New Instrumental Variable Methods for Causal Inference

by

Douglas A. Lehmann

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Professor Yi Li, Co-chair
Research Associate Professor Yun Li, Co-chair
Professor Rajiv Saran
Professor Douglas Schaubel

“We shall not cease from exploration,
and the end of all our exploring
will be to arrive where we started
and know the place for the first time.”

- T.S. Eliot

ACKNOWLEDGEMENTS

I would like to thank my dissertation committee, Professor Yi Li, Research Assistant Professor Yun Li, and Professor Douglas Schaubel, of the Department of Biostatistics; and Professor Rajiv Saran, of the Department of Internal Medicine. I am very grateful for their help, which has guided my research.

Yi has funded me and remained supportive of my work throughout my time as a doctoral student. He has been an excellent mentor and sets a good example for his students. His help and advice during my job hunt has altered the trajectory of my life for the better.

Yun introduced me to instrumental variables research, and deserves all the credit for getting me started in the right direction when I had no idea what I would do. She has spent much time working very closely with me, and I am grateful for the patience she has shown me over the years. Her enthusiasm about my work has been uplifting, and I have always left meetings feeling better than I did going in.

Doug has had honest and straightforward advice whenever I needed it. He has always left me knowing exactly what I had to do next and how to do it. He also taught me most of what I know about working with survival data.

Rajiv has been my clinical mentor. Our collaborative meetings have shaped the applications presented in this dissertation and have taught me much about the complicated issues related to kidney dialysis.

I would also like to thank two close friends and fellow students, John Rice and

Alex Smith, who have had a big impact on my time at Michigan. Having begun the PhD at the same time, John and I faced many of the same challenges together. I am thankful for the times he has worked through my R code when I was stuck, and his unparalleled knowledge of English grammar has helped immensely while writing this dissertation. Alex has made the many late nights in the office enjoyable. He is the first person I go to for help with editing. He also deserves credit for introducing me to healthy eating, backpacking, and the Pomodoro technique for time management, all of which have improved my life greatly.

This research was funded in part with federal funds from the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN276201400001C. Support was also provide by the National Institutes of Health grant 5R01DK070869-09.

The data reported in this dissertation have been supplied by the United States Renal Data System (USRDS), which is funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

An Introduction to Instrumental Variables

The randomized controlled trial (RCT) is considered the “gold standard” for evaluating the effectiveness of a treatment or intervention. Randomizing subjects to a treatment helps to ensure that treatment groups are comparable on both measured and unmeasured covariates. When treatment groups are comparable, effects can be obtained through direct comparisons using standard statistical methods. While this is a major benefit of RCTs, they are not without limitations. They can be costly, and in some cases it is impossible or even unethical to randomize the treatment. Observational data are an increasingly common alternative to RCTs, but come at the cost of removing control over treatment assignment from the hands of the researcher. This gives rise to the possibility of systematic differences between treatment groups. While it may be possible to measure and control for certain covariates, there remains a possibility that treatment groups differ in unmeasured ways that confound the relationship of interest. This is a primary concern in any observational study, and methods that ignore this unmeasured confounding give biased and potentially misleading results.

Instrumental variable (IV) methods are widely used to deal with this issue and are becoming increasingly popular in health and medical research. IV models are able to obtain consistent effect estimates in the presence of unmeasured confounding between the treatment and outcome, but rely on assumptions that are hard to prove and often criticized (Wooldridge, 2001). A key component of an IV analysis is the instrumental variable, or the instrument. The instrument is a variable that influences or encourages individuals toward a particular treatment without directly affecting the outcome. In this sense, the instrument mimics randomization by randomly "assigning" individuals to different likelihoods of receiving the treatment. An instrument must satisfy three basic assumptions (Angrist et al., 1996; Baiocchi et al., 2014). Presented graphically in Figure 1.1, the instrument 1) must be correlated with the treatment, 2) must be randomly assigned, or independent of unmeasured confounders, and 3) cannot directly affect the outcome.

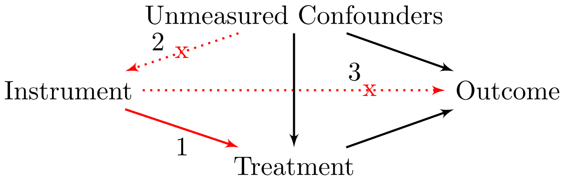


Figure 1.1: Causal diagram depicting the relationship between variables in an instrumental variables analysis, as well as the assumptions imposed on the instrument.

The assumption that the instrument is randomly assigned implies that there are no unmeasured confounders between the instrument and the outcome. Unfortunately, this cannot be verified to hold and the assumption often faces criticism. This assumption is easy to justify when an instrument is based on a truly random process, for example using treatment assignment as an instrument for treatment received in a randomized trial with noncompliance. Without actual randomization, however, finding an instrument that meets this criteria can be a difficult task. In observational studies it is more common to find an instrument that, although not subject to randomization, arguably meets this assumption after controlling for a set of measured instrument-outcome confounders (Garabedian et al., 2014). In other words, the instrument is argued to be conditionally distributed “as good as random.” For example, regional treatment preferences may serve as a reasonable instrument after controlling for patient characteristics such as race, age, education, income, insurance status and comorbidities, geographic characteristics such as rural/urban status and socioeconomic indicators, and provider characteristics such as procedure volume, supply, and profit or teaching status. Garabedian et al. (2014) discuss the most common instruments and potential instrument-outcome confounders associated with each, and emphasize that failing to control for these confounders violates the assumption that the instrument is randomly assigned and can bias estimation.

In Chapter II we propose a weighted IV estimator that controls for measured instrument-outcome confounders using the IV propensity score. The IV propensity score represents the probability that an individual is encouraged, based on their instrument value, toward the treatment. This is different from the more common treatment propensity score, which is the probability that an individual actually receives the treatment. Similar to the treatment propensity score (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004), the IV propensity score balances the distribution of observed covariates across instrument groups while reducing the dimension of the adjustment problem, a major benefit as the number of covariates increases. Unlike the treatment propensity score, which is only useful for addressing measured treatment-outcome confounding, methods based on the IV propensity score can provide consistent effect estimates in the presence of both measured and unmeasured confounders between the treatment and the outcome.

The proposed estimator uses weights that are designed to approximate the probability of being selected into a one-to-one match on the IV propensity score (Frölich, 2007), though we present an extension for approximating k:1 matching designs as well. We therefore refer to the proposed estimator as the IV-matching weight estimator. We show that the IV-matching weight estimator has several benefits over one-to-one IV propensity score matching. These include increased efficiency in estimation, straightforward variance estimation, and speed of computation. The IV-matching weight estimator is further shown to be more efficient than alternative weighting estimators (Tan, 2006), possibly a result of using stable weights that are bounded between 0 and 1.

While we develop this estimator for use with binary outcomes, we present preliminary research on applying the method to time-to-event or survival data. Specifically, we modify the procedure to estimate the difference in mean survival between groups, although other estimands relevant to survival studies may be considered as well. Given the importance of survival data in health and medical research, this work has potential for a broad applications.

Chapter III is concerned with the strength of the correlation between the instrument and the treatment, specifically how this strength can be increased. Instruments with little influence over treatment assignment are termed weak instruments, and there are a number of problems associated with using them. They suffer from greater finite-sample bias and greater variability in estimation (Bound et al., 1995; Wooldridge, 2001). Additionally, results obtained using weak instruments are sensitive to violations of the assumption that the instrument is randomly assigned (Bound et al., 1995; Small and Rosenbaum, 2008; Baiocchi et al., 2010). Given the critical nature of this assumption, the robustness provided by stronger instruments is a major benefit and has motivated recent methods for strengthening the instrument.

We develop a novel method for strengthening the instrument within the IV-matching framework of Baiocchi et al. (2010, 2012). The proposed method involves weighting pairs based on a within-pair measure of instrument strength in a way that increases the strength of the instrument across all pairs. Compared with existing methods for strengthening the instrument (Baiocchi et al., 2010, 2012), the proposed weighting procedure is able to strengthen the instrument without compromising the quality of matches formed. Match quality is a priority in any matching design, since poor match quality can lead to biased and misleading estimated effects. The improved match quality is therefore an important benefit of weighting over existing alternatives.

Finally, in Chapter IV we investigate instrumental variable estimation using strengthened instruments to better understand their properties. Theory suggests that stronger instruments provide for decreased finite-sample bias, increased efficiency in estimation, and results that are more robust to unmeasured instrument-outcome confounders (Bound et al., 1995; Wooldridge, 2001; Small and Rosenbaum, 2008). These benefits have motivated methods for strengthening weak instruments, including those of Baiocchi et al. (2010, 2012) and that proposed in Chapter III. It has yet to be shown, however, that strengthened instruments provide the same benefits as instruments that are naturally stronger. We use the term “strengthened” to refer to any instrument whose strength has been increased by the researcher, and the term “naturally stronger” to refer to instrument that is more correlated with the treatment without the research increasing this correlation.

Our findings suggest that while strengthened instruments provide for more efficient estimation, they are unable to decrease finite-sample bias or improve the robustness to unmeasured instrument-outcome confounders as previously suggested.

We find that methods for strengthening the instrument inadvertently strengthen the relationship between the instrument and any unmeasured instrument-outcome confounders in the process. This important finding has thus far been overlooked in the literature, and has likely led to the belief that strengthened instruments provide estimates that are more robust to unmeasured instrument-outcome confounders. These findings give guidance for future research related to methods for strengthening the instrument.

CHAPTER II

A Weighted Estimator of the Local Average Treatment Effect with Observed Confounders

2.1 Motivation

Instrumental variable (IV) methods are widely used to deal with the selection bias or unmeasured confounding that is often present in observational studies. While IV models can obtain consistent estimates in the presence of this unmeasured confounding, they rely on assumptions that are difficult to verify and often criticized. A key component of an IV analysis is the instrument, a variable that is considered to encourage individuals toward the treatment or control. The instrument is assumed to be correlated with the treatment, have no direct effect on the outcome outside of its effect on the treatment, and be randomly assigned (Angrist et al., 1996; Baiocchi et al., 2014).

The assumption that the instrument is randomly assigned implies that there are no unmeasured instrument-outcome confounders. This assumption is easy to make when an instrument is based on actual randomization, for example using the treatment that a subject is randomly assigned to as an instrument for the treatment a subject ultimately receives in a randomized trial that suffers from noncompliance. Such instruments are rarely available in observational studies, however, and it is more common to find an instrument that meets this requirement only after controlling for a set of measured instrument-outcome confounders. In other words, the instrument is conditionally distributed "as good as random." For example, regional treatment preference may serve as a reasonable instrument after controlling for patient characteristics such as race, age, education, income, insurance status and comorbidities, geographic characteristics such as rural/urban status, socioeconomic indicators, and provider characteristics such as procedure volume, supply, and profit or teaching status. Garabedian et al. (2014) discuss the most commonly used instruments and potential instrument-outcome confounders associated with each, and emphasize that failing to adjust for these can bias estimation.

Instrument-outcome confounders can be adjusted for in several ways. They can be included as covariates in two stage regression models. While two stage least squares is the most common among these, it may be inappropriate for binary outcomes (Bhattacharya et al., 2006). Two stage residual inclusion was proposed in Terza et al. (2008) for use with binary outcomes. Matching on confounders is a common nonparametric alternative to these regression methods, but becomes difficult when there are many confounders or confounders with many discrete levels. Though less common in practice, methods have proposed using the IV propensity score rather than the full set of confounders. These include the inverse probability weighting estimator of Tan (2006), the matching estimator of Frölich (2007), and the weighting and subclassification methods of Cheng and Lin (2013).

The IV propensity score is the probability that an individual is encouraged toward the treatment, as indicated by their instrument value. This is different from the more common treatment propensity score, which represents the probability that an individual actually receives the treatment. Like the treatment propensity score, the IV propensity score balances the distribution of confounders across groups while reducing the dimension of the adjustment problem (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004). Unlike the treatment propensity score, which is only useful for addressing measured treatment-outcome confounding, methods based on the IV propensity score can provide consistent effect estimates in the presence of both measured and unmeasured treatment-outcome confounding.

In this chapter we propose an IV estimator based on the IV propensity score. It is a weighted estimator that uses weights designed to reflect the probability of being selected into a one-to-one IV propensity score match. One-to-one IV propensity score matching involves pairing encouraged subjects to unencouraged subjects with similar scores, usually within a specified range. Often a match cannot be found within this range. Pairing the encouraged subject to an unencouraged subject with a score outside of this range can bias estimation, whereas dropping the encouraged subject reduces sample size which leads to a decrease in efficiency. Furthermore, matching becomes a computationally difficult task as sample size increases. The proposed estimator avoids these pitfalls associated with matching. An additional benefit of weighting estimators is that they allow for straightforward variance estimation, whereas the correlation structures introduced by matching algorithms are difficult to account for when estimating the variance of matching estimators (Austin, 2008, 2009b, 2011a).

We further present two extensions of the proposed estimator that could prove useful in practice. The first is a modification to the weight function to approximate k:1 matching designs. The second is an alternative formulation of the estimator that provides protection against misspecification of the IV propensity score model. Though this requires the additional specification of an outcome model, this double robust estimator will give consistent estimates if at least one of the IV propensity score or outcome models is correctly specified.

The remainder of this chapter is organized as follows. In Section 2.2 we define notation, discuss the IV propensity score, and introduce our proposed estimator. Finite-sample performance is reported through simulations in Section 2.3, and use of the method is illustrated with a real data example in Section 2.4. Future work on applying the method to survival data is presented in Section 2.5, and we conclude with a discussion in Section 2.6.

2.2 Methods

2.2.1 Notation

We define notation using the potential outcomes framework (Rubin, 1974; Neyman, 1923; Angrist et al., 1996). For each of $i = 1, ..., n$ subjects, let $Z_i = 1$ if subject i is encouraged toward the treatment and $Z_i = 0$ if encouraged toward the control. Let $D_i(Z_i)$ indicate treatment received for subject i given their encouragement status, and let $Y_i(Z_i, D_i(Z_i))$ indicate the response for subject i given their encouragement status and treatment value. $D_i(Z_i)$ and $Y_i(Z_i, D_i(Z_i))$ are referred to as potential outcomes. When subject i is encouraged toward the treatment, we observe treatment $D_i(1)$ and response $Y_i(1, D_i(1))$ from subject i, otherwise we observe treatment $D_i(0)$ and response $Y_i(0, D_i(0))$. Our interest is in estimating the parameter

$$\lambda = \frac{E(Y_i(1, D_i(1)) - Y_i(0, D_i(0)))}{E(D_i(1) - D_i(0))}. \tag{2.1}$$

This is the ratio of the instrument's effect on the response to its effect on the treatment, and is often referred to as the local average treatment effect (LATE) (Imbens and Angrist, 1994; Angrist et al., 1996). Rather than an average treatment effect over the entire population, the LATE is interpreted as an average effect over a subgroup of the population known as compliers. Depicted in Table 2.1, compliers are individuals that take the treatment they are encouraged toward, and are one of four population subgroups defined by their response to encouragement.

Table 2.1: Population subgroups defined by the effect of encouragement on treatment. $D(1)$ denotes the treatment a subject will receive if they are encouraged toward treatment, while $D(0)$ denotes the treatment they will receive if they are encouraged toward the control.

		$D(1)$	
		1	0
$D(0)$	1	Always-takers	Defiers
	0	Compliers	Never-takers

The difficulty in estimating λ comes from the fact that we never observe individuals under both states of encouragement, and therefore never observe both of their potential outcomes. The data provides, for example, $E(Y_i(1, D_i(1))|Z_i = 1)$, or the average response under encouragement among the encouraged subjects. This differs from $E(Y_i(1, D_i(1)))$, which is the average response over the entire population if the entire population were encouraged. To recover the expectations in equation (2.1), and to aid in the interpretation of λ, we make the following five assumptions (Angrist et al., 1996):

A1. Stable Unit Treatment Value Assumption (SUTVA). Often known as no interference, SUTVA requires that the outcomes for one subject be unaffected by the encouragement status or treatment assignment of other subjects. This assumption will be violated if spillover effects exist between patients or groups. SUTVA allows us to consider a subjects potential outcomes as a function of their treatment and encouragement, rather than a function of the treatment and encouragement assignments of the entire population.

A2 - Random assignment of the instrument. The instrument is assumed to be randomly assigned, which implies that there are no unmeasured confounders between the instrument and the outcome. This assumption is often made conditional on measured instrument-outcome confounders. It cannot be verified to hold, and weak instruments are especially sensitive to violations of this assumption (Baiocchi et al., 2014; Bound et al., 1995; Small and Rosenbaum, 2008; Staiger and Stock, 1994).

A3 - Exclusion restriction. The instrument is assumed to affect the outcome only through its effect on the treatment. This implies that $Y_i(1, D_i(1) = d) = Y_i(0, D_i(0) = d)$ for all i, d. Since both potential outcomes are never observed for any individual, this assumption cannot be verified to hold.

A4 - Nonzero association between instrument and treatment. The instrument is assumed to be correlated with the treatment. This implies that $E(D_i(1) - D_i(0)) \neq 0$.

A5 - Monotonicity. Monotonicity is the assumption that no individual always does the opposite of what they are encouraged to do. This implies that there are no defiers (Table 2.1) and that $D_i(1) \geq D_i(0)$ for all i.

The SUTVA and random assignment assumptions allow for unbiased estimation of the instrument's effect on the outcome and the treatment, or the numerator and denominator in (2.1). The remaining assumptions give λ a meaningful interpretation. By exclusion restriction, always- and never-takers (Table 2.1) do not contribute to estimation since their treatment values, and therefore their response values, do not vary with encouragement. Monotonicity ensures that the subgroup of defiers is empty, while a nonzero association between the instrument and the treatment ensures that the subgroup of compliers is not empty. With the addition of assumptions these three assumptions, λ can therefore be interpreted as an average treatment effect among the compliers, who are often referred to as "marginal patients." Unlike the average treatment effect, which is applicable to the entire population, λ only applies to subjects that can be encouraged to switched treatment states. Further discussion of these assumptions can be found in Imbens and Angrist (1994), Angrist et al. (1996) or Baiocchi et al. (2014), among many others.

2.2.2 Proposed Estimator

In this section we present our proposed estimator. It is a weighted IV estimator of the local average treatment effect that adjusts for measured instrument-outcome confounders using weights that are based on the IV propensity score. Defined as

$$e(\mathbf{x}) = P(Z = 1|\mathbf{X} = \mathbf{x}), \tag{2.2}$$

the IV propensity score represents the probability of receiving encouragement toward the treatment. This is different from the more common treatment propensity score, which represents the probability of actually receiving the treatment. From the theorems of Rosenbaum and Rubin (1983), we can say that the distribution of covariates $\mathbf{X}$ is balanced across instrument groups conditional on $e(\mathbf{x})$, and if the instrument is independent of unmeasured confounders conditional on $\mathbf{X}$, then it is independent of unmeasured confounders conditional on $e(\mathbf{x})$ as well. Taken together, these two statements imply that conditioning on $e(\mathbf{x})$ is sufficient for adjusting for $\mathbf{X}$.

Define the observed treatment and response values for subject i as $D_i = Z_i D_i(1) + (1 - Z_i)D_i(0)$ and $Y_i = Z_i Y_i(1, D_i(1)) + (1 - Z_i)Y_i(0, D_i(0))$, respectively. Our proposed estimator, which we refer to as the IV-matching weight (IV-MW) estimator, is given as

$$\lambda_{\text{IV-MW}} = \frac{\sum_i W_i Z_i Y_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) Y_i / \sum_i W_i (1 - Z_i)}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)}, \tag{2.3}$$

where weights W_i are defined as

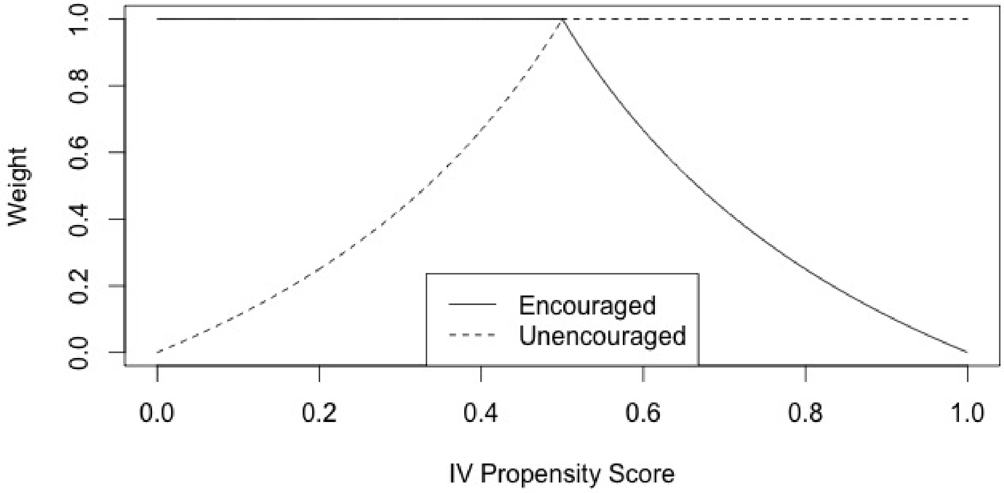
$$W_i = \frac{\min(e_i(\mathbf{x}_i), 1 - e_i(\mathbf{x}_i))}{Z_i e_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \tag{2.4}$$

This weight is similar to the matching weight of Li and Greene (2013), but defined using the IV propensity score rather than the treatment propensity score. The IV-MW estimator is therefore useful for dealing with both measured and unmeasured treatment-outcome confounding, whereas the estimator of Li and Greene (2013) is useful only in the presence of measured treatment-outcome confounding but will not provide consistent estimates if unmeasured treatment-outcome confounders exist.

Weights W_i are referred to as matching weights because they approximate the probability of being selected into a one-to-one match on the IV propensity score. The asymptotic equivalence of the IV-MW estimator with one-to-one matching on the IV propensity score is shown in Appendix A, but we illustrate the idea here with a simple example. Consider a region around IV propensity score $e = 0.1$ with $m = 100$ individuals. From (2.2), there is an expected $me = 10$ encouraged and $m(1-e) = 90$ unencouraged individuals in this region. All encouraged individuals are therefore expected to find a match, and $W = min(0.1, 0.9)/(1 \times 0.1 + 0 \times 0.9) = 1$ for these individuals. However, only 10 of the 90 unencouraged individuals are expected to be matched, and $W = min(0.1, 0.9)/(0 \times 0.1 + 1 \times 0.9) = 1/9$ for these individuals. Figure 2.1 displays the weight assigned to encouraged and unencouraged individuals across the range of IV propensity scores.

Alternatives to the IV-MW estimator proposed in (2.1) include one-to-one IV propensity score matching (Frölich, 2007) and inverse probability weighting using the IV propensity score (Tan, 2006). These two methods are, to the best of our knowledge, the only published alternatives that make use of the IV propensity score to control for measured instrument-outcome confounders. A benefit of the proposed IV-MW estimator over one-to-one IV propensity score matching is that all subjects contribute a fraction of themselves to estimation, avoiding the situation where in-

Figure 2.1: Matching weights for encouraged and unencouraged subjects by IV propensity score.



dividuals are removed after not finding a suitable match. This avoids a decrease in sample size and efficiency. A benefit of the proposed estimator over inverse probability weighting using the IV propensity score is that the weights are bounded between 0 and 1, whereas inverse probability weights can "blow up" near probabilities 0 or 1, causing an increase in the variance of the estimate (Li et al., 2014). We therefore expect the proposed IV-MW estimator to be more efficient than both one-to-one IV propensity score matching and inverse probability weighting using the IV propensity score.

An additional benefit of weighting estimators over matching based estimators is that they allow for straightforward variance estimation. Matching algorithms introduce complicated correlation structures that are difficult to account for when estimating the variance of matching estimators. Often, the matched nature of the data is ignored entirely (Austin, 2008, 2009b, 2011a). Following Lunceford and Davidian (2004) and Li and Greene (2013), a sandwich type variance estimator is obtained using estimating equations

$$\mathbf{0} = \sum_{i=1}^{n} \phi_i(\theta) = \sum_{i=1}^{n} \begin{bmatrix} W_i Z_i (Y_i - \mu_{y1}) \\ W_i(1 - Z_i)(Y_i - \mu_{y0}) \\ W_i Z_i (D_i - \mu_{d1}) \\ W_i(1 - Z_i)(D_i - \mu_{d0}) \\ \mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) \end{bmatrix}, \tag{2.5}$$

where $\theta = (\mu_{y1}, \mu_{y0}, \mu_{d1}, \mu_{d0}, \boldsymbol{\eta}')$, with $\mu_{y1} = E(W_i Z_i Y_i)/E(W_i Z_i)$, $\mu_{y0} = E(W_i(1 - Z_i)Y_i)/E(W_i(1 - Z_i))$ and similar for μ_{d1} and μ_{d0}. $\mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ represent estimating equations for coefficients $\boldsymbol{\eta}$ from the model used to estimate the IV propensity score, often a logistic regression. An estimate of $var(\hat{\theta})$ is obtained as $n^{-1}\hat{A}_n^{-1}\hat{B}_n(\hat{A}_n^T)^{-1}$, where $\hat{A}_n = \sum_{i=1}^n \partial\phi_i(\theta)/\partial\theta|_{\theta=\hat{\theta}}$ and $\hat{B}_n = \sum_{i=1}^n \phi_i(\theta)\phi_i^T(\theta)|_{\theta=\hat{\theta}}$. Applying the multivariate delta method with $g(\theta) = (\mu_{y1} - \mu_{y0})/(\mu_{d1} - \mu_{d0})$, an estimate of $var(\hat{\lambda})$ is obtained as $\nabla g(\theta)^T \hat{var}(\hat{\theta}) \nabla g(\theta)$. This procedure allows for simultaneous estimation of the IV propensity score and λ, and is used for variance estimation for all estimators compared in Sections 2.3 and 2.4. For the IV propensity score matching procedure, this sandwich variance estimate ignores the matched nature of the sample. This typically leads to an overestimated variance (Austin, 2009b, 2011a), though in simulations reported in Section 2.3 the estimated and empirical standard deviations are found to be approximately equal.

Note that W_i is not differentiable everywhere with respect to $\boldsymbol{\eta}$ due to the minimum function in the numerator. To apply this variance estimation procedure, rewrite the weight function as

$$W_i = \frac{e_i(\mathbf{x}_i)I[e_i(\mathbf{x}_i) \leq 0.5] + (1 - e_i(\mathbf{x}_i))I[e_i(\mathbf{x}_i) > 0.5]}{Z_i e_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \tag{2.6}$$

These indicator functions can then be replaced with cumulative distribution functions to create a smooth, differentiable function for W_i (Horowitz, 1992).

In the following two sections we extend the IV-MW estimator in ways that could prove useful in practice. We first adjust the weight function in (2.4) to approximate a k:1 matching algorithm. We then modify the estimator in (2.3) for a double robust IV-MW estimator that protects against misspecification of the IV propensity score model. This requires the additional specification of an outcome model but will give consistent estimates if at least one of the IV propensity score or outcome models is correctly specified.

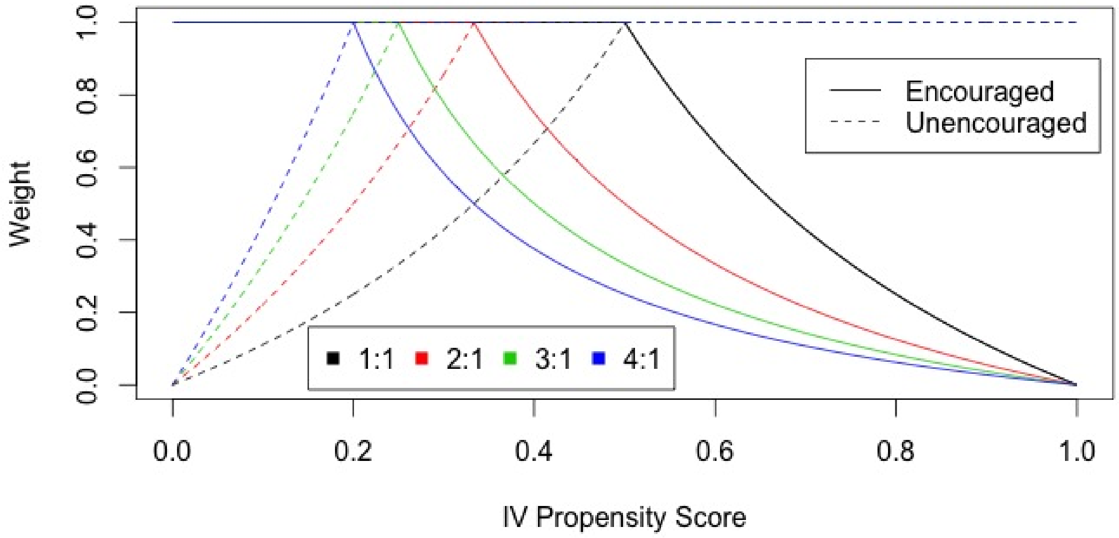
k:1 IV-matching Weights

The weights proposed in (2.4) are designed to approximate a one-to-one match on the IV propensity score. While one-to-one matching is the most common in practice (Austin, 2008), if the pool of unencouraged subjects is large enough we might consider matching multiple unencouraged subjects to each encouraged subject. Increasing the number of unencouraged subjects has the benefit of increasing the sample size, thereby decreasing the variability in estimation. More on k:1 matching using propensity scores, including guidance for the selection of k, can be found in Austin (2010). For approximating k:1 matching designs, we redefine the weights as

$$W_i = \frac{\min(ke_i(\mathbf{x}_i), 1 - e_i(\mathbf{x}_i))}{Z_i k e_i(\mathbf{x}_i) + (1 - Z_i)(1 - e_i(\mathbf{x}_i))}. \tag{2.7}$$

As the number of unencouraged subjects to be matched increases, the probability that they will be selected into a match for any given IV propensity score increases, while decreasing the probability that encouraged subjects will be able to find k individuals to match with. Figure 2.2 displays the weight assigned to encouraged and unencouraged subjects across the range of IV propensity scores for one-, two-, three-, and four-to-one IV propensity score matching.

Figure 2.2: Weights for encouraged and unencouraged subjects across IV propensity scores for 1:1, 2:1, 3:1 and 4:1 matching designs. Weights for 1:1 matching are the same as Figure 2.1



Double Robust IV-MW Estimator

The IV-MW estimator in (2.3) requires correct specification of the IV propensity score model for consistent estimation. In this section, we modify the IV-MW estimator to protect against misspecification of the IV propensity score model. While this requires the additional specification of an outcome model, the double robust (IV-MW$_{DR}$) estimator will provide consistent estimates if at least one of the IV propensity score or outcome models is correctly specified, but does not require correct specification of both.

Let $m_0(\mathbf{X}_i) = E\left\{Y_i|\mathbf{X}_i, Z_i = 0\right\}$ denote the outcome model for the unencouraged group and similarly let $m_1(\mathbf{X}_i)$ the outcome model for the encouraged group. Following from Lunceford and Davidian (2004) and Li and Greene (2013), a double robust version of the IV-MW estimator is given as

$$\lambda_{\text{IV-MW}_{DR}} = \frac{A + B - C}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)}, \tag{2.8}$$

where

$$
\begin{aligned}
A &= \sum_i W_i\{m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i)\}/\sum_i W_i, \\
B &= \sum_i W_i Z_i\{Y_i - m_1(\mathbf{X}_i)\}/\sum_i W_i Z_i, \\
C &= \sum_i W_i(1-Z_i)\{Y_i - m_0(\mathbf{X}_i)\}/\sum_i W_i(1-Z_i).
\end{aligned}
$$

Variance is estimated using the procedure of Section 2.2.2, with estimating equations

$$
\mathbf{0} = \sum_{i=1}^{n} \phi_i(\theta) = \sum_{i=1}^{n} \begin{bmatrix} W_i\{m_1(\mathbf{X}_i) - m_0(\mathbf{X}_i) - \mu_A\} \\ W_i Z_i\{Y_i - m_1(\mathbf{X}_i) - \mu_B\} \\ W_i(1-Z_i)\{Y_i - m_0(\mathbf{X}_i) - \mu_C\} \\ W_i Z_i(D_i - \mu_{d1}) \\ W_i(1-Z_i)(D_i - \mu_{d0}) \\ \mathbf{S}_1(\boldsymbol{\alpha}_1) \\ \mathbf{S}_0(\boldsymbol{\alpha}_0) \\ \mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta}) \end{bmatrix}, \tag{2.9}
$$

where $\theta = (\mu_A, \mu_B, \mu_C, \mu_{d1}, \mu_{d0}, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_0', \boldsymbol{\eta}')$. μ_A, μ_B, μ_C, μ_{d1}, and μ_{d0} correspond to the limits of A, B, C, and the averages in the denominator of (2.8). $\mathbf{S}_1(\boldsymbol{\alpha}_1)$ and $\mathbf{S}_0(\boldsymbol{\alpha}_0)$ represent the estimating equations for the parameters in $m_1(\mathbf{X}_i)$ and $m_0(\mathbf{X}_i)$, and $\mathbf{S}_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ the estimating equations for the parameters in the IV propensity score model.

2.3 Simulation

2.3.1 Setup

In this section we report the results of simulation studies for investigating the finite-sample performance of the proposed IV-MW estimator. We compare it with two alternatives that make use of the IV propensity score: the inverse probability weighting (IV-IPW) estimator of Tan (2006) and one-to-one IV propensity score matching (IV-PSM) (Frölich, 2007). The IV-IPW estimator has the same form as the IV-MW estimator in (2.1), but with the numerator of (2.4) replaced with 1. For the IV-PSM procedure, we match on the logit of the IV propensity score, using an optimal one-to-one match with a caliper of width equal to one fourth the standard deviation of logit of the IV propensity scores. For information about caliper selection, see Cochran and Rubin (1973), Raynor (1983), Rosenbaum and Rubin (1985), or Austin (2011b).

We generate 1,000 datasets with binary outcome, treatment, and instrument for $i = 1, ..., n$ individuals from

$$P(Y_i = 1|D_i, X_{1i}, X_{2i}) = logit^{-1}(\beta D_i + \delta_1 X_{1i} + \delta_2 X_{2i} + \epsilon_i^Y), \tag{2.10}$$

$$P(D_i = 1|Z_i) = logit^{-1}(\gamma_0 + \gamma_1 Z_i + \epsilon_i^D), \tag{2.11}$$

$$P(Z_i = 1|X_{1i}, X_{2i}) = logit^{-1}(\psi_0 + \psi_1 X_{1i} + \psi_2 X_{2_i}). \tag{2.12}$$

(ϵ^Y, ϵ^D) are drawn from a bivariate normal with correlation 0.8 to represent unmeasured treatment-outcome confounding. X_1 and X_2 represent measured instrument-outcome confounders and are randomly drawn from standard normal distributions. β is varied from 0 to 1 and sample size is varied from 500 to 2,000. Results are reported under the following parameter settings: $\psi_0 = \gamma_0 = -1$, $\psi_1 = \delta_1 = -0.25$, $\gamma_1 = 1$ and $\psi_2 = \delta_2 = 0.25$.

2.3.2 Results

Results for estimation and coverage properties of the estimators are reported in Table 2.2. Each of the three estimators are found to be approximately unbiased. This is expected because we have adjusted for instrument-outcome confounders X_1 and X_2 in this simulation, and there are no unmeasured instrument-outcome confounders that would bias these IV methods. Coverage rates for each method are converging to the nominal rate as sample size increases. Both weighted estimators have lower mean squared errors (MSE) than IV-PSM, with the proposed IV-MW estimator achieving the lowest MSE in each scenario. The 2:1 and 3:1 matching scenarios confirm that the IV-MW estimator remains unbiased with the lowest MSE.

Table 2.2: Bias, MSE, and 95% coverage probabilities of IV-MW, IV-IPW, and IV-PSM for estimation of λ. Reported results have been multiplied by 100.

				Weighting						Matching		
				IV-MW			IV-IPW			IV-PSM		
k	N	β	λ	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP
1:1	500	0.0	0.00	-0.39	6.11	97.4	-0.51	6.38	97.2	0.50	12.53	97.2
		0.5	0.12	-0.09	5.88	97.4	-0.38	6.17	97.1	1.05	9.42	98.2
		1.0	0.22	-0.65	5.18	97.9	-0.40	5.50	97.4	-0.13	8.61	98.3
1:1	1000	0.0	0.00	0.16	2.59	96.2	0.27	2.62	96.2	0.69	4.14	96.5
		0.5	0.12	0.42	2.50	96.9	0.67	2.57	96.5	0.70	3.84	97.0
		1.0	0.22	0.31	2.17	97.0	0.81	2.32	96.5	1.19	3.59	96.8
1:1	2000	0.0	0.00	-0.04	1.28	95.6	-0.03	1.32	95.3	0.38	1.85	95.4
		0.5	0.12	0.22	1.15	96.1	0.35	1.19	95.9	0.39	1.61	96.8
		1.0	0.22	-0.71	1.26	95.2	-0.27	1.32	94.6	-0.19	1.87	95.5
2:1	2000	0.00	0.00	-0.10	1.25	96.1	-0.12	1.27	95.7	0.30	1.83	95.8
		0.50	0.12	0.51	1.29	94.8	0.46	1.31	94.7	0.38	1.84	95.1
		1.00	0.22	0.18	1.19	95.0	0.36	1.22	95.4	0.13	1.69	96.0
3:1	2000	0.00	0.00	-0.25	1.27	95.8	-0.31	1.27	96.3	-0.27	1.87	95.8
		0.50	0.12	0.52	1.21	95.0	0.51	1.22	95.2	0.55	1.72	96.4
		1.00	0.22	-0.05	1.16	95.7	-0.09	1.18	95.6	-0.16	1.65	97.5

A comparison of the estimated and empirical standard deviations in Table 2.3 confirms that the sandwich variance estimates (ASD) approximate the empirical standard deviations (ESD) well. Applying the sandwich variance procedure to the IV-PSM procedure requires that we ignore the matched nature of the data. While this typically leads overestimation of the variance (Austin, 2009b, 2011a), this is not seen to be an issue in the simulations reported here.

Additional simulations were performed to study the performance of the double robust IV-MW estimator (IV-MW$_{DR}$) of Section 2.2.2. For these simulations, the

Table 2.3: Comparison of standard deviations obtained empirically (ESD) and using the sandwich variance technique of Section 2.2.2 (ASD). Reported results have been multiplied by 100.

				Weighting				Matching	
				IV-MW		IV-IPW		IV-PSM	
k	N	β	λ	ASD	ESD	ASD	ESD	ASD	ESD
1:1	500	0.0	0.00	24.58	24.71	25.03	25.25	33.01	35.40
		0.5	0.12	23.97	24.26	24.68	24.84	31.22	30.67
		1.0	0.22	23.03	22.74	23.84	23.45	29.56	29.33
1:1	1000	0.0	0.00	16.37	16.08	16.62	16.19	20.06	20.35
		0.5	0.12	16.06	15.79	16.32	16.02	19.76	19.59
		1.0	0.22	15.40	14.74	15.79	15.21	19.03	18.91
1:1	2000	0.0	0.00	11.35	11.30	11.48	11.50	13.66	13.59
		0.5	0.12	11.14	10.73	11.34	10.90	13.45	12.69
		1.0	0.22	10.84	11.20	11.12	11.50	13.20	13.68
2:1	2000	0.0	0.00	11.27	11.20	11.52	11.29	13.74	13.52
		0.5	0.12	11.00	11.35	11.31	11.43	13.42	13.54
		1.0	0.22	10.67	10.90	11.03	11.02	13.07	12.99
3:1	2000	0.0	0.00	11.34	11.26	11.52	11.28	13.71	13.67
		0.5	0.12	11.17	11.00	11.36	11.03	13.44	13.12
		1.0	0.22	10.85	10.78	11.05	10.84	13.09	12.85

generating equations for Y and Z were redefined as

$$P(Y_i = 1|D_i, X_{1i}, X_{2i}) = logit^{-1}(\beta D_i + \delta_1 X_{1i} + \delta_2 X_{2i} + \epsilon_i^Y + X_{1i}X_{2i}), \tag{2.13}$$

$$P(Z_i = 1|X_{1i}, X_{2i}) = logit^{-1}(\psi_0 + \psi_1 X_{1i} + \psi_2 X_{2_i} + X_{1i}X_{2i}). \tag{2.14}$$

The interaction term $X_{1i}X_{2i}$ is ignored to represent an incorrectly specified model.

Results in Table 2.4 confirm that IV-MW$_{DR}$ provides consistent effect estimates and maintains nominal coverage rates if at least one of the outcome or IV propensity score models is correctly specified. The original IV-MW estimator only provides consistent estimates when the IV propensity score model is correctly specified, and its performance suffers greatly when this model is misspecified. An interesting finding in Table 2.4 is that even though IV-MW$_{DR}$ requires the additional specification of an outcome model, its performance is not damaged compared to IV-MW in situations in which the double robust property would not be needed, i.e. when the IV propensity score model is correctly specified. In other words, we do not lose anything, in terms of bias or efficiency, by specifying the outcome model.

Table 2.4: Estimation under correctly and incorrectly specified IV propensity score and outcome models. $n = 1,000$ and $\beta = 1$ for these simulations. Reported results have been multiplied by 100.

Model Specification							
P(Z)	P(Y)	Estimator	% Bias	MSE	ASD	ESD	95% CP
Correct	Correct	IV-MW	-2.50	0.56	7.52	7.47	94.9
		IV-MW$_{DR}$	-2.50	0.56	7.80	7.48	95.5
Correct	Incorrect	IV-MW	-3.12	0.52	7.57	7.18	96.3
		IV-MW$_{DR}$	-3.13	0.52	7.66	7.18	96.2
Incorrect	Correct	IV-MW	84.16	3.86	7.51	7.36	30.1
		IV-MW$_{DR}$	-6.18	0.55	7.66	7.33	95.8
Incorrect	Incorrect	IV-MW	85.14	3.96	7.49	7.56	30.6
		IV-MW$_{DR}$	87.65	4.17	7.57	7.56	29.6

The simulations reported throughout this section demonstrate that the proposed IV-MW estimator performs well compared with alternatives. It provided consistent estimates, achieved the lowest MSE, and maintained approximately nominal coverage in all scenarios reported. Both weighting estimators (IV-MW and IV-IPW) achieved lower MSE than the matching estimator (IV-PSM), a result of allowing the full data to contribute to estimation. The lower MSE for IV-MW compared with IV-IPW could be because IV-IPW weights can “blow up” near probabilities of 0 or 1 (Li et al., 2014). While IV-MW weights are always bounded between 0 and 1, the IV-IPW weights ranged from 1.05 to 20.5 in these simulations. Additionally, the IV-MW$_{DR}$ estimator was able to protect against misspecification of the IV propensity score model, providing consistent effect estimates if at least one of the IV propensity score or outcome models were correctly specified. Both weighting estimators saw computational benefits over IV propensity score matching as well. Using a MacBook Pro with a 2 GHz Intel Core i7 processor, the following times (in seconds) were observed to complete 1,000 simulations for n = 500, 1,000, and 2,000, respectively: IV-MW - 6.3, 9.9, and 23.6, IV-IPW - 6.0, 9.4, and 21.4, IV-PSM - 65.4, 166.1, and 547.9. While computing time was not a limiting factor in these simulations, it quickly becomes one for matching estimators as sample size increases.

2.4 Data Example

We illustrate use of the methods of this chapter with data from the United States Renal Data System (USRDS) to study the association between dialysis session length and mortality among incident hemodialysis patients in the United States. Longer dialysis sessions are thought to decrease mortality by reducing the risk of intradialytic hypotension and better controlling volume excess and serum phosphorous (Daugirdas, 2013), but this relationship is likely confounded. Shorter dialysis sessions are often prescribed to smaller patients, and smaller patients tend to have higher mortality rates, so direct comparisons would likely give biased effect estimates. Many observational studies have found a significant increase in mortality in patients receiving shorter dialysis sessions (Flythe et al., 2013; Saran et al., 2006). A 2002 randomized trial, on the other hand, found no significant relationship between dialysis session length and mortality (Eknoyan et al., 2002), and Brunelli et al. (2010) found longer dialysis sessions to be associated with higher or lower mortality depending on whether the treatment was considered time dependent. These conflicting results suggest that unmeasured treatment-outcome confounding may be present and an IV analysis may provide new and useful insight.

We obtained complete data on 319,168 adults initiating hemodialysis (HD) between January 1, 2010 and December 31, 2013 from the USRDS database. We restricted the analysis to patients on a thrice-weekly dialysis schedule (98% of all incident HD patients in the data were on a thrice-weekly dialysis schedule). We conducted an intention-to-treat analysis, defined the treatment as being prescribed dialysis sessions of four hours or longer, and defined the outcome as death within the first year after initiating dialysis. Mean treatment usage in the hospital service area

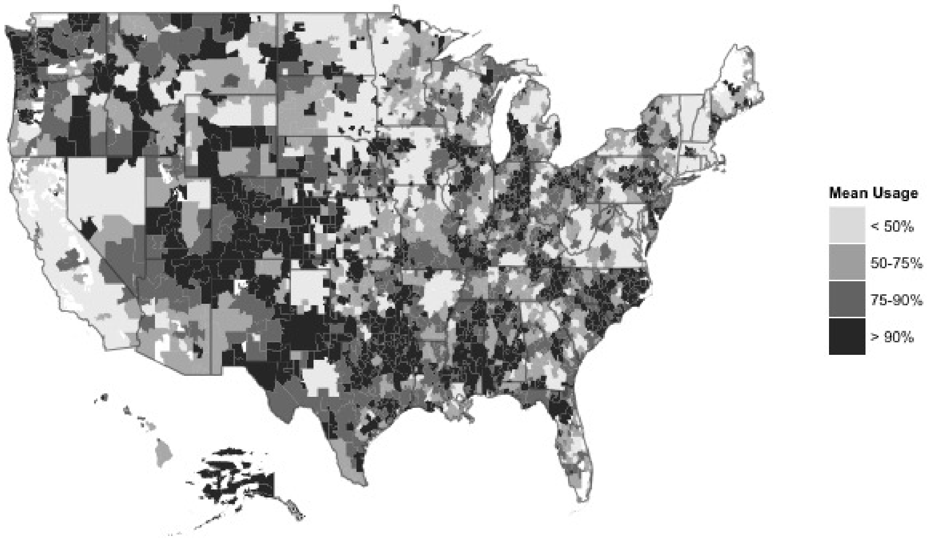
(HSA) from 2007 to 2009 was used as the IV (Figure 2.3). The HSA is a geographic region representing a collection of zip codes with residents that receive most of their healthcare within that region (Dartmouth, 2016). Preference-based instruments such as this one are among the most common in health research (Garabedian et al., 2014), and are thought to measure treatment preferences that are independent of patient level confounders (Brookhart and Schneeweiss, 2007; Li et al., 2015).

Among the 3,336 HSAs in the data, mean treatment usage varied from 0 to 100% with a mean of 74%. The correlation coefficient between the mean treatment usage from 2007-2009 and mean treatment usage from 2010-2013 in an HSA was almost 90%. This indicates that preferences in an HSA are relatively stable through time, and that mean treatment usage in an HSA from 2007-2009 is a strong instrument for treatment in the 2010-2013 data. To fit the methods of this chapter, we dichotomized the instrument by considering HSAs with above average treatment usage to be encouraging subjects toward longer dialysis sessions and HSAs with below average usage to be encouraging their subjects toward shorter dialysis sessions.

The distribution of covariates by treatment and instrument groups is reported in Table 2.5. Patients receiving longer dialysis sessions tend to have higher BMI and are more likely to be male, black, and younger compared with patients receiving shorter dialysis sessions. These patients are also more likely to receive dialysis at for profit facilities in poorer, less educated areas. Comparing across instrument groups greatly improves the balance of covariates. This is evidence that mean treatment usage in an HSA may serve as a valid IV, although some imbalances remain in facility and zip code level variables.

We first fit unadjusted and covariate adjusted logistic regression models to compare with the IV methods of this chapter. These suggest a significant decrease in

Figure 2.3: Distribution of longer dialysis session usage by hospital service area (HSA). Longer dialysis sessions are defined as being prescribed dialysis sessions of four or more hours.



the odds of first year mortality among patients with longer dialysis sessions, with estimated odds ratio and 95% confidence intervals of 0.86 (0.84, 0.87) and 0.95 (0.93, 0.97), respectively. Age, sex, race, ethnicity, BMI, number of comorbidities, access type, profit status of the facility and median income in the zip code were included in the covariate adjusted model. These logistic regressions will be biased if there are confounders between the treatment and the outcome that are not included in the model.

To implement the methods of this chapter, we begin by modeling the IV propensity score, or the probability of being in an HSA with above average usage of longer dialysis sessions. We specify a logistic regression model and include HSA level covariates mean age, BMI, number of comorbidities, percentage of males, blacks, hispanics, and patients without insurance, and median income, as well as patient level covariates age, sex, race, and BMI. Using the estimated IV propensity score, a weight is assigned to each subject for the IV-MW and IV-IPW procedures. For the IV-MW

Table 2.5: Distribution of covariates across treatment (long vs short dialysis sessions) and instrument groups (high vs low treatment usage at HSA level). Reported in the table is the mean and absolute standardized difference, *d*, between groups. An absolute standardized difference of more than 10 is generally considered to indicate an imbalance (Love, 2002).

	Hours on Dialysis			Usage in HSA		
	< 4	≥ 4	d	Low	High	d
Treatment						
4+ hour sessions	-	-	-	60.3%	88.8%	69.1
Outcome						
Death w/in 1st year	23%	21%	6.2	22%	21%	0.8
Patient Level Covariates						
Age	65.9	63.0	19.6	64.3	63.3	6.8
Male	51%	59%	16.7	57%	56%	2.1
BMI	28.0	30.4	28.2	29.4	30.1	8.3
Serum Creatinine	6.4	6.7	1.6	6.7	6.6	0.3
Hemoglobin	10.0	9.9	0.7	9.9	9.9	0.1
Black	23%	31%	16.6	26%	32%	12.4
Hispanic	17%	14%	8.4	17%	11%	10.9
Pre-ESRD 6+ Months	41%	44%	5.6	41%	45%	8.3
Employed	8.3%	9.1%	3.0	9.1%	8.8%	0.9
No Insurance	5%	8%	10.9	6%	8%	10.3
# Comorbidities	2.5	2.6	5.4	2.5	2.6	5.7
Facility Level Covariates						
# Nurses	7.3	7.3	0.3	7.7	7.0	12.2
# Patient techs	8.9	8.8	1.1	9.1	8.5	8.1
# HD stations	20.8	21.7	10.2	21.2	21.7	5.7
For profit	81%	86%	14.9	82%	88%	16.5
Zip Code Level Covariates						
Median income	\$54,551	\$49,286	25.3	\$53,358	\$47,960	26.9
Bachelors degree +	25.5%	22.9%	18.0	24.6%	22.6%	14.9

procedure, this weight is given in (2.4). For the IV-IPW procedure the weight is defined similar to (2.4), but with the numerator replaced with 1. For the IV-PSM procedure, we specified a one-to-one optimal match on the IV propensity score with a caliper of 0.05.

Table 2.6: Instrumental variable estimates and 95% confidence intervals for the effect of longer dialysis sessions on first year mortality. Negative estimates suggest less first year mortality among the patients receiving longer dialysis sessions.

	$\hat{\lambda}$	95% CI
IV-MW	-0.015	(-0.028, -0.002)
IV-IPW	-0.006	(-0.018, 0.006)
IV-PSM	-0.015	(-0.028, 0.001)

The results in Table 2.6 suggest a small protective effect of longer dialysis sessions.

These IV results corroborate those found with the logistic regression models, although the estimated effects appear smaller and are insignificant for the IV-IPW and IV-PSM procedures. These estimates can be interpreted as follows; for $\hat{\lambda} = -0.015$, for example, we expect 1.5 less deaths in the first year for every 100 patients that could be encouraged to take long dialysis sessions. Note that IV-MW and IV-PSM gave similar results, with IV-MW obtaining a narrower confidence interval. This agrees with the idea that the IV-MW estimator is a more efficient approximation to the IV-PSM process.

2.5 Future Work

The IV-MW estimator proposed in this chapter was developed for use with binary outcomes. In this section, we present preliminary work on developing the IV-MW estimator for use with time-to-event or survival data. Given the importance of survival data in public health and medical studies, this work has the potential for broad applications in research. Few IV methods have been extended to survival data thus far. Terza et al. (2008) discuss the use of two stage residual inclusion, an extension of two stage least squares, with a Weibull regression for modeling survival data. Li et al. (2015) develop a two stage estimator of causal effects assuming an additive hazards model.

In this preliminary work, we modify the IV-MW procedure to estimate the difference in restricted mean survival between treatment groups in the presence of unmeasured confounding. We develop the method under both independent and informative censoring schemes. We report simulations to study its performance, and apply it to the data example of Section 2.4.

Restricted mean survival is the expected survival time for an individual over a fixed period of time. While we currently focus on estimating the difference in restricted mean survival, further work on estimating other parameters relevant to survival data is warranted.

Methods

As in Section 2.2.2, define the IV-matching weight for subject i as

$$W_i = \frac{min(e(\mathbf{x}_i), 1 - e(\mathbf{x}_i)}{Z_i e(\mathbf{x}_i) + (1 - Z_i)(1 - e(\mathbf{x}_i))}, \tag{2.15}$$

where $e(\mathbf{x}_i)$ is the IV propensity score conditional on covariates $\mathbf{x}_i$. We assume that we observe event times $Y_i = R_i \wedge C_i$, where R represents the time to response and C represents the time to censoring. Let $\delta_i = I(R_i < C_i)$ indicate observing the response for subject i, and let $t_1^z, t_2^z, ..., t_{k_z}^z$ be distinct event times for instrument groups $Z = 0, 1$. Finally, let D_i indicate treatment received for subject i.

IV-MW with Independent Censoring

We first present the estimator under the assumption that censoring times are independently distributed. Define the weighted number of observed responses in group Z at time t_j^z as

$$d_j^z = \sum_{i=1}^{n} W_i \delta_i I(Y_i = t_j^z, Z_i = z) \tag{2.16}$$

and the weighted number of individuals at risk in group Z at time t_j^z as

$$n_j^z = \sum_{i=1}^{n} W_i I(Y_i \geq t_j^z, Z_i = z). \tag{2.17}$$

Mean survival is then estimated as

$$\hat{\mu}_Y^z = \sum_{t=1}^{t_{k_z}} \hat{S}^z(t), \tag{2.18}$$

where

$$\hat{S}^z(t) = \prod_{j=1}^{t} \left(1 - \frac{d_j^z}{n_j^z}\right). \tag{2.19}$$

This is simply the area under a weighted, or adjusted, Kaplan-Meier curve (Kaplan and Meier, 1958; Xie and Liu, 2005), where each individual is weighted by their IV-matching weight, W_i.

Define the mean treatment usage in group Z as

$$\hat{\mu}_D^z = \frac{1}{n} \sum_{i=1}^{n} W_i D_i I(Z_i = z). \tag{2.20}$$

Finally, define the estimate of interest as

$$\hat{\Delta}_{\text{IV-MW}} = \frac{\hat{\mu}_Y^1 - \hat{\mu}_Y^0}{\hat{\mu}_D^1 - \hat{\mu}_D^0}. \tag{2.21}$$

Similar to the estimator in (2.1), this represents the ratio of the instrument's effect on survival to the instrument's effect on the treatment. Further research on this estimate and how the assumptions of Section 2.2.2 apply here is needed.

IV-MW with Informative Censoring

Assuming that censoring times are distributed independently of survival times is often unrealistic in practice. More often, there are covariates that affect both the censoring and survival time distributions. This is known as informative censoring, and if ignored can lead to biased effect estimates. Robins and Finkelstein (2000) proposed inverse probability of censoring weighting to handle informative censoring. In this section, we show how the estimator developed for independent censoring can be easily extended for use under informative censoring schemes by combining the IV-matching weight with inverse probability of censoring weights.

Assume that we measure covariates $\mathbf{x}_{\mathbf{c},i}$ that affect the censoring and survival time distributions. Define the inverse probability of censoring weights as

$$W_{i,j}^C = \frac{1}{\hat{G}_C(t_j)}, \tag{2.22}$$

where $\hat{G}_C(t_j)$ is a Cox regression model used to model the probability of being censored, given as

$$\hat{G}_C(t_j) = \hat{\lambda}_C(t_j)exp(\hat{\beta}'_C \mathbf{x}_{C,i}), \tag{2.23}$$

where $\hat{\lambda}_C(t_j)$ represents the baseline hazard at time t_j. We fit separate Cox regression models to the censoring times in the $Z = 0$ and $Z = 1$ groups. With these weights, we redefine the weighted number of observed responses and individuals at risk in group Z at time t^z_j as

$$d^z_j = \sum_{i=1}^{n} W_i W^C_{i,j} \delta_i I(Y_i = t^z_j, Z_i = z) \tag{2.24}$$

and

$$n^z_j = \sum_{i=1}^{n} W_i W^C_{i,j} I(Y_i \geq t^z_j, Z_i = z). \tag{2.25}$$

d^z_j and n^z_j in (2.24) and (2.25) are similar to those defined for the independent censoring case in (2.16) and (2.17), though here individuals are weighted by the product of their matching weight W_i and their censoring weight W^C_i, rather than only their matching weight, at each time point. With d^z_j and n^z_j appropriately redefined, $\hat{\mu}^z_Y, \hat{S}^z(t^z_j),$, $\hat{\mu}^z_D$, and $\hat{\Delta}_{\text{IV-MW}}$ are estimated as before.

Simulation

In this section we report simulation results to investigate the performance of the modified IV-MW estimator for estimating the difference in mean 5-year survival. We generate 1,000 datasets, each with n subjects. Instrument Z is generated from a Bernoulli distribution with

$$logit(p(Z = 1)) = \gamma + 0.25x_1 + 0.25x_2, \tag{2.26}$$

where γ, x_1 and x_2 are randomly generated from a $N(0,1)$ distribution. Covariates x_1 and x_2 represent instrument-outcome confounders and will be used in a logistic regression for modeling the IV propensity score. Treatment D is drawn from a Bernoulli distribution as well, with

$$logit(p(D=1)) = -0.5 + Z + 0.5x_u, \tag{2.27}$$

where $x_u \sim N(0,1)$ represents an unmeasured confounder between the treatment and the survival time.

We simulate data under both independent and informative censoring schemes. For independent censoring, censoring times are randomly drawn from a uniform distribution between 1 and 1,825 days (5 years) and survival times are are drawn from an exponential($\lambda/100$) distribution with

$$\lambda = -1.5 + \beta D + 0.25x_1 + 0.25x_2 + 0.5x_u. \tag{2.28}$$

For informative censoring, censoring times are randomly drawn from an exponential($\eta/100$) distribution, with

$$\eta = -3 + x_c, \tag{2.29}$$

where $x_c \sim N(0, \sigma_c^2)$. Survival times are then randomly drawn from an exponential($\lambda/100$) distribution with

$$\lambda = -1.5 + \beta D + 0.25x_1 + 0.25x_2 + 0.5x_u + 0.25x_c. \tag{2.30}$$

Approximately 25% of individuals are censored under these settings.

We set $\beta = 0$ throughout these simulations, so that the treatment has no effect on survival time. We compare estimation of the difference in 5-year restricted mean survival using the following four procedures:

IV-MW: Difference in restricted mean survival is calculated between instrument groups, $Z = 0, 1$, and measured instrument-outcome confounders x_1 and x_2 are adjusted for using the IV-matching weight procedure proposed in this chapter. This method is expected to overcome potential bias arising from both unmeasured confounder x_u and measured instrument-outcome confounders x_1 and x_2.

IV-PSM: Difference in restricted mean survival is calculated between instrument groups, $Z = 0, 1$, and measured instrument-outcome confounders x_1 and x_2 are adjusted for by first matching on the IV propensity score. Similar to IV-MW, this method is expected to overcome potential bias arising from both unmeasured confounder x_u and measured instrument-outcome confounders x_1 and x_2.

IV: Difference in restricted mean survival is calculated between instrument groups, $Z = 0, 1$. While this approach helps overcome bias arising from unmeasured confounder x_u, it will fail to adjust for measured instrument-outcome confounders x_1 and x_2. This violates the assumption that the instrument is randomly assigned, and therefore is expected to give biased estimates.

NAIVE: Difference in restricted mean survival is calculated between treatment groups, $D = 0, 1$. This naive approach is expected to give biased estimates due to unmeasured confounder x_u.

Inverse probability of censoring weights are applied to each procedure for the informative censoring scenario, with covariate x_c used in the Cox models for determining the censoring weights. Confidence intervals are obtained through bootstrapping, though further research related to variance estimation for the estimator is needed.

Simulation results in Table 2.7 show that the IV-MW estimator generally outperforms each of the remaining three estimators. IV-MW and IV-PSM maintained approximately nominal coverage rates, though IV-MW was less biased than IV-PSM

Table 2.7: Estimation of difference in 5-year restricted mean survival (in days) and 95% coverage probabilities for each method. Coverage probabilities are based on bootstrapped confidence intervals with 500 bootstrap iterations. In simulated data, there was no effect of treatment and no difference in mean survival is expected. Results are based on 1,000 simulations.

			IV-MW		IV-PSM		IV		NAIVE	
Censoring	Δ	n	$\hat{\Delta}$	CP	$\hat{\Delta}$	CP	$\hat{\Delta}$	CP	$\hat{\Delta}$	CP
Independent	0	500	-12	94.5	-18	95.5	-202	83.8	-96	53.6
		1,000	-1	95.2	-18	94.1	-192	76.3	-98	23.1
		2,000	-2	93.7	-20	94.2	-196	58.0	-98	4.5
		5,000	-2	94.3	-24	93.6	-194	21.2	-98	0.1
Informative	0	500	15	94.3	-2	93.5	-178	84.1	-98	48.1
		1,000	-9	94.5	-33	94.1	-201	73.2	-98	22.5
		2,000	2	95.1	-30	94.4	-190	56.3	-97	2.9
		5,000	1	95.2	-36	92.5	-192	22.8	-98	0.0

in seven of the eight scenarios compared. As expected, the IV and Naive procedures found severely biased estimates and poor coverage rates. No major differences are seen when comparing methods across the independent and informative censoring scenarios, suggesting that the inverse probability of censoring weights are able to overcome any problems arising from the informative censoring.

Data Example

We reexamine the data example of Section 2.4 using the methods of this section. Previously, we compared mortality and dialysis session length using death within the first year after initiating dialysis. In that analysis, we converted a survival outcome (time to death) into a binary outcome (death within one year). Using the methods of this section, we are able to study mortality and dialysis session length using the original time to death outcome. We censor all individuals at either one, two, or five years and estimate difference in restricted mean survival.

Results in Table 2.8 suggest that patients with longer dialysis sessions generally have longer one-, two-, and five-year restricted mean survival times. The relatively large estimates for the NAIVE procedure could be a result of unmeasured differences

Table 2.8: Estimated difference in restricted mean survival (in days) between patients with longer versus shorter dialysis sessions lengths. 95% confidence intervals were obtained through bootstrapping with 500 bootstrap iterations.

	1 Year		2 Year		5 Year	
Model	$\hat{\Delta}$	(95% CI)	$\hat{\Delta}$	(95% CI)	$\hat{\Delta}$	(95% CI)
IV-MW	5.2	(2.2, 8.2)	11.7	(3.3, 19.0)	28.9	(0.8, 53.7)
IV-PSM	5.6	(2.4, 9.4)	11.4	(3.6, 19.7)	27.1	(-2.8, 53.9)
IV	2.2	(-0.2, 4.4)	2.9	(-2.6, 9.2)	-1.3	(-21.5, 14.4)
NAIVE	6.2	(5.5, 7.1)	17.0	(15.0, 19.0)	71.7	(65.4, 77.7)

between patients receiving longer or shorter dialysis sessions. The three IV models found greatly attenuated effects, and in some cases estimates that are statistically insignificant at the 0.05 level. The similar estimates for the IV-MW and IV-PSM procedures, with IV-MW finding narrower 95% confidence intervals, confirms that IV-MW is a more efficient approximation to the IV-PSM process. Qualitatively, the results Table 2.8 agree with those of Section 2.4, which also suggested that patients with longer dialysis sessions had decreased mortality.

2.6 Discussion

A key assumption in instrumental variable analyses is that the instrument is randomly assigned, which requires that there are no unmeasured confounders between the instrument and the outcome. Unfortunately, unless the instrument is based on actual randomization, this assumption is unlikely to hold without conditioning on a set of known, measured confounders. The researcher must then argue that the instrument is distributed as good as random after controlling for these instrument-outcome confounders. Garabedian et al. (2014) emphasize that the most commonly used instruments have potential instrument-outcome confounders associated with them and that failing to adjust for these confounders can bias estimation.

In this work we developed a weighted IV estimator based on the IV propensity score to adjust for instrument-outcome confounders. The weights reflected the probability that an individual would be selected into a one-to-one IV propensity score match, and modified weights for approximating k:1 matching designs were provided as well. We further presented a double robust version of the estimator that protects against misspecification of the IV propensity score model. Though this required the additional specification of an outcome model, the double robust estimator provided consistent estimates if only one of the outcome or IV propensity score models was correctly specified, but did not require both to be correct.

One-to-one IV propensity score matching involves pairing each encouraged subject to an unencouraged subject with a similar IV propensity scores, often within a specified range. If a match cannot be found within this range, pairing the encouraged subject with an unencouraged subject with a substantially different IV propensity score can bias estimation, whereas removing that subject from the analysis leads a loss of efficiency. The proposed estimator avoids these pitfalls, leading to more efficient estimation as every individual contributes. Additional benefits over matching include straightforward variance estimation and computational efficiency. Through simulation, the proposed estimator was found to outperform alternatives, being equally unbiased with uniformly smaller mean squared errors.

Preliminary work related to extending the IV-MW procedure of this chapter for use with time-to-event or survival data suggested that the IV-MW procedure is useful for obtaining consistent survival estimates in the presence of unmeasured confounding. We discussed estimating the difference in restricted mean survival using the IV-MW procedure and showed that it performed well in simulation. Further work on this topic includes understanding how the assumptions of Section 2.2.2 apply in this context, how to accurately estimate variance of the estimator, and the possibility of estimating other parameters relevant to survival studies. Given the importance of survival data in public health and medical research, extending the IV-MW estimator for use with these data can have broad applications.

Implementation of these methods were illustrated using USRDS data to study the association between dialysis session length and first year mortality among hemodialysis patients in the United States. While longer dialysis sessions are thought to decrease risk of mortality, it is a difficult research question as the relationship between session length and mortality is likely confounded, as smaller patients with higher mortality risk are more likely to be prescribed shorter dialysis sessions. This might explain the lack of consensus among previous studies. Using the IV methods of this article, a small protective effect of longer dialysis sessions was found, suggesting 1.5 fewer first year deaths for every 100 dialysis patients encouraged to switch from shorter to longer dialysis sessions. These findings corroborate the findings from covariate adjusted logistic regression models, although estimates were smaller and insignificant for some IV models. Applying the IV-MW estimator modified for use with survival data similarly suggested that patients with longer dialysis sessions had longer one-, two-, and five-year restricted mean survival times.

In the next chapter we focus on the strength of correlation between the instrument and the treatment, and propose a method for increasing this strength. Instruments with little influence over the treatment are termed weak instruments, and there are several problems associated with them. They suffer from greater finite-sample bias and variability in estimation. Additionally, results obtained using weak instruments are less robust to unmeasured instrument-outcome confounders (Bound et al., 1995; Small and Rosenbaum, 2008). These benefits have motivated methods for increasing the strength of instrumental variables.

CHAPTER III

Strengthening Instrumental Variables Through Weighting

3.1 Motivation

The most common instrumental variables have potential confounders associated with them that, when left unmeasured, violate the assumption that the instrument is randomly assigned and can bias estimation. In the previous chapter we developed an IV estimator that adjusted for measured confounders between the instrument and the outcome. Adjusting for these confounders helps to argue that the instrument is conditionally distributed “as good as random.”

Unfortunately, this assumption cannot be verified to hold and is likely to be criticized even after controlling for a number of measured covariates. It is therefore good to have results that are sufficiently robust to violations of it. One approach to obtaining more robust results is to work with stronger instruments. The strength of the instrument refers to the strength of the relationship between the instrument and the treatment. Instruments that have low correlation with the treatment are referred to as weak instruments and are known to have poor properties (Bound et al., 1995; Small and Rosenbaum, 2008). They have more finite-sample bias and greater variability in estimation. Additionally, results obtained using weak instruments are particularly sensitive to violations of the assumption that the instrument is randomly assigned. These benefits have motivated recent methods for increasing the strength of instrumental variables (Baiocchi et al., 2010, 2012). In this chapter, we propose a novel weighting procedure for strengthening instrumental variables.

The literature relating to weak instrumental variables has primarily focused on detailing the problems and limitations associated with using them. See, for example, Bound et al. (1995), Staiger and Stock (1994), Angrist et al. (1996), Small and Rosenbaum (2008) or Baiocchi et al. (2014). Variable selection methods to select a strong subset among a pool of weak instruments have been proposed in Belloni et al. (2010), Caner and Fan (2010), and Belloni et al. (2012). For working with a single weak instrument, Baiocchi et al. (2010) proposed near-far matching, a novel method to extract a smaller study with a stronger instrument from a larger study (see also Baiocchi et al. (2012) or Zubizarreta et al. (2013)). This matching-based IV methodology aims to construct pairs that are "near" on covariates but "far" in the instrument. In other words, pairs consist of subjects with similar characteristics who have received substantially different amounts of encouragement toward the treatment, with a greater difference indicating a stronger instrument. This difference is increased in the near-far matching procedure through the use of penalties to discourage individuals with similar instrument values from pairing, while allowing a certain number of individuals to be removed from the analysis entirely. This results in a stronger instrument across a smaller number of pairs. One limitation of near-far matching is that it may strengthen the instrument at the cost of match quality.

We propose weighted IV-matching, an alternative for strengthening the instrument within this IV-matching framework. Rather than using penalties to discourage individuals that receive similar encouragement from pairing, we propose strengthening the instrument after matches have been formed through weighting, with a pair's weight being a function of the instrument within that pair. A fundamental difference between these two techniques is the stage at which the instrument is strengthened. Weighted IV-matching strengthens the instrument after matches have been formed, allowing the matching algorithm to focus on creating good matches with similar covariate values. Near-far matching, on the other hand, strengthens the instrument and matches on covariates simultaneously, requiring the algorithm to share priority between these two goals. This generally leads to better quality matches for weighted IV-matching, a major benefit since failing to properly match on important covariates may lead to bias in estimation.

We illustrate these methods with a comparison of hemodialysis (HD) and peritoneal dialysis (PD) on six-month mortality among patients with end stage renal disease (ESRD) using data from the United States Renal Data System (USRDS). PD has several benefits over HD, including cost benefits, an improved quality of life, and the preservation of residual renal function (Marrón et al., 2008; Tam, 2009; Goodlad and Brown, 2013). Despite this, PD remains underutilized in the United States (Jiwakanon et al., 2010). One explanation for this may be a lack of consensus regarding the effect of PD on patient survival. A randomized trial investigating this question was stopped early due to insufficient enrollment (Korevaar et al., 2003). Many observational studies have suggested that PD is associated with decreased mortality, though results are often conflicting (Heaf et al., 2002; Vonesh et al., 2006; Weinhandl et al., 2010; Mehrotra et al., 2011; Kim et al., 2014; Kumar et al., 2014). Complicating comparisons of HD and PD patients is a strong selection bias, with PD patients tending to be younger and healthier than HD patients. Studies have dealt with this issue by measuring and controlling for important confounders, but to our knowledge none have addressed the possibility of unmeasured confounding that likely remains. We define PD as the treatment and consider a binary outcome for six-month survival. The focus on six-month survival is to study the influence of initial dialysis modality on early mortality, which tends to be high for dialysis patients. Studying early mortality can provide guidance for selecting the initial dialysis modality in order to reduce this early mortality. See, for example, Noordzij and Jager (2012), Sinnakirouchenan and Holley (2011), or Heaf et al. (2002).

A possible instrument in the data is the mean PD usage at the facility level. Instruments based on mean treatment usage in a geographic region, facility, or other group are often called preference-based instruments (Brookhart and Schneeweiss, 2007; Li et al., 2015), because it is believed that these groups may have preferences that at least partially override both measured and unmeasured patient characteristics when making treatment decisions. In other words, facilities with high PD usage are more likely to "encourage" their patients towards PD than those with low usage. Preference-based instruments are among the most commonly used instruments in practice (Garabedian et al., 2014), and methods to improve them may have broad applications.

The remainder of this chapter is organized as follows. In Section 3.2 we outline the proposed weighted IV-matching procedure and briefly compare it to near-far matching. Inference and sensitivity are discussed in Section 3.3. The finite sample performance of these methods are compared in Section 3.4 through simulation, and they are illustrated with a data analysis in Section 3.5. We conclude with a discussion in Section 3.6.

3.2 Methods

We begin this section with an outline of the IV matching framework presented in Baiocchi et al. (2010, 2012) and then propose a weighting procedure for strengthening the instrument within this framework. We briefly compare the proposed weighting procedure with near-far matching and highlight key differences.

With a preference-based instrument, two rounds of matching are implemented (Baiocchi et al., 2012). In the context of our motivating data example, an optimal non-bipartite matching algorithm first pairs facilities (Derigs, 1988; Lu et al., 2011). After facilities have been paired, the instrument is dichotomized into encouraging and unencouraging. This is done by comparing instrument values within each facility pair and considering the facility with the higher value to be an encouraging facility and the other to be an unencouraging facility. An optimal bipartite matching algorithm then pairs patients at the PD encouraging facility with patients in the other. This results in I pairs of two subjects with similar patient and facility characteristics that received different levels of encouragement toward PD. Instrument strength can be assessed by the average difference, or separation, of this encouragement across pairs. For example, the instrument is considered stronger in a study in which the average encouraged and unencouraged subjects were treated at facilities with 85% and 30% treatment usage compared to one with average treatment usage of 60% and 45%.

Creating a stronger instrument in this framework is thus equivalent to increasing this separation. We propose increasing this separation by assigning more weight to pairs that are more influenced by the instrument. Specifically, we propose weighting by the probability that the encouraged subject receives the treatment while the unencouraged subject receives the control. This can be thought of as the probability that a pair "complies" with encouragement, and giving more weight to pairs more likely to comply creates a stronger instrument across all pairs. Without loss of generality, assume that subject j in pair i was treated at the encouraging facility and subject j' at the unencouraging facility, with $Z_{ij} = 1$ indicating encouragement and $Z_{ij'} = 0$ indicating unencouragement. Let D_{ij} indicate treatment received. The weight for pair i is then defined as

$$w_i = P(D_{ij} = 1|Z_{ij} = 1)P(D_{ij'} = 0|Z_{ij'} = 0). \tag{3.1}$$

Similar to separation of the instrument, this probability is a measure of instrument strength, though rather than an average across all pairs it is a measure of the influence of the instrument within pair i. A stronger instrument is created when more weight is given to pairs in which the instrument has more influence over treatment. This has the effect of redistributing the data in a way that highlights "good" pairs that are more influenced by the instrument and increasing separation of the instrument in the process.

In practice, the probabilities in equation (3.1) are unlikely to be known but will need to be estimated. Using facility level mean PD usage as the instrument, $P(D_{ij} = 1|Z_{ij} = 1)$ is estimated by the mean PD usage at the encouraging facility, while $P(D_{ij'} = 0|Z_{ij'} = 0)$ is estimated with one minus the mean PD usage at the unencouraging facility. Weights can be standardized to maintain the effective sample size and statistical power if necessary.

The near-far matching procedure of Baiocchi et al. (2010, 2012) forces separation of the instrument in the matching process. This is done in the first round by adding a penalty to the distance measure between facilities whose instrument values are within a certain threshold, and allowing a certain number to be removed. This requires the matching algorithm to pair facilities with similar covariates and enforce separation of encouragement simultaneously, and generates an implicit tradeoff. A large penalty will dominate the distance used to reflect similarity on covariates, thereby increasing instrument separation but at the expense of match quality, whereas a small penalty may get overshadowed by the covariate distance, leading to better matches, but with less separation. Removing a number of facilities serves to alleviate some of the damage to match quality that arises when requiring the matching algorithm to share priority between creating good matches and enforcing instrument separation.

A fundamental difference between weighted IV-matching and near-far matching is the stage in which the instrument is strengthened. Weighted IV-matching strengthens the instrument after matches have been formed, which allows the matching algorithm to focus solely on creating good matches with similar covariate values. Near-far matching, on the other hand, strengthens the instrument in the matching process, forcing the algorithm to balance creating good matches and enforcing separation of the instrument. This difference highlights a theme that we will see when comparing the performance of these two methods; in a tradeoff between match quality and instrument strength, weighted IV-matching tends to favor match quality while near-far matching tends to favor instrument strength. Strength in either of these areas has implications on the resulting analysis.

3.3 Inference

3.3.1 Notation

We return to the potential outcomes notation presented in Section 2.2 of Chapter II for defining causal effects. Let $Z_{ij} = 1$ if subject j in pair i is encouraged toward treatment, $Z_{ij} = 0$ otherwise. Let $D_{ij}(Z_{ij})$ indicate treatment received for subject j in pair i given their encouragement, and let $Y_{ij}(Z_{ij}, D_{ij}(Z_{ij}))$ indicate mortality. $D_{ij}(Z_{ij})$ and $Y_{ij}(Z_{ij}, D_{ij}(Z_{ij}))$ are referred to as a subjects “potential outcomes.”

For encouraged subjects, with $Z_{ij} = 1$, we observe treatment $D_{ij}(1)$ and response $Y_{ij}(1, D_{ij}(1))$. Similarly for unencouraged subjects, we observe $D_{ij}(0)$ and response $Y_{ij}(0, D_{ij}(0))$. Our interest lies in estimating the parameter

$$\lambda = \frac{\sum_i \sum_j (Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0)))}{\sum_i \sum_j (D_{ij}(1) - D_{ij}(0))}. \tag{3.2}$$

This parameter is often referred to as the local average treatment effect (Imbens and Angrist, 1994; Angrist et al., 1996). In contrast to an average treatment effect, which is applicable to the entire population, the local effect is interpreted as an average treatment effect among a subgroup of the population known as "compliers." Depicted in Table 2.1, compliers are individuals that will take the treatment that they are encouraged to take.

3.3.2 Assumptions

Unfortunately, subjects are never observed under both states of encouragement, and we thus never observe both $Y_{ij}(1, D_{ij}(1))$ and $Y_{ij}(0, D_{ij}(0))$ or both $D_{ij}(1)$ and $D_{ij}(0)$ and must estimate λ from the data. We impose the following five assumptions to aid us in estimation (Angrist et al., 1996; Baiocchi et al., 2014). We list them briefly here, but a more thorough discussion of these assumptions can be found in Section 2.2.

A1. Stable Unit Treatment Value Assumption (SUTVA).

A2. Random assignment of the instrument.

A3. Exclusion Restriction.

A4. Nonzero association between instrument and treatment.

A5. Monotonicity.

Assumptions *A1* and *A2* allow for unbiased estimation of λ. Adding assumptions it A3-A5 restricts the applicability of λ to the subgroup of compliers (Table 2.1).

Further discussion of these assumptions can be found in Section 2.2, as well as Imbens and Angrist (1994), Angrist et al. (1996), Baiocchi et al. (2014).

3.3.3 Estimation

Denote the observed response and treatment for subject j in pair i as $Y_{ij} = Z_{ij}Y(1, D_{ij}(1)) + (1 - Z_{ij})Y(0, D_{ij}(0))$ and $D_{ij} = Z_{ij}D_{ij}(1) + (1 - Z_{ij})D_{ij}(0)$, respecively. Estimate λ as

$$\hat{\lambda} = \frac{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} [Z_{ij}Y_{ij} - (1 - Z_{ij})Y_{ij}]}{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} [Z_{ij}D_{ij} - (1 - Z_{ij})D_{ij}]}. \tag{3.3}$$

For inferences regarding λ, Baiocchi et al. (2010) developed an asymptotically valid test for the null hypothesis $H_0^{(\lambda)}$. $H_0^{(\lambda)}$ is true under many population distributions, and therefore is a composite null hypothesis. The size of a test for a composite null is the supremum over all null hypotheses in the composite null, and a test is considered valid if it has size less than or equal to its nominal level. Using statistics

$$\begin{aligned} T(\lambda_0) &= \frac{1}{I}\sum_{i=1}^{I} \hat{w}_i \left[\sum_{j=1}^{2} Z_{ij}(Y_{ij} - \lambda_0 D_{ij}) - \sum_{j=1}^{2}(1 - Z_{ij})(Y_{ij} - \lambda_0 D_{ij})\right] \\ &= \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0) \end{aligned}$$

and

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}[V_i(\lambda_0) - T(\lambda_0)]^2,$$

we can test $H_0^{(\lambda)}$ by comparing $T(\lambda_0)/S(\lambda_0)$ to a standard normal cumulative distribution for large I. Inverting this test and solving for $T(\lambda_0)/S(\lambda_0) = 0$ and $\pm$ 1.96 provides an estimate and 95% confidence interval for λ. A detailed discussion of this statistic, its distribution, and related issues can be found in Baiocchi et al. (2010).

This inference procedure provides a confidence interval for the estimate but unfortunately it does not provide a standard error. To obtain a standard error estimate, we implement the sandwich variance procedure of Section 2.2. In Sections 3.4 and 3.5, intervals and coverage results will be based on the permutation inference procedure.

3.4 Simulation

In this section we compare the finite sample performance of three IV-matching techniques through simulation. The standard IV-match (IVM) uses the full data and makes no attempt to strengthen the instrument, while weighted IV-matching (WIVM) and near-far matching (NFM) will strengthen the instrument as described in section 3.2. For the NFM procedure, we add a penalty to the distance between facilities if their instruments are within a distance equal to the interquartile range of instrument values. As in Baiocchi et al. (2010), we specify a penalty function that begins at 0 and increases exponentially a pairs instrument values become closer, and allow 50% of facilities to be removed during the matching process.

3.4.1 Setup

One thousand datasets are generated containing $i = 1, ..., 200$ facilities with $j = 1, ..., 40$ subjects at each. Binary treatment D and binary outcome Y are randomly assigned with

$$P(D_{ij} = 1) = logit^{-1}(\gamma_i + \alpha X_{1,i} + \delta X_{2,ij} + \nu_{ij}), \tag{3.4}$$

$$P(Y_{ij} = 1) = logit^{-1}(\beta D_{ij} + \alpha X_{1,i} + \delta X_{2,ij} + \epsilon_{ij}). \tag{3.5}$$

$\gamma_i \sim N(0, 1)$ represents a facility effect. Standard normal covariates $X_{1,i}$ and $X_{2,ij}$ represent observed confounders and are used for matching. $X_{1,i}$ is a facility level confounder and $X_{2,ij}$ is a patient level confounder. Coefficients α, δ, and β represent the effects of X_1, X_2, and D, respectively. Unobserved confounding is created by generating $(\nu_{ij}, \epsilon_{ij})$ as bivariate normal with correlation $\rho = .75$. The proportion of treated individuals at each facilities serves as the instrument.

To obtain the "true" local average treatment effect that we wish to estimate, or λ in (3.2), we need counterfactual treatments and responses for every individual. These are not easily obtained under the current setup, since γ, not encouragement, is in equation (3.4). Furthermore, we do not know which counterfactual state an individual will be considered to have been observed in until after matching, since subjects are determined to have been observed in an encouraging or unencouraging facility by comparing instrument values within pairs. Despite this caveat, suitable counterfactuals can be obtained in the following way.

Consider patients treated at facilities with $\gamma_i > 0$ to be observed in the encouragement state, while those at facilities with $\gamma_i \leq 0$ to be observed in the unencouragement state. For individuals in the encouragement state, we have $D_{ij} = D_{ij}(1)$ and $Y_{ij} = Y_{ij}(1, D_{ij}(1))$ from Equations (3.4) and (3.5). For counterfactuals, sample a γ from the unencouragement group and denote it γ^*. $D_{ij}(0)$ is then obtained using equation (3.4) with $P(D_{ij} = 1) = P(D_{ij}(0) = 1) = logit^{-1}(\gamma_i^* + \alpha X_{1,i} + \delta X_{2,ij} + \nu_{ij})$ and $Y_{ij}(0, D_{ij}(0))$ is obtained using equation (3.5) with $P(Y_{ij} = 1) = P(Y_{ij}(0, D_{ij}(0))) = logit^{-1}(\beta D_{ij}(0) + \alpha X_{1,i} + \delta X_{2,ij} + \epsilon_{ij})$. Counterfactuals for patients observed in the unencouragement state can be obtained similarly. After obtaining $D_{ij}(1)$, $D_{ij}(0)$, $Y_{ij}(1, D_{ij}(1))$, and $Y_{ij}(0, D_{ij}(0))$, these are plugged into Equation (3.2) for the true effect, λ.

3.4.2 Results

Instrument Strength

The present work is motivated by the desire to strengthen the instrument by increasing the separation of encouragement within pairs. Table 3.1 shows that both

WIVM and NFM were able to do so, increasing the standardized difference in encouragement by approximately 25% and 65%, respectively. All things being equal, the stronger instrument is preferred. Looking at match quality in the next section, however, we will see that all things are not equal.

Table 3.1: Separation of encouragement within pairs. Reported is the mean treatment usage at unencouraging facilities ($\bar{Z}_U$), encouraging facilities ($\bar{Z}_E$), and the standardized difference between them, calculated as St Diff $= 100(\bar{Z}_E - \bar{Z}_U)/\sqrt{.5(s^2_{Z_E} + s^2_{Z_U})}$, where $s^2_{Z_E}$ and $s^2_{Z_U}$ are sample variances of mean treatment usage in each group. Results are based on 1,000 simulations.

	$(\bar{Z}_U, \bar{Z}_E)$	St Diff
IVM	(37%, 62%)	141
WIVM	(35%, 65%)	175
NFM	(30%, 70%)	232

Match Quality

Table 3.2 reports balance of covariates X_1 and X_2 as indicated by the standardized difference within pairs. The WIVM procedure produced consistently better covariate balance than the NFM procedure. The particularly poor balance of facility level X_1 under the NFM procedure shows that introducing penalties to the match negatively impacted the ability to properly match on X_1 in the first round.

Table 3.2: Covariate balance as reported by the standardized differences in covariates X_1 and X_2 within pairs. Results based on 1,000 simulations.

	α	δ	IVM	WIVM	NFM
X_1	0	0	0.01	0.01	0.34
	0.25	0.25	0.15	0.14	18.01
	0.50	0.50	0.14	0.16	36.10
X_2	0	0	0.01	0.02	0.10
	0.25	0.25	0.58	0.68	1.02
	0.50	0.50	1.35	1.57	2.10

The pattern seen in Tables 3.1 and 3.2 shows a tradeoff of instrument strength and match quality between WIVM and NFM. WIVM allows the matching algorithm to focus entirely on matching on covariates, and strengthens the instrument through weighting after the matches have been formed. NFM, on the other hand, incorporates penalties into the match to enforce separation of the instrument, requiring the matching algorithm to share priority between matching on covariates and strengthening the instrument. A large penalty might dominate the distance used for matching and diminish the ability to properly match on covariates. In the tradeoff between instrument strength and match quality, WIVM is willing to trade less instrument strength for higher quality matches, while NFM is willing to trade lower quality matches for a stronger instrument.

Estimation and Coverage

Table 3.3 presents estimation and coverage results under increasing magnitudes of observed confounding. When α and δ are zero and matching on X_1 and X_2 is trivial, each method is nearly unbiased and maintains nominal coverage. WIVM and NFM achieved lower mean squared error than IVM, which is one benefit associated with stronger instruments (Wooldridge, 2001). As α and δ increase and matching on X_1 and X_2 becomes more important, the performance of IVM and WIVM remain mostly unchanged. NFM, on the other hand, sees a large increase in bias and mean squared errors and low coverage rates. The deterioration of performance for NFM as α and δ increase can be attributed to its inability to properly match on X_1.

Table 3.3: Bias, mean squared error (MSE), and 95% coverage probabilities (CP) for estimation of λ. Bias and MSE are multiplied by 1,000. Coverage probabilities are based on confidence intervals obtained using the permutation inference procedure Section 3.3 and Baiocchi et al. (2010). Results are based on 1,000 simulations.

					IVM			WIVM			NFM	
α	δ	β	λ	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP
0	0	0.0	0.0	4.6	2.3	94.3	4.5	1.6	93.9	2.5	1.7	94.2
		0.6	0.14	1.4	2.0	94.3	1.3	1.4	95.1	0.3	1.4	95.6
		1.0	0.23	4.6	1.9	94.6	3.7	1.4	95.2	2.5	1.4	95.0
0.25	0.25	0.0	0.0	3.0	2.1	94.8	4.9	1.5	94.6	29.2	2.5	84.9
		0.6	0.14	4.9	1.9	95.2	4.9	1.5	95.2	25.8	2.3	87.6
		1.0	0.23	4.4	1.8	95.0	4.7	1.3	96.0	26.4	2.2	86.5
0.50	0.50	0.0	0.0	8.7	2.4	93.6	9.7	1.7	94.2	93.9	10.5	28.3
		0.6	0.14	8.7	2.3	94.3	7.9	1.7	93.6	88.6	9.7	30.9
		1.0	0.23	4.0	2.0	93.7	3.7	1.5	93.8	78.6	7.8	38.3

3.5 Data Example

In this section we illustrate the use of IV-matching (IVM), weighted IV-matching (WIVM), and near-far matching (NFM) with a study comparing mortality in the first six months between patients receiving hemodialysis (HD) or peritoneal dialysis (PD) as treatment for end stage renal disease. Complete information on 164,195 adults initiating dialysis for the first time between January 1, 2010 and December 31, 2013 was obtained from the United States Renal Data System. The analysis was restricted to patients being treated at dialysis facilities with at least ten patients that used both HD and PD during the study period. The analysis was conducted as intention-to-treat, with treatment defined as the modality prescribed at the onset of dialysis.

The instrument, facility mean PD usage, was calculated using data from 2007-2009 to avoid correlation with patient level confounders in the 2010-2013 period that was used for the analysis. The instrument varied greatly across facilities, ranging from 0 to 100% with a mean of 9.8%. The correlation coefficient between a facilities 2007-2009 and 2010-2013 PD usage was 0.68, indicating that facility preferences toward PD are relatively stable through time, and that a facilities 2007-2009 PD usage is a useful predictor of their 2010-2013 usage.

Figure 3.1 and Table B.1 of Appendix B confirm the belief that patients treated with PD are generally healthier than those treated with HD. On average, they are six years younger, receive more pre-ESRD care, suffer from less comorbidities, and are more likely to be employed than HD patients. Additionally, facilities with higher PD usage tend to be larger, as indicated by the higher number of nurses, social workers, and hemodialysis stations. Since these factors could be related to unmeasured confounders that affect patient outcomes, it is important to control for these variables when matching.

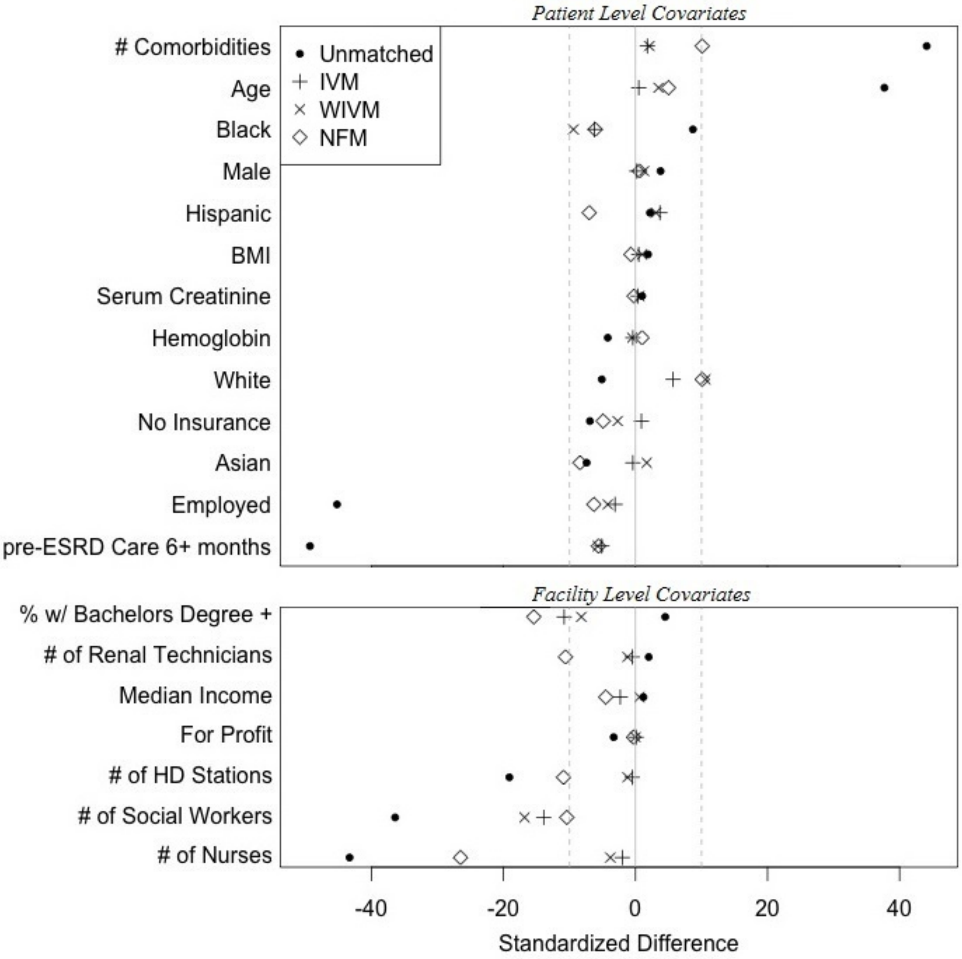
We follow the two round matching procedure described in Section 3.2 for constructing matches. An optimal non-bipartite match first pairs facilities based on facility level covariates. The facility in each pair with the greater mean PD usage is considered to be an encouraging facility, while the other is considered unencouraging. Within each of these pairs, an optimal bipartite match then pairs patients from the encouraging facility with patients in the unencouraging facility.

For the first round facility level match, we defined the distance between facilities using a Mahalanobis distance based on the facility covariates in Figure 3.1 and Tables B.1 or B.2 of Appendix B. For the NFM procedure, a penalty was added to this distance if facilities instrument values were within 14% of each other (the interquartile range), and 50% of facilities were allowed to pair with sinks and be removed from the analysis. For the second round patient level match, we matched on a prognostic score based on the patient level covariates in Figure 3.1 and Tables B.1 or B.2 of Appendix B. For the WIVM procedure, a weight was assigned to each pair based on Equation (3.1), where probabilities were estimated using the instrument, facility mean PD usage from 2007-2009.

Of the 164,195 patients, 128,700 were paired using the IVM and WIVM procedure, while 67,904 were paired using the NFM procedure. The average unencouraged and encouraged patient was treated at a facility with PD usage from 2007-2009 of 4.7% and 15.3% using the IVM procedure, 6.3% and 27.8% using the WIVM procedure, and 3.8% and 25.3% using the NFM procedure. For WIVM and NFM, the increased separation corresponds with roughly a 100% increase in the standardized difference in encouragement, with neither procedure performing notably better than the other in terms of instrument strength.

Covariate balance after matching is presented in Figure 3.1 as well as Table B.2 of Appendix B. Covariate balance is improved compared with pre-matching under each of the three methods. IVM and WIVM, however, generally resulted in better balance than NFM, particularly for facility level covariates where NFM seems to struggle. These results are similar to those seen in the simulations of Section 3.4.

Figure 3.1: Covariate balance before and after matching as indicated by the standardized differences within pairs. Dashed grey lines are at ±10. Standardized differences larger than this have been suggested to represent an imbalance (Normand et al., 2001).



Estimation results reported in Table 3.4 indicate that PD has a protective effect on mortality in the first six months. For example, $\hat{\lambda} = -0.09$ suggests that for every 100 subjects that are encouraged to switch from HD to PD, there are nine fewer

Table 3.4: Estimate and 95% confidence interval for the local average treatment effect. Estimated effect represents the expected decrease in death for patients that could be encouraged to switch from hemodialysis to peritoneal dialysis.

	$\hat{\lambda}$	95% CI
IVM	-0.09	(-0.14, 0.03)
WIVM	-0.09	(-0.15, -0.06)
NFM	-0.07	(-0.10, -0.04)

deaths in the first six months. Both WIVM and NFM decreased the width of the confidence interval associated with λ compared to IVM, with NFM leading to the narrowest interval.

3.6 Discussion

Weak instrumental variables present many problems to an IV analysis, including greater finite-sample bias and greater variability in estimation. Results obtained using weak instruments have also been shown to be less robust to unmeasured instrument-outcome confounders that violate the assumption that the instrument is randomly assigned (Bound et al., 1995; Small and Rosenbaum, 2008). These benefits have motivated recent methods for strengthening the instrument (Baiocchi et al., 2010, 2012).

In this chapter, we proposed a weighting procedure for building a stronger instrument in the IV-matching framework (Baiocchi et al., 2010, 2012). The key idea is that the data can be redistributed through weighting to highlight pairs in a way that increases the strength of the instrument. The proposed weights were based on the probability that a pair complies with encouragement, or that the encouraged subject in a pair receives the treatment while the unencouraged subject receives the control. Other weights could be considered as long as more weight is assigned to pairs that are more influenced by the instrument. In future work we are considering the possibility of an “optimal” weight, perhaps subject to a constraint on covariate balance.

Compared with existing methods for strengthening the instrument, weighting is able to strengthen the instrument without compromising match quality. This is because weights are applied after matches have been formed, as opposed to methods that strengthen the instrument simultaneously with matching. This is a major strength of the proposed method since failing to properly match on important covariates can bias estimation.

Using data from the United States Renal Data System, methods were illustrated in a study comparing mortality in the first six months between patients receiving hemodialysis or peritoneal dialysis as treatment for end state renal disease. The proposed weighting procedure was able to strengthen the instrument while maintaining good match quality. A protective effect of peritoneal dialysis was found, suggesting that we expect nine fewer deaths for every 100 patients that could be encouraged to switch from hemodialysis to peritoneal dialysis.

The methods discussed in this chapter have been motivated by the desire to achieve the benefits associated with stronger instruments. These include decreased finite-sample bias, greater efficiency, and improved robustness to violations of the assumption that the instrument is randomly assigned. It has yet to be shown, however, that a strengthened instrument can provide the same benefits as an instrument that naturally has high correlation with the treatment. We undertake this task in the next chapter, where we investigate the performance of strengthened instrumental variables to better understand their properties.

CHAPTER IV

Properties of Strengthened Instrumental Variables

4.1 Motivation

In the previous chapter we discussed methods for increasing the strength of an instrumental variable. Methods for strengthening the instrument have been motivated by the desire to capitalize on the advantages of using stronger instruments. These advantages include decreased finite-sample bias and greater efficiency in estimation (Bound et al., 1995; Angrist et al., 1996; Wooldridge, 2001). Additionally, results obtained using stronger instruments are more robust to violations of the assumption that the instrument is randomly assigned (Small and Rosenbaum, 2008; Baiocchi et al., 2010). It has yet to be shown, however, that strengthened instruments provide the same benefits as instruments that are naturally stronger. We use the term "strengthened" to refer to an instrument whose correlation with the treatment has been increased by the researcher and the term "naturally stronger" to refer to an instrument that is more highly correlated with the treatment without any effort made by the researcher to increase this correlation. The focus of this chapter is to study estimation with strengthened instruments to better understand their properties and how they compare with instruments that are naturally stronger.

To illustrate the benefits of using stronger instruments, suppose that we wish to estimate β in the linear model

$$Y = \beta_0 + \beta D + \epsilon. \tag{4.1}$$

If any unmeasured confounders exist between the treatment D and outcome Y they will go into the error term, ϵ. This induces correlation between the treatment and the error term, which then biases estimates of β when (4.1) is modeled directly. This bias arising from unmeasured treatment-outcome confounding is the problem an IV analysis hopes to overcome.

Now assume that we have an instrument, Z, that is related to the treatment as

$$D = \psi_0 + \psi_1 Z + \nu, \tag{4.2}$$

where $\psi_1 \neq 0$ by assumption. Given a random sample $\{(D_i, Y_i, Z_i) : i = 1, 2, ..., n\}$, the IV estimator of β is defined as

$$\hat{\beta}_{IV} = \left(n^{-1} \sum_{i=1}^{n} Z_i D_i \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} Z_i Y_i \right). \tag{4.3}$$

From Bound et al. (1995) or Wooldridge (2001), the probability limit of this estimator can be written as

$$\text{plim}(\hat{\beta}_{IV}) = \beta + \frac{Corr(Z, \epsilon)}{Corr(Z, D)} \frac{\sigma_\epsilon}{\sigma_D}, \tag{4.4}$$

where σ_ϵ and σ_D are the standard deviations of the error and the treatment, respectively. Equation (4.4) shows that correlation between the instrument Z and error term ϵ biases the IV estimate of β.

An important insight from (4.4) is that stronger instruments, or those that have greater correlation with the treatment, lessen the damage to estimation caused by correlation between Z and ϵ. Correlation between Z and ϵ can arise in two ways. It can be a finite-sample issue, since correlation is never exactly zero in any finite-sample. It can also result from unmeasured instrument-outcome confounders. Just as unmeasured treatment-outcome confounders induce correlation between D and ϵ that biases estimates of β when (4.1) is modeled directly, unmeasured instrument-outcome confounders induce correlation between Z and ϵ that biases IV estimates of β. Stronger instruments therefore decrease finite-sample bias and increase robustness unmeasured instrument-outcome confounders that violate the assumption that the instrument is randomly assigned (Bound et al., 1995; Angrist et al., 1996).

The improved robustness to unmeasured instrument-outcome confounders is especially important. Finite-sample bias deceases toward zero as sample size increases and is often not a major issue in observational studies. Bias arising from unmeasured instrument-outcome confounders, however, cannot be eliminated through increasing sample sizes (Small and Rosenbaum, 2008). Additionally, we cannot guarantee that unmeasured instrument-outcome confounders do not exist and the assumption may face criticism. Stronger instruments are therefore one way to increase the credibility of the results of an IV analysis.

Recent methods have been proposed for strengthening the instrument in IV-matching designs (Baiocchi et al., 2010, 2012). It has been taken for granted, however, that strengthened instruments provide the same benefits as instruments that are naturally stronger. In this chapter, we study estimation with strengthened instruments to determine their properties and how they compare with instruments that are naturally stronger. The remainder of this chapter is organized as follows. In Section 4.2 we introduce matching in IV analyses and discuss two recently proposed methods for strengthening the instrument in IV-matching designs. Inference is discussed in Section 4.3, along with a sensitivity analysis for assessing sensitivity to unmeasured instrument-outcome confounders. We study the performance of strengthened instruments through simulations in Section 4.4, and conclude with a discussion in Section 4.5.

4.2 Methods

Throughout this section, we will assume that we measure covariates $\mathbf{X}$, binary treatment D, and response Y for each of n individuals. We further assume that we measure a continuous instrument that can be dichotomized to indicate "enouragement" toward the treatment. The term encouragement is used to mean that individuals with instrument $Z = 1$ are more likely to receive treatment than individuals with $Z = 0$, and not necessarily that individuals received actual or physical encouragement.

Matching in an IV analysis with a binary instrument aims to pair individuals that have been encouraged toward the treatment with individuals that look similar on relevant covariates, but have been encouraged toward the control. After matching, the match quality can be assessed through covariate balance (Love, 2002; Austin, 2009a) and instrument strength can be assessed by the average difference, or separation, of the instrument within pairs (Baiocchi et al., 2010, 2012). Stronger instruments are associated with greater separation of the instrument. We therefore hope to pair individuals that are similar on relevant covariates but received a large difference in encouragement toward the treatment.

In the following sections we present three strategies for matching in an IV analysis: a standard matching procedure which we refer to as IV-matching, the weighted IV-matching procedure of Chapter III, and the near-far matching procedure of Baiocchi et al. (2010, 2012). IV-matching involves matching on covariates $\mathbf{X}$, but takes the strength of the instrument as given and does not strengthen it. The other two procedures match on covariates $\mathbf{X}$ and increase the separation of the instrument within pairs, thereby strengthening the instrument. Weighted IV-matching increases separation by assigning more weight to pairs that are more influenced by the instrument, while near-far matching increases separation by discouraging individuals with similar instrument values from ever pairing.

4.2.1 IV-Matching

IV-matching begins by defining a distance between each individual based on covariates $\mathbf{X}$. An optimal nonbipartite match (Derigs, 1988; Lu et al., 2011) then pairs individuals such that the sum of these distances over all pairs is minimized. After matching, the instrument is dichotomized into encouragement by comparing instrument values within each pair, and considering the individual in each pair with the higher instrument value to have been encouraged toward treatment. Let i denote pair and let j and j' denote subjects within pair i. Assign $Z_{ij} = 1$ and $Z_{ij'} = 0$ if subject j in pair i had a higher instrument value than subject $j' \neq j$. Assign $Z_{ij} = 0$ and $Z_{ij'} = 1$ otherwise. Note that $Z_{ij} + Z_{ij'} = 1$ for all i, so that each pair has one encouraged and one unencouraged subject.

4.2.2 Weighted IV-Matching

Weighted IV-matching follows the same steps as IV-matching for constructing pairs based on covariates $\mathbf{X}$ and assigning encouragement. To strengthen the instrument, a weight is assigned to each pair after matching based on the influence of the instrument within that pair. In Chapter III, we proposed weighting pair i by

$$w_i = P(D_{ij} = 1 | Z_{ij} = 1) P(D_{ij'} = 0 | Z_{ij'} = 0), \tag{4.5}$$

which is designed to reflect the probability that the encouraged subject receives the treatment while the unencouraged subject receives the control. This is a measure of the instrument's influence over the treatment within pair i, and a strengthened instrument is created when more weight is assigned to pairs in which the instrument has more influence over the treatment. This strategy for strengthening the instrument effectively redistributes the data through weights to highlight portions associated with greater instrument strength. This weight, as well as the weighted IV-matching procedure, is discussed throughout Chapter III.

4.2.3 Near-far Matching

The near-far matching procedure of Baiocchi et al. (2010, 2012) strengthens the instrument with a modification to the matching phase of the IV-matching process. When defining the distance between individuals based on covariates $\mathbf{X}$, a penalty is added to the distance between individuals whose instrument values are within a pre-specified range of each other. This penalty discourages the matching algorithm from pairing those individuals. As a result, pairs are more likely to consist of individuals with instrument values suitably far apart. In addition to these penalties, sinks are added to the match to allow a pre-specified number of individuals to be optimally removed from the analysis (Lu et al., 2001). This alleviates some of the damage to match quality caused by adding penalties to the distances used for matching. The instrument is again dichotomized into encouragement by comparing instrument values within each pair, and considering the individual in each pair with the higher instrument value to have been encouraged toward treatment.

After implementing one of these three matching procedures, we have I pairs that were matched on covariates $\mathbf{X}$. One subject in each pair is considered to have been encouraged toward the treatment and the other toward the control. In the next section, we discuss estimating an effect of treatment. We also present a sensitivity analysis for assessing robustness to unmeasured instrument-outcome confounders that violate the assumption that the instrument is randomly assigned.

4.3 Inference

4.3.1 Estimation

We implement the estimation and inference procedures of Section 3.2 for estimating the local average treatment effect, defined here as

$$\beta_{IV} = \frac{\sum_i \sum_j \left(Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))\right)}{\sum_i \sum_j \left(D_{ij}(1) - D_{ij}(0)\right)}, \tag{4.6}$$

where $Y_{ij}(1, D_{ij}(1))$ and $Y_{ij}(0, D_{ij}(0))$ denote the potential responses for subject j in pair i and $D_{ij}(1)$ and $D_{ij}(0)$ denote the potential treatments. Under assumptions *A1-A5* of Section 2.2, this parameter is interpreted as an average treatment effect among the subgroup of the population known as compliers (Table 2.1). Compliers are individuals that can be encouraged to switch treatment states and are often referred to as "marginal patients." More discussion of this parameter can be found in Chapter II, as well as Imbens and Angrist (1994), Angrist et al. (1996), or Baiocchi et al. (2014).

Let $Y_{ij} = Z_{ij}Y_{ij}(1, D_{ij}(1)) + (1 - Z_{ij})Y_{ij}(0, D_{ij}(0))$ be the observed response and $D_{ij} = Z_{ij}D_{ij}(1) + (1 - Z_{ij})D_{ij}(0)$ the observed treatment for subject j in pair i. We estimate β_{IV} after matching as

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} \left[Z_{ij}Y_{ij} - (1 - Z_{ij})Y_{ij}\right]}{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} \left[Z_{ij}D_{ij} - (1 - Z_{ij})D_{ij}\right]}, \tag{4.7}$$

where $\hat{w}_i = 1$ for the IV-matching and near-far matching procedures. We follow the permutation inference procedure of Baiocchi et al. (2010), outlined in Section 3.2, for obtaining confidence intervals for $\hat{\beta}_{IV}$.

4.3.2 Sensitivity Analysis

The improved robustness to violations of the assumption that the instrument is randomly assigned is a key benefit of working with stronger instruments and has been the primary motivation behind methods for strengthening the instrument. Random assignment of the instrument implies that there are no unmeasured instrument-outcome confounders. Unfortunately, this assertion cannot be verified and is often criticized. Being robust to unmeasured instrument-outcome confounders can therefore increase the credibility of the results from an IV analysis.

In this section we discuss a sensitivity analysis that provides guidance as to how robust estimates are to violations of this assumption. This sensitivity analysis is outlined in Rosenbaum (2002) and applied in instrumental variable settings in Baiocchi et al. (2010, 2012). The goal of this analysis is to determine how far an instrument can deviate from being randomly assigned before the qualitative results of the study are altered. We can think of this as determining how large or how strong an unmeasured instrument-outcome confounder would need to be to explain what appears to be a significant treatment effect.

Following Rosenbaum (2002), we assume that within pair i matched on covariates $\mathbf{X}$, subjects j and $j^{'}$ differ in their odds of receiving encouragement by at most a factor of $\Gamma \geq 1$, where

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \leq \Gamma \text{ for all } i, j, j' \text{ with } \mathbf{X}_{ij} = \mathbf{X}_{ij'} \tag{4.8}$$

and $\pi_{ij} = P(Z_{ij} = 1|\mathbf{X}_{ij})$. When the instrument is randomly assigned and each subject has equal odds of receiving encouragement, $\pi_{ij} = \pi_{ij^{'}}$ and $\Gamma = 1$. As random assignment of the instrument is increasingly violated, these probabilities diverge and Γ increases.

The sensitivity analysis is conducted by using Γ in inference procedures to obtain bounds on the p-value associated with testing the hypothesis that $\beta_{IV} = 0$. For matched pairs with a continuous response, we can do this using Wilcoxon's signed rank test for testing the association between encouragement and the outcome (Small and Rosenbaum, 2008). This involves taking the difference in response values between encouraged and unencouraged subjects for each of the I pairs, ranking these differences, and summing the ranks for pairs in which the encouraged subject had the higher response value. This sum is then compared to two normal distributions with expectation $pI(I+1)/2$ and variance $p(1-p)I(I+1)(2I+1)/6$ to obtain p-values, where $p = 1/(1+\Gamma)$ is used for obtaining a lower bound and $p = \Gamma/(1+\Gamma)$ is used for an upper bound. This is repeated for increasing values of Γ. The largest deviation from random assignment that can be sustained is given by the largest Γ value in which the upper bound for the p-value remains less than 0.05, with larger deviations indicating results that would require a larger unmeasured instrument-outcome confounder to explain them. Notice that when $\Gamma = 1$ the expectation and variance reduce to the usual expectation and variance for Wilcoxon's signed-rank test (Lehmann, 1975). For applications of this sensitivity analysis with other test statistics, see Rosenbaum (2002), Baiocchi et al. (2010, 2012).

The parameter Γ has the advantage of being a univariate measure for quantifying a deviation from random assignment, but its magnitude is not easily interpreted in the context of the problem. To help with interpretations, Rosenbaum and Silber (2009) present a mapping of Γ to two components as

$$\Gamma = \frac{\Delta\Lambda + 1}{\Delta + \Lambda}, \tag{4.9}$$

where Δ represents the effect of an unmeasured confounder on the instrument and Λ the effect of that unmeasured confounder on the response. For example, an unmeasured confounder that triples the odds of receiving encouragement ($\Delta = 3$) while doubling the odds of having the higher response value ($\Lambda = 2$) corresponds with a Γ value of $(3 \cdot 2 + 1)/(3 + 2) = 1.4$. This mapping of Γ allows the sensitivity analysis to remain relatively simple while allowing the researcher to better understand and interpret the results in a meaningful way and in the context of the problem.

4.4 Simulations

4.4.1 Setup

In this section we report simulation results to study the properties of strengthened instruments. We compare estimation under four matching procedures. Two involve pairing individuals using the IV-matching procedure of Section 4.2.1, matching on covariates without strengthening the instrument. One of these matches will use a relatively weak instrument (IVM-I) while the other will use an instrument that is naturally stronger (IVM-II), allowing for comparisons when all else is equal except the strength of the instrument. The remaining two matches will use the relatively weak instrument but strengthen it, one using the weighted IV-matching (WIVM) procedure described in Section 4.2.2 and one using the near-far matching (NFM) procedure described in Section 4.2.3. We refer to these four matches as using weaker (IVM-I), stronger (IVM-II), and strengthened (WIVM, NFM) instruments.

We generate 1,000 datasets from

$$Y_i = \beta D_i + X_i + \epsilon_i^Y + \alpha U_i, \tag{4.10}$$

$$logit(P(D_i = 1)) = Z_{1i} + Z_{2i} + \epsilon_i^D, \tag{4.11}$$

$$Z_{1i} = \gamma_{1i} + \alpha U_i, \tag{4.12}$$

$$Z_{2i} = \gamma_{2i}, \tag{4.13}$$

where Y represents the response and D the treatment for each of $i = 1, ..., 4,000$ individuals. Covariate X and random effects γ_1 and γ_2 are randomly generated from standard normal distributions. Errors (ϵ^Y, ϵ^D) are generated from a bivariate normal with correlation 0.75 to represent unmeasured treatment-outcome confounding. Unmeasured instrument-outcome confounder U is generated from a Bernoulli distribution with probability 0.5. The parameter α is used to control the strength of U and is varied from 0 to 1. When $\alpha > 0$, the presence of U in the generating equations for Y and Z violates the assumption that the instrument is randomly assigned. We estimate the effect of the treatment on the response, β, which we set to 0 throughout these simulations.

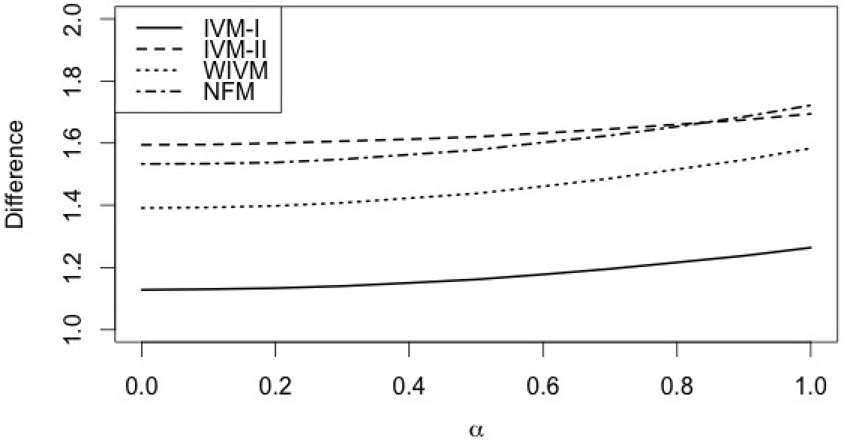
To represent the weaker instrument for IVM-I, the instrument is defined as $Z_i = Z_{1i} = \gamma_{1i} + \alpha U_i$. For IVM-II, the stronger instrument is defined as $Z_i = Z_{1i} + Z_{2i} = \gamma_{1i} + \gamma_{2i} + \alpha U_i$. These two instruments capture approximately 10% and 22% of the variation in the treatment, respectively, while holding everything else constant. WIVM and NFM will use the weaker instrument but strengthen it via the methods described in Sections 4.2.2 and 4.2.3. This setup allows for a comparison of weaker instruments (IVM-I) with instruments that are naturally stronger (IVM-II) and instruments that have been strengthened by the researcher (WIVM, NFM).

4.4.2 Results

Instrument Strength

Figure 4.1 reports instrument strength as indicated by separation of the instrument. Greater separation is considered to correspond with greater instrument strength. Results confirm that IVM-II, WIVM, and NFM are using a stronger instrument than IVM-I as expected. Both WIVM and NFM were able to strengthen the weak instrument, with NFM achieving separation of the instrument roughly equivalent to that of IVM-II. Separation of the instrument increases slightly as unmeasured instrument-outcome confounding increases, a result of the the instrument taking more extreme values as α increases.

Figure 4.1: Separation of the instrument for each method by magnitude of unmeasured instrument-outcome confounding. Displayed is the average difference in instrument values between encouraged and unencouraged individuals. Greater differences are considered to correspond with greater instrument strength.

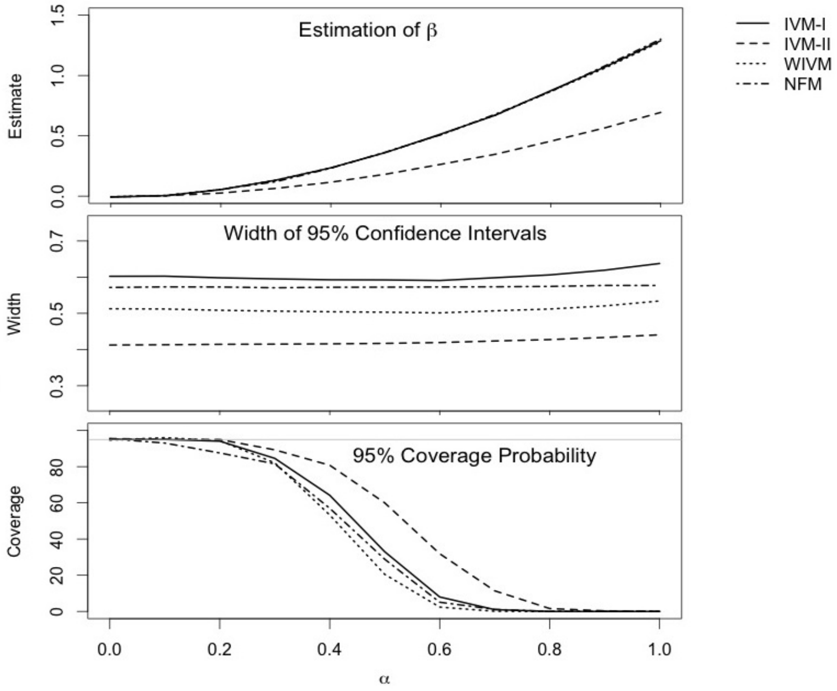


Estimation

Figure 4.2 reports estimates of β (top), the width of 95% confidence intervals (middle), and 95% coverage probabilities (bottom) for each method. Since β is set to 0, estimates of β can also be considered the bias in estimation. Note that the lines for IVM-I, WIVM, and NFM are overlapping in the plot of the estimates. We see that as the level of unmeasured instrument-outcome confounding increases, as measured by α, the bias increases as well for each method. The bias for IVM-II is decreased compared with the bias for IVM-I at every level of $\alpha > 0$. WIVM and NFM, on the other hand, are equally biased compared with IVM-I. The decreased bias for IVM-II allows it to maintain higher 95% coverage probability than either of the remaining three methods, though each method has coverage decreasing to 0% as unmeasured instrument-outcome confounding increases. The plot of confidence interval widths shows that IVM-II, WIVM, and NFM were more efficient than IVM-I, with IVM-II leading to the narrowest intervals of the four methods.

These results reveal an important difference between strengthened (WIVM, NFM)

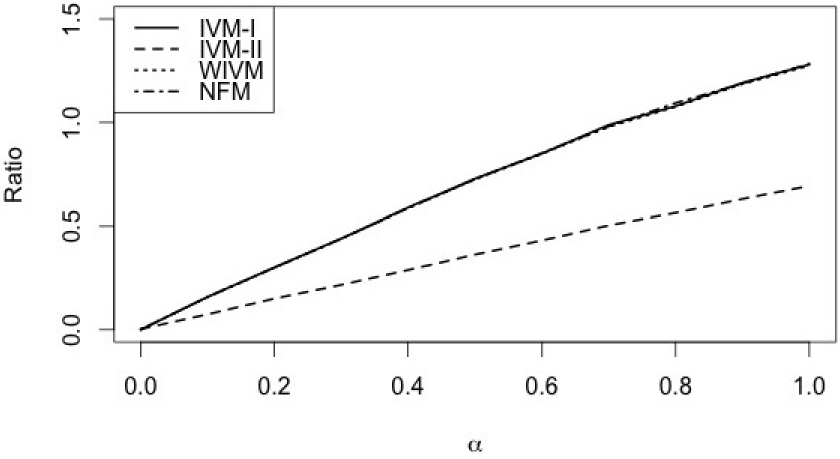
Figure 4.2: Estimation of β, width of confidence intervals, and 95% coverage probabilities under an increasing magnitude (α) of unmeasured instrument-outcome confounding. Note that the lines for IVM-I, WIVM, and NFM are overlapping in the top plot.



and naturally stronger instruments (IVM-II). Theoretical results suggest that greater correlation between the instrument and the treatment decreases the bias arising from unmeasured instrument-outcome confounding and increases efficiency in estimation. IVM-II had both of these properties while using the naturally stronger instrument. WIVM and NFM, however, provided for more efficient estimation but failed to decrease the bias caused by unmeasured instrument-outcome confounder U.

We investigate this issue further by estimating the ratio of the correlation between the instrument and unmeasured instrument-outcome confounder U to the correlation between the instrument and the treatment. This ratio is displayed by α in Figure 4.3. Notice that the lines for IVM-I, WIVM, and NFM are again overlapping, while that for IVM-II is decreased. The overlapping of WIVM and NFM with IVM-I indicates that although WIVM and NFM were able to increase the correlation between the instrument and the treatment, the correlation between the instrument and the unmeasured instrument-outcome confounder U was increased proportionally. This proportional increase in the correlation between the instrument and U explains why WIVM and NFM did not see a decrease in bias in Figure 4.2 like IVM-II did.

Figure 4.3: Ratio of the correlation between the instrument and unmeasured instrument-outcome confounder U to the correlation between the instrument and the treatment. Note that lines for IVM-I, WIVM, and NFM are overlapping.



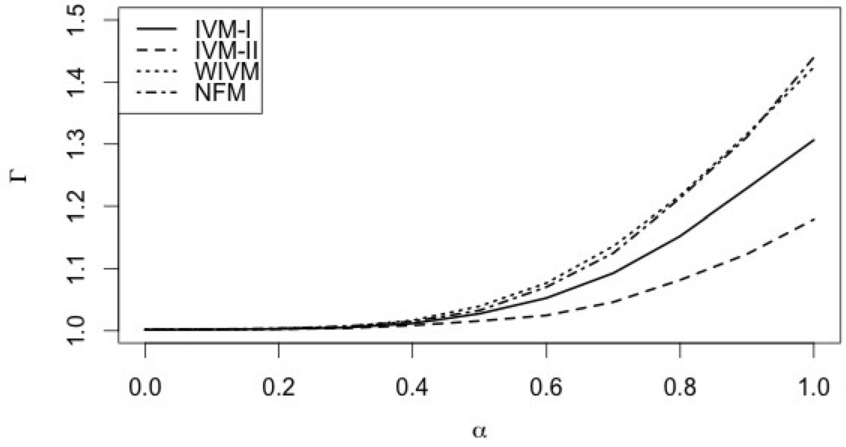
The increase in the strength of unmeasured instrument-outcome confounding resulting from methods that strengthen the instrument has thus far been overlooked in the literature. When doing so, results from the sensitivity analysis presented in the following section can be misinterpreted to suggest improved robustness to unmeasured instrument-outcome confounding when, as Figures 4.2 and 4.3 make clear, there is no improvement. We believe this to be why strengthened instruments have previously been suggested to be more robust to unmeasured instrument-outcome confounders.

Sensitivity

Figure 4.4 displays the results of the sensitivity analysis of Section 4.3.2. The vertical axis, Γ, measures the deviation from random assignment that our estimates are considered robust to. Γ can be related to the size of an unmeasured instrument-outcome confounder as in (4.9), telling us how large one would need to be to explain what appears to be a significant effect of the treatment.

Results suggest that estimates obtained using WIVM and NFM would require a larger unmeasured instrument-outcome confounder to explain them compared with IVM-I, while estimates obtained using IVM-II would require a smaller one. These results need to be interpreted with caution. Applications of this sensitivity analysis typically consider larger values of Γ to correspond with more robust results (Rosenbaum, 2002; Small and Rosenbaum, 2008; Baiocchi et al., 2010, 2012), but we see here that this is not always true. This sensitivity analysis tells us how strong an unmeasured instrument-outcome confounder would need to be to explain a significant effect. This is often used interchangeably with robustness to unmeasured confounders but it is quite different. Consider, for example, the sensitivity results for WIVM and NFM. Both methods found larger Γ values in the sensitivity analysis, suggesting that estimates obtained using WIVM and NFM would require a stronger unmeasured instrument-outcome confounder to explain them compared to estimates obtained using IVM-I or IVM-II. While this sounds like improved robustness, results in Figure 4.3 showed that WIVM and NFM actually increased the strength of unmeasured instrument-outcome confounder U, and results in Figure 4.2 showed that results were no less biased in the presence of U. If the increase in the strength of U caused by the WIVM and NFM procedures is ignored, Γ can easily be misinterpreted to suggest improved robustness when there is none.

Figure 4.4: Results of sensitivity analysis for assessing robustness to unmeasured instrument-outcome confounding as measured by Γ. A larger Γ indicates that a larger unmeasured instrument-outcome confounder would be required to explain a significant effect estimate.



4.5 Discussion

Instrumental variable methods are increasingly used in health and medical research. Unfortunately, instrumental variable analyses rely on assumptions that are difficult to verify and often criticized. One way to increase the credibility of the results of an instrumental variable analysis is to work with stronger instruments, or instruments that are highly correlated with the treatment. Benefits of using stronger instruments include a decrease in finite-sample bias, increased efficiency in estimation, and improved robustness to unmeasured instrument-outcome confounders that violate the assumption that the instrument is randomly assigned. Motivated by the desire to capitalize on these benefits, recent methods have been proposed to increase the strength of a weak instrumental variable. It has been taken for granted, however, that a weak instrument that has been strengthened provides these same benefits.

In this chapter, we investigated estimation with strengthened instruments to better understand their properties and how they compare with instruments that are naturally stronger. Our findings revealed important differences between the two.

Specifically, we found that while strengthened instruments are able to increase efficiency in estimation, they do not lead to a decrease in finite-sample bias or improve the robustness to unmeasured instrument-outcome confounders. We found that methods for strengthening the instrument additionally strengthen the relationship between the instrument and any unmeasured instrument-outcome confounders, which offsets any potential decrease in bias or improved robustness.

The increase in the strength of unmeasured instrument-outcome confounding that results from methods that strengthen the instrument is an important finding that has been overlooked in the literature thus far. Ignoring this issue leads to misleading sensitivity results, and is likely why strengthened instruments have been suggested to improve robustness to unmeasured instrument-outcome confounders in the same way that naturally stronger instruments do. This is a major shortcoming of strengthened instruments, since improved robustness to unmeasured instrument-outcome confounding is arguably the most important benefit of working with stronger instruments. These findings suggests that strengthened instruments should not be considered equal to instruments that are naturally stronger.

Results of this work can give guidance for future research related to strengthening instrumental variables. One important takeaway from these findings is that improving robustness to unmeasured instrument-outcome confounders via strengthening the instrument will require the development of methods that do not increase the strength of unmeasured instrument-outcome confounding in the process. These findings highlight that strengthened instruments appear most useful for decreasing the variability in estimation. If future research on strengthening the instrument is framed in the context of decreasing variability rather than improving robustness to unmeasured instrument-outcome confounding, this may pave the way for the development of more efficient IV methods.

CHAPTER V

Conclusion

In observational studies, unmeasured differences between treatment groups often confound the relationship of interest. Instrumental variable (IV) methods can give consistent effect estimates in the presence of this unmeasured confounding, and are becoming increasingly popular in health and medical research. This dissertation has focused on the development of new IV methods, with applications to studies comparing mortality among patients receiving dialysis as treatment for end stage renald disease.

In Chapter II, we developed a weighted IV estimator that adjusts for measured instrument-outcome confounders through the IV propensity score. The weights were designed to reflect the probability of being selected into a one-to-one match. Advantages of weighting over matching include increased efficiency, straightforward variance estimation, and ease of computation. Through simulation, the estimator was shown to be more efficient than both matching and alternative weighted estimators. Use of the estimator was illustrated in a study comparing the relationship between mortality and dialysis session length among hemodialysis patients. Future work related to applying this estimator to time-to-event or survival data was also presented.

In Chapter III, we developed a weighting procedure for increasing the strength of the instrument when matching. Compared with existing methods, this weighting procedure strengthened the instrument without compromising match quality. This is a major advantage of the proposed method, as poor match quality can bias estimation. Methods were illustrated with a study comparing early mortality in hemodialysis and peritoneal dialysis patients.

In Chapter IV, we compared estimation with strengthened instruments to estimation with instruments that are naturally stronger. Methods for strengthening the instrument have been motivated by the benefits of using stronger instruments, including decreased finite-sample bias, increased efficiency, and results that are more robust to unmeasured instrument-outcome confounders. We found that strengthened instruments were unable to provide these same benefits, as has been previously suggested. Our findings indicated that while strengthened instruments provide for more efficient estimation, they do not lead to a decrease in finite-sample bias or improve the robustness to unmeasured instrument-outcome confounders. We found that methods for strengthening the instrument inadvertently strengthen unmeasured instrument-outcome confounders in the process. This important issue has thus far been overlooked in the literature, which has led to the misbelief that strengthened instruments improve robustness to unmeasured instrument-outcome confounders. These findings give guidance for future research related to strengthening the instrument.

APPENDICES

APPENDIX A

Asymptotic equivalence of IV-MW and IV-PSM estimators

In this section we show that the IV-MW and IV-PSM estimators have the same limit as $n \to \infty$. Following Li and Greene (2013), we will assume that the IV propensity score takes finitely many values c_k for $k = 1, ..., K$ with $c_k \in (0, 1)$. This assumption is to allow exact matching on the IV propensity score and avoid unnecessary complications of working with other matching algorithms. For the IV-PSM estimator we assume one-to-one exact matching without replacement on the IV propensity score. Additionally, we simplify the notation of section 2.2, letting $Y(1, D_i(1)) = Y_i^1$, $Y_i(0, D_i(0)) = Y_i^0$, $D_i(1) = D_i^1$, $D_i(0) = D_i^0$, $Y_i = Z_i Y_i^1 + (1 - Z_i)Y_i^0$, $D_i = Z_i D_i^1 + (1 - Z_i)D_i^0$, and $e_i(\mathbf{x}_i) = e_i$. We further denote $P(e_i = c_k) = \tau_k$, with $\sum_k \tau_k = 1$.

We begin with the IV-MW estimator, defined as

$$\begin{aligned}\lambda_{\text{IV-MW}} &= \frac{\sum_i W_i Z_i Y_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) Y_i / \sum_i W_i (1 - Z_i)}{\sum_i W_i Z_i D_i / \sum_i W_i Z_i - \sum_i W_i (1 - Z_i) D_i / \sum_i W_i (1 - Z_i)} \\ &\equiv \frac{A/F - B/G}{C/F - D/G}.\end{aligned}$$

The limit for A is

$$\begin{aligned} n^{-1}\sum_i W_i Z_i Y_i &\to_p E\{W_i Z_i Y_i^1\} \\ &= E\left\{E(\frac{min(e_i, 1-e_i)}{e_i} I(Z_i = 1) Y_i^1 | \mathbf{x}_i)\right\} \\ &= E\{min(e_i, 1-e_i) E(Y_i^1|\mathbf{x}_i)\}. \end{aligned}$$

Similarly, the limits for B, C, and D are given by

$$\begin{aligned} n^{-1}\sum_i W_i(1-Z_i)Y_i &\to_p E\{min(e_i, 1-e_i)E(Y_i^0|\mathbf{x}_i)\}, \\ n^{-1}\sum_i W_i Z_i D_i &\to_p E\{min(e_i, 1-e_i)E(D_i^1|\mathbf{x}_i)\}, \\ n^{-1}\sum_i W_i(1-Z_i)D_i &\to_p E\{min(e_i, 1-e_i)E(D_i^0|\mathbf{x}_i)\}. \end{aligned}$$

Taking the limit of F and G gives

$$\begin{aligned} n^{-1}\sum_i W_i Z_i &\to_p E\{W_i Z_i\} \\ &= E\left\{\frac{min(e_i, 1-e_i)}{e_i} I(Z_i = 1)\right\} \\ &= E\{min(e_i, 1-e_i)\} \end{aligned}$$

and

$$n^{-1}\sum_i W_i(1-Z_i) \quad \to_p \quad E\{min(e_i, 1-e_i)\}$$

Combining these and reducing, the limit of the IV-MW as $n \to \infty$ is given as

$$\hat{\lambda}_{\text{IV-MW}} \to_p \frac{E\{min(e_i, 1-e_i)(E(Y_i^1|\mathbf{x}_i) - E(Y_i^0|\mathbf{x}_i))\}}{E\{min(e_i, 1-e_i)(E(D_i^1|\mathbf{x}_i) - E(D_i^0|\mathbf{x}_i))\}}.$$

Next we consider the IV-PSM estimator, which we write as

$$\hat{\lambda}_{\text{IV-PSM}} = \frac{\left\{\frac{\sum_k \sum_i Y_i I(i \in S_{1k})}{\sum_k \sum_i I(i \in S_{1k})}\right\} - \left\{\frac{\sum_k \sum_i Y_i I(i \in S_{0k})}{\sum_k \sum_i I(i \in S_{0k})}\right\}}{\left\{\frac{\sum_k \sum_i D_i I(i \in S_{1k})}{\sum_k \sum_i I(i \in S_{1k})}\right\} - \left\{\frac{\sum_k \sum_i D_i I(i \in S_{0k})}{\sum_k \sum_i I(i \in S_{0k})}\right\}} \equiv \frac{A/F - B/G}{C/F - D/G},$$

where S_{1k} and S_{0k} represent the sets of encouraged and unencouraged subjects matched at c_k, respectively. The limit of A is then

$$\begin{aligned}
n^{-1}\sum_k\sum_i Y_i I(i \in S_{1k}) &= n^{-1}\sum_k\sum_i Y_i^1 I(i \in S_{1k}) \\
&\to_p E\left\{\sum_k Y_i^1 I(i \in S_{1k})\right\} \\
&= E\left\{E(Y_i^1|\mathbf{x}_i)E(\sum_k I(i \in S_{1k})|\mathbf{x}_i)\right\} \\
&= E\left\{E(Y_i^1|\mathbf{x}_i)\sum_k \tau_k e_i \frac{min(e_i, 1-e_i)}{e_i}\right\} \\
&= E\{min(e_i, 1-e_i)E(Y_i^1|\mathbf{x}_i)\}.
\end{aligned}$$

Similarly, the limits for B, C, and D are given as

$$\begin{aligned}
n^{-1}\sum_k\sum_i Y_i I(i \in S_{0k}) &\to_p E\{min(e_i, 1-e_i)E(Y_i^0|\mathbf{x}_i)\}, \\
n^{-1}\sum_k\sum_i D_i I(i \in S_{1k}) &\to_p E\{min(e_i, 1-e_i)E(D_i^1|\mathbf{x}_i)\}, \\
n^{-1}\sum_k\sum_i D_i I(i \in S_{0k}) &\to_p E\{min(e_i, 1-e_i)E(D_i^0|\mathbf{x}_i)\}.
\end{aligned}$$

Finally, for F we have

$$\begin{aligned}
n^{-1}\sum_k\sum_i I(i \in S_{1k}) &\to_p E\left\{\sum_k I(i \in S_{1k})\right\} \\
&= E\left\{min(e_i, 1-e_i)\sum_k \tau_k\right\} \\
&= E\{min(e_i, 1-e_i)\},
\end{aligned}$$

and similarly for G

$$n^{-1}\sum_k\sum_i I(i \in S_{0k}) \to_p E\{min(e_i, 1-e_i)\}.$$

Combining everything and reducing, the limit of the IV-PSM estimator as $n \to \infty$ is found to be

$$\hat{\lambda}_{\text{IV-PSM}} \to_p = \frac{E\{min(e_i, 1-e_i)(E(Y_i^1|\mathbf{x}_i) - E(Y_i^0|\mathbf{x}_i))\}}{E\{min(e_i, 1-e_i)(E(D_i^1|\mathbf{x}_i) - E(D_i^0|\mathbf{x}_i))\}},$$

which is the same as that of the IV-MW estimator.

APPENDIX B

Full list of covariates used in data example of Chapter III

Table B.1. Summary of covariates before matching. Patient level covariates are compared across dialysis modality and facility level covariates are compared across first and fourth quartile of the PD usage.

Patient Covariates	HD	PD	St Diff
N	142,737	21,458	-
Outcome			
Death w/in 6 months	14%	4%	35.7
Covariates			
Age	64	58	37.7
Male	57%	55%	3.8
Bmi	29.6	29.5	1.9
6+ months pre-ESRD care	45%	69%	-49.3
# of comorbidities	2.4	1.9	44.1
Hemoglobin	9.9	10.6	-4.2
Serum creatinine	6.6	6.4	1.0
No insurance	7%	8%	-6.9
White	68%	71%	-5.1
Black	26%	22%	8.7
Asian	4%	5%	-7.4
Hispanic	13%	12%	2.2
Employed	9%	26%	-45.2
Facility Covariates	Q 1	Q 4	St Diff
Instrument			
PD usage	3%	30%	-208
Covariates			
For profit	85%	86%	-3.3
# of nurses	6.7	8.7	-43.3
# of technicians	8.2	8.1	2.0
# of social workers	0.8	1.1	-36.4
# of HD stations	20.3	21.9	-19.1
Median income	$51,086	$50,850	1.2
Bachelors degree +	23.7%	23.4%	4.5

Table B.2. Summary of covariates after matching, by matching algorithm. U and E correspond to patients considered to have been treated at PD unencouraging (U) and PD encouraging (E) facilities.

	IVM (64,350 pairs)			WIVM (64,350 pairs)			NFM (33,702 pairs)		
	U	E	St Diff	U	E	St Diff	U	E	St Diff
Instrument									
Facility % PD 2007-09	4.7%	15.3%	-96.3	6.3%	27.8%	-194.8	3.8%	25.3%	-195.2
Treatment									
PD	10.0%	16.3%	-18.7	11.5%	23.5%	-35.7	9.0%	23.8%	-44.1
Outcome									
Died w/in 6 months	11.9%	11.3%	1.7	11.7%	10.7%	3.3	11.8%	10.8%	3.1
Patient Covariates									
Age	62.8	62.7	0.5	62.6	62.1	3.5	63.0	62.2	5.0
Male	57.0%	56.9%	0.2	57.3%	56.7%	1.3	57.0%	56.7%	0.6
BMI	30.5	30.4	0.5	30.5	30.2	1.0	31.0	31.2	-0.7
6+ mos pre-ESRD care	47.8%	50.3%	-5.1	50.2%	53.1%	-5.7	50.4%	53.2%	-5.6
# of comorbidities	2.5	2.4	1.8	2.5	2.4	2.1	2.6	2.4	10.1
Hemoglobin	9.9	10.0	-0.4	9.9	10.0	-0.5	10.0	9.9	1.0
Serum Creatinine	6.6	6.5	0.4	6.7	6.5	0.6	6.5	6.5	0.2
No insurance	7.1%	6.9%	0.9	6.8%	7.5%	-2.7	6.1%	7.4%	-4.9
White	68.6%	65.9%	5.7	68.9%	63.9%	10.6	68.8%	64.1%	10.1
Black	25.3%	28.1%	-6.2	25.1%	29.2%	-9.4	26.7%	29.4%	-6.1
Asian	3.8%	3.8%	0.4	4.1%	3.8%	1.7	2.6%	4.2%	-8.4
Hispanic	13.8%	12.6%	3.7	13.4%	12.4%	3.1	9.9%	12.2%	-7.0
Employed	11.4%	12.4%	-3.1	12.3%	13.7%	-4.2	11.6%	13.6%	-6.3
Facility Covariates									
For profit	84.3	84.3	0.1	81.1	81.1	0.0	83.3	83.4	-0.3
# of nurses	9.1	9.2	-2.0	10.0	10.2	-3.8	9.2	10.7	-26.5
# of technicians	9.8	9.9	-0.5	9.7	9.8	1.2	9.0	9.7	-10.6
# of social workers	1.1	1.3	-13.9	1.1	1.4	-16.8	1.1	1.3	-10.4
# of HD stations	24.0	24.0	-0.5	24.2	24.4	-1.2	23.5	24.6	-10.9
Median income	$50,874	$51,343	-2.32	$50,618	$50,496	0.6	$50,470	$51,368	-4.5
Bachelors degree +	23.4	25.0	-10.8	24.0	25.1	-8.2	23.5	25.7	-15.4

BIBLIOGRAPHY

BIBLIOGRAPHY

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* **27,** 2037–2049.

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* **28,** 3083–3107.

Austin, P. C. (2009b). Type i error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics* **5,**.

Austin, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology* **172,** 1092–1097.

Austin, P. C. (2011a). Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in Medicine* **30,** 1292–1301.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* **10,** 150–161.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine* **33,** 2297–2340.

Baiocchi, M., Small, D., Lorch, S., and Rosenbaum, P. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* **105,** 1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012). Near/far matching: a study design approach to instrumental variables. *Health Services and Outcomes Research Methodology* **12,** 237–253.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80,** 2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2010). Lasso methods for gaussian instrumental variables models. *arXiv:1012.1297* .

Bhattacharya, J., Goldman, D., and McCaffrey, D. (2006). Estimating probit models with self-selected treatments. *Statistics in Medicine* **25,** 389–413.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association* **90,** 443–450.

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics* **3,**.

Brunelli, S. M., Chertow, G. M., Ankers, E. D., Lowrie, E. G., and Thadhani, R. (2010). Shorter dialysis times are associated with higher mortality among incident hemodialysis patients. *Kidney international* **77,** 630–636.

Caner, M. and Fan, Q. (2010). The adaptive lasso method for instrumental variable selection. Technical report, Working Paper, North Carolina State University.

Cheng, J. and Lin, W. (2013). Understanding causal effects in observational studies with instrumental propensity scores. *Joint Statistical Meeting* .

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *SankhyÄ: The Indian Journal of Statistics, Series A (1961-2002)* **35,** 417–446.

Dartmouth (2016). *The Dartmouth Atlas of Healthcare.* Trustees of Dartmouth College.

Daugirdas, J. T. (2013). Dialysis time, survival, and dose-targeting bias. *Kidney International* **83,** 9–13.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research* **13,** 225–261.

Eknoyan, G., Beck, G. J., Cheung, A. K., Daugirdas, J. T., Greene, T., Kusek, J. W., Allon, M., Bailey, J., Delmez, J. A., Depner, T. A., et al. (2002). Effect of dialysis dose and membrane flux in maintenance hemodialysis. *New England Journal of Medicine* **347,** 2010–2019.

Flythe, J. E., Curhan, G. C., and Brunelli, S. M. (2013). Shorter length dialysis sessions are associated with increased mortality, independent of body weight. *Kidney International* **83,** 104–113.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics* **139,** 35–75.

Garabedian, L. F., Chu, P., Toh, S., Zaslavsky, A. M., and Soumerai, S. B. (2014). Potential bias of instrumental variable analyses for observational comparative effectiveness researchpotential bias of instrumental variable analyses for observational cer. *Annals of Internal Medicine* **161,** 131–138.

Goodlad, C. and Brown, E. (2013). The role of peritoneal dialysis in modern renal replacement therapy. *Postgraduate Medical Journal* **89,** 584–590.

Heaf, J. G., Løkkegaard, H., and Madsen, M. (2002). Initial survival advantage of peritoneal dialysis relative to haemodialysis. *Nephrology Dialysis Transplantation* **17,** 112–117.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60,** 505–531.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62,** pp. 467–475.

Jiwakanon, S., Chiu, Y.-W., Kalantar-Zadeh, K., and Mehrotra, R. (2010). Peritoneal dialysis: an underutilized modality. *Current opinion in nephrology and hypertension* **19,** 573–577.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Kim, H., Kim, K. H., Park, K., Kang, S.-W., Yoo, T.-H., Ahn, S. V., Ahn, H. S., Hann, H. J., Lee, S., Ryu, J.-H., Kim, S.-J., Kang, D.-H., Choi, K. B., and Ryu, D.-R. (2014). A population-based approach indicates an overall higher patient mortality with peritoneal dialysis compared to hemodialysis in korea. *Kidney International* **86,** 991–1000.

Korevaar, J. C., Feith, G., Dekker, F. W., van Manen, J. G., Boeschoten, E. W., Bossuyt, P. M., and T Krediet, R. (2003). Effect of starting with hemodialysis compared with peritoneal dialysis in patients new on dialysis treatment: a randomized controlled trial. *Kidney International* **64,** 2222–2228.

Kumar, V. A., Sidell, M. A., Jones, J. P., and Vonesh, E. F. (2014). Survival of propensity matched incident peritoneal and hemodialysis patients in a united states health care system. *Kidney International* **86,** 1016–1022.

Lehmann, E. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2014). Balancing covariates via propensity score weighting. *arXiv:1404.1785* .

Li, J., Fine, J., and Brookhart, A. (2015). Instrumental variable additive hazards models. *Biometrics* **71,** 122–130.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics* **9,** 215–234.

Li, Y., Lee, Y., Wolfe, R. A., Morgenstern, H., Zhang, J., Port, F. K., and Robinson, B. M. (2015). On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication. *Statistics in Medicine* **34,** 1150–1168.

Love, T. (2002). Displaying covariate balance after adjustment for selection bias. In *Joint Statistical Meetings*.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician* **65,**.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* **96,** 1245–1253.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960.

Marrón, B., Remón, C., Pérez-Fontán, M., Quirós, P., and Ortíz, A. (2008). Benefits of preserving residual renal function in peritoneal dialysis. *Kidney International* **73,** S42–S51.

Mehrotra, R., Chiu, Y.-W., Kalantar-Zadeh, K., Bargman, J., and Vonesh, E. (2011). Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. *Archives of internal medicine* **171,** 110.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science* **5,** 463–480.

Noordzij, M. and Jager, K. (2012). Survival comparisons between haemodialysis and peritoneal dialysis. *Nephrology Dialysis Transplantation* .

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* **54,** 387–398.

Raynor, W. J. (1983). Caliper pair-matching on a continuous variable in case-control studies. *Communications in Statistics - Theory and Methods* **12,** 1499–1509.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics* **56,** 779–788.

Rosenbaum, P. (2002). *Observational studies.* Springer.

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

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39,** 33–38.

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* **104,**.

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

Saran, R., Bragg-Gresham, J., Levin, N., Twardowski, Z., Wizemann, V., Saito, A., Kimata, N., Gillespie, B., Combe, C., Bommer, J., et al. (2006). Longer treatment time and slower ultrafiltration in hemodialysis: associations with reduced mortality in the dopps. *Kidney International* **69,** 1222–1228.

Sinnakirouchenan, R. and Holley, J. L. (2011). Peritoneal dialysis versus hemodialysis: risks, benefits, and access issues. *Advances in Chronic Kidney Disease* **18,** 428–432.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association* **103,** 924–933.

Staiger, D. O. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. *Econometrica* .

Tam, P. (2009). Peritoneal dialysis and preservation of residual renal function. *Peritoneal Dialysis International* **29,** S108–S110.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* **101,** 1607–1618.

Terza, J., Basu, A., and Rathouz, P. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics* **27,** 531–543.

Vonesh, E., Snyder, J., Foley, R., and Collins, A. (2006). Mortality studies comparing peritoneal dialysis and hemodialysis: what do they tell us? *Kidney International* **70,** S3–S11.

Weinhandl, E. D., Foley, R. N., Gilbertson, D. T., Arneson, T. J., Snyder, J. J., and Collins, A. J. (2010). Propensity-matched mortality comparison of incident hemodialysis and peritoneal dialysis patients. *Journal of the American Society of Nephrology* **21,** 499–506.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press.

Xie, J. and Liu, C. (2005). Adjusted kaplan–meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* **24,** 3089–3110.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics* **7,** 25–50.