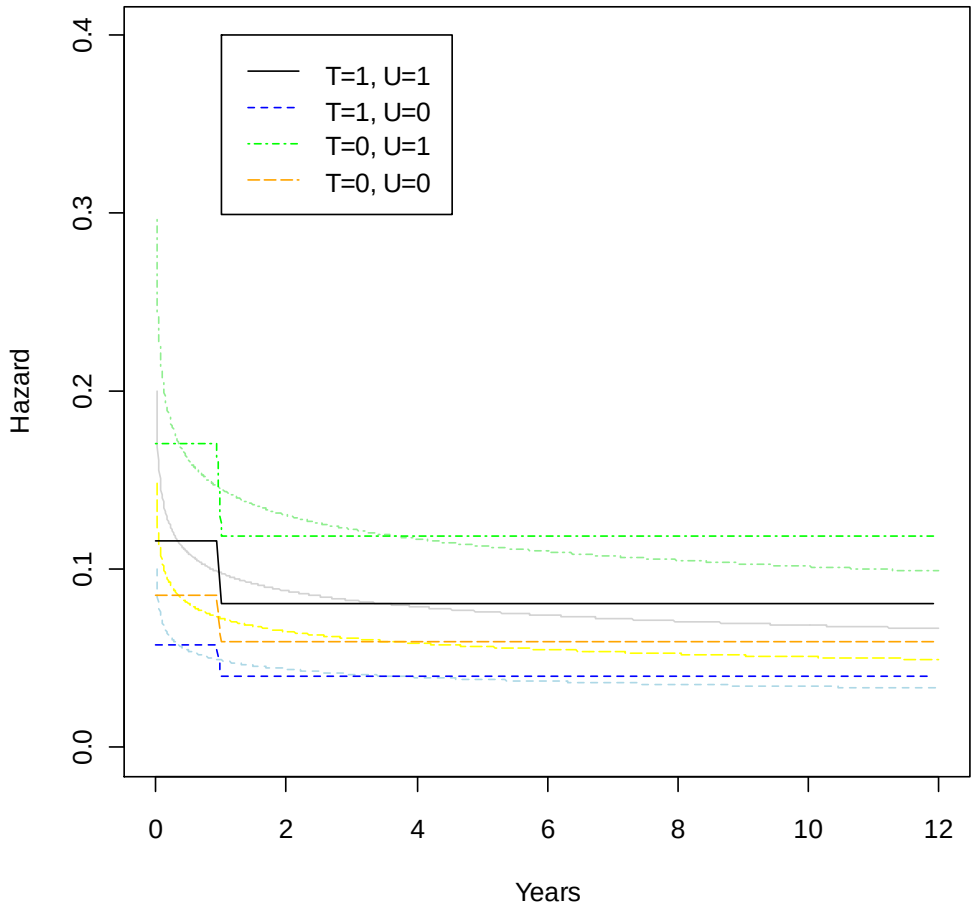
The results show that Wald method in model **c** gives a good estimate of the true value with reasonable large standard error. The two-stage likelihood-based IVA in model **d** also provides a point estimator for the true value. However, the standard error is very large, so that the treatment effect is statistically insignificant. This is because we use two-piecewise constant hazard curve to approximate the Weibull curve in model **d**. In model **c**, a Cox proportional hazard model is used. If we fit constant hazard curve with more than two pieces, the standard error should be smaller.

Figure 4.8.1 Histogram of $\hat{\beta}_{1IV}$ from two-stage likelihood-based IVA – fit Weibull curve with two-piecewise constant hazard curve



Note: Dotted line indicates the true value.

Figure 4.8.2 Hazard function based on estimated parameters from two-stage likelihood-based IVA – two-piecewise constant hazard model vs Weibull distribution



A histogram of the estimated treatment effects is given in Figure 4.8.1. In addition, estimated hazard functions from the Weibull distribution and the two-piecewise constant hazard function are illustrated in Figure 4.8.2.

Since the two-piecewise constant hazard function is not sufficient to approximate the Weibull distributed data, we increase the number of pieces in the hazard function to four or five. Because the simulation takes a considerable amount of time, we generate an additional set of Weibull distributed data with size of 10,000.

We fit this dataset with four-piecewise and five-piecewise constant hazard functions separately. We obtain the MLEs of parameters and the standard errors of the MLEs are from the hessian matrix directly.

Table 4.8.3 Fit Weibull curve with four-piecewise and five-piecewise constant hazard functions

True value	**Wald method**	**Four-piecewise constant hazard** $\tau=(0,1,3,6,12)$	**Five-piecewise constant hazard** $\tau=(0,1,2,4,6,12)$
		$\hat{\beta}_{01IV}=-2.36$ SE=0.07	$\hat{\beta}_{01IV}=-2.36$ SE=0.07
		$\hat{\beta}_{02IV}=-2.61$ SE=0.07	$\hat{\beta}_{02IV}=-2.54$ SE=0.07
		$\hat{\beta}_{03IV}=-2.77$ SE=0.07	$\hat{\beta}_{03IV}=-2.73$ SE=0.07
		$\hat{\beta}_{04IV}=-2.94$ SE=0.08	$\hat{\beta}_{04IV}=-2.77$ SE=0.08
			$\hat{\beta}_{05IV}=-2.94$ SE=0.08
$\beta_1=-0.41$	$\hat{\beta}_{1Wald}=-0.51$ SE=0.17	$\hat{\beta}_{1IV}=-0.59$ SE=0.22	$\hat{\beta}_{1IV}=-0.58$ SE=0.22
$\beta_2=0.69$			

$\tau=(0,1,3,6,12)$: $\tau_0=0,\tau_1=1,\tau_2=3,\tau_3=6,\tau_4=12.$

$\tau=(0,1,2,4,6,12)$: $\tau_0=0,\tau_1=1,\tau_2=2,\tau_3=4,\tau_4=6,\tau_5=12.$

Results in Table 4.8.3 shows the standard error is reduced with the four-piecewise constant hazard function.

To further examine the standard error from the four-piecewise constant hazard function, we generate seven sets of Weibull distributed data with size of 10,000. The

estimated parameter means and their empirical standard errors are given in Table 4.8.4.

Table 4.8.4 Fit Weibull curve with four-piecewise constant hazard functions

$\hat{\beta}_{01IV} = -2.45$	$\hat{\beta}_{02IV} = -2.72$	$\hat{\beta}_{03IV} = -2.82$	$\hat{\beta}_{04IV} = -2.97$	$\hat{\beta}_{1IV} = -0.42$
SE=0.08	SE=0.08	SE=0.07	SE=0.11	SE=0.23

The standard error for the estimated treatment effect is 0.24. The four-piecewise constant hazard function improves the approximation of the Weibull distribution compared to the two-piecewise constant hazard function.

4.9 Example of using two-stage likelihood-based IVA in survival analysis

A subset of the SEER/Medicare database is used in our example. The database includes a cohort study of men with localized prostate cancer who received Medicare. Table 4.9.1 gives the number of patients within the combination of treatment and health service areas for moderately differentiated prostate cancer and poorly differentiated prostate cancer. The health service areas are classified as PADT high usage areas and PADT low usage areas. The classifications of the health service areas serve as an instrumental variable.

Table 4.9.1 Frequency counts of patients with prostate cancer

Treatment	Health Service Areas by PADT usage	Death (%)	Censored (%)	Total
Moderately differentiated localized prostate cancer				
PADT	High	203 (10%)	1855 (90%)	2058
PADT	Low	156 (10%)	1346 (90%)	1502
Conservative Management	High	129 (6%)	2091 (94%)	2220
Conservative Management	Low	428 (10%)	3968 (90%)	4396
Total		916 (9%)	9260 (91%)	**10176**
Poorly differentiated localized prostate cancer				
PADT	High	177 (17%)	881 (83%)	1058
PADT	Low	154 (20%)	603 (80%)	757
Conservative Management	High	61 (16%)	314 (84%)	375
Conservative Management	Low	145 (20%)	576 (80%)	721
Total		537 (18%)	2374 (82%)	**2911**

There are a total of 10,176 patients diagnosed with moderately differentiated localized prostate cancer, and 2911 patients diagnosed with poorly differentiated localized prostate cancer. Their survival status is followed up to about 13 years. By the end of 13 years, the prostate cancer-specific mortality is 9% among the patients with moderately differentiated localized prostate cancer, and 18% among the patients with poorly differentiated localized prostate cancer. Kaplan-Meier survival curves for high PADT usage areas and low PADT usage areas are plotted in Figures 4.9.1 and 4.9.2. For moderately differentiated prostate cancer, 48% of the patients from high PADT usage areas received PADT, and 25% of the patients from low PADT usage areas received PADT. For poorly differentiated prostate cancer, 74% and 51% of the

patients received PADT among high PADT usage areas and low PADT usage areas, respectively.

Figure 4.9.1 Kaplan-Meier survival curve – moderately differentiated prostate cancer

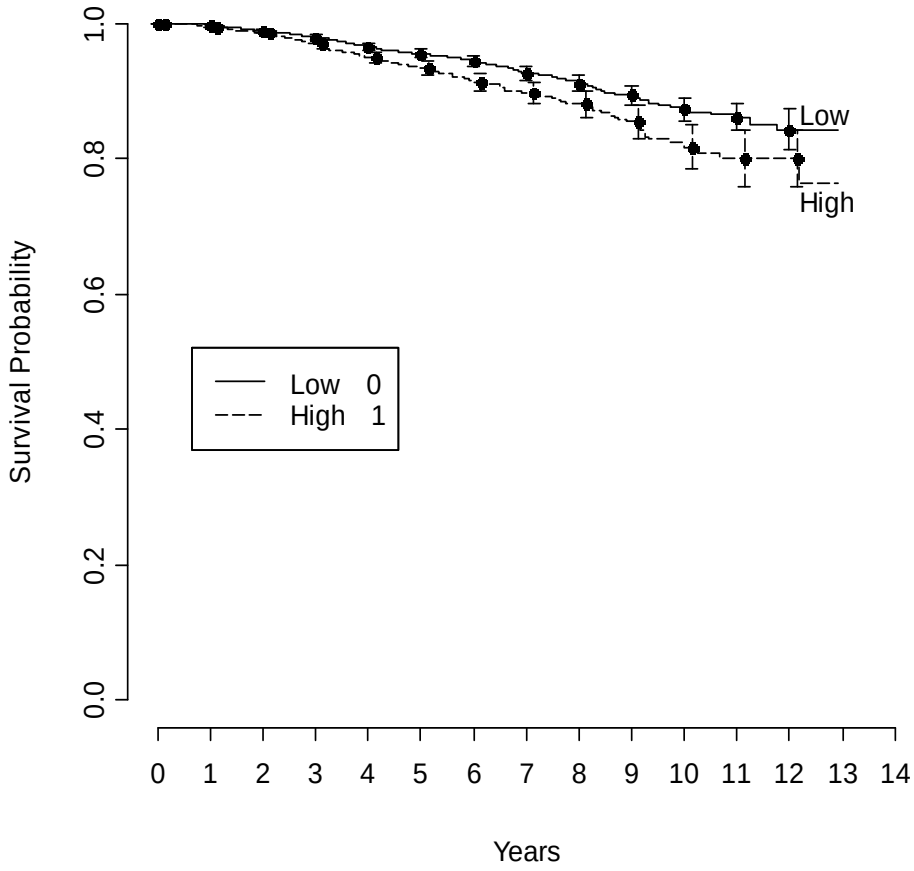
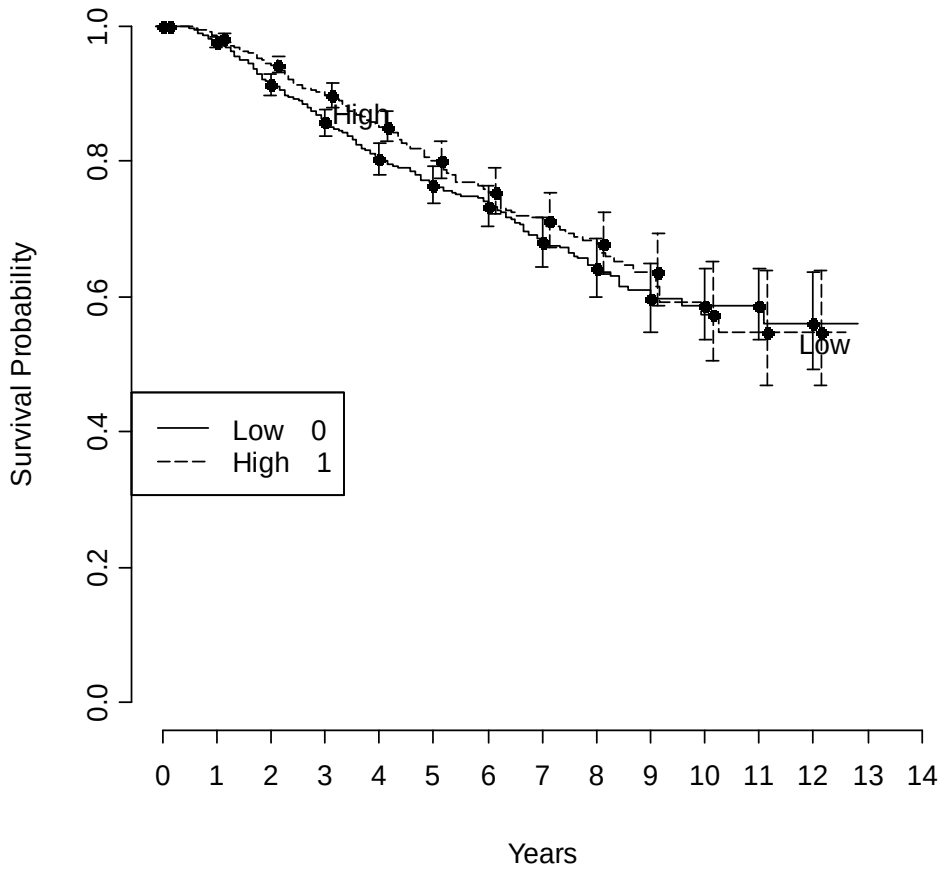


Figure 4.9.2 Kaplan-Meier survival curve – poorly differentiated prostate cancer



We make the same assumptions as in the simulation in section 4.6. We assumed that the overall high PSA rate is 0.2 among the population. The odds of obtaining PADT among patients with high PSA is assumed being 21 times the odds of obtaining PADT among patients with normal PSA. Now that $\alpha_2 = \log 21 = 3.04$, we use patients information on whether they lived in PADT usage areas as one of the predictors, and apply the first stage likelihood function on the data. We obtain that the estimated odds ratio of receiving PADT is 3.08 for patients who lived in high PADT usage areas compared to the patients who lived in low PADT usage areas. A six-

piecewise constant hazard model is fitted in the second stage, and the likelihood function (4.2.8) with $J = 6$ is applied to estimate the treatment effect. Again, as in section 4.6, the hazard ratio for patients with high PSA compared to the patients with normal PSA is assumed to be 2. The MLEs of parameters obtained from two-stage likelihood-based model are listed in Tables 4.9.2 to 4.9.5. Tables 4.9.2 and 4.9.3 give MLEs from the first stage, i.e., treatment status is regressed on PADT usage areas and PSA index. Tables 4.9.4 and 4.9.5 give MLEs from the second stage.

Table 4.9.2 Estimated parameters in the first stage of the two-stage likelihood-based IVA model, moderately differentiated localized prostate cancer

Odds of PADT at Low PADT usage areas and Normal PSA (95% CI)	
$\hat{\alpha}_0$ **(log of odds)**	$e^{\hat{\alpha}_0}$ **(odds)**
-1.909 (-1.997, -1.821)	0.148 (0.136, 0.162)
Odds Ratio of High PADT usage areas vs Low PADT usage areas (95% CI)	
$\hat{\alpha}_1$ **(log of odds ratio)**	$e^{\hat{\alpha}_1}$ **(odds ratio)**
1.377 (1.262, 1.492)	3.963 (3.532, 4.447)

Table 4.9.3 Estimated parameters in the first stage of the two-stage likelihood-based IVA model, poorly differentiated localized prostate cancer

Odds of PADT at Low PADT usage areas and Normal PSA (95% CI)	
$\hat{\alpha}_0$ **(log of odds)**	$e^{\hat{\alpha}_0}$ **(odds)**
-0.379 (-0.503, -0.254)	0.685 (0.605, 0.775)
Odds Ratio of High PADT usage areas vs Low PADT usage areas (95% CI)	
$\hat{\alpha}_1$ **(log of odds ratio)**	$e^{\hat{\alpha}_1}$ **(odds ratio)**
1.125 (0.947, 1.303)	3.08 (2.578, 3.680)

Table 4.9.4 Estimated hazards and hazard ratios in the second stage of the two-stage likelihood-based IVA model, moderately differentiated localized prostate cancer

Baseline Hazards (95% CI)	$\hat{\beta}_{0j}$ **(log of hazard)**	$e^{\hat{\beta}_{0j}}$ **(hazard)**
0-2 years	-6.276 (-7.158, -5.394)	0.002 (0.001, 0.005)
2-4 years	-5.491 (-6.347, -4.634)	0.004 (0.002, 0.010)
4-6 years	-5.481 (-6.310, -4.652)	0.004 (0.002, 0.010)
6-8 years	-5.142 (-5.948, -4.337)	0.006 (0.003, 0.013)
8-10 years	-4.691 (-5.461, -3.920)	0.009 (0.004, 0.020)
10-13 years	-5.101 (-6.010, -4.191)	0.006 (0.002, 0.015)
Hazard Ratio (95% CI)	$\hat{\beta}_1$ **(log of hazard ratio)**	$e^{\hat{\beta}_1}$ **(hazard ratio)**
PADT vs CM	1.714 (0.639, 2.790)	5.553 (1.895, 16.275)

Table 4.9.5 Estimated hazards and hazard ratios in the second stage of the two-stage likelihood-based IVA model, poorly differentiated localized prostate cancer

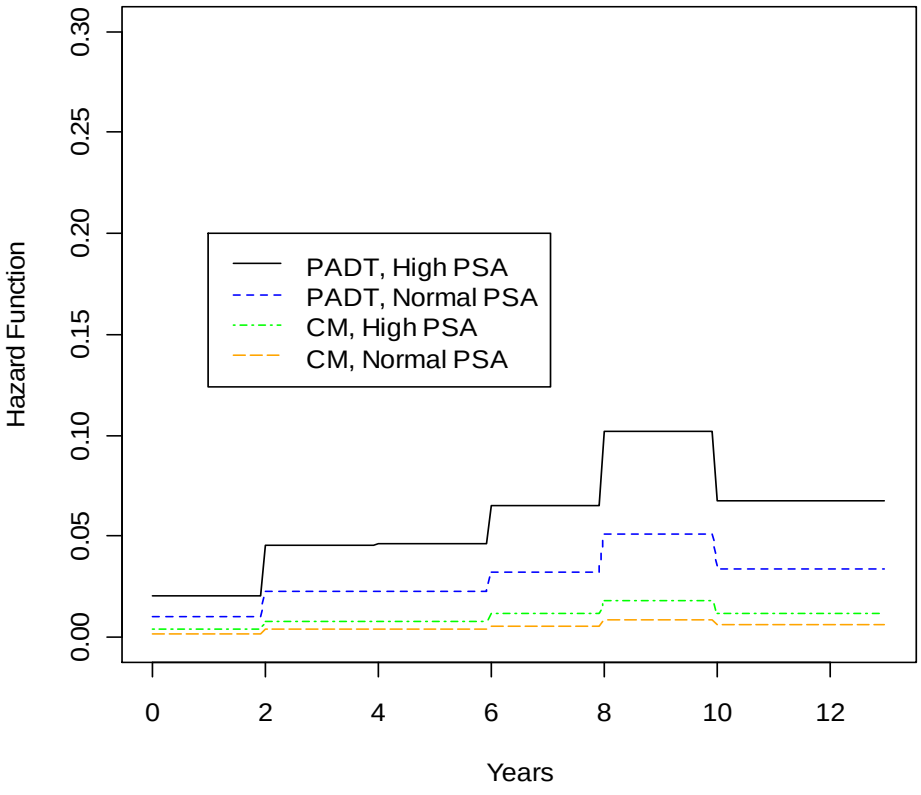
Baseline Hazards (95% CI)	$\hat{\beta}_{0j}$ **(log of hazard)**	$e^{\hat{\beta}_{0j}}$ **(hazard)**
0-2 years	-3.060 (-3.441, -2.679)	0.047 (0.032, 0.069)
2-4 years	-2.596 (-2.992, -2.201)	0.075 (0.050, 0.111)
4-6 years	-2.670 (-3.107, -2.233)	0.069 (0.045, 0.107)
6-8 years	-2.502 (-2.993, -2.012)	0.082 (0.050, 0.134)
8-10 years	-2.417 (-3.018, -1.815)	0.089 (0.049, 0.163)
10-13 years	-3.384 (-4.834, -1.935)	0.034 (0.008, 0.144)
Hazard Ratio (95% CI)	$\hat{\beta}_1$ **(log of hazard ratio)**	$e^{\hat{\beta}_1}$ **(hazard ratio)**
PADT vs CM	-0.725 (-1.418, -0.032)	0.484 (0.242, 0.968)

The results show that PADT plays significant roles in both moderately differentiated localized prostate cancer patients and poorly differentiated localized

prostate cancer patients, but in opposite directions. PADT is harmful to the patients with moderately differentiated localized prostate cancer compared to conservative management. The hazard ratio is 5.553 with 95% confidence interval of (1.895, 16.275). On the other hand, PADT benefits the patients with poorly differentiated localized prostate cancer. Compared to conservative management, the hazard ratio is 0.484 with 95% confidence interval of (0.242, 0.968).

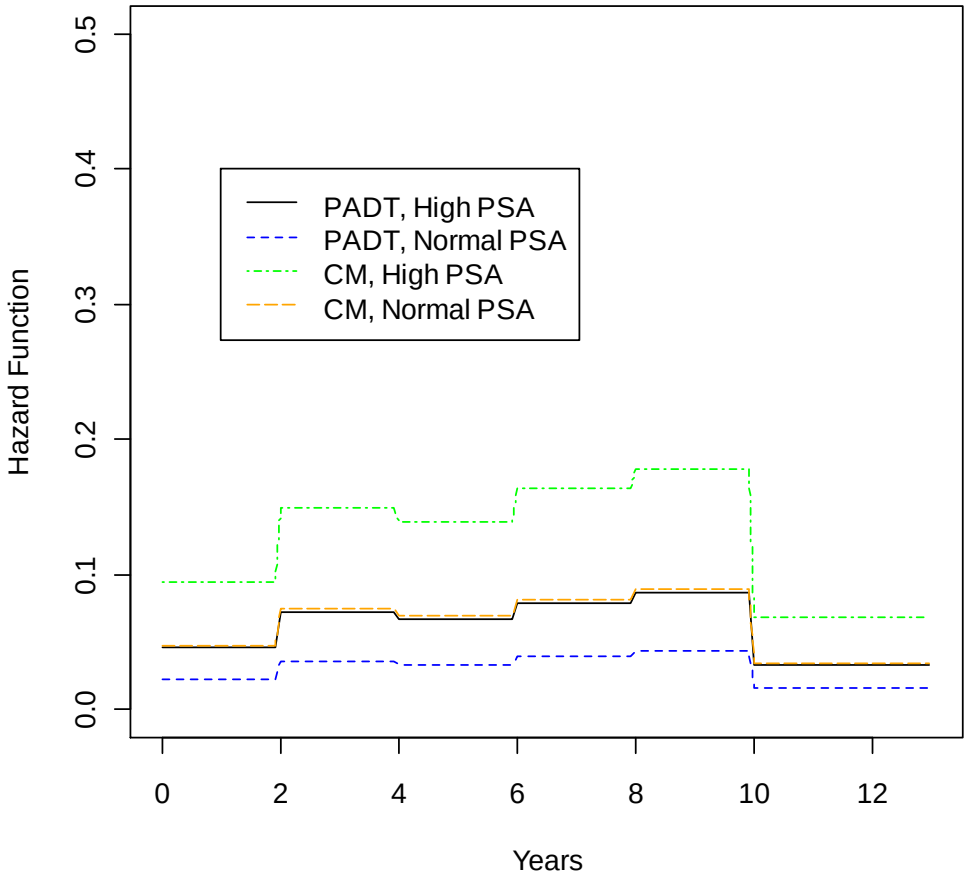
The hazard function for patients with moderately differentiated prostate cancer is drawn in Figure 4.9.3. Patients being treated with PADT and having high PSA experience the highest hazards among the four groups. Patients being treated with conservative management and having normal PSA experience the least hazards. Hazards reach the highest values for all four groups from year eight to ten.

Figure 4.9.3 Hazard function – moderately differentiated prostate cancer



The hazard function for patients with poorly differentiated prostate cancer is drawn in Figure 4.9.4. By contrast, patients receiving conservative management and having high PSA experience the highest hazards among the four groups. Patients being treated with PADT and having normal PSA experience the least hazards. Hazards are relatively high for all four groups between year two and year ten compared to years before two or after ten.

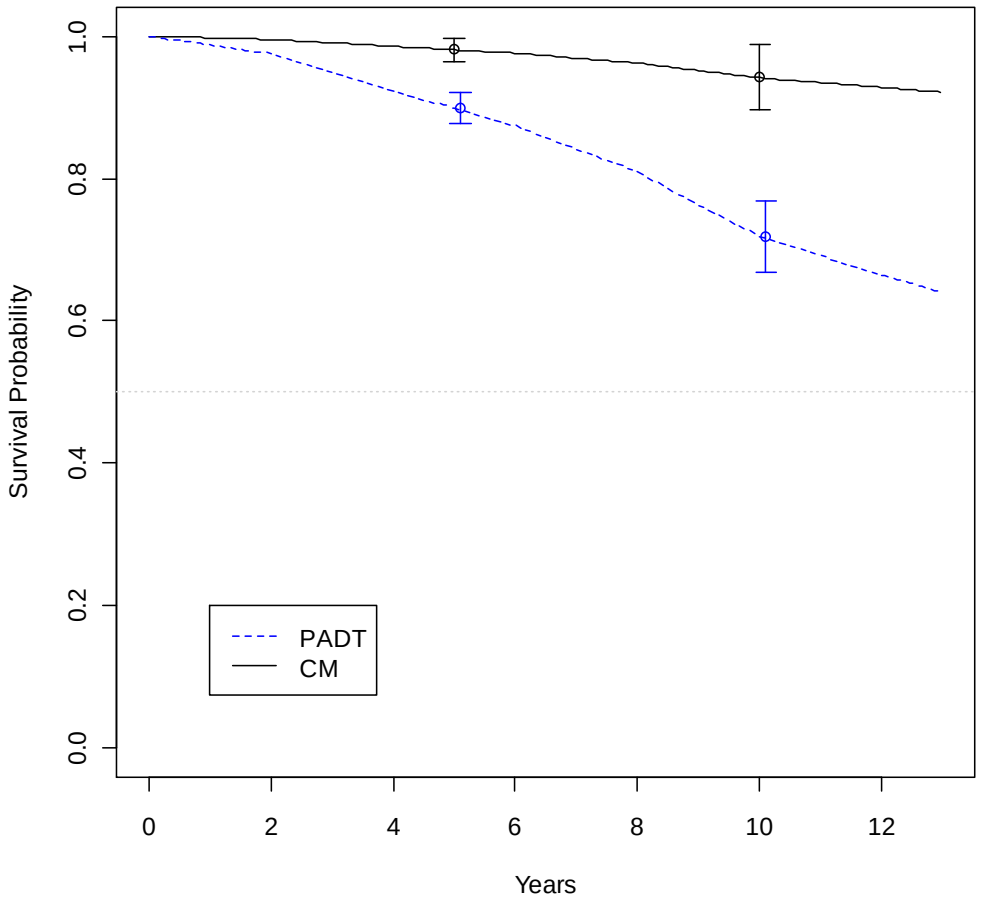
Figure 4.9.4 Hazard function – poorly differentiated prostate cancer



Estimated marginal survival functions are calculated using the maximum likelihood estimators $\hat{\underset{\sim}{\beta}}_{IV}$ in Table 4.9.4 and Table 4.9.5 assuming $\pi_0 = 0.8$ for the probability of $U = 0$ and $\pi_1 = 0.2$ for the probability of $U = 1$. They are plotted in

Figures 4.9.5 and 4.9.6. 95% confidence intervals for the 5-year and 10-year survival probabilities given in the plots are obtained from the delta method.

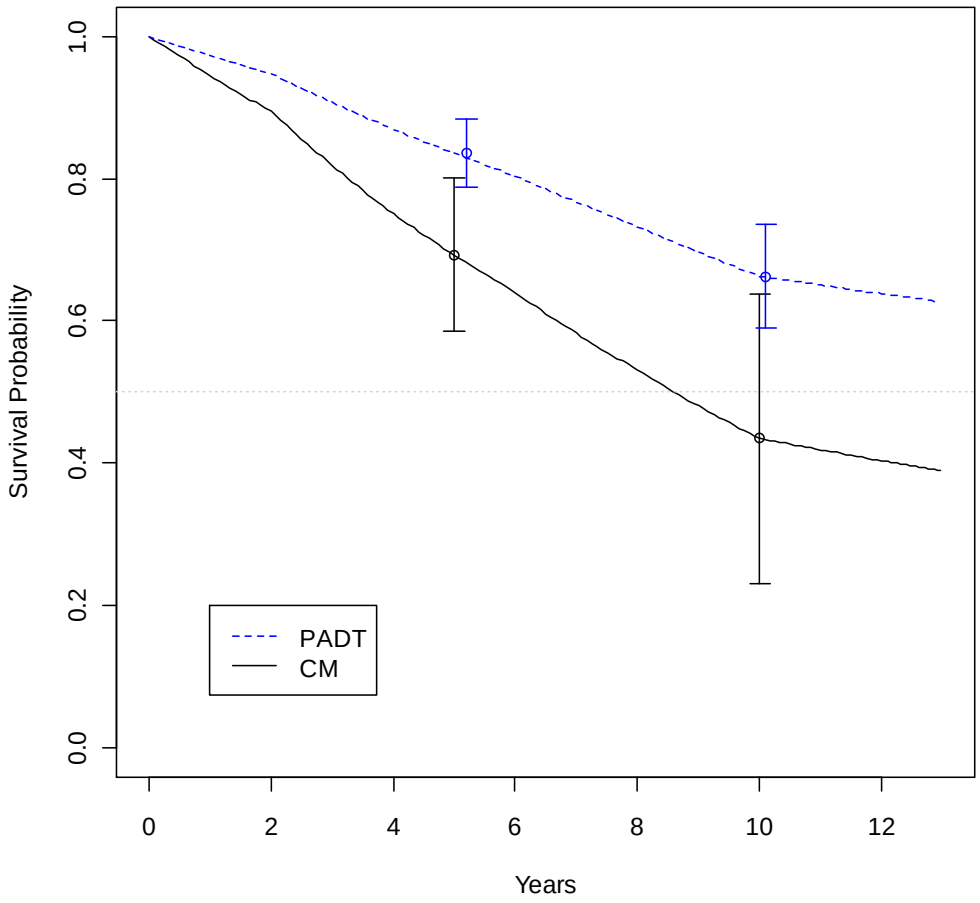
Figure 4.9.5 Survival probability – moderately differentiated prostate cancer



In general, patients with moderately differentiated prostate cancer have significantly higher survival probabilities if they receive conservative management rather than PADT. The 5-year survival rate is 98% in CM group versus 90% in PADT group. The 10-year survival rate is 94% in CM group, but only 72% in PADT group. Both 5-year and 10-year survival rates are significantly different between the two treatment groups as indicated by estimated rate ratio or rate difference and their

confidence intervals in Table 4.9.7. PADT not only increases the cost, but also increases the mortality among patients with moderately differentiated prostate cancer.

Figure 4.9.6 Survival probability – poorly differentiated prostate cancer



By contrast, the estimated marginal survival functions from patients with poorly differentiated prostate cancer show a reversed result. PADT seems to increase the survival probability among the patients with poorly differentiated prostate cancer compared to CM. The 5-year survival rate is 84% in PADT group versus 69% in CM group. The 10-year survival rate is 66% in PADT group versus 44% in CM group. However, the 95% confidence intervals for estimated survival rate ratio or estimated

survival rate difference do not support the conclusion of significant differences between the two treatment groups in 5-year or 10-year survival rates.

Table 4.9.6 Estimated survival probability with two-stage likelihood-based IVA model using marginal survival function

	PADT	**CM**
Moderately differentiated localized prostate cancer		
5-year survival probability (95% CI)	0.898 (0.876, 0.920)	0.981 (0.965, 0.997)
10-year survival probability (95% CI)	0.719 (0.669, 0.769)	0.942 (0.896, 0.987)
Poorly differentiated localized prostate cancer		
5-year survival probability (95% CI)	0.836 (0.788, 0.883)	0.693 (0.585, 0.800)
10-year survival probability (95% CI)	0.663 (0.589, 0.736)	0.435 (0.231, 0.639)

Table 4.9.7 Estimated survival rate ratio and rate difference between PADT and CM using marginal survival function

	Rate Ratio (PADT vs CM)	**Rate Difference (PADT vs CM)**
Moderately differentiated localized prostate cancer		
5-year survival probability (95% CI)	0.916 (0.893, 0.940)	-0.082 (-0.105, -0.060)
10-year survival probability (95% CI)	0.764 (0.662, 0.881)	-0.223 (-0.310, -0.135)
Poorly differentiated localized prostate cancer		
5-year survival probability (95% CI)	1.206 (0.985, 1.477)	0.143 (-0.012, 0.298)
10-year survival probability (95% CI)	1.523 (0.980, 2.368)	0.228 (-0.029, 0.485)

Estimates of the hazard ratio from the two-stage likelihood-based IVA model are compared to those estimates from the Wald type method. That is, we fit data using a Cox proportional hazard regression model with the PADT usage areas as a predictor, and then, adjust the coefficient for the percentage of compliers. A Cox proportional hazard regression model including only the treatment is investigated as well. As expected, this model gives biased estimates of hazard ratio. Estimates of the hazard ratio from the two-stage likelihood-based IVA method using 6-piecewise constant hazard model in the second stage are consistent with the estimators from the Wald type method (Table 4.9.8).

Table 4.9.8 Estimated hazard ratios from comparative models

	Moderately differentiated localized prostate cancer		**Poorly differentiated localized prostate cancer**	
PADT vs CM	$\hat{\beta}_1$ **(SE)**	**Hazard Ratio (95% CI)**	$\hat{\beta}_1$ **(SE)**	**Hazard Ratio (95% CI)**
Two-stage likelihood-based IVA	1.71 (0.55)	5.553 (1.895, 16.275)	-0.72 (0.35)	0.484 (0.242, 0.968)
Wald method	1.68 (0.38)	5.345 (2.540, 11.248)	-0.79 (0.38)	0.454 (0.214, 0.964)
Cox regression model on treatment only	0.90 (0.09)	2.467 (2.088, 2.915)	0.10 (0.09)	1.103 (0.927, 1.313)

4.10 Sensitivity analysis

So far, the results from the two-stage likelihood-based IVA model are based on the assumption that hazard ratio of high PSA versus normal PSA is 2. We examine the sensitivity of this assumption on the estimation of treatment effect. Using data from patients with poorly differentiated prostate cancer, we estimate PADT effect

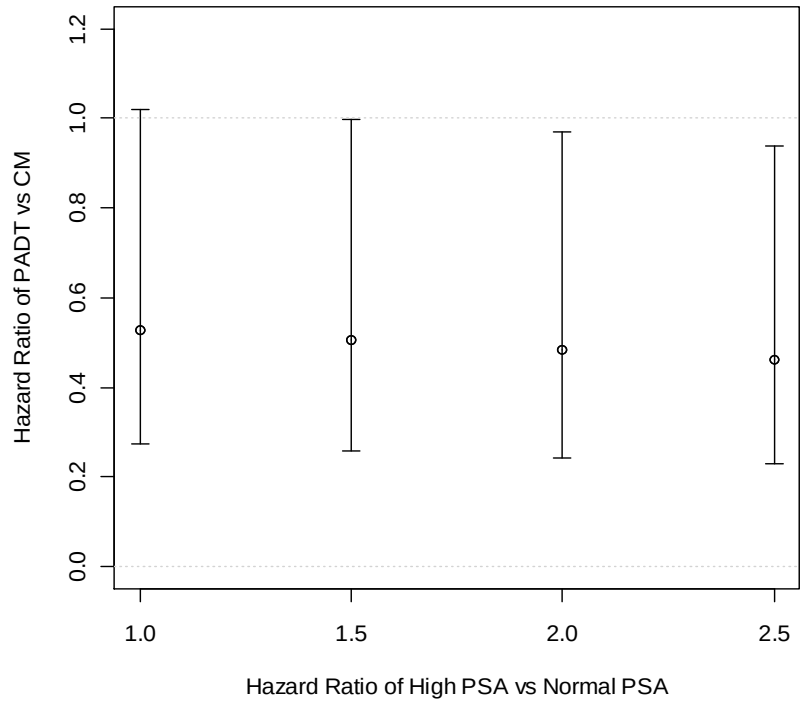
when the hazard ratio of high PSA versus normal PSA is 1, 1.5, 2, or 2.5. The hazard ratios of PADT versus CM as well as their 95% confidence intervals are given in Table 4.10.1, and plotted in Figure 4.10.1. When the hazard ratio of high PSA versus normal PSA is 1, it implies that the level of PSA has no impact on the survival outcome. Under this assumption, the survival benefit from PADT is no longer significant. For other values of 1.5, 2, or 2.5, the treatment effect does not have large fluctuations.

A sensitivity analysis on the values of α_2, which defines the relationship between treatment received and PSA values, can be done in the same way. We will conduct this analysis in our future research.

Table 4.10.1 Treatment effect vs PSA effect

Poorly differentiated localized prostate cancer	
Hazard ratio of high PSA vs normal PSA	Hazard ratio of PADT vs CM (95% CI)
1	0.529 (0.274, 1.020)
1.5	0.507 (0.258,0.997)
2	0.484 (0.242, 0.968)
2.5	0.463 (0.228, 0.939)

Figure 4.10.1 Treatment effect vs PSA effect



4.11 Discussion

It has been repeatedly demonstrated that the results from the two-stage likelihood-based IVA and the Wald method are highly positively correlated. The Wald method provides unbiased causal effect estimator only for an identity link function or a log link function in GLM. For a logit link function in GLM or a Cox proportional hazard model (Gail, 1986), Wald's estimator is biased. In the two-stage likelihood-based IVA model, MLEs are obtained to estimate the causal effect. Therefore, both stages can be any forms of the nonlinear equations. The likelihood-based estimates allow for explicitly accounting for the unknown confounding variable, and permit sensitivity analyses of key assumptions. The method can be extended to accommodate interaction terms of the confounder with the treatment, outcome, or both. In principle, the method also will allow for more complex models

of the effect of an instrument on outcome, such as accommodating continuous and/or multiple instrumental variables.

There are a few disadvantages for the two-stage likelihood-based IVA model. In order to obtain the MLEs of the treatment effect, we have to make assumptions on the values of α_2, β_2, and the distribution of the unobserved confounder. In our future research, we wish to find a better method to weaken these assumptions.

In our examples of the two-stage likelihood-based IVA model, we didn't make any adjustment for the covariates such as age, gender, or race. Accommodating other covariates in the model could be complicated and may change the values of α_2 or β_2 in the assumptions. Further exploratory analyses will be needed in our future research.

Another disadvantage of the two-stage likelihood-based IVA model is that the convergence of the model takes considerable amount of time. We will explore more efficient computational methods in our future research.

Chapter 5

Optimal Sample Size for Subunit of an Instrument

One of the advantages of two-stage likelihood-based IVA is that the instrumental variables can be in any form, continuous, binary, or categorical. However, in Rubin's causal model, the instrumental variables are strictly binary. Many instrumental variables are binary in nature, such as, water supply company in the epidemic of cholera in London, treatment assignment in clinical trial, and draft lottery status in the Vietnam War. When the instrumental variable is not binary, we have to dichotomize it. This raises issues about how best to define the two subgroups, which we address in this chapter.

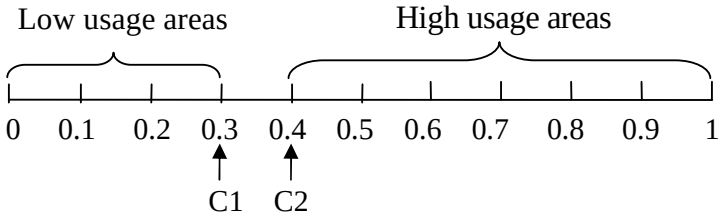
5.1 Defining binary instrument values

In the study of PADT among men with localized prostate cancer, the PADT high and low usage areas are arbitrarily defined based on the percentage use of PADT among target patients within each health service area. Percentage use of PADT is a continuous variable. If these percentages accurately reflect the local medical practice convention, we are able to follow the steps below to dichotomize the continuous variable into a binary instrumental variable.

Step 1: Define the scope of a health service area. A state, a single zip code area, or a large hospital can all be treated as a single health service area. A well defined health service area should contain enough sample size and have a homogenous usage rate of PADT among doctors.

Step 2: Calculate the percentage use of PADT among target patients for each health service area.

Step 3: Select reasonable cut-off points and define the ranges of the percentage for high PADT usage areas and low PADT usage areas. For example, in the diagram below, health service areas with less than or equal to 30% of the patients receiving PADT are defined as low usage areas. Health service areas with greater than or equal to 40% of the patients receiving PADT are defined as high usage areas. C1 and C2 are the cut-off points.

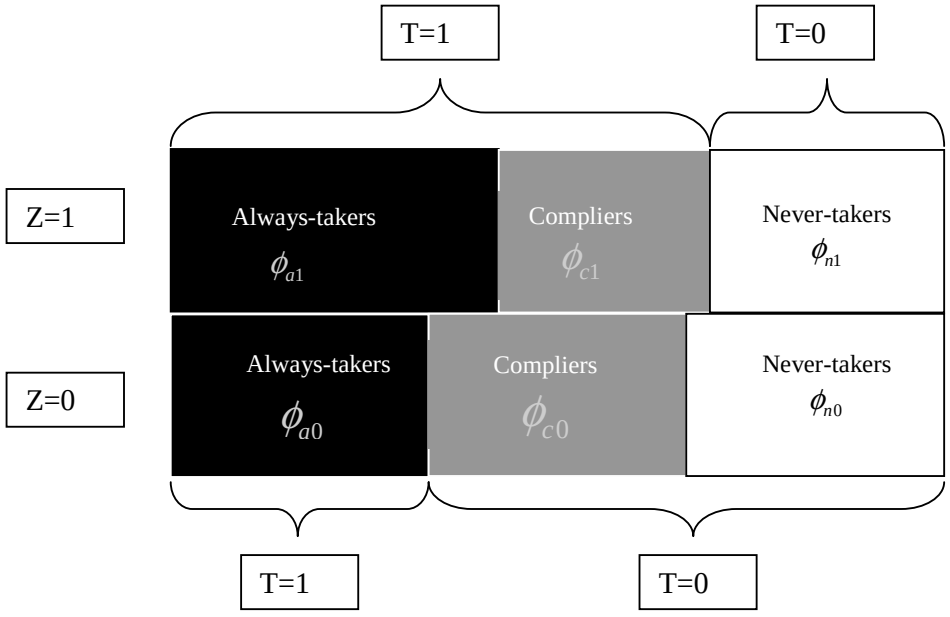


5.2 Examine assumption of random assignment

When the sample size of a health service area is not large enough, the percentage of PADT use in that health service area may not reflect the actual medical practice convention, since it can be affected by other factors such as PSA. We have mentioned in section 3.6 that the PSA screening test result confounds PADT usage. High PSA values cause high percentage use of PADT in both high PADT usage health service areas and low PADT usage health service areas. We assume that the results of PSA test are evenly distributed across geographic regions with a binomial distribution. The mean of the distribution is 0.2. When the sample size is large enough for any single health service area as mentioned at step 1 in section 5.1, we are

expecting the same distribution of the PSA across health service areas. However, when the health service areas are small and the number of patients is low, the sample distribution of the PSA results can be positively correlated with the observed PADT usage rates from those small health service areas. It turns out that the observed differentiated probabilities of PADT usage across the health service areas are not only caused by their geographic regions but also by the differentiated proportions of patients having high values on the PSA test, a violation of the assumption of the independence between instrumental variable and unobserved confounder. In terms of the Rubin causal model, the second assumption of random assignment is violated. Diagram 5.2.1 illustrates that the principal strata are no longer comparable due to the violation of random assignment. This results in biased treatment effect estimates.

Diagram 5.2.1



In order to estimate the bias caused by the positive correlation between PSA results and PADT usage rates, we make a few assumptions to run a simple demonstration.

Assumption 1: All health service areas have the same sample size. Let n denote the total number of localized prostate cancer patients in a single health service area.

Assumption 2: All patients with high PSA values receive PADT.

Assumption 3: For the simplicity of current calculation, the 10-year mortality of patients with high PSA values is assumed to be 100%. We may use 90% of the mortality for future work.

The outcome distribution of Y_i is a mixture distribution from 3 types of patients, patients with high PSA values, patients with normal PSA values and receiving PADT, and patients with normal PSA values and not receiving non-PADT .

$$f_{Z=1}(y) = P_{PSA=1,T=1,Z=1} f_{PSA=1}(y) + P_{PSA=0,T=1,Z=1} f_{PSA=0,T=1}(y) + P_{PSA=0,T=0,Z=1} f_{PSA=0,T=0}(y) \qquad (5.2.1)$$

$$f_{Z=0}(y) = P_{PSA=1,T=1,Z=0} f_{PSA=1}(y) + P_{PSA=0,T=1,Z=0} f_{PSA=0,T=1}(y) + P_{PSA=0,T=0,Z=0} f_{PSA=0,T=0}(y) \qquad (5.2.2)$$

where P represents the percentage of patients in a health service area. For example, $P_{PSA=0,T=1,Z=1}$ represents the percentage of patients in a high PADT usage area with normal PSA values and receiving PADT. The constraints of the equations (5.2.1) and (5.2.2) are:

$$P_{PSA=1,T=1,Z=1} + P_{PSA=0,T=1,Z=1} + P_{PSA=0,T=0,Z=1} = 1 \quad \text{and}$$

$$P_{PSA=1,T=1,Z=0} + P_{PSA=0,T=1,Z=0} + P_{PSA=0,T=0,Z=0} = 1$$

Equations 5.2.1 and 5.2.2 are analogous to equations (3.2.1) and (3.2.2) in section 3.2. By using the mixture distribution, the sample average 10-year mortalities could be obtained from equations (5.2.3) and (5.2.4).

$$\bar{Y}_{Z=1} = \hat{P}_{PSA=1,T=1,Z=1} \cdot \bar{Y}_{PSA=1} + \hat{P}_{PSA=0,T=1,Z=1} \cdot \bar{Y}_{PSA=0,T=1} + \hat{P}_{PSA=0,T=0,Z=1} \cdot \bar{Y}_{PSA=0,T=0} \qquad (5.2.3)$$

$$\bar{Y}_{Z=0} = \hat{P}_{PSA=1,T=1,Z=0} \cdot \bar{Y}_{PSA=1} + \hat{P}_{PSA=0,T=1,Z=0} \cdot \bar{Y}_{PSA=0,T=1} + \hat{P}_{PSA=0,T=0,Z=0} \cdot \bar{Y}_{PSA=0,T=0} \quad (5.2.4)$$

Therefore,

$$\begin{aligned} &\bar{Y}_{Z=1} - \bar{Y}_{Z=0} \\ &= \left[\left(\hat{P}_{PSA=1,T=1,Z=1} - \hat{P}_{PSA=1,T=1,Z=0} \right) \cdot \bar{Y}_{PSA=1} + \left(\hat{P}_{PSA=0,T=1,Z=1} - \hat{P}_{PSA=0,T=1,Z=0} \right) \cdot \bar{Y}_{PSA=0,T=1} \right. \\ &\quad \left. + \left(\hat{P}_{PSA=0,T=0,Z=1} - \hat{P}_{PSA=0,T=0,Z=0} \right) \cdot \bar{Y}_{PSA=0,T=0} \right] \end{aligned} \quad (5.2.5)$$

where $\hat{P}_{PSA=1,T=1,Z=1}$, $\hat{P}_{PSA=1,T=1,Z=0}$, $\hat{P}_{PSA=0,T=1,Z=1}$, $\hat{P}_{PSA=0,T=1,Z=0}$, $\hat{P}_{PSA=0,T=0,Z=1}$, and $\hat{P}_{PSA=0,T=0,Z=0}$ are estimators of the true percentages .

For a single health service area of size n, let U be the total number of patients with high PSA values and who therefore receive PADT, and let V be the total number of patients with low PSA values and thus receive PADT by chance. The sum of the two, S, is the total number of patients who receive PADT within the health service area.

$$S = U + V$$

Assumption 4: The total number of patients who receive PADT within a health service area, S, follows a binomial distribution with a mean of $n \cdot p_{T=1}$, and a variance of $n \cdot p_{T=1} \cdot (1 - p_{T=1})$. Furthermore, $p_{T=1}$ is assumed to follow a beta distribution with a mean of $\frac{\alpha}{\alpha+\beta}$, and a variance of $\frac{\alpha \cdot \beta}{(\alpha+\beta+1) \cdot (\alpha+\beta)^2}$.

$f_{T=1}(S) \sim binomial(S; n, p_{T=1})$ where $f_{T=1}(p_{T=1}) \sim beta(p_{T=1}; \alpha, \beta)$

$$\begin{aligned} f_{T=1}(S, p_{T=1}) &= f_{T=1}(S \mid p_{T=1}) \cdot f_{T=1}(p_{T=1}) \\ &= binomial(S; n, p_{T=1}) \cdot beta(p_{T=1}; \alpha, \beta) \end{aligned} \quad (5.2.6)$$

Assuming the overall usage rate of PADT for a high PADT usage health service area, $P_{T=1,Z=1}$, is greater than or equal to C2, the total number of patients who receive PADT in that high PADT usage area is distributed as:

$$f_{T=1,Z=1}(S)=\frac{\int_{C2}^{1} f_{T=1}(S,p_{T=1})\cdot dp_{T=1}}{\sum_{S=C2\cdot n}^{n}\int_{C2}^{1} f_{T=1}(S,p_{T=1})\cdot dp_{T=1}}= \frac{\int_{C2}^{1}\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\}\cdot dp_{T=1}}{\sum_{S=C2\cdot n}^{n}\int_{C2}^{1}\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\}\cdot dp_{T=1}} \tag{5.2.7}$$

Assuming the overall usage rate of PADT for low usage health service areas, $P_{T=1,Z=0}$, is less than or equal to C1, the total number of patients who receive PADT in that low usage area is distributed as:

$$f_{T=1,Z=0}(S)=\frac{\int_{0}^{C1} f_{T=1}(S,p_{T=1})\cdot dp_{T=1}}{\sum_{S=0}^{C1\cdot n}\int_{0}^{C1} f_{T=1}(S,p_{T=1})\cdot dp_{T=1}}= \frac{\int_{0}^{C1}\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\}\cdot dp_{T=1}}{\sum_{S=0}^{C1\cdot n}\int_{0}^{C1}\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\}\cdot dp_{T=1}} \tag{5.2.8}$$

Assumption 5: The number of patients with high PSA values within a health service area follows a binomial distribution with a mean of $n\cdot p_{PSA=1}$ and a variance of $n\cdot p_{PSA=1}\cdot(1-p_{PSA=1})$, given the total number of patients, S, receiving PADT in that area.

$$f_{PSA=1,Z=1}(U|S)=\frac{binomial(U;n,p_{PSA=1})}{\sum_{U=0}^{S} binomial(U;n,p_{PSA=1})} \tag{5.2.9}$$

From assumptions 4 and 5, the joint distribution of U and S from a high PADT usage health service area is:

$$f_{T=1,Z=1}(U,S) = f_{T=1,Z=1}(S)\cdot f_{PSA=1,Z=1}(U|S)$$

$$= \frac{\int_{C2}^{1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}{\sum_{S=ceiling(C2\cdot n)}^{n}\int_{C2}^{1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}\cdot\left\{\frac{binomial(U;n,p_{PSA=1})}{\sum_{U=0}^{S}binomial(U;n,p_{PSA=1})}\right\} \quad (5.2.10)$$

The joint distribution of U and S from a low PADT usage health service area is:

$$f_{T=1,Z=0}(U,S) = f_{T=1,Z=0}(S)\cdot f_{PSA=1,Z=0}(U|S)$$

$$= \frac{\int_{0}^{C1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}{\sum_{S=0}^{floor(C1\cdot n)}\int_{0}^{C1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}\cdot\left\{\frac{binomial(U;n,p_{PSA=1})}{\sum_{U=0}^{S}binomial(U;n,p_{PSA=1})}\right\} \quad (5.2.11)$$

Using a Jacobian transformation, the joint distribution of V and S from a high PADT usage health service area is:

$$f_{T=1,Z=1}(V,S) = f_{T=1,Z=1}(S)\cdot f_{PSA=1,Z=1}(V|S)$$

$$= \frac{\int_{C2}^{1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}{\sum_{S=ceiling(C2\cdot n)}^{n}\int_{C2}^{1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}\cdot\left\{\frac{binomial(S-V;n,p_{PSA=1})}{\sum_{V=0}^{S}binomial(S-V;n,p_{PSA=1})}\right\} \quad (5.2.12)$$

The joint distribution of V and S from a low PADT usage health service area is:

$$f_{T=1,Z=1}(V,S) = f_{T=1,Z=1}(S)\cdot f_{PSA=1,Z=1}(V|S)$$

$$= \frac{\int_{0}^{C1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}{\sum_{S=0}^{floor(C1\cdot n)}\int_{0}^{C1}\left\{binomial(S;n,p_{T=1})\cdot beta(p_{T=1};\alpha,\beta)\right\}\cdot dp}\cdot\left\{\frac{binomial(S-V;n,p_{PSA=1})}{\sum_{V=0}^{S}binomial(S-V;n,p_{PSA=1})}\right\} \quad (5.2.13)$$

The expected difference is:

$$
\begin{aligned}
&E\left(\bar{Y}_{Z=1}-\bar{Y}_{Z=0}\right)\\
&=\left[E\left(\hat{P}_{PSA=1,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=1,T=1,Z=0}\right)\right]\cdot E\left(\bar{Y}_{PSA=1}\right)+\left[E\left(\hat{P}_{PSA=0,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot E\left(\bar{Y}_{PSA=0,T=1}\right)\\
&\quad+\left[E\left(\hat{P}_{PSA=0,T=0,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=0,Z=0}\right)\right]\cdot E\left(\bar{Y}_{PSA=0,T=0}\right)\\
&=\left[E\left(\hat{P}_{PSA=1,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=1,T=1,Z=0}\right)\right]\cdot E\left(\bar{Y}_{PSA=1}\right)+\left[E\left(\hat{P}_{PSA=0,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot E\left(\bar{Y}_{PSA=0,T=1}\right)\\
&\quad+\left[\left(1-E\left(\hat{P}_{PSA=1,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=1}\right)\right)-\left(1-E\left(\hat{P}_{PSA=1,T=1,Z=0}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right)\right]\cdot E\left(\bar{Y}_{PSA=0,T=0}\right)\\
&=\frac{1}{n}\left[\left(\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{U=0}^{S}U\cdot f_{T=1,Z=1}(U,S)-\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}U\cdot f_{T=1,Z=0}(U,S)\right)\cdot E\left(\bar{Y}_{PSA=1}\right)+\right.\\
&\quad\left(\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{V=0}^{S}V\cdot f_{T=1,Z=1}(V,S)-\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}V\cdot f_{T=1,Z=0}(V,S)\right)\cdot E\left(\bar{Y}_{PSA=0,T=1}\right)+\\
&\quad\left(1-\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{U=0}^{S}U\cdot f_{T=1,Z=1}(U,S)-\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{V=0}^{S}V\cdot f_{T=1,Z=1}(V,S)\right)\cdot E\left(\bar{Y}_{PSA=0,T=0}\right)-\\
&\quad\left.\left(1-\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}U\cdot f_{T=1,Z=0}(U,S)-\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}V\cdot f_{T=1,Z=0}(V,S)\right)\cdot E\left(\bar{Y}_{PSA=0,T=0}\right)\right]
\end{aligned}
\tag{5.2.14}
$$

The PADT usage rates among the target population are assumed to be independent of the expected outcomes. The variance of the two independent variables is:

$$
\begin{aligned}
&\operatorname{var}(a\cdot b)=E\left(a^2\cdot b^2\right)-\left[E(a)\cdot E(b)\right]^2\\
&=E\left(a^2\right)\cdot E\left(b^2\right)-\left[E(a)\right]^2\cdot\left[E(b)\right]^2\\
&=\left\{E\left(a^2\right)-\left[E(a)\right]^2\right\}\cdot\left\{E\left(b^2\right)-\left[E(b)\right]^2\right\}+E\left(a^2\right)\cdot\left[E(b)\right]^2+E\left(b^2\right)\cdot\left[E(a)\right]^2-2\left[E(a)\cdot E(b)\right]^2\\
&=\operatorname{var}(a)\cdot\operatorname{var}(b)+\operatorname{var}(a)\cdot\left[E(b)\right]^2+\operatorname{var}(b)\cdot\left[E(a)\right]^2
\end{aligned}
\tag{5.2.15}
$$

Applying equation (5.2.15), the variance of the expectation in difference is:

$$
\begin{aligned}
&\operatorname{var}\left(\bar{Y}_{Z=1}-\bar{Y}_{Z=0}\right)\\
&=\left[\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=0}\right)\right]\cdot\operatorname{var}\left(\bar{Y}_{PSA=1,T=1}\right)+\left[\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=0}\right)\right]\cdot\left[E\left(\bar{Y}_{PSA=1,T=1}\right)\right]^2\\
&+\left[E\left(\hat{P}_{PSA=1,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=1,T=1,Z=0}\right)\right]^2\cdot\operatorname{var}\left(\bar{Y}_{PSA=1,T=1}\right)+\\
&\left[\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot\operatorname{var}\left(\bar{Y}_{PSA=0,T=1}\right)+\left[\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot\left[E\left(\bar{Y}_{PSA=0,T=1}\right)\right]^2\\
&+\left[E\left(\hat{P}_{PSA=0,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]^2\cdot\operatorname{var}\left(\bar{Y}_{PSA=0,T=1}\right)+\\
&\left[\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=0}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot\operatorname{var}\left(\bar{Y}_{PSA=0,T=0}\right)\\
&+\left[\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=0}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=1}\right)+\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right]\cdot\left[E\left(\bar{Y}_{PSA=0,T=0}\right)\right]^2\\
&+\left[\left(1-E\left(\hat{P}_{PSA=1,T=1,Z=1}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=1}\right)\right)-\left(1-E\left(\hat{P}_{PSA=1,T=1,Z=0}\right)-E\left(\hat{P}_{PSA=0,T=1,Z=0}\right)\right)\right]^2\cdot\operatorname{var}\left(\bar{Y}_{PSA=0,T=0}\right)
\end{aligned}
\tag{5.2.16}
$$

where

$$
\begin{aligned}
\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=1}\right)&=\frac{1}{n^2}\cdot\left\{\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{U=0}^{S}U^2\cdot f_{T=1,Z=1}(U,S)-\left(\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{U=0}^{S}U\cdot f_{T=1,Z=1}(U,S)\right)^2\right\}\\
\operatorname{var}\left(\hat{P}_{PSA=1,T=1,Z=0}\right)&=\frac{1}{n^2}\cdot\left\{\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}U^2\cdot f_{T=1,Z=0}(U,S)-\left(\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}U\cdot f_{T=1,Z=0}(U,S)\right)^2\right\}\\
\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=1}\right)&=\frac{1}{n^2}\cdot\left\{\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{V=0}^{S}V^2\cdot f_{T=1,Z=1}(V,S)-\left(\sum_{S=ceiling(C2\cdot n)}^{n}\sum_{V=0}^{S}V\cdot f_{T=1,Z=1}(V,S)\right)^2\right\}\\
\operatorname{var}\left(\hat{P}_{PSA=0,T=1,Z=0}\right)&=\frac{1}{n^2}\cdot\left\{\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}V^2\cdot f_{T=1,Z=0}(V,S)-\left(\sum_{S=0}^{floor(C1\cdot n)}\sum_{U=0}^{S}V\cdot f_{T=1,Z=0}(V,S)\right)^2\right\}
\end{aligned}
\tag{5.2.17}
$$

$$
\hat{\beta}_{IV}=\frac{\bar{Y}_{Z=1}-\bar{Y}_{Z=0}}{\left[\left(\hat{P}_{PSA=1,T=1,Z=1}+\hat{P}_{PSA=0,T=1,Z=1}\right)-\left(\hat{P}_{PSA=1,T=1,Z=0}+\hat{P}_{PSA=0,T=1,Z=0}\right)\right]}
\tag{5.2.18}
$$

When the sample size gets large enough, the expected IV estimand can be approximated by:

$$
E\left(\hat{\beta}_{IV}\right)\approx\frac{E\left(\bar{Y}_{Z=1}-\bar{Y}_{Z=0}\right)}{E\left[\left(\hat{P}_{PSA=1,T=1,Z=1}+\hat{P}_{PSA=0,T=1,Z=1}\right)-\left(\hat{P}_{PSA=1,T=1,Z=0}+\hat{P}_{PSA=0,T=1,Z=0}\right)\right]}
\tag{5.2.19}
$$

However, there is no closed form solution for the variance of the IV estimand. Simulation will be used in the future to estimate the variance of the IV estimand.

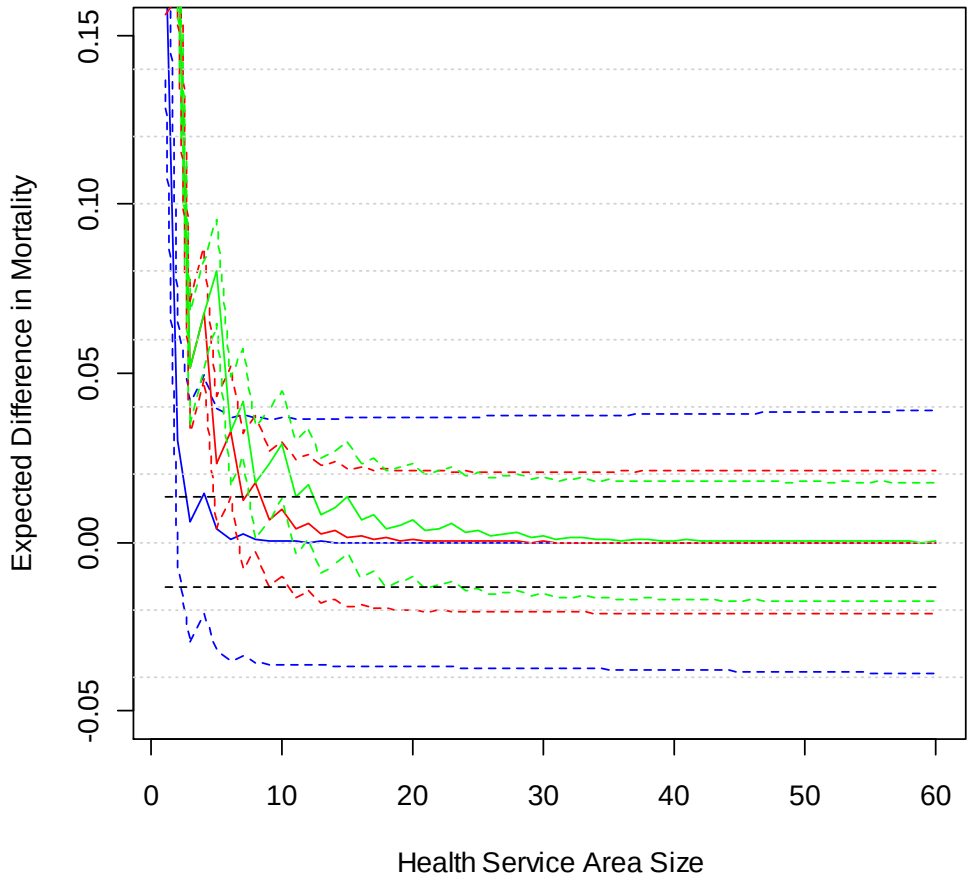
In Figure 5.2.1, the bias of the mean difference in 10-year mortality between PADT and CM is examined when there is truly no difference between the two groups. Our estimation is based on a total of 10,000 patients. The number of patients within each health service area receiving PADT is assumed to have a binomial distribution with a mean which follows a beta distribution, beta(2,3). The 10-year mortality for patients with normal PSA values is assumed to be 0.1 ignoring their treatment status. Patients with high PSA values receive PADT, and their 10-year mortality is assumed to be 100%. The expected differences between PADT and CM patients in 10-year mortality and their respective 95% confidence intervals are plotted against the single health service area size in Figure 5.2.1. The PADT usage rates from each health service area are sorted from the lowest to the highest. The plot in blue is from the top one tenth high usage areas and the bottom one tenth low usage areas. The rest of the four fifths health service areas with the PADT usage rates between the highest and lowest usage areas are discarded. The plot in red is from the top one third high usage areas and the bottom one third low usage areas. The rest of the one third middle areas are discarded. The plot in green is from the top one half high usage areas and the bottom one half low usage areas, i.e., all health service areas are included in our analysis. The two black dashed lines outline the 95% confidence interval for the true difference of 0 without using instrumental variable analysis, i.e.,

$$\text{var}\left(\bar{Y}_{T=1}-\bar{Y}_{T=0}\right)=\text{var}\left(\bar{Y}_{T=1}\right)+\text{var}\left(\bar{Y}_{T=0}\right),\ \text{and}\quad 95\%CI=0\pm 1.96\cdot SE\left(\bar{Y}_{T=1}-\bar{Y}_{T=0}\right).$$

In Figure 5.2.1, it is obvious that when all service areas are used in the analysis, a small size of the single health service area can cause the largest bias in treatment effect estimation. On the other hand, when service areas falling in the

middle quantiles are discarded, the variance of the estimator becomes very large because of the reduction in overall sample size. An example is as the one that only health service areas from one tenth of the top and bottom quantiles are used. When we use the top and bottom one third of the data, the estimator of the difference becomes consistent after single health service area reaches a size of 30, and the variance isn't enlarged much compared to the one without discarding any of health services areas. If information of PSA is available for all patients, direct comparison between groups of PADT and CM would be feasible, and the 95% confidence interval would the narrowest as shown in black dashed lines. It is apparent that the IVA method is less precise.

Figure 5.2.1 Estimated mean difference when no effect of PADT in mortality compared to CM



Note: 1. Data used in quantiles: blue 2/10; red 2/3; green 1.
2. Solid lines: expected mean difference; Dashed lines: 95% CI

Chapter 6

Future Research

6.1 Exploration in the two-stage likelihood-based IVA model

In order to apply our two-stage likelihood-based IVA model, we need to make assumptions about the distribution of the unobserved confounder, the association between the treatment status and the unobserved confounder at stage one, and the association between the outcome and the unobserved confounder at stage two. In this thesis, we used information from external sources to establish these assumptions, and we conducted a sensitivity analysis on β_2 to assess the impact of changes in the assumptions on the final results. We wish to improve this model by weakening these assumptions.

Besides the sensitivity analysis on β_2, we plan to conduct more sensitivity analyses on the values of α_2, and the mean of confounder μ_U. We wish to test the validity of these assumptions and examine the effects of these assumptions cause.

We also plan to develop ways to add covariates such as age, cancer stage, race, marital status, and economic indicators in the two-stage likelihood-based IVA model in addition to the PSA. Since the model takes a day to converge for a six-piecewise constant hazard model at stage two, we will have to explore more efficient computational methods.

6.2 Instrumental variable analysis with clustered data

In the prostate cancer example, health service areas vary widely in size. For example, Detroit and Los Angeles are very large as compared to others. Therefore, they could have a dominant effect on the IVA if each patient is treated as an independent observation. How can we reduce the weight of these large areas? One way could be to use a linear mixed model or a generalized linear mixed model with "area" as the clustering factor, that is to treat area as a random effect. For survival outcomes, those are known as frailty models. Patients from the same health service area are assumed to be correlated with respect to treatment and outcome. Patients from different health service areas are independent of each other with respect to treatment and outcome. In linear models, the two-stage instrumental variable analysis can be expressed as:

$$T_{ij} = \alpha_0 + \alpha_1 \cdot Z_i + \alpha_2 \cdot U_{ij} + C_j + \upsilon_{ij} \tag{6.2.1}$$

$$Y_{ij} = \beta_0 + \beta_1 \cdot T_{ij} + \beta_2 \cdot U_{ij} + C_j + \varepsilon_{ij} \tag{6.2.2}$$

where j represents the jth health area, and i represents the ith subject within the jth health area. C_j represents the random effect from the health areas and follows a normal distribution with a mean 0 and a variance τ^2. υ_{ij} is distributed as normal $\left(0, \omega^2\right)$, and ε_{ij} is distributed as normal $\left(0, \sigma^2\right)$. The cluster data structure is found in Table 6.2.1.

Table 6.2.1 Cluster data structure

Cluster	Subject	Instrumental Variable	Confounder	Treatment	Response
1	1	Z_{11}	U_{11}	T_{11}	Y_{11}
1	2	Z_{21}	U_{21}	T_{21}	Y_{21}
⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	$Z_{n_1 1}$	$U_{n_1 1}$	$T_{n_1 1}$	$Y_{n_1 1}$
2	1	Z_{12}	U_{12}	T_{12}	Y_{12}
⋮	⋮	⋮	⋮	⋮	⋮
2	n_2	$Z_{n_2 2}$	$U_{n_2 2}$	$T_{n_2 2}$	$Y_{n_2 2}$
⋮	⋮	⋮	⋮	⋮	⋮
$j-1$	1	$Z_{1\,j-1}$	$U_{1\,j-1}$	$T_{1\,j-1}$	$Y_{1\,j-1}$
⋮	⋮	⋮	⋮	⋮	⋮
$j-1$	n_{j-1}	$Z_{n_{j-1} j-1}$	$U_{n_{j-1} j-1}$	$T_{n_{j-1} j-1}$	$Y_{n_{j-1} j-1}$
j	1	Z_{1j}	U_{1j}	T_{1j}	Y_{1j}
⋮	⋮	⋮	⋮	⋮	⋮
j	n_j	$Z_{n_j j}$	$U_{n_j j}$	$T_{n_j j}$	$Y_{n_j j}$

The correlation matrix for outcome is:

$$corr(Y)=\begin{bmatrix} 1 & \cdots & \rho_Y & & & & & & \\ \vdots & \ddots & \vdots & & 0 & & & 0 & \\ \rho_Y & \cdots & 1 & & & & & & \\ & & & 1 & \cdots & \rho_Y & & & \\ & 0 & & \vdots & \ddots & \vdots & & 0 & \\ & & & \rho_Y & \cdots & 1 & & & \\ & & & & & & 1 & \cdots & \rho_Y \\ & 0 & & & 0 & & \vdots & \ddots & \vdots \\ & & & & & & \rho_Y & \cdots & 1 \end{bmatrix}$$

Let $\bar{Y}_j$ be the sample mean of outcome from health area j.

$\bar{Y}_j = \hat{p}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{n}$, where n is the total number of patients in health area j.

$$\mathrm{var}\left(\hat{p}_j\right) = \frac{n-1}{n} \cdot \rho_Y \cdot \hat{p}_j \cdot \left(1-\hat{p}_j\right) + \frac{\hat{p}_j \cdot \left(1-\hat{p}_j\right)}{n} \tag{6.2.3}$$

After the random effect from health areas are taken into account, the overall mean from all health areas is calculated as a weighted mean of each health area. The weight is $\frac{1}{\mathrm{var}\left(\hat{p}_j\right)}$. We have

$$\bar{Y}_{ij} = \sum_{j=1}^{J} \frac{\hat{p}_j}{\mathrm{var}\left(\hat{p}_j\right)} = \sum_{j=1}^{J} \frac{1}{\frac{n-1}{n} \cdot \rho_Y \cdot \left(1-\hat{p}_j\right) + \frac{1}{n} \cdot \left(1-\hat{p}_j\right)} = \sum_{j=1}^{J} \frac{1}{\rho_Y \cdot \left(1-\hat{p}_j\right) + \frac{1}{n} \cdot \left(1-\hat{p}_j\right) \cdot \left(1-\rho_Y\right)}, \tag{6.2.4}$$

where J is the total number of health areas. When n is small, $\mathrm{var}\left(\hat{p}_j\right)$ tends to be large, and when n is large, $\mathrm{var}\left(\hat{p}_j\right)$ tends to be small. Therefore, larger health service areas put more weight on the overall mean of outcome. However, when the size of a health service area becomes extremely large, the weight approaches a constant, so that even very large clusters do not dominate the estimate of the mean.

$$\lim_{n\to\infty} \mathrm{var}\left(\hat{p}_j\right) = \rho_Y \cdot \hat{p}_j \cdot \left(1-\hat{p}_j\right) \tag{6.2.5}$$

The correlation matrix for treatment status is:

$$corr\left(T\right) = \begin{bmatrix} 1 & \cdots & \rho_T & & & & & & \\ \vdots & \ddots & \vdots & & 0 & & & 0 & \\ \rho_T & \cdots & 1 & & & & & & \\ & & & 1 & \cdots & \rho_T & & & \\ & 0 & & \vdots & \ddots & \vdots & & 0 & \\ & & & \rho_T & \cdots & 1 & & & \\ & & & & & & 1 & \cdots & \rho_T \\ & 0 & & & 0 & & \vdots & \ddots & \vdots \\ & & & & & & \rho_T & \cdots & 1 \end{bmatrix}$$

The potential effect of clustering at stage one of the two-stage likelihood-based IVA model needs further study.

References

Angrist, Joshua D., Imbens, Guido W., and Rubin, Donald B. (1996): Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91: 444-454.*

Cutting Edge Information (2006): Clinical operations: accelerating trials, allocating resources and measuring performance. *http://www.lifesciencesworld.com/news/view*. Accessed on June 1, 2009.

Dawid, A. Philip (2003): Causal inference using influence diagrams: the problem of partial compliance. *Highly Structured Stochastic Systems, Oxford University Press:45-81.*

Dehejia, Rajeev H., and Wahba, Sadek (2002): Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84(1):151-161.*

Didelez, Vanessa, and Sheehan, Nuala (2007): Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research, 16:309-330.*

Didelez, Vanessa, Meng, Sha, and Sheehan, Nuala (2010): Assumptions of IV methods for observational epidemiology. *Statistical Science, 25(1):22-40.*

Dunn, Graham, Maracy, Mohammad, and Tomenson, Barbara (2005): Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research, 14:369-395.*

Durbin, James (1954): Errors in variables. *Review of the International Statistical Institute, 22:23-32.*

Earle, Craig C., Tsai, Jerry S., Gelber, Richard D., Weinstein, Milton C., Neumann, Peter J., and Weeks, Jane C. (2001): Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. *Journal of Clinical Oncology, 19(4):1064-1070.*

Foster, Michael E. (1997): Instrumental variables for logistic regression: an illustration. *Social Science Research, 26:487-504.*

Gail, Mitchell H. (1986): Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. *Modern Statistical Methods in Chronic Disease Epidemiology (edited by Moolgavkar, H. S. and Prentice, L. R.), 3-18,* Wiley, New York.

Gail, Mitchell H., Wieand, Sam, and Piantadosi, S. (1984): Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71:431-444.*

Goodman, Melody S., Li, Yi, and Tiwari, Ram C. (2011): Detecting Multiple Change Points in Piecewise Constant Hazard Functions. *Journal of Applied Statistics, First article.* Published by Taylor & Francis.

Greenland, Sander, Robins, James M., and Pearl, Judea (1999): Confounding and Collapsibility in Causal Inference. *Statistical Science, 14(1):29-46.*

Greenland, Sander (2000): An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology, 29:722-729.*

Hansen, L. P. (1982): Large sample properties of generalized method of moments estimators. *Econometrica, 34:646-660.*

Hansen, L. P. (1985): A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics, 30:203-238.*

Hansen, L. P., and Singleton, K. J. (1982): Generalized instrumental variables estimators of nonlinear rational expectations models. *Econometrica, 50:1269-1286*.

Hearst, Norman, Newman, Thomas B., and Hulley, Stephen B. (1986): Delayed effects of the military draft on mortality: a randomized natural experiment. *New England Journal of Medicine, 314:620-624.*

Holford, Theodore R. (1980): The Analysis of Rates and of Survivorship Using Log-Linear Models. *Biometrics, 36:299-305.*

Imbens, Guido W., and Angrist, Joshua D. (1994): Identification and estimation of local average treatment effects. *Econometrica, 62(2):467-475.*

Johnston, K.M., Gustafson, P., Levy, A.R., and Grootendorst, P. (2008): Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine, 27:1539-1556.*

Kelejian, Harry H. (1971): Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association, 66(334):373-374.*

Laird, Nan, and Olivier, Donald (1981): Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. *Journal of the American Statistical Association, 76(374):231-240.*

LaLonde, Robert (1986): Evaluating the econometric evaluations of training programs. *American Economic Review, 76(4): 604-620.*

Lu-Yao, Grace L., Albertsen, C. Peter, Moore, Dirk F., Shih, Weichung, Lin, Yong, DiPaola, Robert S., and Yao, Siu-Long (2008): Survival following primary androgen deprivation therapy among men with localized prostate cancer. *JAMA, 300(2):173-181.*

Mullahy, John (1997): Instrumental variable estimation of count data models: Application to models of cigarette smoking behaviour. *Review of Economics and Statistics, 79(4):586-593.*

National Heart Lung and Blood Institute (2008): *http://www.nhlbi.gov/health/dci/diseases/hemophilia_what.html.*

Pearl, Judea (2009): Chapter 3: Causal diagrams and the identification of causal effects. *Causality, second edition*, Cambridge University Press.

Rosenbaum, Paul R. (1983): The central role of the propensity score in observational studies for causal effects. *Biometrika, 70(1):41-55.*

Shahinian, Vahakn B., Kuo, Yong-fang, Freeman, Jean L., Orihuela, Eduardo, and Goodwin, James S. (2005): Increasing use of gonadotropin-releasing hormone agonists for the treatment of localized prostate carcinoma. *Cancer, 103(8):1615-1624.*

Stukel, Therese A., Fisher, Elliott S., Wennberg, David E., Alter, David A., Gottlieb, Daniel J., and Vermeulen, Marian J. (2007): Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *The Journal of the American Medical Association, 297(3):278-285.*

Valenta, Zdenek, and Weissfeld, Lisa (2002): Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model. *Statistics in Medicine, 21:717-727.*

Vinga, Susana, and Almeida, Jonas S. (2004): Rényi continuous entropy of DNA sequences. *Journal of Theoretical Biology, 231(3):377-388.*

Wald, Abraham (1940): The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics, 11(3):284-300.*

Walke, Rainer (2010): Example for a Piecewise Constant Hazard Data Simulation in R. *Technical reports of the Max Planck Institute for Demographic Research.*

Wikipedia (2010): The Free Encyclopedia: Rubin Causal Model: *http://en.wikipedia.org/wiki/Rubin_Causal_Model.* Accessed on September 18, 2010.

Zeliadt, Steven B., Potosky, Arnold L., Penson, David F., and Etzioni, Ruth (2006): Survival benefit associated with adjuvant androgen deprivation therapy combined with radiotherapy for high- and low-risk patients with nonmetastatic prostate cancer. *International Journal of Radiation OncologyBiologyPhysics, 66(2):395-402.*

Zhang, Junni (2004): Chapter 8: Causal inference with instrumental variables. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives (edited by Gelman, Andrew, and Meng, Xiao-Li),* John Wiley & Sons.