Near/Far Matching - A Nonparametric Instrumental Variables Technique for Binary Outcomes

Mike Baiocchi†, Paul Rosenbaum PhD†, Dylan Small PhD† and Scott Lorch MD*
†University of Pennsylvania
*The Children’s Hospital of Pennsylvania

June 30, 2009

1Mike Baiocchi is a PhD Candidate, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6340 (E-mail: mbaiocch@wharton.upenn.edu). The authors are grateful to Marshall Joffe for valuable comments on a draft and to Matt White for insightful discussion. Corinne Fager also deserves special thanks for preparing, maintaining and updating the data sets.
Abstract

Instrumental variables (IV) is a framework for making causal inferences about the effect of a treatment based on an observational study in which there are unmeasured confounding variables. The most common form of IV estimation is two-stage least squares (2SLS), which works well when the outcome of interest is continuous. However, in many policy settings the objective is to estimate the effect of a treatment on a binary outcome. We propose a nonparametric matching technique - near/far matching - which is capable of estimating population level treatment effects when the outcome is binary. We provide a test statistic with standard errors. Our method also allows us to manipulate the strength of the instrument. We illustrate our method using a study of neonatal intensive care units (NICUs) treatment effect on premature babies (preemies) born in Pennsylvania.

Keywords: causal inference; econometrics; matching; instrumental variables; binary outcomes; strength of instrument
1. The treatment of interest, instruments and outline

1.1 The treatment of premature babies at NICUs

Neonatal intensive care units (NICUs) provide specialized care geared toward treating babies born prematurely (preemies).\(^1\) NICUs are usually part of a larger general hospital, though some are part of separate facilities entirely devoted to treating children. The notion is that specialized units can be highly trained and the physical layout of the environment can be constructed to provide high levels of care to preemies. Currently, the American Academy of Pediatrics recognizes 6 different levels of neonatal intensive care units, corresponding to their technical expertise and capability of caring for more medically complex infants. The levels, in ascending order, are 1, 2, 3A, 3B, 3C, 3D and regional centers.

Regionalized perinatal systems were developed in the 1970s when neonatal intensive care first began to save infants with a birth weight under 1500 grams. The goal for this system was to provide access to costly but potentially life-saving medical technology to all infants regardless of where they lived, even when the supply of neonatologists was too low and the cost of running a neonatal intensive care unit was too high to place these units in every hospital. High technology neonatal care was centralized in designated ”regional perinatal centers” where all high-risk mothers and infants were transferred. These centers also coordinated transport, educational services, and data evaluation for their geographic area. However, in the 1990s, neonatal intensive care services began to diffuse from regional centers to community hospitals, in response to wishes from patients and obstetricians to deliver infants closer to the home of patients.

Previous studies suggest that the outcomes of premature infants - specifically, in-hospital mortality - is improved if they deliver at higher level facilities [Phibbs and Ciufentes]. However, an empirical fact is that a large percent of preemies end up be-

\(^1\)Typically defined as babies born after no more than 37 weeks of gestation.
ing treated at type 1 NICUs. Although studies from five states in the United States [Powell (1995), Bode (2001), Yeast (1998), Shlossman (1997) and Menard (1998)] did not show any change in regional neonatal mortality rates after de-regionalization began - that is, the delivery of infranats at lower level NICUs increased - an outstanding question is: what would happen if all preemies were to be treated at a higher level NICU? It is unclear whether or not the current situation - a large percentage of preemies being treated in type 1 programs - is demonstrably worse than having all preemies attending high level NICUs. The policy question must focus on the cost and benefits of a policy shift.

In this paper we will focus on the benefit of attending a high level NICU: what is the percent reduction in mortality if all preemies were treated at a high level NICU rather than a low level NICU? To simplify this question for this paper, we consider the treatment to be "high level NICUs" - defined as levels 3A-3D and regional centers which deliver a large number of preemies per year - and the control to be "low level NICUs" - defined as all other NICUs.

Our study is similar to the seminal work of McClellan, McNeil and Newhouse (1994) in that we are most concerned about an unobserved selection process by which subjects with higher probabilities of mortality also have higher probabilities of attending the higher quality facilities. Analogous to McClellan, McNeil and Newhouse (1994), the instrument we use is the relative travel time to the different treatment facilities from the geographical centroid of the mother's home zip code. Relative travel time is defined as the travel time, from the center of the mother's home zip code, to the entrance of the closest high level NICU minus the travel time to the closest low level NICU. If a mother lives close to a high level NICU then the relative travel time will be positive. If a mother lives equidistant it will be zero. If she lives closest to a low level NICU it will be negative.

It is a little bit incorrect refer to level 1 NICUs as "low NICUs" because level 1 facilities are often considered to not be NICUs at all.
1.2 Using an instrument

A fundamental problem in making inferences about the causal effect of a treatment based on observational data is the potential presence of unmeasured confounding variables. Instrumental variables (IV) provides a framework for overcoming this problem. The method requires a valid IV, which is a variable that is associated with the treatment, is independent of the unmeasured confounding variables and has no direct effect on the outcome. The fundamental idea of the IV technique is to extract variation in the treatment that is unrelated to the unmeasured confounding variables and then use this variation to estimate the causal effect of the treatment. In this paper we use relative travel time to treatment facilities as our instrument.

The logic behind using relative travel time as an instrument is like so: When a woman selects where to live she does not consider her relative travel time to different level NICUs. When the woman becomes pregnant she has an increased probability of establishing a relationship with the proximal facility, regardless of the level, because she is not anticipating having a preemie. Proximity as a leading determinate in choosing a facility has been discussed in Phibbs (1993). During the first several months of treatment the woman goes to the selected facility to receive prenatal checkups. When the mother goes into labor prematurely she is likely to present at the hospital that she has developed a relationship with. Thus the women randomly assign themselves to be more or less likely to deliver in a high level NICU. A challenge to this logic is discussed in §6.5.

In epidemiological and health policy settings, the outcome of interest is often binary (e.g., mortality). The common technique of two-stage least squares (2SLS) is designed for continuous outcomes and has problems estimating treatment effects when the outcome of interest is binary. Bhattacharya (2006), through a simulation study, shows that 2SLS can introduce high levels of bias in the estimation of treatment effects when the observed probability of the outcome is up against the parameter
space (i.e., \( \bar{Y} \) is near 0 or 1). For a highly informative discussion of using 2SLS in settings with binary outcomes see Angrist (2001).

Led by analogy, some practitioners try using a two-stage method by replacing the second stage with a probit or logistic regression (and if the treatment is binary, the first stage as well). But this is fraught with problems, as described in Davidson and MacKinnon (1993, §7.6). See §3 for more discussion on the problems associated with two-stage methods.

The problem of estimating the treatment effect with binary treatment and binary outcomes is common enough that several methods have been developed - Rassen et al (2008). Many of the methods proposed to deal with binary outcome in an IV setting have been model based, maximum likelihood procedures. An excellent discussion of such a method can also be found in Bhattacharya (2006), with both an analytical comparison of methods and simulation study. In developing our technique, near/far matching, we make use of the IV framework but break from the parametric/model-based maximum likelihood approach.

In the common matching approach, the goal is to find subjects which are similar in their covariates. An ideal situation would be to observe the outcome under treatment and the outcome under the control on the same individual and then compare the difference between outcomes. Often, in the real world, giving a subject the treatment precludes giving the same subject the control, thus for any subject we can only observe the outcome given either the treatment or the control. By finding subjects who are very similar in the observed covariates we are getting closer to the idealized case of observing the outcomes given the treatment and the control for the same subject. Through matching we get closer to the idealized case, but we do not reach exact matches because there may be unobserved covariates that differ between subjects. Our proposed procedure, near/far matching, addresses the unobserved covariates issue.
1.3 An overview of near/far matching

Near/far matching can reasonably described by two criteria: First, we want to match subjects who are ”near” in their covariate space (i.e., the subjects share similar, or exact, values of their observed covariates). Second, we want to match subjects who are ”far” in their values of the instrument (i.e., the encouragement to receive the treatment is high for one subject in the match and low for others in the match). In the IV framework the treatment is allowed to be assigned by selection on either observed or unobserved covariates, but the instrumental variable itself is randomly assigned, perhaps conditional on observed covariates.

One additional issue with our example of near/far matching is the continuous nature of the instrument. The instrumental variable is a mothers’s distance to a treatment facility; thus every subject receives some level of encouragement. The idea being that: if one lives closer to a facility providing high intensity treatment then one is more likely to receive the high intensity treatment. In contrast, if one lives far from a facility providing high intensity treatment then one is less likely to receive the high intensity treatment. Because our instrument is travel distance to the treatment facility each subject has some level of encouragement to take the treatment. Therefore we do not operate under the more usual bipartite matching framework. That is, we cannot divide the subjects into two groups - ”encouraged” and ”unencouraged” before constructing the matched groups. We must use some form of nonbipartite matching. The schema we use to approach this problem was developed for dose matching in Lu et al. (2001).

Our method has several features applied researchers may find attractive. First, the method is capable of dealing with both continuous and binary outcomes as well as continuous and discrete instruments. Second, the method we propose allows the researcher to be agnostic about the underlying data generating function. Third, our method allows the researcher to manipulate the strength of the instrument. Fourth,
a common issue in complicated policy questions, where selection into the treatment is a concern, is that there may be little overlap in the covariates between the treatment and the control groups. This is particularly important because model based approaches, while appealing in many ways, tend to obscure the fact that extrapolation may be occurring because of the full confounding of a covariate with the treatment or control groups. For other authors’ discussion of this particular issue see Grieve et al. (2008) and Hansen (2008). Near/far matching, because it is based on matching, makes nonoverlap issues apparent. The commonly accepted reporting of balance make researchers and reviewers aware of the nonoverlap issues. While matching will not solve the problem of nonoverlap, it forces the researcher to confront the issue and facilitates finding a subset of the subjects on which an analysis may be performed. Fifth, near/far matching lends itself to a clear interpretation of the results and process by which the results were obtained. The general outline for this interpretation follows the common ”natural experiment” narrative but, because we are matching, it is easier to see our way through to a more traditional clinical trial sense of ”experiment.”

1.4 Outline of paper

Our paper is organized as follows. In §2, we describe the classic model based approach for IV and inference techniques. In §3, we describe the difficulties with binary outcomes in the IV framework. In §4, we develop the underlying test statistics for our method. In §5, we show the connection between near/far matching and an estimator already in the literature. In §6, we illustrate our method on a motivating example - estimating the treatment effect of neonatal intensive care units (NICUs). In §7, we provide conclusions and discussion.

2 Instrumental Variables Model

In this section, we describe an additive, linear, constant effect causal model and explain how valid IVs enable identification of the model. For defining causal effects,
we use the potential outcomes approach (Neyman, 1923; Rubin, 1974).

2.1 The Classic Instrumental Variables Regression Model

Let $y$ denote an outcome and $t$ denote a treatment variable that an intervention could in principle alter. For example, a simplified model of McClellan, McNeil and Newhouse’s 1994 study could be described as follows: $y$ is mortality (i.e., dead/alive), $t$ is intensity of treatment (i.e., high vs. low) and the intervention that could alter $t$ is for an individual to choose to receive a high intensity treatment. Let $y_i^{(t^*)}$ denote the outcome that would be observed for unit $i$ if unit $i$’s level of $t$ was set equal to $t^*$. We assume that the potential outcomes for unit $i$ depend only on the level of $t$ set for unit $i$ and not on the levels of $t$ set for other units – this is called the stable unit treatment value assumption by Rubin (1986). Let $y_{i}^{obs} := y_i$ and $t_{i}^{obs} := t_i$ denote the observed values of $y$ and $t$ for unit $i$. Each unit has a vector of potential outcomes, one for treatment and one for control, but we observe only one potential outcome, $y_i = y_i^{(t_i)}$. An additive, linear constant effect causal model for the potential outcomes as in Holland (1988) is

$$y_i^{(t^*)} = y_i^{(0)} + \beta t^*.$$  

(1)

Our parameter of interest is $\beta = y_i^{(t=1)} - y_i^{(t=0)}$, the causal effect of going from control to treatment. One way to estimate $\beta$ is ordinary least squares (OLS) regression of $y^{obs}$ on $t^{obs}$. The OLS coefficient on $t$, $\hat{\beta}_{OLS}$, has probability limit $\beta + Cov(t_{i}^{obs}, y_{i}^{(0)})/Var(t_{i}^{obs})$. If $t_{i}^{obs}$ were randomly assigned, then $Cov(t_{i}^{obs}, y_{i}^{(0)})$ would equal 0 and $\hat{\beta}_{OLS}$ would be consistent. But in an observational study, often $Cov(t_{i}^{obs}, y_{i}^{(0)}) \neq 0$ and $\hat{\beta}_{OLS}$ is inconsistent. One strategy to address this problem is to collect data on all confounding variables $q$ and then to regress $y^{obs}$ on $t^{obs}$ and $q$. If $t^{obs}$ is conditionally independent of $y^{(0)}$ given $q$ (i.e., the mechanism of assigning treatments $t^{obs}$ is ignorable per Rubin (1978)) and the regression function is specified correctly, this strategy produces a consistent estimate of $\beta$. However, it is often difficult to know and/or collect all confounding variables $q$. 

IV regression is another strategy for estimating $\beta$ in (1). A variable is a valid IV $z_i$ if it satisfies (see Angrist et al., 1996 and Tan, 2006):

(A1) $z_i$ is associated with the observed treatment $t_i^{obs}$; and

(A2) $z_i$ is independent of $\{y_i^{(t^*)}, t^* \in T\}$ where $T$ is the set of possible values of $t$; note that under model (1), this is equivalent to $z_i$ being independent of $y_i^{(0)}$.

The basic idea of IV regression is to use $z$ to extract variation in $t^{obs}$ that is independent of the confounding variables and to only use this part of the variation in $t^{obs}$ to estimate the causal relationship between $t$ and $y$. Assumption (A1) is needed to be able to use $z$ to extract variation in $t^{obs}$. (A2) is needed for the variation in $t^{obs}$ extracted from variation in $z$ to be independent of the confounding variables - that is, for the new variation to be "exogenous."

### 2.2 The Encouragement Design

An example of the usefulness of IVs is the encouragement design (Holland, 1988). An encouragement design is used when we want to estimate the causal effect of a treatment $t$ that we cannot control, but we can control (or observe from a natural experiment) the variable $z$ which, depending on its level, encourage or does not encourage a unit to have the treatment for $t_{obs}$. If the levels of the encouragement variable $z$ are randomly assigned (or the mechanism of assigning the encouragement variable is ignorable) and encouragement, in and of itself, has no direct effect on the outcome, then $z$ is a valid IV (Holland, 1988; Angrist et al., 1996). Our case study is a good example of an encouragement design in which it is assumed that the mechanism of assigning distance to the nearest NICU providing the high intensity treatment is ignorable, and that the closer a mother is to the NICU the more encouraged the mother is to receive treatment at the proximal facility. Further we assume that the assignment mechanisms have no direct effect on the preemies’ probability of death. The unobserved covariate which is most problematic is the preemie’s probability of mortality - the researchers would have preferred to have a treating physician’s as-
essment of the preemie’s mortality risk upon arrival at the facility, perhaps in the form of a probability of death without treatment. The selection problem may be that the NICU providing high intensity treatment are also receiving sicker preemies. Conditional upon knowing the true severity of the preemie’s condition (rather than just proxies) upon arrival at the treating facility, the treatment may be ignorable and the treatment effect may be recoverable.

As is usual in IV settings, it is possible to contest the necessary assumptions for the proposed instrument. Perhaps travel distance does not cause variation in treatment; it may be that patients with a severe condition may be willing to travel no matter the distance, therefore the instrument does not cause variation in the treatment and (A1) would not hold. This issue can be checked in the data (there is an association between the instrument and the treatment). But should the instrument be found to be deficient, improving the instrument may be challenging. Also, it is possible that the instrument is still associated with the unobserved severity of the patient. Imagine that the high intensity hospitals are located in urban centers. Further, imagine that the urban centers are highly polluted areas (think Manchester circa the Industrial Revolution) and one of the pollutants, which is present in the city is not present in the country, exacerbates severe acute myocardial infarctions. Now there is a causal pathway from the instrument through the problematic unobserved variable which causes variation in the outcome. In this situation the treatment effect is nonidentifiable. While addressing (A1) is difficult there is an approach to improve the chances of (A2) holding.

In order for assumption (A2) to be plausible, it is often necessary to condition on a vector of covariates \( \mathbf{x}_i \) (Tan, 2006). For example, in our case study we condition on gestational age, birth weight, congenital disorders and several other variables because these variables may be associated with both potential mortality outcomes \( y_i^{(r)} \) and distance from nearest hospital providing high intensity treatment. Conditioning on
\( x_i \) can also increase the efficiency of the IV regression estimator. We call the variables in \( x_i \) the included exogenous variables. Similar conditions to (A1) and (A2) can be derived whereby the instrument is conditionally valid (Angrist et al., 1996). We will refer to these conditions as (A1’) and (A2’). For the rest of this paper, we assume assumptions (A1’) and (A2’) hold for the proposed IV.

3. Methods for binary outcomes in the IV framework

In this section we discuss the common methods used for binary outcomes in the IV framework. We follow the excellent development of these issues described by Bhattacharya(2006).

3.1 The linear probability model 2SLS approach

Angrist (2001) suggests using linear probability models for both stages of the two-stage approach, essentially ignoring the dichotomous nature of the outcome and treatment. The clearest advantage of this method is its simplicity - executing a 2SLS in most statistical packages is trivial. In Bhattacharya (2006) the authors’ simulations show that this approach may be reasonable if the expected value of the outcome is in the range of 0.5. But their simulation shows severe bias in the estimation of the treatment effect as the expected value moves toward the edge of the parameter space - i.e., 0 or 1. This result is not too surprising given that the linear probability model performs the best with moderate values but is unconstrained and allows estimated probabilities above 1 and below 0.

3.2 Maximum likelihood methods

The approach Bhattacharya (2006), and others, have taken is to specify a model that takes into account the correlation between the selection to the treatment and the outcome. They then use a maximum likelihood approach to estimate the parameters of the model. If the underlying model is correct and all assumptions hold then these approaches have certain advantages. But there are several issues one may have with this approach. Most issues are practical and may prove to be obstacles for applied
researchers: (1) appropriate model diagnostics have not been extensively developed for such a model; (2) maximizing over a function with several parameters is tricky and it is not clear that convergence to the global maximum is guaranteed; (3) separation of covariates across the treatment and control is not immediately apparent (e.g., if only men received the treatment); in a model based approach, the usual accompanying diagnostics (e.g., t-statistics and residual plots) may not clearly expose confounding of the treatment and covariates; (4) obtaining standard errors for $\beta$ requires either the delta method or bootstrapping and if bootstrapping one should consider the issue of converging to the global maximum. While these issues are by no means insurmountable, they may prove to be steep barriers for many applied researchers. They surely act as pitfalls for the unaware researcher.

4. The Effect Ratio - A Nonparametric Potential Outcomes Statistic

In this section we break from the model-based approaches discussed above and introduce the effect ratio as a nonparametric population level treatment effect of interest. Near/far matching uses the effect ratio for inference.

4.1 The Effect Ratio

The effect ratio is predicated on matched sets. The language in this section allows for multiple controls being matched to one treatment, though in our motivating example we will match one treatment to one control.

Assume there are $I$ matched sets, $i = 1, \ldots, I$, and set $i$ contains $n_i \geq 2$ subjects, $j = 1, \ldots, n_i$, with one treated subject and $n_i - 1$ controls. The matched sets were formed by matching for an observed covariate $x_{ij}$ but may have failed to control an unobserved covariate $u_{ij}$. Said in a slightly different way, $x_{ij} = x_{ik}$ for all $i,j,k$, but possibly $u_{ij} \neq u_{ik}$. Write $N = \sum n_i$. If the $j^{th}$ subject in set $i$ receives the treatment, write $Z_{ij} = 1$. If the $j^{th}$ subject receives the control, write $Z_{ij} = 0$. Thus $1 = \sum_{j=1}^{n_i} Z_{ij}$ for $i = 1, \ldots, I$.

For any outcome, each subject has two potential responses, one seen under treat-
ment $Z_{ij} = 1$, the other seen under control, $Z_{ij} = 0$; this is in keeping with Neyman(1923) and Rubin(1974)’s potential outcomes framework. Here, there are two responses, $(y_{ij}^{(Z=1)}, y_{ij}^{(Z=0)})$ and $(t_{ij}^{(Z=1)}, t_{ij}^{(Z=0)})$ where $y_{ij}^{(Z=1)}$ and $t_{ij}^{(Z=1)}$ are observed from the $j^{th}$ subject in set $i$ under encouragement, $Z_{ij} = 1$, while $y_{ij}^{(Z=0)}$ and $t_{ij}^{(Z=0)}$ are observed from this subject if unencouraged, $Z_{ij} = 0$. To simplify the notation we will write: $y_{ij}^{(Z=1)} := y_{ij}^{(1)}$. The effect of the encouragement on a subject’s outcome, $y_{ij}^{(1)} - y_{ij}^{(0)}$, or treatment selection, $t_{ij}^{(1)} - t_{ij}^{(0)}$, are not observed for any subject. However, $Y_{ij} = Z_{ij}y_{ij}^{(1)} + (1 - Z_{ij})y_{ij}^{(0)}$, $T_{ij} = Z_{ij}t_{ij}^{(1)} + (1 - Z_{ij})t_{ij}^{(0)}$, and $Z_{ij}$ are observed from every subject. Let $F = \{(y_{ij}^{(1)}, y_{ij}^{(0)}, t_{ij}^{(1)}, t_{ij}^{(0)}, x_{ij}, u_{ij}), i = 1, ..., I, j = 1, ..., n_i\}$. Think of $Y_{ij}$ as the observed outcome variable, $T_{ij}$ as the observed treatment, $Z_{ij}$ as whether the subject was encouraged to take the treatment or control.

Write $|A|$ for the cardinality of a finite set $A$. Let $Z = (Z_{11}, Z_{12}, ..., Z_{In_i})^T$, let $\Omega$ be the set containing the $|\Omega| = \prod_{i=1}^{I} n_i$ possible values $z$ of $Z$, so $z \in \Omega$ if $z = (z_{11}, z_{12}, ..., z_{In_i})^T$. In a randomized experiment, $Z$ is picked at random from $\Omega$ so $P(Z = z|F) = 1/|\Omega|$ for each $z \in \Omega$.

The effect ratio, $\lambda$, is the parameter

$$\lambda = \frac{\sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^{(1)} - y_{ij}^{(0)}}{\sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} t_{ij}^{(1)} - t_{ij}^{(0)}}$$

(2)

where it is implicitly assumed that $0 \neq \sum_{j=1}^{n_i} t_{ij}^{(1)} - t_{ij}^{(0)}$. Here $\lambda$ is unknown and is a function of $F$. The effect ratio has the interpretation of “an average change of one unit in the treatment of the population causes a $\lambda$ unit change in the average outcome.” Rearranging (2) yields

$$\sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^{(1)} - \lambda t_{ij}^{(1)} = \sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^{(0)} - \lambda t_{ij}^{(0)}$$

(3)

4.2 Inference about an Effect Ratio in a Randomized Experiment
Consider the null hypothesis, \( H_0 : \lambda = \lambda_0 \). Define

\[
\tilde{T}(\lambda_0) = \sum_{i=1}^{I} \left\{ \sum_{j=1}^{n_i} Z_{ij}(Y_{ij} - \lambda_0 T_{ij}) - \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (1 - Z_{ij})(Y_{ij} - \lambda_0 T_{ij}) \right\}
\]

\[
= \sum_{i=1}^{I} \left\{ \sum_{j=1}^{n_i} Z_{ij}(y_{ij}^{(1)} - \lambda_0 t_{ij}^{(1)}) - \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (1 - Z_{ij})(y_{ij}^{(0)} - \lambda_0 t_{ij}^{(0)}) \right\}
\]

\[
= \sum_{i=1}^{I} V_i(\lambda_0)
\]

where \( V_i(\lambda_0) = \sum_{j=1}^{n_i} Z_{ij}(y_{ij}^{(1)} - \lambda_0 t_{ij}^{(1)}) - \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (1 - Z_{ij})(y_{ij}^{(0)} - \lambda_0 t_{ij}^{(0)}) \). Note that we use the notation \( \tilde{T} \) to denote the test statistic of interest and to differentiate this from \( T_{ij} \) the observed value of the treatment.

In a randomized experiment, \( E(Z_{ij}) = 1/n_i \), so that, using (3),

\[
E(\tilde{T}(\lambda_0)|F) = \sum_{i=1}^{I} \frac{1}{n_i} \left\{ (y_{ij}^{(1)} - \lambda_0 t_{ij}^{(1)}) - \sum_{j=1}^{n_i} (y_{ij}^{(0)} - \lambda_0 t_{ij}^{(0)}) \right\}
\]

\[
= (\lambda - \lambda_0) \sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} (t_{ij}^{(1)} - t_{ij}^{(0)})
\]

so that, under the null hypothesis, \( E(\tilde{T}(\lambda_0)|F) = 0. \)

Now,

\[
\text{var}(V_i(\lambda_0)|F) = E(V_i(\lambda_0)^2|F) - [E(V_i(\lambda_0)|F)]^2 \leq E(V_i(\lambda_0)^2|F)
\]

(4)

so that, in a randomized experiment,

\[
\text{var}(V_i(\lambda_0)|F) \leq \sum_{i=1}^{I} E(V_i(\lambda_0)^2|F) = E(S^2(\lambda_0)|F)
\]

(5)
where
\[ S^2(\lambda_0) = \sum_{i=1}^{I} V_i(\lambda_0)^2 \]

Moreover, in a randomized experiment, if the effects \( y_{ij}^{(1)} - y_{ij}^{(0)} \) or \( t_{ij}^{(1)} - t_{ij}^{(0)} \), were proportional,
\[ y_{ij}^{(1)} - y_{ij}^{(0)} = \lambda(t_{ij}^{(1)} - t_{ij}^{(0)}) \]
for \( i = 1, \ldots, I, j = 1, \ldots, n_i \), then \( E(V_i(\lambda)|F) = 0 \) for \( i = 1, \ldots, I \) and there is equality in (3) and (4), so in this special case \( \text{var}(T(\lambda_0)|F) = E(S^2(\lambda_0)|F) \).

Under the null hypothesis, \( H_0 : \lambda = \lambda_0 \) when \( I \) is large, the proposal is to approximate the randomization tail probability, \( P(T(\lambda_0)/S(\lambda_0) \leq k|F) \), by \( 1 - \Phi(k) \), where \( \Phi(\cdot) \) is the standard Normal cumulative distribution. A bit of care is required in considering the sampling distribution of the tail probability as we let \( I \to \infty \) under the randomization distribution. Let \( F_I = \{(y_{ij}^{(1)}, y_{ij}^{(0)}, t_{ij}^{(1)}, t_{ij}^{(0)}, x_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, \ldots, n_i\} \) describe the first \( I \) matched sets, and let \( \lambda_I \) be the corresponding effect ratio for the first \( I \) matched sets, which is a function of \( F_I \). First, we assume that \( F_I \) grows as \( I \to \infty \) in such a way that both \( \rho_I = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^{(1)} - y_{ij}^{(0)} \) and \( \delta_I = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{n_i} \sum_{j=1}^{n_i} t_{ij}^{(1)} - t_{ij}^{(0)} \) tend to limits \( \bar{\rho} \) and \( \bar{\delta} \) with \( \bar{\delta} > 0 \), with the consequence that \( \lambda_I \) tends to a limit, say \( \lambda_I \to \bar{\lambda} \). In the randomization distribution, given \( F_I \), the contribution \( V_i(\lambda_0) \) from the matched set \( i \) is random only because \((Z_{i1}, \ldots, Z_{in_i})^T \) is random, taking the \( n_i \) possible values \((1, 0, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, 0, \ldots, 1)\) with equal probabilities \( 1/n_i \), with mutual independence between the \( I \) matched sets. So the \( V_i(\lambda_0) \) are independent but not identically distributed. For \( \bar{T}(\bar{\lambda}_0)/\sqrt{I} \) to converge in distribution to a Normal distribution with expectation zero via a central limit theorem, say Theorem 9.2 in Breiman (1968, p. 186), as \( i \) increases, the individual \( V_i(\bar{\lambda}) \) must become neither degenerate nor dominant. Specifically, with \( \vartheta_i = E(|V_i(\bar{\lambda})|^3|F_I) < \infty \), Breiman requires \( \limsup(\sum_{i=1}^{I} \vartheta_i)/\sum_{i=1}^{I} \text{var}[V_i(\bar{\lambda})|F_I])^{3/2} = 0 \). In the application in the current paper \( y_{ij}^{(1)}, y_{ij}^{(0)}, t_{ij}^{(1)}, t_{ij}^{(0)} \) and \( n_i \) are bounded.
and not degenerate, which is sufficient for Breiman’s condition. Because strict inequality is possible in (4), this central limit theorem implies $P(T(\bar{\lambda})/S(\bar{\lambda}) \geq k|F_I)$ is tending to something no greater than $1 - \Phi(k)$, so that rejecting $H_0 : \lambda = \lambda_0$ when $\bar{T}(\lambda_0)/S(\lambda_0) \geq \Phi^{-1}(1 - \alpha)$ will, for sufficiently large $I$, reject a true null hypothesis with probability below or close to $\alpha$.

5. Near/Far Matching and the Anderson-Rubin IV Estimator

In this section we show that the Anderson-Rubin Test, developed in Anderson(1949), is a special case of our new method.

5.1 The Anderson-Rubin IV estimator

Consider the following instrumental variable model:

\[ Y_i = \beta T_i + \pi'X_i + u_i, \]
\[ Z_i \parallel u_i \]
\[ u_i \sim N(0, \sigma_u^2) \]

where the $u_i$ are iid. Anderson and Rubin proposed the following test of $H_0 : \beta = \beta_0$:

Consider the linear regression of $Y_i - \beta_0 T_i$ on $X_i$ and $Z_i$,

\[ E^*[Y_i - \beta_0 T_i | X_i, Z_i] = \omega'X_i + \tau Z_i, \quad (6) \]

where $E^*$ denotes the linear projection. If $\beta = \beta_0$ then $\tau = 0$. Thus, Anderson and Rubin proposed testing whether $\tau$ is 0 in (6) to test $H_0 : \beta = \beta_0$.

5.2 Near/far matching and the Anderson-Rubin estimator

Now, consider our setting with matched pairs and response $Y_{ij}$, treatment variable $T_{ij}$ and instrument $Z_{ij}$. Then (6) is the linear regression of $Y_{ij} - \beta T_{ij}$ on dummy variables for each matched pair and $Z_{ij}$. Using standard results for panel data linear regression, the sum of squares for this regression is the same as the sum of squares for the regression of $Y_{ij} - \beta T_{ij} - (\bar{Y}_i - \beta \bar{T}_i)$ on $(Z_i - \bar{Z}_i)$ where $\bar{A}_i = (A_{i1} - A_{i2})/2$ for
any variable A. If the coefficient $\tau$ is assumed to be 0, then the sum of squares is:

$$\sum_{i=1}^{I} \sum_{j=1}^{2} [Y_{ij} - \beta T_{ij} - (\tilde{Y}_i - \beta \tilde{T}_i)]^2$$

$$= \sum_{i=1}^{I} 2[\frac{Y^{(1)}_i}{2} - \beta \frac{T^{(1)}_i}{2} - (\frac{Y^{(0)}_i}{2} - \beta \frac{T^{(0)}_i}{2})]^2$$

$$= \frac{1}{2} \sum_{i=1}^{I} V_i^2$$

Now consider testing $\tau = 0$ in (6) by a score test. For a linear regression with normal errors, $Y_i = \alpha' X_i + u_i, i = 1, ..., N$ the score test statistic $Z_s$ for testing $\alpha_k = 0$ is the following, where $\hat{\alpha}^*$ is the restricted MLE when $\alpha_k$ is set to 0:

$$Z_s = \frac{\frac{\partial \log L(\hat{\alpha}^*)}{\partial \alpha_k}}{\sqrt{\text{Var}(\frac{\partial \log L(\hat{\alpha}^*)}{\partial \alpha_k})}}$$

$$= \frac{\sum_{i=1}^{n} (Y_i - (\hat{\alpha}^*)' X_i)(-X_{ik})}{\sqrt{\text{Var}(\sum_{i=1}^{n} (Y_i - (\hat{\alpha}^*)' X_i)(-X_{ik}))}}$$

For testing $\tau = 0$ in (5) the numerator of $Z_s$ is $\sum_{i=1}^{I} [-\frac{Y^{(1)}_i}{2} - \alpha \frac{T^{(1)}_i}{2} - (\frac{Y^{(0)}_i}{2} - \alpha \frac{T^{(0)}_i}{2})] = -\frac{\sum_{i=1}^{I} V_i}{2\sigma^2}$, where $Y^{(1)}_i, T^{(1)}_i$ are the response and treatment of the encouraged unit in pair $i$ and $Y^{(0)}_i, T^{(0)}_i$ are the response and treatment of the unencouraged unit in pair $i$.

Note that when $\beta = \beta_0$ and the model we are considering holds, $V_i = u^{(1)}_i - u^{(0)}_i$.
and the $V_i$ are iid $N(0, 2\sigma_u^2)$. Thus,

\[
\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} V_i^2}{2}
\]

\[
Z_s = \frac{-\sum_{i=1}^{I} V_i}{2\sigma_u} \sqrt{\frac{I}{2\sigma_u^2}}
\]

\[
= -\frac{-\sum_{i=1}^{I} V_i}{\sqrt{2\sigma_u}} \sqrt{\frac{I}{2\sigma_u^2}}
\]

\[
= -\frac{-\sum_{i=1}^{I} V_i}{\sqrt{\sum_{i=1}^{I} V_i^2}}.
\]

Therefore, $Z_s$ is equal to the negative of our statistic and the confidence interval based on the Anderson-Rubin test using a score statistic is equivalent to our confidence interval. Note that the Anderson-Rubin test is usually based on the F-statistic rather than the score statistic so would not be exactly equivalent to our test.

6 Illustrative Example: Neonatal Intensive Care Units

In this section we use near/far matching to estimate the treatment effect of high level vs. low level neonatal intensive care units. See §6.2 for a technical description of near/far matching. Much of the methodology of this section finds its roots in Lu et al. (2001)’s paper on matching with doses.

6.1 The Data

All preemies born in hospitals in the state of Pennsylvania in the years 1995-2004 and the first six months of 2005 are recorded in the data set. We have data on a little over 200,000 preemies.

We use data from the hospital UB-92 form linked to birth and death certificates. Each hospital submits a UB-92 form to the state for each course of care. This form records fifteen to twenty-five fields for principal diagnoses and procedures that occurred during the hospital stay, as well as length of stay and discharge status. Information about congenital disorders are also included in these forms. Congenital
disorders were considered a measurement of illness severity because NICUs cannot influence whether an infant is born with these disorders, but the presence of these conditions increases the risk of death. From the birth and death certificates of the infant, we also have information on birth weight, gestational age, mother’s education level, race/ethnicity, prenatal visits, parity, and whether the infant was born as part of a multiple gestation birth. Using the mother’s zip code contained on both the birth certificate and the hospital UB-92 form, we also have information from the US Census on several socioeconomic variables, including median income, median home value, percent renting, percent below poverty line, percent completed high school and percent completed college.

The level of the NICU where the infant delivered is coded as a binary variable - i.e., 1 if the hospital delivers more than 50 preemies and is of level 3A-3D or regional; the variable is coded as 0 if it is any other hospital. Using ArcView software from ESRI, Inc. (Redlands, CA), the travel time from the geographical centroid of the mother’s zip code to all hospitals in the state are calculated. The instrumental variable of interest in this paper is the relative travel time between the closest high level NICU and the closest low level NICU. The higher this variable, the more distance a woman would have to travel to deliver at a high level NICU versus a low level NICU, which is typically the closest hospital to that woman’s zip code.

The data set contains 10.5 years of births in one of the most populous states in the United States. Because of the size of our data set we will approach our problem with several techniques meant to deal with large data sets. Future research is required for approaches in which the techniques we use are not an option.

6.2 The technical aspect of near/far matching

The instrument in this case, relative travel time, is a quantity that all subjects in the study have. Hence, our environment for matching is not the typical "treated-to-control" matching, in which only subjects who were in the control can be matched
to those in the treated. We cannot break the subjects into two disjoint groups and then match between the groups. Instead, each subject could be matched to any other subject. Thus we cannot use the typical matching algorithms and we need to carefully consider the criteria for selecting our matches so as to best match in the covariates yet posses the greatest difference in the instrument.

The type of matching we are performing is commonly referred to as nonbipartite matching (Papadimitriou and Steiglitz (1998, sec 11.3)). Most directly, the work in this paper follows from the work of Lu et al (2001) on dose matching. Their ”dose” is our instrument. In addition to the theory they develop, we also use code developed for their project. All of the nonbipartite matching is based on Derigs (1988)’s work on finding optimal matches in a nonbipartite scheme.

As in Rubin (1980) we use the Mahalanobis distance of the covariates of the subjects to produce scalar measurements of the difference between any pair of subjects. If there are N subjects in the study then one can think of creating an $N \times N$ distance matrix, call this matrix $M$. Entry in $M_{ij}$ is the Mahalanobis distance between the $i^{th}$ and the $j^{th}$ subjects. If we were only interested in finding pairs of preemies who are most similar in their observed covariates we would stop here and put the distance matrix $M$ into the Derigs algorithm. Covariates for each pair would be more balanced, thus the bias from these observed covariates would be reduced. However, in an observational study, we are concerned about the unobserved covariates as well. A good balance on the observed covariates does not imply a good balance on the unobserved covariates. However, in near/far matching, we have access to an instrumental variable that randomizes the treatment and control and thus the unobserved covariates. Without near/far matching, within a given matched pair there may not be much difference between the value of their instrument. If both preemies within a matched pair were to have the same value of the instrument, then there is no difference in the amount of encouragement to take the treatment. This pair would be noninformative
because the random variation that comes from the instrument is nonexistent. We want to create pairs in which one preemie lived very close to a high level NICU and the other lived very close to a low level NICU.

Obtaining the most informative matches will require considering the instrument. The strategy taken in this work is to do a second pass over $M$. The first step is to construct a distance matrix containing the Mahalanobis distance summary of the covariate distance between each pair. The second step is to add a penalty to the distance matrix for any pair which has a difference in instrument below a prescribed threshold. For example, consider the $i^{th}$ and the $j^{th}$ subjects. After the first pass, $M_{ij}$ is a number reporting the difference in the covariates of the two subjects. In the next pass over $M$, if subjects $i$ and $j$ both live close to a high level NICU then we add some additional amount to $M_{ij}$. We add to $M_{ij}$ because we want the nonbipartite matching algorithm to be less inclined to match the pair because they are uninformative, because they are both encouraged to go to the high level NICU. If $i$ were close to a low level NICU and $j$ were close to a high level NICU then we would not change $M_{ij}$ after the first pass.

Consider $M$ after the first pass. $M$ is a summary of how different each pair of preemies is from one another in their covariates. Now consider $M$ after the second pass. $M$ is a summary of the covariates and the difference in the instruments. $M_{ij}$ takes on large values if the $i^{th}$ and the $j^{th}$ preemies are very different in their covariates and very similar in their relative travel time. $M_{ij}$ is at its smallest when the covariates are very similar but the instrument takes on very different values. After the second pass, the distance matrix $M$ will force the nonbipartite matching algorithm to match the pairs which are near in their covariates and far in their instrument values. This produces a matching with low bias due to difference in covariates between the groups and the most informative matches because there is a larger discrepancy between the instrument (i.e., encouragement).
Note that near-far matching is not predicated on nonbipartite matching. It works in the dichotomous instrument setting as well. In this setting the "far" part of near/far matching becomes trivial: we create two groups - encouraged and unencouraged based on the instrument. These two groups are as far from one another in the instrument as possible and thus the matching program becomes bipartite.

In the continuous instrument case, we can control the strength of the instrument by increasing the threshold. Note that in the second stage of designing the distance matrix, $M$, we use thresholding to add a penalty to $M_{ij}$ if the subjects $i$ and $j$ are too similar in their instrument values. That is, we penalize pairs that have similar values of the instrument by adding distance to $M_{ij}$ in the distance matrix (even though the preemies may be similar in their covariates). As long as the instrument’s strength increases monotonically, increasing the value of the threshold will increase the strength of the instrument.

6.3 Implementing near/far matching

As mentioned earlier we have a little over 200,000 preemies from the 10.5 years of data. This is an overwhelming number of subjects to consider at one time. To make the problem more manageable we create subsets of the data and match within those sets. For example, we first subset within year - that is, only preemies born in 1995 are matched to preemies matched within 1995. The next level of partitioning we do is to group preemies by birth weight and gestational age. Medical professionals believe that birth weight and gestational age are two of the most influential variables in determining mortality [Hack (1995), Fanaroff (1995), Stevenson (1998) and Anonymous (1993)], so by subsetting and only matching within these groups we create pair matches that are very close in birth weight and gestational age.

To improve the quality of the matches we also make use of sinks, or extra phantom preemies. Sinks allow the most unique preemies in our data set to be pulled out of the analysis. Suppose there is a preemie whose covariates are not similar to any of
the other preemies in the data set. If the algorithm is required to match this preemie to another preemie then it will produce a mismatched pair - a perfectly reasonable looking preemie will have to be matched to this very dissimilar preemie. But, instead of matching this unique looking preemie to a regular preemie, we can match the dissimilar preemie to a sink. If the dissimilar preemie is matched to a sink then the perfectly reasonable preemie is free to pair off with another preemie that is more similar, thus improving the balance of the covariates across all of the matches. We decrease the number of preemies included in the analysis in order to improve the quality of the matches.

Implementing sinks in the matching process is relatively simple. The idea here is that if we have the $N \times N$ distance matrix $M$ created as described above, we can add additional rows and columns to it. These additional rows and columns are "phantom" preemies who are not actually in our data set, the reason we are adding them to the data set is that these sinks can be matched to some of the more difficult preemies to match, therefore improving the overall quality of the match. To see how this works, say we start with the $N \times N$ distance matrix $M$ and want to add $s$ sinks. We would add $s$ rows and $s$ columns to $M$ so we would now have an $(N+s) \times (N+s)$ matrix. The sinks are allowed to be as close as possible to the actual preemies, that is $M_{ij} = 0$ for $i = 1, 2, ..., N$ and $j = N + 1, N + 2, ..., N + s$ and (because $M$ is symmetric) $M_{ij} = 0$ for $i = N + 1, N + 2, ..., N + s$ and $j = 1, 2, ..., N$. But a sink is not allowed to be matched to another sink, that is $M_{ij} = \infty$ for $i = N + 1, N + 2, ..., N + s$ and $j = N + 1, N + 2, ..., N + s$. The sinks are as close as possible to all of the actual preemies, so an optimal matching will place the most difficult to match preemies in a pair with a sink instead of another preemie. Once a preemie is matched to a sink it is removed from the analysis. This improves the quality of the covariate balance between the encouraged and unencouraged groups.

There are certain limitations to our implementation. By constructing subsets we
lose the continuous nature of birth weight and gestational age. Within one subset, we only allow preemies with a birth weights between 2,500g to 3,000g to be matched. Physiologically there maybe no real difference between a preemie that weighs 2,999g and a preemie weighing 3,001g. But because we are subsetting, the 2,999g is allowed to be matched to a 2,500g but not the 3,001g preemie. We are also ”losing” data by allowing preemies to be matched to sinks. It seems troubling at first that we are losing information on this preemies. But there are benefits to the use of sinks. By using sinks we are decreasing the bias in observed covariates between the two groups. The difference between the covariates of the two groups is lowered as we increase the percent of sinks. This problem, not being able to include some preemies in our analysis, is in analogy to clinical trials. In clinical trials subjects with highly unusual values of covariates are often excluded from the trial. By using sinks we lose the ability to comment on highly atypical preemies experience in NICUs, which is the price of decreased bias due to covariate imbalance. This tradeoff causes some pause when thinking about the generalizability of the inference, but it is highly probable that we are drawing from a more heterogeneous population in our observational study than the one in a clinical trial would be allowed to draw from.

6.4 Assessing Balance

The goal of matching is to achieve balance in the observed covariates between the encouraged and the unencouraged groups. We have been discussing balance, but what constitutes a good balance? The question of balance assessment is an open area of research - Rubin and Thomas (2000).

Tables 1, 2 and 3 (which you can find at the end of this document) summarize the means of the covariates for different subgroups of the preemies. Table 1 shows the difference in covariates between the preemies who were delivered at a high level NICU versus the preemies delivered at the low level NICUs. The preemies delivered at the high level NICUs tended to have lower birth weights and were born at a
lower gestational age. These two variables are commonly identified by physicians as associated with the patient’s mortality risk. This implies that there may be serious unobserved, medically significant, covariates which are also unbalanced between the two groups.

Table 2 shows the balance after near/far matching. The birth weight and gestational age between the two groups are very similar. Also, given that near/far matching uses an instrument, we now have reason to believe that unobserved covariates have been randomized to some degree. (See §6.6 for a discussion of Table 3 and how to interpret it.)

To assess the quality of several covariates, we will consider quantile-quantile plots (Q-Q plots) for birth weight, gestational age and the percentage of households renting their home in the mother’s residential zip code. Q-Q plots are a way to compare the distribution of two groups of variables: (1) preemies who attended high level NICUs vs. low level NICUs (before near/far matching) and (2) the preemies we assigned to the encouraged vs. unencouraged groups (after near/far matching). The closer the dots in the plots lie to the line, the better the quality of the balance.

Figure 1: Q-Q plots of birth weight and gestational age before near/far matching.
From Figure 1 it can be seen that the distribution of birth weights and gestational ages deviate noticeably from one another - that is, the data veer away from the line. Also, because gestational age is reported as an integer in the birth certificate data, it appears as a stairway shape in this plot.

Figure 2: Q-Q plots of birth weight and gestational age after near/far matching.

The distribution of birth weight and gestational age after the near/far matching is shown in Figure 2. Both variables have a greater degree of agreement in the covariates' balance. These two covariates we forced to match very closely, but some of the other variables we were not as easily matched. For example, the percentage of households renting their home in the mother’s zip code was difficult to match.

From Figure 3, it is apparent that rent balance between the two groups is not as well matched as gestational age and birth weight. While not perfect, this match is the worst of the covariates and may not be as highly associated with mortality as other covariates.

The common rule of thumb is to match balance such that the covariates do not differ more than .2 of a standard deviation. This rule is used for covariates not consider highly important. The highly important variables - gestational age, birth
6.5 Generalizability

We should consider the generalizability of our results. Suppose one-hundred percent of preemies with a life-threatening congenital anomaly are sent to high level NICUs. In this case we have no variation due to the instrument for this subset of the preemies. We would not have any of the data needed to build a contrafactual model of what would have happened had a preemie with a life-threatening congenital anomaly gone to a low level NICU. Therefore our estimate would not including such preemies. Our estimate would not be answering the question of ”what would happen if there were no high level NICUs” or ”what is the sum total of the treatment effect of high level NICUs” because we would be unable to comment on what would happen if preemies with a life-threatening congenital anomaly were to be delivered at low level NICUs. If these preemies tend to have worse outcomes if they are delivered at low level NICUs then we would underestimate the total mortality-lowering effect of high level NICUs. While we would like to say something about the effect for this particular
subset of preemies, the responsible statistician must refrain. Statistical procedures are designed to deal with noise, not with silence.

By using sinks we are pealing off the most unique preemies. The preemie who has no peer with similar covariates is matched to a sink. This is analogous to the constraints researchers experiences in clinical trials; subjects too dissimilar from the general public are excluded from the trial. This exclusion could be due to aberrant size or age. In the clinical trial this could be due to the health of the subject. There are ethical reasons and practical reasons a subject might not be included in a study. As we are dealing with an observational data set we do not encounter many of these constraints resulting in a sample being drawn from a larger set of the population. In our case, the set we are drawing from is the entire population of the state. Thus we are able to sample from a larger set than a clinical trial may have been able to sample from. This is an advantage. The penalty for being able to sample from this set is that we must use an instrument which is inherently less reliable than a researcher-controlled randomization process.

6.6 Results

There is reason to believe that this instrument works well for the medically significant covariates. As you can see in Table 1 (all tables are at end of document) the average birth weight of preemies delivered in high level NICUs is lower than those born in low level NICUs - 2,518g to 2,743g, which is almost a third of a standard deviation difference. Medically speaking, heavier babies tend to thrive better than their smaller counterparts. This is similarly the case with gestational age. Preemies born in high level NICUs were, on average, 34.79 weeks of gestation versus 35.81 weeks at low level NICUs. The distributions of both variables suggest there is some appreciable level of selection before the implementation of near/far matching. But consider Table 3, which compares the covariates of preemies across the quartiles of the instrument (i.e., the first quartile are preemies with mother’s living closest to a high
level NICU, the fourth quartile are closest to low level NICUs). The birth weights and gestational ages are quite similar across the different quartiles of the instrument. This is reassuring because the instrument appears to randomize the medically significant variables which are observed. This fact supports the belief that the unobserved medically significant variables are also being randomly assigned. This would make assumption (A2') more believable.

Also seen in Table 3 is that as we go from the lowest quartile of the instrument to the highest quartile of the instrument we see a monotonic decrease in the percentage of preemies delivered at high level NICUs. Recall that the instrument is difference in travel time - a large positive number means the closest high level NICU is farther away than the closest low level NICU. As the instrument goes from low to high we see an increase in the encouragement for a mother to deliver at the high level NICU. Thus assumption (A1') is believable.

Interestingly, other covariates seem to be affected by the instrument. For example, the average income is high, then lower, then highest, then low again as distance from the closest high level NICU increases. The key to understanding this is that most high level NICUs are located in urban environments and what we are seeing can be roughly be described as area-level socioeconomic status of the center of the city, outer rim of the city, suburbs and rural communities. We have several variables which measure socioeconomic status (SES), so we must be careful to monitor these variables in the matching stages because the instrument may not completely randomize their distribution or is random only conditionally.3 For our purposes, it is not too problematic that SES is not being randomized because the typical causal pathways suggested by the medical literature have any effect on preemie mortality.

3Note that the instrument need not be "random" in the sense of a coin flip or draws from a box. In fact, the process by which the value of the instrument is assigned may be entirely deterministic. The requirement for a good instrument is better understood as: the assignment mechanism is uncorrelated with the causal pathway of interest. The assignment process for the instrument values must be, in some sense, orthogonal to the causal process of interest.
passing through birth weight or gestational age.

We started with a little more than 200,000 preemies born in the 10.5 years of our data set. We matched roughly half of those preemies to sinks in order to improve the balance between the two groups, resulting in 49,953 matched preemies. A summary of the matching between the encouraged and unencouraged can be found in Table 2. We estimate a point estimate of a 0.939% decrease in mortality if all preemies were to be treated in high level NICUs rather than low level NICUs (remember: this is an estimate for only the preemies sensitive to the instrument that we have used to estimate the effect - i.e., the compliant preemies). We calculate a 95% confidence interval of $(-1.271\%, -0.606\%)$.

What is interesting in this result is that this estimate is just for the preemies affected by the instrument (see §6.5). Thus, our estimate of a decrease in mortality of 0.939% is for preemies for which there is no consensus on the necessity to treat them at a high level facility. A policy intervention may have the desirable consequence of decreasing preemie mortality.

7. Conclusions and Discussion

In this paper we have introduced a nonparametric instrumental variable technique which is capable of estimating a treatment effect even when the outcome and treatment variables are binary. The test statistic we propose is the effect ratio - which can be thought of as a population level estimate of the treatment’s effect on the outcome. We use matching to create the most informative matches to estimate the effect ratio. We call our method near/far matching in order to emphasize the criteria we are looking for for the most informative matches - near in the covariate space and far in the instrument space.

Our method has several features that applied researchers may find to be to their liking. The first is that this is a nonparametric approach, meaning inference is not predicated on a particular model specification. The second is that we have constructed
standard errors and can therefore invert the hypothesis test to construct a confidence interval, without making appeals to bootstrap or the delta method. A third advantage is, in the nonbipartite setting, we have a procedure to control the strength of the instrument. By increasing the threshold a pair must be separated in the instrument, the researcher can create a stronger instrument. Nothing comes for free, by increasing the threshold the researcher must either be willing to accept worse balance in the covariates or perhaps be willing to match more of the subjects to sinks and thereby lower the number of pairs included in the inference.

Our implementation of near/far matching is based on matching. Matching has a few advantages. The first advantage is that matching forces the researcher to be aware of imbalance in the covariates between the treated and control. In model-based settings, an unaware research is quite likely to miss the fact that there is nonoverlap in the covariates - that is, a covariate may be confounded with the treatment, rendering it impossible to attributed the difference in outcomes to the variation in treatment or the variation in the confounded covariate. Often, in observational studies nonoverlap is a serious issue that must be addressed. A second, and perhaps less practical issue, is that the near/far matching framework lends itself nicely to the ”natural experiment” paradigm. The ”natural experiment” description of observational studies is usually described like so: Just by accident, through some intervention by nature, subjects were randomized to either the treatment or the control. In the near/far matching, we are attempting to find pairs of subject who were very similar in their covariates but very different in their instrument - thus we are locating, within our observational data set, the experiment which was embedded in the real world. Hidden, somewhere, in all of those preemies was an experiment that could have taken place in a laboratory - where the randomization was carried about by assigning expectant mothers to live in different parts of the state. Our job was just to find the experiment which nature hid. This may seem a silly point to make, but this method’s interpretability may
afford it acceptance in research areas more comfortable with the more traditional, lab-based experimental design methods. Instead of having to justify a particular model specification we make appeals to the validity of our instrument and the randomization inherent within.

This particular implementation of near/far matching, using Derigs’s nonbipartite matching, may certainly be improved upon. Faster algorithms and more precise definitions of the optimality of matched pairs may be proposed. In future research we anticipate using better tuned implementation procedures. Another area for improvement is to consider other forms of separating the instrument. We follow in the Lu et al. (2001) dose matching framework, which uses a form of thresholding to penalize pairs too similar in their dose. But other schemes may, in fact, create better overall matches. Perhaps a form of inverse weighting of the difference in travel time would have worked better in our setting. This is an issue we may return to in future research.

References


Phibbs CS, Mark DH, Luft HS, Peltzmanrennie DJ, Garnick DW, et al. (1993)
Rassen, J., Schneeweiss, S., Glynn, R., Mittleman, M., and Brookhart, M., (2008),


*Journal of the American Statistical Association*, 90, 151-156.