Estimating Individual Causal Effects

Estimating Individual Causal Effects

A dissertation presented by

Patrick Kenneth Lam

to

the Department of Government

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of Political Science

Harvard University
Cambridge, Massachusetts
September 2013

Dissertation Advisor: Professor Gary King Patrick K. Lam

Estimating Individual Causal Effects

Abstract

Most empirical work focuses on the estimation of average treatment effects (ATE). In this dissertation, I argue for a different way of thinking about causal inference by estimating individual causal effects (ICEs). I argue that focusing on estimating ICEs allows for a more precise and clear understanding of causal inference, reconciles the difference between what the researcher is interested in and what the researcher estimates, allows the researcher to explore and discover treatment effect heterogeneity, bridges the quantitative-qualitative divide, and allows for easy estimation of any other causal estimand.

The framework I develop for estimating ICEs starts from the potential outcomes framework and then combines existing methods for matching in causal inference with a Bayesian model to impute missing potential outcomes. Researchers can use the resulting posteriors for the ICEs to derive the posterior for any other causal estimand by simple aggregation. In my dissertation, I first lay out the basic framework and estimation strategy. I then compare various models via simulation to test the effectiveness in recovering the true ICEs. Finally, I apply the model for estimating ICEs to two applications: a randomized field experiment on monitoring corruption from Olken (2007) and an experiment on the effectiveness of job training programs. I show the flexibility of the model in estimating ICEs for different types of outcome and treatment variables as well as with two-stage models using instrumental variables. I also show the various ways one can use the model to detect treatment heterogeneity and estimate a large number of different causal estimands.

Contents

List of Tables

List of Figures

Acknowledgments

It is impossible to adequately thank the people who have been instrumental in getting me to where I am today. First and foremost, my committee deserves my utmost gratitude.

Gary King is everything that you would want in an advisor and chair. From reading drafts to providing research opportunities and career advice to building the amazing environment that is IQSS, I can't thank him enough. I look forward to many more years of collaborating with him.

Jim Alt has been there from the start, first as Director of Graduate Studies and then as a member of my committee. His input and contributions to my graduate career have been enormous.

When I first entered graduate school, I had no intention of being a methodologist. Adam Glynn taught the first class in the methods sequence in the department, and I haven't looked back ever since. I thank him for constantly allowing me to walk into his office randomly and answering all my questions or walking through the math with me. His vast knowledge in statistics and causal inference has been invaluable in shaping my thoughts.

Every committee needs a taskmaster to keep you on track and check up with you from time to time. Arthur Spirling played that role for me, and for that I am grateful. Not only were his insights and contributions to my career valuable, but he writes some of the coolest papers I've seen and I can only hope that some of his ideas rubbed off on me.

I have also benefited tremendously from simply being around smart and wonderful people at Harvard. Don Rubin's work has probably been the single biggest influence on my dissertation. I was fortunate to be able to take multiple courses with him and pick his brain. I thank him for entertaining my ideas when they were not fully developed and offering suggestions. Luke Miratrix also influenced my thinking in its formation stages. In graduate school, the people who ultimately influence your work and shape your career the most are your fellow graduate students. For being the best intellectual peers, I thank Matt Blackwell, Andrew Coe, Andy Eggers, Ben Goodrich, Justin Grimmer, Jenn Larson, Rich Nielsen, Iain Osgood, Maya Sen, Miya Woolfalk, and many others.

Thom Wall deserves my utmost appreciation as the department graduate program administrator. He keeps everything running smoothly and has been a constant presence from my first visit to Harvard during admit week to the end of the process submitting my degree application and dissertation. I don't know what the department would do without him. For institutional and financial support, I thank the Harvard Institute for Quantitative Social Science, the Department of Government, and the Graduate School of Arts and Sciences. For various great conversations and giving me the opportunity to apply my skills to the real world, I thank Alexis Diamond at the IFC.

I would be remiss not to thank the many influences during my undergraduate time at UCLA. It was at UCLA that I first learned about political science and statistics and I would certainly not be where I am today without them. Ron Rogowski was my undergraduate advisor and an amazing one at that. Without him, I would have never made it to graduate school. He was an enormous influence when I was an undergraduate and even during my grad school years. I am lucky to consider him a friend to this day. For opening my eyes to the world of statistics in political science, I thank John Zaller and Marisa Kellam. For first introducing me to the world of statistics outside of OLS, I thank James Honaker. He has been a constant influence in my methods training and thinking even to this day. Other influences and collaborators from UCLA include Drew Linzer, Brian Min, Hans Noel, and Mike Thies.

For providing the job training data used in this dissertation, I thank Paolo Frumento. For making his corruption monitoring data so easily available that I did not even need to ask, I thank Ben Olken. And for the tedious task of reading and editing this dissertation, I thank my friends Hillary Chu and Kyle Peyton.

I don't know where I would be without my friends and my parents. I'd like to thank them for being a constant source of support and encouragement.

Chapter 1

A Framework for Estimating Individual Causal Effects

1.1 Introduction

What is the effect of political institutions on economic growth? Does UN intervention shorten the length of wars? Do job training programs increase wages and employment prospects? Does aspirin lower blood pressure? Researchers and scholars in every facet of industry and science grapple with causal questions all the time, using randomized studies and/or observational data to answer these questions. Almost always, the answers come in the following form: "there is a positive/negative causal effect[1] of the treatment on the outcome *on average*." Almost all research focuses on estimating the average causal effect, which is defined as the average of all the causal effects for every individual.[2] Yet in almost all cases, the average causal effect is not a specific causal effect for any one individual. Thus, there is a strong disconnect between what researchers generally measure (the average causal effect) and their actual quantity of interest (the causal

[1] I use "treatment effect" and "causal effect" interchangeably throughout.

[2] I use the terms individual, observation, and unit interchangeably throughout.

effect for person i or country j).

The causal literature is quite clear on the difference between the average and individual-level causal effects. Under the potential outcomes framework, which dates back to Neyman but was formally defined and popularized under the "Rubin Causal Model" (Rubin 1974), let W be a binary treatment variable taking on a value of 1 if a unit receives treatment and 0 if it receives control. The potential outcomes $Y(1)$ and $Y(0)$ represent the unit's outcome if it had received either treatment or control. The **individual causal effect (ICE)** for individual i is simply the difference between its potential outcomes under treatment and control.

$$\tau_i = Y_i(1) - Y_i(0)$$

Since at most one of the potential outcomes for each unit is observed, one cannot observe the causal effect of the treatment on the outcome. Rubin (1978) and Holland (1986) refer to this as the *fundamental problem of causal inference.*

Almost every causal inference introduction begins with the ICE, yet quickly moves on to ways of identifying the average treatment effect (ATE).

$$\begin{aligned}\tau_{ATE} &= E[Y(1) - Y(0)] \\ &= E[Y(1)] - E[Y(0)]\end{aligned}$$

The ATE is easier to identify because one only needs to identify the means of the marginal distributions of the two potential outcomes. Standard regression techniques that are widely used and easy to implement have made the ATE the default quantity of interest. However, I argue that the focus on the ATE and various other average effects, while easier to estimate, loses a lot of potential information about treatment effect heterogeneity and has important implications for both research and policy. In this paper, I present a unified framework for estimating individual causal effects using many of the same tools already in place for estimating ATEs. I argue for a reorientation of the causal inference literature back toward estimating individual causal effects and expound upon the benefits of such an approach.

1.2 The Case for ICEs

Consider the following two statements:

- The treatment effect of W on Y is $\hat{\tau}$.
- Our model predicts that an increase of one unit of W increases Y by $\hat{\beta}$.

Variations of both statements are standard ways of describing causal effects in studies where a treatment variable W purportedly affects an outcome of interest Y. Whether the treatment effect is estimated from a regression model, from an experimental design, or from other forms of estimation, the estimate is usually some average treatment effect, yet the language is often unclear as to the units of interest. In the two statements above, $\hat{\tau}$ and $\hat{\beta}$ are average treatment effects, but it is important to note what average treatment effects represent. An ATE is not the effect of treatment on any one individual in the data (in most cases). An ATE is not the effect of treatment on a hypothetical individual with a given set of covariates. An ATE is not the effect of treatment for an average individual. Strictly speaking, an ATE is simply the average of all the individual effects for the individuals in the data. By reframing ATEs and other causal quantities in terms of aggregations of individual effects, estimating ICEs can ***allow for a more precise and clear understanding of causal inference***. Possible confusion over what ATEs represent can be cleared up by referring to them as average effects of certain groups of individuals.

Often times, there is a temptation to apply the ATE to individuals of interest, such as in the case of using a regression coefficient to predict outcomes for future or counterfactual observations. There is a disconnect between what researchers are interested in, which is the effect for certain individuals or groups of individuals, and what researchers estimate, which is an average effect. For example, academics may be interested in explaining the effect of treatment in certain individuals, while policymakers may be interested in predicting the treatment effect for certain individuals. Rarely are researchers actually interested in “the average effect” per se. Estimating ICEs can ***reconcile the difference between what researchers estimate and what they are interested in***. Average effects only apply to individuals if researchers make the assumption of a constant treatment effect across individuals, which is a strong and usually unrealistic assumption. This leads to another point of emphasis between ICE estimation and ATE estimation, which is the ability of the former to examine treatment effect heterogeneity.

Consider the following study in Table 1.1 of a binary treatment indicator W on outcome Y with six observations. In Table 1.1a, the data are presented in a traditional setup where Y denotes the observed outcomes. In Table 1.1b, the same data are now presented in the form of potential outcomes. The question marks represent unobserved data, so one can think about causal inference as simply a missing data problem where the missing data are the unobserved potential outcomes for each unit i. A standard causal inference study would proceed to estimate the ATE with mild assumptions simply as

Table 1.1: A Study with Six Observations

i	W_i	Y_i
1	1	15
2	0	10
3	0	15
4	1	8
5	1	10
6	0	8

(a) Data

i	W_i	$Y_i(1)$	$Y_i(0)$
1	1	15	?
2	0	?	10
3	0	?	15
4	1	8	?
5	1	10	?
6	0	?	8

(b) Data with Potential Outcomes

$$\begin{aligned}\hat{\tau}_{ATE} &= \bar{Y}_t - \bar{Y}_c \\ &= 11 - 11 \\ &= 0\end{aligned}$$

where $\bar{Y}_t$ and $\bar{Y}_c$ denote the average observed outcomes for individuals receiving treatment and control respectively. The researcher would then note that the treatment has no effect. In a completely randomized experiment, this estimate is an unbiased estimate of the ATE since it is assumed that the observed potential outcomes are a random sample from the marginal distributions of the potential outcomes.

Now consider the same study in two different hypothesized worlds depicted in Table 1.2. In both scenarios, the missing potential outcomes are filled in (*italicized*) by drawing from the observed potential outcomes. The ATE remains the same as above in both cases. The last column of both tables contain the ICEs (**bolded**). If the researcher proceeded by estimating the ATE, the estimate would be unbiased and equal to 0 in both cases. However, the two scenarios are dramatically different. In Table 1.2a, the treatment has no effect for every individual. In Table 1.2b, the treatment has a large positive effect for some and a large negative effect for others. One may be tempted to conclude that an ATE of 0 implies the first scenario, but the second scenario is just as likely. With any given ATE value, there are an infinite number of ways in which the ICEs can aggregate to the same ATE. When estimating an ATE, researchers cannot say anything about effects for specific individuals or groups of individuals without further assumptions. Often, researchers use language that implies a constant effect for all individuals when only the ATE is estimated. In the presence of treatment effect heterogeneity, the ATE is a misleading quantity that hides much of what goes on in the data. By looking directly at ICEs, researchers can ***explore and discover treatment effect heterogeneity*** in a straightforward manner and explore any potential outliers or different underlying causal mechanisms amongst individuals or groups of individuals. The heterogeneity of treatment effects across individuals has important implications for research and policy-making.

Table 1.2: Two Different Scenarios with Identical Average Treatment Effects

i	W_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	15	*15*	**0**
2	0	*10*	10	**0**
3	0	*15*	15	**0**
4	1	8	*8*	**0**
5	1	10	*10*	**0**
6	0	*8*	8	**0**

(a) Treatment Has No Effect for Everybody

i	W_i	$Y_i(1)$	$Y_i(0)$	τ_i
1	1	15	*10*	**5**
2	0	*15*	10	**5**
3	0	*8*	15	**-7**
4	1	8	*15*	**-7**
5	1	10	*8*	**2**
6	0	*10*	8	**2**

(b) Treatment Helps Some and Hurts Some

Estimating ICEs also allows researchers to ***bridge the divide between quantitative and qualitative studies*** that exists in many areas of social science. As King, Keohane and Verba (1994) note, "the same logic of inference underlies both good quantitative and good qualitative research designs," yet there is still a disconnect between quantitative and qualitative scholars over which type of study is better and which results are more reliable. Part of the disconnect exists because quantitative studies use large N statistical analyses to estimate causal effects whereas qualitative studies focus more on causal mechanisms and look closely at a small number of cases. I argue that another part of the disconnect stems from the different estimands and claims that each type of study attempts to make. Quantitative studies tend to collect data for a large N population, estimate average effects, and then implicitly attempt to apply the average effects to explain individual cases. Qualitative studies collect data for a small n sample, estimate individual or small n average effects, and implicitly attempt to generalize to the entire population. Each side estimates a different estimand, yet both attempt to address general average and specific individual effects. The results can often be dissatisfying to both sides, which leads to a divide. By estimating ICEs, quantitative researchers can speak directly to qualitative researchers about treatment effects on individual cases without sacrificing the ability to estimate average effects.

Although the inability to observe individual causal effects is the "fundamental problem of causal inference", one point that is seldom addressed is that the ICEs are fundamental to causal inference. If one can observe or estimate the ICEs, then any other causal estimand can be observed or estimated with very little effort. Thus, an additional benefit of focusing on estimating ICEs is that ***once the ICEs are estimated, the researcher can estimate any other causal effect by simply aggregating the ICEs***. Typically, if the researcher wants to estimate multiple causal estimands, he would have to develop a new model for each. By focusing on estimating the fundamental quantity in causal inference, researchers are able to estimate an unlimited number of other estimands by simple aggregation.

I have argued that there are at least five benefits to focusing on estimating ICEs rather than ATEs.

1. ICEs allow for a more precise and clear understanding of causal inference
2. ICEs reconcile the difference between the quantity in which the researcher is interested and the quantity the researcher estimates
3. ICEs allow researchers to explore and discover treatment effect heterogeneity
4. ICEs bridge the quantitative-qualitative divide
5. ICEs allow for easy estimation of every other causal estimand

Estimating ICEs, however, entails a cost because they are unidentified and much harder to estimate correctly. I argue that one can borrow existing techniques and frameworks in the causal inference and missing data literature to tackle the problem of estimation.

1.3 Existing Approaches to Causal Inference

Consider the typical situation in data analysis where there is a sample of N units indexed by i sampled from a large or infinite population.[3] Each unit i receives treatment W_i, where $W_i = 1$ indicates i received treatment and $W_i = 0$ indicates i received control. Each unit also has potential outcomes $Y_i(1)$ and $Y_i(0)$, where Y_i is the observed potential outcome depending on the value of W_i. Each unit also has a set of pretreatment covariates X_i which are assumed to be exogenous. Two basic assumptions are often needed to estimate causal effects:

Assumption 1: Ignorability of Treatment Assignment

$$(Y(1), Y(0)) \perp W|X$$

This assumption is satisfied with random assignment of treatment or when X contains all pretreatment confounders that affect both W and the potential outcomes $Y(1), Y(0)$. Along with the ignorability of treatment assignment usually comes an assumption that $0 \leq P(W|X) \leq 1$, namely that there is positive probability of treatment for any X. The second important assumption is SUTVA.

Assumption 2: Stable Unit Treatment Value Assumption (SUTVA)

1. treatment assignment for one unit does not affect the potential outcomes of another (no interference or spillover effect):

$$(Y_i(1), Y_i(0)) \perp W_j, \;\; \forall i \neq j$$

2. only one version of each treatment possible for each unit

With this basic setup, I now review the causal inference literature and different approaches used to estimate different causal estimands.

[3] The causal inference literature often uses the words sample, population, and superpopulation in different applications. Generally speaking, the sample is drawn from a population of a given size. Sometimes, the population is the sample, in which case the population is drawn from a larger superpopulation. For simplicity, I will generally refer to the data as the sample drawn from a very large or infinite population, but one can also think of the framework as a sample drawn from a superpopulation if the size of the sample is very close or equal to the size of the population.

1.3.1 Average Treatment Effects

Imbens (2004) provides an in-depth review of the literature of estimating average treatment effects, which I briefly review here. The most basic average treatment effect (ATE) that researchers estimate is simply

$$\tau^p_{ATE} \quad = \quad E[Y(1) - Y(0)]$$

The expectation here is over the population that the sample was drawn from. The more accurate definition for this estimand is the population average treatment effect (PATE), which differs from the sample treatment effect (SATE).

$$\tau^s_{ATE} \quad = \quad \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$$

Since all the information known about PATE is captured in SATE, an estimator for SATE is the best and often a good estimator for PATE. Assuming that the sample is a random or representative sample from the population, the difference between SATE and PATE is in the variance of the estimates. Even if all the potential outcomes for the sample were observed, the potential outcomes for units not in the sample are not observed, so the variance needs to be adjusted upward for PATE. In most cases, researchers are interested in the population estimands, although the sample estimands can be of interest in situations where the sample is not representative of the population. In reviewing the causal inference literature, I ignore the differences between the sample and population versions of the estimands, assuming that researchers are estimating sample estimands with possible adjustments to estimate population estimands.

If treatment assignment is randomized or plausibly randomized such as in an experiment, then researchers can estimate the ATE by a simple difference in means,

$$\hat{\tau}_{ATE} = \bar{Y}_t - \bar{Y}_c$$

where Y_t and Y_c denote outcomes for observations that received treatment and control respectively.

Regression Approaches

Short of treatment assignment randomization, researchers need to condition on the set of confounders X to estimate the ATE. Perhaps the most common class of methods to condition on X is the class of regression estimators, which uses some functional form to estimate the average potential outcomes $\mu_{(w)}(x)$ given $X = x$ for $w = 1, 0$. The general form of the regression estimator averages over the empirical distributions of the covariates for treatment and control groups:

$$\hat{\tau}_{ATE,reg} = \frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)]$$

Many regression estimators impose a functional form for $\hat{\mu}_w(X_i)$ and possibly a parametric distribution for Y. The common linear model imposes a linear relationship between X and $\mu_w(X)$

$$\begin{aligned} \mu_{(w)}(X_i) &= \alpha + \tau W_i + \beta' X_i \\ Y_i &= \alpha + \tau W_i + \beta' X_i + \epsilon_i \end{aligned}$$

and estimates the parameters by ordinary least squares. Generalized linear models (McCullagh and Nelder 1989) specify the relationship between X and $\mu_{(w)}(x)$ through a linear functional form and a link function $g(\cdot)$ and also impose a parametric distribution $f(\cdot)$ on Y.

$$\begin{aligned} g(\mu_{(w)}(X_i)) &= \alpha + \tau W_i + \beta' X_i \\ Y_i &\sim f(\cdot|\mu_{(w)}(X_i)) \end{aligned}$$

Other regression models, such as kernel regression, generalized additive models, smoothing splines, local polynomial regression, are semiparametric or nonparametric and relax the parametric and linearity assumptions. The literatures on these models is enormous and I will not review it here (see Hastie, Tibshirani and Friedman (2009) for a extensive introduction). Although regression is commonly used to reduce bias and increase precision in estimating ATEs, it can actually lead to more bias when the functional form of the covariates is specified incorrectly. In the case where there is little covariate overlap between treatment and control groups, regression results can be very dependent on model specifications.

Matching Approaches

Another way to condition on X is to use matching methods, which first appeared in the early 20th century but was not developed theoretically until the 1970s (Rubin 1973*a*,*b*). Unlike regression methods, matching methods rely less on functional form and model assumptions. The goal of matching is to approximate a randomized experiment by matching individuals from treatment and control groups with similar covariate profiles. Observations that do not have overlap in covariates are removed from the matched sample to avoid extrapolation. In the ideal matching scenario, each treatment observation would be matched with one or more control observations with the same exact values on all the covariates and/or vice versa. The average treatment effect would then be calculated by differencing the treatment and control outcomes from this matched sample. This exact matching approach may be feasible in the case of a small number of discrete covariates. However, if there are continuous covariates and/or as the number of covariates increases, exact matching is not feasible in a finite sample because of the curse of dimensionality. Numerous matching methods have been developed to match observations in the hopes of achieving covariate balance across treatment and control groups. Stuart (2010) provides a comprehensive overview of the current matching methods developed. One point to note is that when estimating average treatment effects, the only requirement is that the distributions of the covariates for the treated and control groups be similar in the matched sample, which is less restrictive than requiring close or exact matches on all the variables for all observations. Researchers can then combine the matched sample with regression analysis to adjust for remaining imbalance after matching. Using the two methods in combination also helps to form "doubly robust" estimators which are less sensitive to misspecifications in either the matching method or regression model (Rubin 1973*b*, 1979; Ho et al. 2007).

Researchers who use matching methods to estimate average treatment effects do so by matching each treatment observation to one or more control observations and each control observation to one or more treatment observations. Often researchers are interested in another causal estimand, the average treatment effect for the treated (ATT):

$$\tau_{ATT} = \frac{1}{N_t} \sum_{i:W_i=1} [Y_i(1) - Y_i(0)]$$

where $N_t = \sum_{i=1}^{N} W_i$ is the number of treated units. From a computational standpoint, estimating the ATT is simpler since the researcher only needs to match the treated units with control units and does not have to match the control units with treated units or worry about whether the best matches for one imply best matches for the other.. From a policy and academic standpoint, since the treatment effect is of interest, it may be more appropriate to only look at units that actually received the treatment. Depending on the nature of treatment assignment, the treated group may be qualitatively different than the control group and thus important to look at separately. Although less common, the average treatment effect for the controls (ATC) may also be of interest:

$$\tau_{ATC} = \frac{1}{N_c} \sum_{i:W_i=0} [Y_i(1) - Y_i(0)]$$

where $N_c = \sum_{i=1}^{N}(1 - W_i)$ is the number of control units. If ATT = ATC, then the ATE = ATT = ATC. Also note that when using matching, observations which do not have good matches may be discarded, which changes the quantity of interest being estimated.

When implementing a matching method, the researcher has to make several choices. At each step, there are many options that the researcher can choose from. The factors to consider in any matching method are:

1. **The variables to include in the matching**

   Since matching is a way of conditioning on confounding variables to satisfy ignorability of treatment assignment, the researcher should include all pre-treatment variables that affect treatment assignment and the outcome. However, the curse of dimensionality almost certainly implies that covariate balance will be harder to achieve as the number of variables to match increases. With many variables to match on, improving balance on one variable may very well decrease balance on another and increase the bias of the estimate. Researchers may have to make choices on which variables to prioritize or choose matching methods that put different weights on different variables (Diamond and Sekhon 2013). One type of variable that generally should not be included is any variable that is affected by the treatment. Including these post-treatment variables can result in bias in the estimate (Rosenbaum 1984).

2. **The measure of closeness between observations**

When balancing on multiple variables (especially in the presence of continuous variables), the curse of dimensionality makes it difficult to determine how "close" observations are on the covariates. Researchers need to determine a measure of distance between observations and also decide how to match observations given their distances. In the ideal case of exact matching, observations are matched if all their covariate values are the same. Exact matching is rarely feasible, but one strategy is to coarsen the covariates and match exactly on the coarsened variables (Iacus, King and Porro 2012). Another strategy is to define distance between observations by a one-dimensional balancing score for each observation that summarizes the information in the covariates. Some examples of balancing scores include the Euclidean distance, the Mahalanobis distance, propensity scores (Rosenbaum and Rubin 1983), and prognostic scores (Hansen 2008).

Once distance is defined, researchers must then choose how to convert the distances into matches. One option is to do nearest-neighbor matching, where each treated observation is matched with its closest neighbors. The algorithm for nearest-neighbor matching may be greedy, with each observation choosing its matches in order, or optimal, taking into account all possible matches and minimizing a global distance measure. Note that greedy algorithms depend on the order of the observations. Another option is to divide the observations into a number of subclasses based on their distance measures, where each subclass contains at least one treatment and one control observation. Observations in the same subclass are then matched. Deciding the number and boundaries of the subclasses themselves is another choice for the researcher. The researchers can define these directly, indirectly as in the case of coarsened exact matching (Iacus, King and Porro 2012), or through an algorithm as in the case of full matching (Rosenbaum 1991). A third option is to match using the whole set of observations but weighting the observations by their distance measure as in Imbens (2000). Hainmueller (2012) uses entropy balancing to derive weights, optimizing balance on the sample moments of the covariate distributions. Note that all the options can be considered as weighted matching, where the first two options put weights of either 0 or 1 on every observation. One can also combine any of these options with calipers, which place restrictions on the distances for acceptable matches.

3. **The number of observations to serve as the "donor pool"**

For each observation, the researcher chooses to match it with one or more "donor" observations that received the opposite treatment. When matching each treated observation with control observations, all the control observations represent the donor population from which the researcher chooses M of them for the donor pool. The size of the donor pool in M-to-1 matching is often an arbitrary choice by the researcher. The most common choice is 1-to-1 matching where the closest observation on the distance measured is chosen. The choice of M is often a trade-off between bias and variance. In the case of an unlimited pool of exact matches, increasing M reduces the variance of the estimate by including more observations with more information. However, in practice, there are rarely exact matches, so increasing M results in matching on observations that are farther away on the distance measure. This decrease in variance by increasing M comes at the cost of increasing bias from matching on less similar observations (Rosenbaum and Rubin 1985). M is also often chosen indirectly, such as in methods using strata or subclassification where M is determined by the number of donor observations in the subclass or in methods using weights where the number of donor observations is determined by the weights. M can also be allowed to vary, which may reduce bias even further (Ming and Rosenbaum 2000).

4. **Whether to match with or without replacement**

When the number of possible donor pool observations is relatively small, researchers have an option to match with replacement. Matching with replacement reuses observations in multiple donor pools such that certain observations may be matched more than once. Matching with replacement can reduce bias since it usually results in better quality donor pools. However, the outcome analysis should take into account the fact that observations are used multiple times. The number of unique donor observations used should also be monitored so that the results are not dependent on using information from a small number of the donor population.

5. **How to check covariate balance to determine the success of the matching**

    The goal of matching is to create a matched dataset with similar distributions in the covariates for the treated and control groups. Therefore, to verify that the matching worked properly, the researcher must assess covariate balance in the matched sample such that $\tilde{p}(X|W = 1) = \tilde{p}(X|W = 0)$ where $\tilde{p}$ is the empirical distribution. Ideally one would like to examine the multivariate distributions of the covariates for the treatment and control groups. However, comparing multivariate distributions becomes difficult as the dimensions increase. Although some have suggested using multivariate imbalance measures (such as the $\mathcal{L}1$ statistic), most applications look at the marginal empirical distributions of the covariates and check balance on the moments (such as the standardized means) of the distributions. Others visualize balance graphically with Q-Q plots or plots of the different moments of the distributions. Other ways to check balance include running hypothesis tests to test whether the marginal distributions of the treated and control group are the same, although Imai, King and Stuart (2008) argue rightly that what matters is the in-sample balance rather than out-of-sample population balance. There are also certain matching methods that allow the researcher to define the level of imbalance ex-ante, thus constraining the post-matching imbalance to a certain level.

Matching methods have become increasingly popular in the causal inference literature because of its ability to mimic randomized experiments and its lesser reliance on parametric modeling assumptions. I revisit many of these matching methods in more detail and discuss how matching methods can be used to estimate individual causal effects.

Other Approaches

Besides matching and regression, other approaches exist that try to identify average treatment effects, usually by leveraging aspects of the data or external circumstances to approximate random assignment of treatment. For example, one can use natural experiments where a treatment has been pseudo-randomized by nature. Another approach is to use instrumental variables analysis, where the researcher has a randomized or plausibly ignorable instrumental variable that is correlated with the treatment variable of interest. Finally, regression discontinuity designs attempt to leverage sharp discontinuities in the treatment variable to conduct analyses as if the treatment has been randomized for units near the discontinuity.

A lesser known but potentially powerful modeling approach to estimate average treatment effects is with Bayesian methods. Rubin (1978) introduces a general Bayesian framework for estimating treatment effects. One way to use Bayesian methods is to model the potential outcomes directly. Another way is to estimate regression models using priors on the regression coefficients to weight the importance of various covariates. Although very little has been done on integrating matching methods and Bayesian approaches, I argue for using a Bayesian framework with matching methods to estimate individual causal effects.

1.3.2 Treatment Effect Heterogeneity

Treatment effect heterogeneity exists when there are varying average treatment effects for various subgroups of the inferential population. Treatment effect heterogeneity is an important topic in many fields, especially in the medical sciences where a treatment may help some patients but hurt others (Kravitz, Duan and Braslow 2004; Rothwell 2005). Political scientists are also increasingly interested in treatment effect heterogeneity with substantive implications (Feller and Holmes 2009; Arceneaux and Nickerson 2009; Gaines and Kuklinski 2011; Imai and Strauss 2011). In the presence of treatment effect heterogeneity, estimating a simple average treatment effect may mask important differences in treatment effects as I demonstrated above. The most common way to test for treatment effect heterogeneity is to estimate the average treatment effect for different subgroups of the sample using any of the methods described above. The subgroups are defined by the specific covariates and the average treatment effect within a subgroup is commonly known as the **conditional average treatment effect (CATE)**:

$$\tau_{CATE,x} = E[Y(1) - Y(0)|X = x]$$

where x denotes the covariate values of the subgroup. Treatment effect heterogeneity occurs when the CATEs differ for different subgroups. However, two general sets of complications arise when estimating multiple CATEs: 1) small sample sizes and limited power and 2) multiple testing problems and arbitrarily defined subgroups.

Recall that for any statistical test, the power of the test is inversely related to the sample size. When testing for effects within subgroups in the same dataset, the sample size is usually significantly smaller than the size of the original dataset N. This is especially true in clinical trials, where N is usually small to begin with. Unless the subgroup treatment effects are quite large, standard statistical tests often fail to detect effects in subgroups (Pocock et al. 2002). One solution to the problem of small sample sizes in subgroup analyses is to use interaction terms where the variable defining the subgroups is interacted with the treatment indicator. Although the use of interaction terms better captures the extent of the information in the data and uses the data more efficiently, the estimators used are still usually limited by the need to appeal to large sample properties, while the subgroup analyses rely on smaller and smaller samples.

Ironically, many existing subgroup analyses are also susceptible to a second complication of multiple testing problems and arbitrarily defined subgroups. When looking for treatment effect heterogeneity, the researcher often tests for significant effects over multiple subgroups defined by the covariates. With multiple tests, the probability of a false positive is greatly inflated and can lead to misleading results (Lagakos 2006). Crump et al. (2008) develop nonparametric tests for the null of no treatment effect heterogeneity, which bypass the multiple testing problem but fail to specify exactly which subgroups have heterogeneity. In addition to the multiple testing problem, the choice of subgroups to examine for treatment effect heterogeneity is often left to the researcher, which creates potential validity and incentive compatibility concerns. Subgroups can be chosen either arbitrarily or with some substantive theory in mind. They can be prespecified before the experiment or chosen post-hoc. Recent data mining techniques have been developed to remove the choice of subgroups from the researcher's control by using learning algorithms to search through the space of treatment-covariate interactions to detect statistically significant effects (Green and Kern 2012; Imai and Ratkovic 2013).

The literature on subgroup analysis and treatment effect heterogeneity is relatively small compared to the literature on estimating ATEs. When testing for treatment effect heterogeneity, it is sometimes unclear whether the quantities of interest are the CATEs themselves or the differences in CATEs. Estimating CATEs often seems to boil down to estimating ATEs on smaller randomly chosen subsets of data. The estimators themselves often rely on large sample approximations that may not even hold in the larger full dataset. Matching techniques that often work well in estimating ATEs are seldom used in estimating CATEs. The interpretations of the interaction terms in the treatment effect heterogeneity setting may also be tricky, especially when the covariate that is interacted is more complicated than a binary variable. Other scholars have approached the topic differently by developing bounds for the proportion of the population that has treatment effect heterogeneity (Gadbury, Iyer and Albert 2004). I argue that an even easier way to examine treatment effect heterogeneity is to estimate the individual causal effects themselves, bypassing the need for complicated interaction models and testing at the subgroup level.

1.3.3 Individual Causal Effects

The literature on estimating individual causal effects is substantially smaller than either the literature on estimating ATEs or treatment effect heterogeneity, which mirrors the lack of attention scholars have paid to the topic. The simplest way to estimate an ICE is to estimate a general model for ATEs and predict the individual effects based on that model. For example, in medicine, researchers suggest calculating the baseline disease risk for any individual patient based on covariates and an existing model and then calculate the effect of treatment on that patient using the overall effect from a clinical trial (Dorresteijn et al. 2011). A second approach to estimate ICEs requires multiple datapoints over time, usually one or more "pre-treatment" datapoints and one or more "post-treatment" datapoints. The simplest example would be a crossover design, where individuals are randomized to one treatment at time t and another at time $t + 1$. In this case, the individual would act as both treatment and control observations. However, strong assumptions about time-period effects and treatment carry-over effects across time need to be made. Steyer (2005) proposes a more general model involving multiple pre-treatment and post-treatment observations to measure the "latent" true expected outcomes. Abadie, Diamond and Hainmueller (2010) introduce the use of synthetic controls to estimate the treatment effect for a single unit with time-series data. The synthetic controls are created by comparing and weighting all the control units with the unit that received treatment and calibrating based on the outcome variables for the time periods before the unit received the treatment.

Recent work has focused on using both Bayesian methods and matching methods developed for estimating ATEs and adapting them to estimate ICEs. As Abadie and Imbens (2006) put it, any matching estimator simply "imputes the missing potential outcomes." Rubin and Waterman (2006) use propensity score matching to create “clones” for each treated unit in order to estimate ICEs, although their approach does not include any uncertainty estimates. An (2010) suggests that using a Bayesian propensity score estimator can incorporate uncertainty over the matching procedure to estimate individual effects. Rubin (2005) presents a general framework in which missing potential outcomes can be imputed by drawing from the posterior predictive distribution of potential outcomes in any Bayesian model. Pattanayak, Rubin and Zell (2012) stratify treatment and control observations using estimated propensity scores and then use a Bayesian model within each strata to estimate ICEs. Gutman and Rubin (2012) develop imputation methods using subclassification and splines with knots at the borders of the subclasses to impute the missing potential outcomes. Finally, Jin and Rubin (2008) assume that the potential outcomes $Y(1)$ and $Y(0)$ are correlated by the parameter ρ and test the sensitivity of the causal effects to different values of ρ. In the next section, I introduce a flexible general framework to estimating ICEs that builds on many of these studies, using both Bayesian methods and a wide variety of matching methods.

1.4 Estimating Individual Causal Effects

One reason why ICEs are not estimated or points of focus is that ICEs are not identified in the data without further assumptions. Suppose that for an individual i, one posits that the ICE can be -1000, 0, or 9999.8. Statistical identification requires that the data and our estimation method tell us which of the three values is more likely to be true. However, since one does not observe the missing potential outcome, the data cannot give us any more information about the ICE for individual i. Given that identification is impossible, I argue that one should estimate ICEs by deriving a range of plausible values for the ICEs given information from other observations in the data. I use a Bayesian framework which gives us a posterior distribution of our ICEs based on information from the data and our prior beliefs rather than an identified point estimate.

The approach I use builds on a Bayesian framework for imputing missing potential outcomes first introduced by Rubin (1978), with similarities to the approach used in Pattanayak, Rubin and Zell (2012). As before, let W_i denote a binary treatment assignment indicator for unit i with an observed outcome Y_i and a vector of pre-treatment covariates X_i. Define Y_i^{mis} to be the unobserved potential outcome for unit i:

$$Y_i^{mis} = \begin{cases} Y_i(1) & \text{if } W_i = 0 \\ Y_i(0) & \text{if } W_i = 1 \end{cases}$$

Let τ_i be the individual causal effect for unit i:

$$\tau_i = \begin{cases} Y_i^{mis} - Y_i & \text{if } W_i = 0 \\ Y_i - Y_i^{mis} & \text{if } W_i = 1 \end{cases}$$

which I can rewrite simply as

$$\tau_i = W_i(Y_i - Y_i^{mis}) + (1 - W_i)(Y_i^{mis} - Y_i)$$

Since τ_i is a deterministic function of Y_i^{mis} and the observed data, I can calculate τ_i by simply imputing Y_i^{mis}. Our uncertainty around τ_i also comes only from our uncertainty around Y_i^{mis} since Y_i is observed.

To start, recall the most basic framework found in many regression models used in the social sciences (e.g. generalized linear models). In a typical regression setup, Y is a random variable that follows some probability distribution defined by a set of parameters θ conditional on covariates X and treatment W.

$$Y_i \quad \sim \quad f(\cdot|\theta_i, X_i, W_i)$$

The parameter vector θ_i includes the mean of Y_i, μ_i, which is usually parameterized as a function of the regression coefficients β, and possibly some ancillary parameters ϕ. I then estimate β in our regression model and derive average causal effects, since β is not subscripted by i. Note that typical regression models do not reference the missing potential outcomes, although one could use the regression model to predict the missing potential outcomes.

In my framework for estimating τ_i, I take a slightly different approach to modeling the data. Suppose instead that our data is a finite sample of size N drawn from the following data generating process:

$$\left.\begin{array}{rcl} Y_i & = & h(X_i^{(p)}) \\ Y_i^{mis} & = & h(X_i^{(p)}, \tau_i) \end{array}\right. \quad \text{for } W_i = 0$$

$$\left.\begin{array}{rcl} Y_i & = & h(X_i^{(p)}, \tau_i) \\ Y_i^{mis} & = & h(X_i^{(p)}) \end{array}\right. \quad \text{for } W_i = 1$$

where $X_i^{(p)}$ is the set of all prognostic variables (variables that predict the outcome) including any confounding variables and $h(\cdot)$ is some unknown function. First, note that the framework is restricted to the finite sample and one can only estimate individual causal effects for units in the data. Looking only at the finite sample allows us to appeal to a Bayesian setup. Also, the idea of individual causal effects is fundamentally restricted to the sample since individuals only appear in the data, and not in some superpopulation. I also assume that if the data generating process repeated multiple times under the same exact conditions, τ_i remains constant for i. Second, the potential outcomes are fixed and completely determined by $X_i^{(p)}$ and W, which are also fixed. In theory, if every single variable that affects the outcome can be measured, one could predict the outcome perfectly.[4] In practice, only a very small subset of $X_i^{(p)}$ is observed. Partition $X_i^{(p)}$ into a set of observed covariates, X_i, and a set of unobserved covariates, $X_i^{(u)}$.

$$X_i^{(p)} = \{X_i, X_i^{(u)}\}$$

If X_i contains at least all the variables that makes treatment assignment ignorable, then the ignorability

[4] The approach I am taking to the data generating process is that any outcome can be predicted perfectly by observing the complete set of prognostic variables and knowing the functional form. Philosophically, this argument may conflict with the traditional statistical idea of randomness and unpredictability. In practice, the two approaches are the same since the full set of prognostic variables is never observed and I proceed by modeling the outcomes as random. However, I take this approach to make the two points. First, since the quantity of interest is the individual causal effect, I want to stress that the subscript i takes on a special meaning that is specific to that individual. Therefore, i can be modeled and predicted completely in theory. Second, I want to make the point that including more prognostic variables can give us more information about the missing potential outcome.

assumption gives us

$$
\begin{aligned}
(Y(1), Y(0)) &\perp W|X \\
\tau &\perp W|X \\
X^{(u)} &\perp W|X
\end{aligned}
$$

Note that the assumption that τ is independent of treatment assignment conditional on X implies that one can use information from the opposite treatment group to inform the missing potential outcome for i. In a simple example, assume that observation i is treated and observation j is control and they have the same value for $Y(0)$. If, for example, the ICEs were systematically larger for those assigned control, then using information from j would overestimate τ_i.

These ignorability statements imply some type of randomness in the data. I assume that conditional on the observed X, the unobserved $X^{(u)}$ are essentially random across treatment and control observations. The randomness is then modeled with the following:

$$
Y_i^{mis} \quad \sim \quad f(\cdot|\theta_i^{mis}, X_i, W_i)
$$

where θ_i^{mis} represents the distributional mean of the outcomes conditional on the observed X_i.[5] Simply put, observations with the same values of X_i and W_i are randomly drawn from a common distribution, conditional on Assumptions 1 and 2 being satisfied. Strictly speaking, θ_i^{mis} should be denoted as θ_{X_i,W_i}^{mis}, which indicates that it is the mean of the missing potential outcome and that observations with the same observed covariate vector and treatment status as i have the same mean. I use θ_i^{mis} to simplify notation.

Consider an observation j where $X_j = X_i$ and $W_j = 1 - W_i$. Then this implies that Y_i^{mis} and Y_j are

[5] For some distributions, there may be ancillary parameters in addition to the mean. In that case, θ_i^{mis} would be a vector of parameters. For the sake of notational convenience and simplicity, I assume that $f(\cdot)$ is parameterized solely by the mean for now.

modeled as generated from the same distribution:

$$\begin{aligned} Y_i^{mis} &\sim f(\cdot|\theta_i^{mis}, X_i, W_i) \\ Y_j &\sim f(\cdot|\theta_i^{mis}, X_j = X_i, W_j = 1 - W_i) \end{aligned}$$

This suggests that if i is a treated observation, one can use observed outcomes for control observations with the same value on X as i to estimate θ_i^{mis}. This also implies that one can model the data generating process for the observed data as

$$\begin{aligned} Y_i &\sim f(\cdot|\theta_i^{obs}, X_i, W_i) \\ \theta_i^{obs} &= W_i(\theta_i^{mis} + \tau_i) + (1 - W_i)(\theta_i^{mis} - \tau_i) \end{aligned}$$

However, because I assume that Y_i is fixed and observed, θ_i^{obs} is not an interesting parameter and is not estimated. θ_i^{mis}, the mean of the distribution for the missing potential outcome, is the key parameter of interest in this framework. The stochastic nature of the outcomes reflects the contributions of the unmeasured prognostic variables, which are assumed to be independent of treatment assignment. In other words, each potential outcome for any individual i is a deterministic function of observed and unobserved prognostic covariates. Then θ_i^{mis} is estimated by matching to create observations that are considered to be similar on the observed covariates X:

$$\theta_i^{mis} = m(X_i, W_i, Y)$$

where $m(\cdot)$ is a matching estimator. The assumption made with this setup is that the potential outcomes are independent conditional on X_i. That is, Y_i gives no extra information about Y_i^{mis} and vice versa.

There is a slight difference between my framework and other approaches to causal inference as to where the randomness occurs in the dataset. Most approaches make appeals to superpopulations and estimate population parameters. Units are assumed to be drawn from these superpopulations. For example, one common approach is to assume that W, X and Y are all random variables (Rubin 2005, 2008). Abadie and Imbens (2006) on the other hand assume that the triplet $\{Y, W, X\}$ is drawn at random. In my approach, there is no superpopulation and the only randomness comes from the unknown $X^{(u)}$. I am strictly interested in estimands in the observed sample. If $X^{(p)}$ was fully observed, there would be no randomness and all the parameters can be calculated. Although my framework can be adjusted and applied to other approaches or appeal to superpopulations, I make explicit the notion that randomness in Y^{mis} comes only from not observing $X^{(u)}$. In practice, there is very little difference between my assumption about the source of randomness and the typical setup. For example, one can think of $X^{(u)}$ as simply the error term ϵ in linear regression models.

This framework involves two steps: a matching step to estimate θ_i^{mis} and an imputation step to get an imputed value of Y_i^{mis} accounting for the unobserved prognostic covariates. Each step is also characterized by a type of uncertainty that eventually propagates to uncertainty around τ_i. The matching step has *estimation uncertainty* and the imputation step has *fundamental uncertainty* (King, Tomz and Wittenberg 2000). Estimation uncertainty refers to the uncertainty in estimating θ_i^{mis}, which encompasses uncertainty over the parameters of the matching procedure, uncertainty due to finite sample size, and possibly even uncertainty over the choice of the matching procedure itself. Estimation uncertainty is a function of the variation in outcomes and the size of the donor pool. Fundamental uncertainty is usually described as randomness or chance events that affect the outcome but is not included in the set of conditioning variables. In other words, fundamental uncertainty reflects the influence of our unmeasured prognostic variables. All things being equal, conditioning on more variables that affect the outcome should reduce fundamental uncertainty. I introduce various ways to perform the matching step in the next section, borrowing from many existing techniques in the causal inference literature. Both matching and imputation steps are then incorporated into a general Bayesian model. I then test the performance of the various techniques for estimating τ_i via simulation.

1.4.1 The Matching Step

To estimate τ_i, I first need to conduct matching N times to estimate θ_i^{mis} for all i in the data. Let $D_j^{(i)}$ be a binary variable that denotes whether or not an observation j is in the donor pool for observation i when estimating τ_i:[6]

$$D_j^{(i)} = \begin{cases} 1 & \text{if } W_j \neq W_i \text{ \& } j \text{ is a match to } i \\ 0 & \text{otherwise.} \end{cases}$$

The size of the donor pool for observation i is simply $\sum_{j=1}^{N} D_j^{(i)}$. In matching procedures where observations can be weighted donors, $D_j^{(i)}$ acts as the donor weight and can take on any value between 0 and 1.[7] The matching step involves defining $D_j^{(i)}$ by choosing a set of donor observations that are similar to i on the conditioning variables X_i. I then use the observed outcomes in the donor pool to estimate θ_i^{mis}.

In an ideal world, one can expand X_i to include all prognostic covariates measured without error and the observations in the donor pool would be exact matches to i on all X_i. There would be no estimation or fundamental uncertainty and Y_i^{mis} can be imputed exactly. However, in practice, finite sample sizes, a large number of prognostic covariates, many of which are unobserved, and/or the presence of continuous covariates precludes the possibility of exact matching on all prognostic covariates. Instead, I use matching procedures to define $D_j^{(i)}$ and calculate the mean of the donor pool as

$$\bar{Y}_{D^{(i)}} = \frac{\sum_{j=1}^{N} Y_j D_j^{(i)}}{\sum_{j=1}^{N} D_j^{(i)}}$$

I then use a Bayesian model (described below) to combine $\bar{Y}_{D^{(i)}}$ and a prior to estimate θ_i^{mis}.

The decisions made with respect to the selection of the matching procedure mirrors the choices usually made when using matching to estimate average treatment effects. In this case, since the quantity of interest is the individual causal effect, the goal is no longer simply distributional balance across treatment and control observations. Instead, one needs to create a donor pool that is as close to i on X_i as possible.

The following choices must be made with respect to our matching algorithm:

- **The set of conditioning variables** X: All confounding variables should be conditioned on to satisfy

[6] In the case of exact matching, $D_j^{(i)}$ denotes whether j and i are exact matches ($X_j = X_i$). Since most methods researchers use are approximate matching methods, $D_j^{(i)}$ is random even if W and X are fixed. One should conceptually think about $D_j^{(i)}$ as an indicator for whether or not $X_j \approx X_i$.

[7] For non-binary donor weights, some of the equations below must be adjusted. For now, I assume that $D_j^{(i)}$ only takes on a value of 0 or 1.

the ignorability of treatment assignment and causal effect independence assumptions. In addition, other prognostic variables should also be conditioned on to improve the efficiency of the estimates and possibly reduce bias (Rubin and Thomas 2000; Pocock et al. 2002). However, with limited sample sizes and small donor pools, there is a tradeoff between finding good matches and conditioning on more variables, akin to a bias-variance tradeoff. Researchers should prioritize conditioning on confounders that are also highly predictive of the outcome.

- **Size of the donor pool**: The size of the donor pool, $M = \sum_{j=1}^{N} D_j^{(i)}$ is chosen either directly or indirectly by the researcher and may vary across i. By definition, increasing the size of the donor pool results in the inclusion of matches that are either worse or about the same in terms of similarity to i on X_i. This results in a more efficient estimate of θ_i^{mis}, but may also introduce more bias due to the inclusion of poorer quality matches.

- **Matching with or without replacement**: Since the quantity of interest is at the individual level, reusing matches for multiple ICEs does not pose any problems and leverages better information. Matching with replacement is ideal and may be necessary for small sample sizes.

- **Weighting donor observations**: By default, in most matching applications, donor observations each receive a weight of 1, implying that all donors are equally good matches. Expanding the size of the donor pool likely results in matches that are poorer matches, so the researcher can choose to downweight donors as a way to reduce the influence of poor matches on the estimate. This also reduces the effective size of the donor pool and incorporates greater uncertainty in the presence of poorer matches.

- **Definition of closeness**: Since in most cases, exact matching is impossible, choosing the definition of closeness between matches is probably the most important task. One can choose amongst a myriad of dimension-reducing balancing scores, although exact matching should be used when possible. A mix of exact matching and balancing scores is also feasible.

- **What to do with unmatched observations**: For some observations, it is likely that the predefined criteria produces no matches for the donor pool. For estimating individual causal effects, discarding unmatched observations means not estimating a causal effect for that individual. When aggregating to average effects, discarding observations changes the quantity of interest. The researcher can force matches by relaxing some of the matching criteria imposed.

Short of exact matching, it is unclear which matching procedure performs the best a priori in estimating θ_i^{mis}. I consider a few options that are prevalent in the matching literature, adapting and combining some of them to try to gain efficiency and reduce bias. I then test the performance of each of these options via simulation. My Bayesian model also includes an option to incorporate uncertainty around any parameters within a specific matching procedure or uncertainty over the matching procedure itself. The matching procedures that I consider are:

- nearest neighbor matching on the Mahalanobis distance
- nearest neighbor matching on the predictive mean (often used in the missing data imputation literature)
- nearest neighbor matching on the propensity score
- subclassification on the propensity score

While there are numerous matching procedures to consider, I focus on these four methods because they are relatively easy to estimate and understand, they allow for all observations to be matched, and they have been used extensively by researchers. For each matching procedure, I match N times, once for each observation in the data. I match with replacement in the sense that an observation can be a part of more than one of the N donor pools, but each observation may only be used once per pool. I also test each procedure using multiple donor pool sizes, varying the choice of donor pool size.

Although the authors of the various procedures have demonstrated the performance of their procedures in estimating average treatment effects, none of the procedures attain the ideal of exact matching. The procedures are simply a means to achieve covariate balance, where the distributions of the covariates are similar across treatment and control groups. In this case, since the comparison is between a single observation and a donor pool, the analogue to balance is simply whether the donor pool observations are exact matches to i. Deviations from exact matches creates bias in what is known as the matching discrepancy. Abadie and Imbens (2006) argue that the bias from the matching discrepancy may be negligible for ATEs when matching on a scalar or when the number of observations is large. It is unclear how the bias from the matching discrepancy affects the estimates of θ_i^{mis} and τ_i. Apart from the matching discrepancy, there may also be bias because of the estimation uncertainty around θ_i^{mis}, for which a Bayesian model accounts, as described below.

1.4.2 The Imputation Step

Since estimating τ_i is essentially a missing data problem where Y_i^{mis} is missing, the methods used are very similar to multiple imputation to deal with missing data (Rubin 1987; Little and Rubin 1987). Once I get an estimate of θ_i^{mis}, I need to fill in a value for Y_i^{mis}, denoted by $\tilde{Y}_i^{mis}$ to calculate $\tilde{\tau}_i$.[8] The imputation step is necessary to account for some fundamental uncertainty associated with $X_i^{(u)}$ that the matching does not account for. Y_i^{mis} should be imputed with values consistent with the observed Y values, so $\tilde{Y}_i^{mis}$ should be binary for binary Y and continuous for continuous Y. Recall that Y_i^{mis} was assumed to be drawn from some distribution $f(\cdot)$ conditional on observed covariates:

$$Y_i^{mis} \quad \sim \quad f(\cdot|\theta_i^{mis}, X_i, W_i)$$

For the imputation, I use a parametric approach that follows Rubin (2008) by drawing a value of $\tilde{Y}_i^{mis}$ from its posterior predictive distribution and repeating the process multiple times for each i. I end up with many imputed $\tilde{Y}_i^{mis}$ for each i, which forms a posterior predictive distribution that characterizes both estimation and fundamental uncertainty. I then use that posterior predictive distribution to calculate a posterior distribution for τ_i. The performance of parametric imputation likely depends on how accurately θ_i^{mis} is estimated as well as the size of the donor pool.

1.4.3 A Bayesian Model for Estimating τ_i

The general method I introduce is very simple with the following steps:

1. Choose a matching procedure.

2. For each i, use the matching procedure to create a donor pool.

[8] The $\sim$ above a parameter refers to a simulated draw of that parameter from its posterior distribution.

3. Impute the missing potential outcome Y_i^{mis} using the donor pool and an assumed parametric distribution.

4. Calculate τ_i from the observed and imputed missing potential outcomes.

5. Repeat 2-4 for all i.

6. Repeat 1-5 many times for uncertainty.

I incorporate these steps into a Bayesian model for a coherent and statistically principled framework. The Bayesian model also allows for inclusion of priors when qualitative knowledge exists on any specific observations, although in general, I use uniform priors so that the results approximate those that may be derived from a non-Bayesian framework. The Bayesian model accounts for both estimation and fundamental uncertainty using Markov Chain Monte Carlo (MCMC) methods to simulate from the posterior distribution of the parameters.

Let θ denote the vector of parameters to be estimated. At the most general level, θ includes the vector of θ_i^{mis}, parameters from the matching procedure which are denoted by $\theta_{\mathcal{M}}$, and possibly the choice of matching procedure, denoted by $\mathcal{M}$.[9] Although $\mathcal{M}$ is treated as a parameter, the data tells us nothing about $\mathcal{M}$ so the marginal posterior is equal to the prior for $\mathcal{M}$. $\mathcal{M}$ is simply included here as an option to reflect the researcher's uncertainty over the best or "correct" matching specification.

The typical Bayesian posterior is expressed as

$$p(\theta|Y, X, W) \propto p(Y|\theta, X, W)p(\theta)$$

Since W is independent of the potential outcomes through the ignorability assumption and X is independent of the potential outcomes conditional on θ_i^{mis}, I suppress W and X from the conditioning set for notational simplicity.

[9] For example, $\mathcal{M}$ can be nearest neighbor 3-to-1 propensity score matching, in which case $\theta_{\mathcal{M}}$ are the coefficients in the propensity score equation. The researcher can vary $\mathcal{M}$ by choosing a different number of donor observations, changing how the distance metric is defined, or changing the set of matching variables.

The idea behind this model is simple. Because the observed Y_i is fixed when i is the individual of interest, the only randomness comes from Y_i^{mis}. Nature randomly generates observations that come from the same distribution as Y_i^{mis}. The goal of the matching is to determine which of the observed observations comes from this distribution parameterized by θ_i^{mis}. The posterior is roughly translated into

$$p(\theta|Y, X, W) \propto \prod_{j=1}^{N} \{p(Y_j|\theta, X, W) \text{ given } j \text{ is a match for } i\} \times \text{priors}$$

The model is composed of two parts. The first part is a matching part to find the posterior for the parameters $\theta_{\mathcal{M}}$. The second part finds the posterior for θ^{mis}. I consider the matching part to be largely independent of the second part conditional on finding the observations matched. That is, once one knows which observations are matches, θ^{mis} is independent of $\theta_{\mathcal{M}}$. Depending on the matching procedure, the matching parameters may or may not appear in the likelihood.[10] For simplicity and generality, I restrict my discussion of the likelihood term to simply focus on the likelihood for θ^{mis} assuming that the matching parameters are given.

Likelihood

The likelihood requires specifying the distribution that generated the data. Recall that our matching procedure is intended to generate a set of donor observations with "the same" values of X such that the donor observations are drawn from the same distribution as Y^{mis}. Now suppose one observes N binary variables $D^{(i)}$ (one variable for each i), which are indicators for whether j is a good match for i. Denote the set of $D^{(i)}$ variables as D. Then the likelihood[11] becomes

$$\begin{aligned} \mathcal{L}_{comp}(\theta^{mis}|Y, D) &= p(Y, D|\theta) \\ &= p(Y|D, \theta^{mis})p(D|\theta) \end{aligned}$$

[10] For example, in Mahalanobis or propensity score matching, the outcome is not used so the matching parameters do not appear in the likelihood for Y. For predictive mean matching, the outcome is used. One can choose to model $\theta_{\mathcal{M}}$ separately or jointly with θ^{mis}. The process I describe models them separately by doing the matching independently first.

[11] Again I assume that the matching parameters are estimated separately and given.

This likelihood is known as the *complete data likelihood* as it refers to the likelihood if one were to observe the complete set of data including D. The distribution in the second term of the complete data likelihood is determined by matching:

$$p(D|\theta) \quad = \quad \prod_{i=1}^{N}\prod_{j=1}^{N} p(D_j^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})$$

The first term in the complete data likelihood specifies the sampling distribution for the donor observations for each θ_i^{mis}:

$$\begin{aligned} p(Y|D, \theta^{mis}) &= \prod_{i=1}^{N}\prod_{j=1}^{N} \left[p(Y_j|\theta_i^{mis})\right]^{D_j^{(i)}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N} \left[f(\cdot|\theta_i^{mis})\right]^{D_j^{(i)}} \end{aligned}$$

Since Y_i is assumed fixed and not modeled when estimating τ_i, this piece of the likelihood implies there is randomness only when an observation is used as a donor. The complete data likelihood is then rewritten as

$$\begin{aligned} \mathcal{L}_{comp}(\theta^{mis}|Y, D) &= \prod_{i=1}^{N}\prod_{j=1}^{N} \Big[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}}, \mathcal{M}) + \\ &\qquad p(Y_j|\theta_j^{other})\left(1 - p(D_j^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})\right)\Big] \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N} \Big[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}}, \mathcal{M})\Big]^{D_j^{(i)}} \end{aligned}$$

In the first equation, θ_j^{other} simply refers to the fact that if j is not a match for i, then it is drawn from some other distribution that is not of interest. Therefore, the second term of the first equation drops out since non-matches do not contribute information to θ^{mis}. In all the likelihoods, the product over all i's indicates the full set of ICEs for every observation in the data.

The observed data likelihood simply integrates over our missing D:

$$\mathcal{L}_{obs}(\theta^{mis}, Y) = \int p(Y, D|\theta^{mis})\, dD$$

The integral is generally mathematically intractable but one can simulate from the posterior via MCMC methods. The Bayesian model presented here uses the data augmentation algorithm of Tanner and Wong (1987). The original posterior is augmented with D to make computation more tractable.

Priors

All Bayesian models require specifying a prior distribution over all the parameters in the model. In this case, a prior is needed for θ_i^{mis}, $\theta_{\mathcal{M}}$, and $\mathcal{M}$. I assume that the parameters are independent a priori.

$$\begin{aligned} p(\theta) &= p(\theta^{(1)})p(\theta^{(2)})\ldots \\ &= p(\theta_{\mathcal{M}})p(\mathcal{M})p(\theta_i^{mis})\ldots p(\theta_N^{mis}) \end{aligned}$$

For $\theta_{\mathcal{M}}$ and θ_i^{mis}, I generally use uninformative priors although one could incorporate qualitative knowledge into the priors. The choice of a prior for $\mathcal{M}$ boils down to which matching procedures one wants to consider. Since the data gives no information about the "best" matching procedure, the prior completely dominates the posterior for $\mathcal{M}$. If the researcher only wants to use one matching procedure as is typical in the causal inference literature, then the prior over $\mathcal{M}$ is essentially a spike prior. More research needs to be done on the influence of priors in my model on estimating individual causal effects.

Simulating from the Posterior via MCMC

I can simulate from the posterior of τ_i by using a Gibbs sampler, embedding the matching step within the sampler, and then drawing from the posterior predictive distribution (PPD) and calculating τ_i. For the Gibbs sampler, I draw from the full conditional distributions of the parameters conditional on the other parameters. The steps to simulate from the posterior of τ_i are:

1. $\mathcal{M}$ refers to any specification within the matching procedure. This can include any specification such as donor pool size, distance metric, or even the complete matching procedure itself. This leads to an important flexibility that my model allows, namely that I can simulate over the uncertainty of

MCMC Algorithm for the Posterior of τ_i

```
Repeat the following $n_{sim}$ times:[a]

Gibbs Sampler:
    1. Draw a matching procedure $\tilde{\mathcal{M}}$ from $p(\mathcal{M})$.
    2. Draw $\tilde{\theta}_{\mathcal{M}}$ from $p(\theta_{\mathcal{M}}|Y, X, W, D, \theta^{mis}, \mathcal{M})$.

for ($i$ in 1:$N$){
    3. Determine $\tilde{D}^{(i)}$ from matching procedure. (matching step)
    4. Draw $\tilde{\theta}_i^{mis}$ to estimate $\theta_i^{mis}$.
}

Draw from PPD and Calculate $\tau_i$:
for ($i$ in 1:$N$){
    5. Draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\tilde{\theta}_i^{mis})$. (imputation step)
    6. Calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$.
}
```

[a] Each draw of a parameter should be conditional on the current or previous draws of the other parameters. I have suppressed the iteration notation for aesthetic purposes.

which matching procedure or which specifications within the matching procedure to choose. The data and other parameters do not generally give any information about model specification, so the full conditional is

$$p(\mathcal{M}|Y, X, W, \theta_{\mathcal{M}}, D, \theta^{mis}) = p(\mathcal{M})$$

which means that uncertainty over $\mathcal{M}$ is driven completely by the prior.[12] This flexibility is still useful in the case where the researcher is equally unsure about the various matching procedures and/or the number of observations in the donor pool, in which case he would put a uniform prior over the various permutations and incorporate that uncertainty within the simulation. In essence, allowing for uncertainty over $\mathcal{M}$ is similar to Bayesian model averaging approaches prevalent in the literature (Raftery 1995; Montgomery and Nyhan 2010). One important caveat is that $\mathcal{M}$ should produce a set

[12] The assumption that there is no information inherent in the data to distinguish between matching procedures is a simplifying assumption. One can imagine that the data provides information on which matching procedures are "better" by evaluating empirical balance in the covariates under each procedure and sampling the procedures probabilistically depending on the balance measure. More research into the feasibility of such approaches should be done.

of matches for the same individuals all the time or else the quantities of interest are unclear. The researcher may also simply choose to use one matching procedure, in which case $p(\mathcal{M})$ is a spike prior.

2. $\theta_{\mathcal{M}}$ represents possible parameters in the matching procedure. One example would be the coefficients in a model to estimate a propensity score or prognostic score. Not all matching procedures have parameters to be estimated, so step 2 may be skipped. The full conditional is

$$p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M}, D, \theta^{mis}) = p(\theta_{\mathcal{M}}|Y, X, W, \mathcal{M})$$

because $\theta_{\mathcal{M}}$ is estimated from the observed data and only depends on the data and the matching procedure used.

3. $D^{(i)}$ is calculated directly from the first two steps. $\mathcal{M}$ and $\theta_{\mathcal{M}}$ determine the rules by which an observation is considered a match so once $\mathcal{M}$ and $\theta_{\mathcal{M}}$ are known, D_i is completely determined. The other parameters do not affect $D^{(i)}$, so the full conditional can be thought of as

$$p(D^{(i)}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M}, \theta^{mis}) = p(D^{(i)}|Y, X, W, \theta_{\mathcal{M}}, \mathcal{M})$$

where the full conditional is a spike. Any uncertainty or randomness over $D^{(i)}$ is simply a function of uncertainty over $\mathcal{M}$ and/or $\theta_{\mathcal{M}}$. I also consider each $D^{(i)}$ to be independent so that an observation can be a donor for multiple donor pools.

4. θ_i^{mis} is finally estimated from the matched sample. Conditional on D_i, estimating θ_i^{mis} requires simply estimating the mean from a sample consisting of the donor pool. In most cases, if conjugate priors are chosen, then the full conditionals are also conjugates where

$$p(\theta_i^{mis}|Y, X, W, D, \theta_{\mathcal{M}}, \mathcal{M}) \quad = \quad p(\theta_i^{mis}|Y, D^{(i)})$$

What was previously an intractable posterior for θ_i^{mis} becomes incredibly easy to simulate from with the augmentation of D. Once the donor pool is known, it is simply a matter of modeling the donor pool. The draws of $\tilde{\theta}_i^{mis}$ form the posterior distribution of θ_i^{mis} and capture the estimation uncertainty.

5. After simulating n_{sim} values of $\tilde{\theta}_i^{mis}$ from the posterior, I impute Y_i^{mis} by drawing one $\tilde{Y}_i^{mis}$ for each

$\tilde{\theta}_i^{mis}$ from the posterior predictive distribution

$$p(Y_i^{mis}|Y) = \int p(Y_i^{mis}|\theta)p(\theta|Y)d\theta$$

Simply put, the model uses each draw of $\tilde{\theta}_i^{mis}$ and predicts a value of Y_i^{mis} by drawing from $f(\cdot|\tilde{\theta}_i^{mis})$. While the estimation uncertainty is captured by the n_{sim} draws of $\tilde{\theta}_i^{mis}$, the fundamental uncertainty is captured by the sampling in this step.

6. Drawing from the posterior of τ_i is straightforward given that there is a deterministic relationship between τ_i, Y_i, and Y_i^{mis}. Let the posterior distribution for τ_i be

$$p(\tau_i|Y) = \int p(\tau_i|Y_i^{mis}, Y)p(Y_i^{mis}|Y)dY_i^{mis}$$

where $p(\tau_i|Y_i^{mis}, Y)$ is a spike distribution. Since I have simulations from $p(Y_i^{mis}|Y)$, the posterior of τ_i can be simulated simply by taking each draw of $\tilde{Y}_i^{mis}$ and calculating

$$\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$$

Note that in the algorithm, steps 3-6 are conducted separately for each i. Although in practice, the steps may be done altogether for all i, I choose to characterize the i's separately for both pedagogical and substantive purposes. One should consider each τ_i as a separate estimand estimated separately to avoid criticisms of multiple testing and cherry-picking specific ICEs. Theoretically, one should think of this framework as conducting N separate studies to estimate N different causal effects. For each study, imagine a dataset consisting only of observation i and all observations j where $j \neq i$ and $W_j = W_i$. In this framework, each observation may be used as a donor observation for multiple pools. When estimating ATEs, researchers who match with replacement must reweight the donor observations to reflect the correct number of observations in the data. In the case of estimating ICEs, no reweighting is necessary from a conceptual standpoint since the N ICEs are estimated in "separate" studies. However, if certain observations are used as donors many times, the multiple testing problem may be exacerbated, especially if the repeat donors are outliers. Overall, it is still unclear how including observations in multiple donor pools affects the estimates of the variances of the ICEs.

1.4.4 Comparison to Existing Approaches

I see a few contributions of the framework and model I have proposed. In addition to calling attention to focusing on ICEs in general, my model combines the ideas of matching and Bayesian analysis to estimate different causal quantities of interest. My model is flexible in the choice of matching and also allows for exploration and discovery of different treatment effects and treatment effect heterogeneity.

The approach I use to estimate ICEs bears many similarities to existing frameworks. I now discuss the similarities between my approach and the approach laid out by Rubin first in Rubin (1978) and then discussed in Rubin (2008) and most recently extended in Pattanayak, Rubin and Zell (2012), hereafter known as PRZ.[13] While none of the papers explicitly discuss individual causal effects as a quantity of interest, they all allow for the imputation of missing potential outcomes using Bayesian methods, which is also a characteristic of my approach. I argue that although there are subtle differences between my approach and the Rubin approach, my framework can be described as a generalization of the Rubin framework.

The first difference between the two approaches is in the data generating process and defining what is random. The Rubin approach assumes that Y, W, and X are all realizations from random variables whereas I assume that W and X are fixed and Y is only random because of unmeasured prognostic variables. I see the distinction between the two approaches on this point to be negligible. The idea of unmeasured prognostic variables leading to random outcomes is not incompatible with the Rubin approach. Furthermore, both approaches place great importance on the assumptions of ignorability of treatment assignment and SUTVA. My approach also allows for conditioning on non-confounding prognostic variables to improve the imputations of the missing potential outcomes. Since the estimand in the Rubin approach is an average treatment effect, including non-confounding prognostic variables is less important although in many cases, it can lead to more efficient estimates.

A second difference between the two approaches is that the Rubin approach models the observed outcomes whereas I keep the observed outcomes fixed. On the surface, this may seem like a big difference. But in reality, the difference is mostly in the framing of the problem rather than any substantive differences. The

[13] The PRZ approach has a very specific model and specific quantities of interest that are applicable to their data and question. I describe the PRZ approach in very general terms and discuss how the general PRZ setup compares to my framework.

Rubin approach estimates θ_T and θ_C, which are the means of the distributions of treated and control units, from the observed treated and observed control units respectively. PRZ go one step further by stratifying observations either by their propensity scores or by existing substantive strata and estimating a separate pair of θ for each strata. The Rubin approach then draws the missing potential outcomes from distributions centered at θ_T and θ_C. This is exactly the same approach that I use. For a missing $Y_i(0)$ outcome, θ_i^{mis} is estimated from a donor pool of control observations deemed to be good matches. Similarly, for a missing $Y_i(1)$ outcome, θ_i^{mis} is estimated from a donor pool of treated observations deemed to be good matches.

The difference is that each observation has a separate θ_i^{mis} to impute its missing potential outcome. In the earlier Rubin approaches, there are only two θ's, a θ_C to impute for treated units and a θ_T to impute for control units. PRZ allows for more flexibility by having strata-specific θ's. My approach basically generalizes PRZ by allowing each i to have its own individual strata.

To see this more clearly, suppose that there is a strata consisting of two treatment units, T_1 and T_2, and two control units C_1 and C_2. All four units are deemed to be good matches for each other, so assume ignorability of treatment assignment. In the PRZ approach, one would impute the missing $Y(0)$ for T_1 and T_2 with $\tilde{\theta}_C$ estimated from C_1 and C_2. Similarly, one would impute the missing $Y(1)$ for C_1 and C_2 with $\tilde{\theta}_T$ estimated from T_1 and T_2. Under my approach, the missing outcome for T_1 is imputed from $\tilde{\theta}_{T_1}$ estimated from C_1 and C_2, the missing outcome for T_2 is also imputed from the same $\tilde{\theta}_{T_2}$ estimated from C_1 and C_2 where $\tilde{\theta}_{T_1} = \tilde{\theta}_{T_2}$ and the missing outcomes for C_1 and C_2 are imputed from the $\tilde{\theta}_{C_1}$ and $\tilde{\theta}_{C_2}$ estimated from T_1 and T_2, where $\tilde{\theta}_{C_1} = \tilde{\theta}_{C_2}$. The two approaches are exactly the same assuming that my matching procedure produces the same strata. However, my approach is more generalizable in that the researcher can implement a matching procedure that does not restrict the donor pool to be within the same strata. T_1 can have a donor pool of C_1 and C_2 whereas T_2 can have a donor pool of C_1, C_2, and C_3.

This brings us to a third difference between my approach and the existing Rubin approach, namely that my framework allows for matching and uncertainty in the matching procedure and matching parameters. In PRZ, the strata are assumed to be exogenously defined or estimated beforehand with propensity score stratification. Once the strata are defined, they cannot be changed and the donor pool stays constant. My approach allows for multiple matching procedures and uncertainty within each matching procedure to characterize uncertainty about which observations constitute the correct donor pools. In approximate matching methods, this uncertainty certainly exists amongst researchers.

The Bayesian model I have proposed is unique in a couple ways. First, my model is explicit in that the quantity of interest is the individual causal effects. Most models estimate average treatment effects and consider the individual effects only indirectly if at all. Second, the data generating process I propose is slightly unconventional. Third, it embeds a relatively non-parametric matching step in the imputation of Y^{mis}. Finally, it allows for uncertainty over parameters within the matching procedure or uncertainty over which matching procedure to choose itself. As with any Bayesian model, the model is sensitive to choice of priors and convergence is not guaranteed in finite time. However, the ability to use priors also has the advantage of incorporating substantive information or restricting the range of possible values to help overcome sample size issues.

1.5 Other Quantities of Interest

Once the posterior for the individual causal effects is obtained, any sample estimand can be calculated rather easily by aggregating subsets of individual causal effects. For example, the posterior of the sample average treatment effect can be obtained by averaging the set of draws of τ_i for all i at each iteration of the Markov chain. Similarly, my approach allows for discovery and exploration of treatment effect heterogeneity by averaging over subsets of τ_i, such as averaging the draws for the τ_i for treated individuals to get the posterior of the sample ATT, averaging over draws for subsets of individuals with certain covariate values to get the sample CATE, etc. The researcher can graphically visualize heterogeneity by plotting the ICEs against various covariates. One can also ask questions such as the probability that the sample CATE is greater for individuals with $X = a$ versus individuals with $X = b$ for any values a and b simply by differencing the posterior draws. Obtaining posterior draws for τ_i for every individual in the sample allows for almost limitless possibilities to examine treatment effect heterogeneity.

Although various sample estimands are easy to calculate with this framework, it is unclear how one would estimate population or super-population estimands under my framework. Recall that the model assumes a finite sample and a Bayesian framework. It imputes the missing potential outcome for each individual in the sample while allowing the observed outcome to be fixed and unmodeled. The framework does not extend easily to super-population estimands because both potential outcomes are missing for individuals not in the sample. One way to get at super-population estimands may be to use bootstrapping. For each bootstrapped sample, calculate the estimands using the estimated posterior for the bootstrapped individuals and repeat to obtain a posterior over the super-population estimand. However, this process assumes that our sample is completely representative of the super-population. More specifically, it assumes that every other individual not observed in the super-population is exactly the same as an individual in our observed dataset. Furthermore, the bootstrap process almost certainly underestimates the uncertainty around super-population estimates because of the fixed and unmodeled potential outcome in the model. Generally speaking, the idea of estimating individual causal effects and estimating super-population estimands are contradictory in the sense that a super-population by definition contains nameless and exchangeable individuals whereas individual causal effects involve specific individuals in the dataset. For these reasons, I restrict the framework to estimating sample estimands of interest.

Another related issue is whether or not my framework allows for out-of-sample predictions or predictions for future observations. For reasons similar to those for estimating super-population effects, out-of-sample predictions are not straightforward. One can reasonably predict the treatment effect for an out-of-sample observation by finding and using the results for an in-sample observation with a similar covariate profile. For out-of-sample observations where no in-sample observations match reasonably well, the data does not give much information and a parametric model is needed. However, I argue that the same issues of model dependence for prediction occur in any other estimation framework. My model uses all available information in the data.

1.6 Applications and Extensions

The framework I have introduced is flexible enough to be applied to many situations and can be extended in various ways. Some applications and extensions to consider include:

- **Binary treatment with any type of outcome variable**: The simplest situation that I apply the model to is a dataset with a binary treatment variable, various covariates, and an outcome variable of any type. The outcome can be continuous or discrete, and treatment should be ignorable given the observed covariates.

- **Non-binary treatment**: The framework can be easily extended to non-binary treatment variables by retaining a linearity assumption. Instead of two potential outcomes, each individual has possibly an infinite number of potential outcomes. However, by assuming a linear relationship between the treatment and the outcome, one only needs to impute one missing potential outcome and extrapolate the rest by assumption. The linearity assumption also allows researchers to use individuals with significantly different treatment values to impute the same missing potential outcome.

- **Missing data in the covariates**: Since the model uses a Bayesian framework, one can easily incorporate imputation of missing data in the covariates via any of the existing multiple imputation techniques prevalent in the missing data literature.

- **Two-stage models**: The two related topics of treatment non-compliance and instrumental variables can be incorporated into the model via existing techniques. For example, one can model treatment non-compliance via principal stratification (Frangakis and Rubin 2002) by applying ICEs into the first stage of a two-stage model and incorporating existing Bayesian models (Imbens and Rubin 1997) into the sampler. The researcher can then use the principal stratifications from the first stage to calculate ICEs in the second stage. The framework can also be used to test the monotonocity assumption in instrumental variables models by estimating individual causal effects in the first stage.

- **Time-series cross-sectional/panel/multiple measurements data**: The framework can also handle data where individuals are measured repeatedly over time. Multiple measurements of outcomes and/or covariates and treatment give the researcher more information to match on and impute with. One would simply need to model the time component and decide on how to incorporate the extra information into the framework.

The next chapter tests various aspects of my framework via simulation to see how well various methods can recover individual causal effects. I then present various applications of my framework to real data and questions of interest to academics and policymakers in the general social science world.

Chapter 2

A Simulation Study

To test the ability of my model and the various matching procedures to recover individual causal effects and other quantities of interest, I conduct a simulation study to compare multiple methods. The simulation study generates toy data with the individual causal effects known and I evaluate the ability of the various matching methods to recover the ICEs on several evaluation criteria. I consider both continuous and binary dependent variables and evaluate the performance of the different matching methods as well as the different choices researchers must make with regard to the number of matches and the number of conditioning variables to include. The simulations suggest that in general, predictive mean matching seems to outperform other matching methods in recovering the ICEs.

2.1 Methods to be Compared

Recall the MCMC algorithm for the posterior of τ_i from before restated below. The simulations test various choices of $\mathcal{M}$ in step 1 of the algorithm. The choice of $\mathcal{M}$ consists of choosing a matching method, the number of matches used, and the set of variables to match on. To test the performance of different specifications of $\mathcal{M}$, I hold $\mathcal{M}$ constant each time, with the exception of possibly a random choice of the number of matches to use. Thus, in the simulation study, step 1 of the sampler is the same for each iteration within a single specification with the exception of specifications with random number of matches M. In those specifications, the number of matches varies across iterations but stays constant across i within the same iteration.

```
MCMC Algorithm for the Posterior of $\tau_i$

Repeat the following $n_{sim}$ times:

Gibbs Sampler:
    1. Draw a matching procedure $\tilde{\mathcal{M}}$ from $p(\mathcal{M})$.
    2. Draw $\tilde{\theta}_{\mathcal{M}}$ from $p(\theta_{\mathcal{M}}|Y, X, W, D, \theta^{mis}, \mathcal{M})$.

for ($i$ in 1:$N$){
    3. Determine $\tilde{D}^{(i)}$ from matching procedure. (matching step)
    4. Draw $\tilde{\theta}_i^{mis}$ to estimate $\theta_i^{mis}$.
}

Draw from PPD and Calculate $\tau_i$:
for ($i$ in 1:$N$){
    5. Draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\tilde{\theta}_i^{mis})$. (imputation step)
    6. Calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$.
}
```

I define matching method to be the specification of the distance metric used and the method of picking matches given the distance metric. The four matching methods I consider are

1. **Mahalanobis matching**: The first distance metric I consider is the (squared) Mahalanobis distance metric used in Rubin (1980). The Mahalanobis distance between two observations with covariate values X_1 and X_2 is

$$\Delta_M(x_1, x_2) = \sqrt{(X_1 - X_2)^T S^{-1} (X_1 - X_2)}$$

   where S^{-1} is the sample covariance matrix of X. For τ_i, I calculate the squared Mahalanobis distance between X_i and X_j, $\forall W_i \neq W_j$ and then use the M nearest neighbors as matches. Unlike the remaining matching methods, Mahalanobis matching is model-free in the sense that it only looks at the in-sample covariate distances rather than imposing a parametric model.

2. **predictive mean matching**: Since the goal of estimating ICEs is to impute the missing potential outcomes with matching, one way to do this is to first model the means of the observed outcomes in the treatment and control groups and match based on the model. Let i index any treated observation. Then the best imputation of $Y_i(0)$ is likely to come from control observations with observed outcomes that are closest to $Y_i(0)$. Denote Y_c and X_c as the observed outcomes and covariates for the control group and Y_t and X_t as the analogous for the treatment group. Since $Y_i(0)$ is unobserved, I first make a best guess of $Y_i(0)$ by modeling the outcomes for the control group with a linear regression of Y_c on X_c.[1] Let θ_c denote the vector of parameters (β_c, σ_c^2) from this regression. I then calculate a predictive mean score for all observations as

$$\tilde{\mu}_{(c)} = X\tilde{\beta}_c$$

Note that $\tilde{\mu}_{(c)}$ is calculated for all observations and the subscript refers only to the fact that the predictive mean score is calculated from $\tilde{\beta}_c$. For treated observation i, use the M nearest neighbor control observations on $\tilde{\mu}_{(c)}$ as its matches.[2] $\tilde{\mu}_{i,(c)}$ basically serves as our best initial guess of $Y_i(0)$ based on a regression model.

Now let j index any control observation. To estimate τ_j, I do predictive mean matching with a similar process. Regress Y_t on X_t to get an estimate of θ_t, which consists of (β_t, σ_t^2). Calculate another predictive mean score for all observations as

$$\tilde{\mu}_{(t)} = X\tilde{\beta}_t$$

For control observation j, use the M nearest neighbor treated observations on $\tilde{\mu}_{(t)}$ as its matches. $\tilde{\mu}_{i,(t)}$ serves as the initial guess of the missing $Y_j(1)$. In essence, one can think of this process as conducting predictive mean matching twice with the treatment indicators reversed the second time.

Within the MCMC algorithm, predictive mean matching involves drawing $\theta_{\mathcal{M}} = \{\beta_t, \beta_c, \sigma_t^2, \sigma_c^2\}$ in the

[1] For now, I assume that the covariates enter the regression linearly without any interactions or polynomials.

[2] One can also match $\tilde{\mu}_{(c)}$ with the actual observed control outcomes although it will be more difficult to differentiate between good matches with discrete outcome variables.

second step with a Gaussian linear regression. For priors, I use

$$\begin{aligned}\beta_w &\sim \text{improper uniform}\\ \sigma_w^2 &\sim \mathcal{IG}\left(\frac{0.001}{2}, \frac{0.001}{2}\right)\end{aligned}$$

for $w = t, c$. Since these parameters only depend on $\mathcal{M}$ and the observed data, the full conditionals to draw from are simply the conditional distributions in a Gaussian linear regression.

$$\begin{aligned}\beta_w|\sigma_w^2, Y_w, X_w, \mathcal{M} &\sim \mathcal{N}\left(m^*, V^*\right)\\ V^* &= \left(X_w'(\sigma_w^2 I)^{-1}X\right)^{-1}\\ m^* &= V^*(X_w'(\sigma_w^2 I)^{-1}Y_w)\end{aligned}$$

$$\begin{aligned}\sigma_w^2|\beta_w, Y_w, X_w, \mathcal{M} &\sim \mathcal{IG}\left(\frac{\nu^*}{2}, \frac{\delta^*}{2}\right)\\ \nu^* &= n_w + 0.001\\ \delta^* &= (Y_w - X_w\beta_w)'(Y_w - X_w\beta_w) + 0.001\end{aligned}$$

for $w = t, c$ where n_w is the number of observations in treatment group w. Step 3 of the algorithm uses the draw of $\theta_{\mathcal{M}}$ at each iteration to find matches for each observation through the predictive mean matching process described. The benefit of predictive mean matching is that the distance measure is most directly related to the quantity of interest of the missing potential outcomes. With a large enough sample, predictive mean matching should produce balance between an observation and its matches since observations with the same observed covariate values should have the same predictive mean up to some degree of randomness. Predictive mean matching reverses the process by assuming that observations with similar predictive means should have similar observed covariate values.

3. **propensity score matching**: The propensity score is defined as the conditional probability of being assigned to treatment given a vector of covariates X. Under randomized treatment assignment, the propensity score should be a known function whereas in observational studies, the propensity score is unknown and must be estimated. The propensity score reduces the dimensions of X down to a scalar and Rosenbaum and Rubin (1985) show that adjusting for the propensity score is sufficient

for producing unbiased estimates of treatment effects. Furthermore, they show that adjusting for the sample estimate of the propensity score can produce balance on the covariates in the sample.

Define the propensity score for observation i as

$$e_i = P(W_i = 1|X_i)$$

I estimate the propensity scores for all observations using a logistic regression within the MCMC algorithm. In step 2 of the algorithm, let $\theta_{\mathcal{M}}$ be the coefficients β from a Bayesian logistic regression of W on X.[3] Our estimated propensity scores take the form

$$\tilde{e}_i = \frac{1}{1 + \exp(-X_i \tilde{\beta})}$$

Note that the propensity scores are a function of draws from the posterior of the regression. For each draw of $\tilde{\beta}$, I calculate a propensity score $\tilde{e}_i(X_i)$ for each individual. Since the propensity scores are unknown and estimated, this incorporates uncertainty over the propensity scores, an approach similar to that in An (2010). For matching, I use nearest neighbor matching on the linear propensity score

$$\ln\left(\frac{\tilde{e}_i}{1 - \tilde{e}_i}\right) = X_i \tilde{\beta}$$

which has been found effective for reducing bias in the matching literature (Rubin 2001). For each observation i, matches are produced by taking the M observations in the opposite treatment group with the closest linear propensity score. Observations may be used as donors to multiple other observations, but can only be used once for any particular observation.

Within the MCMC algorithm, estimating a logistic regression in step 2 requires embedding a Metropolis-Hastings step. I use an improper uniform prior on β and a random walk Metropolis algorithm.

4. **subclassification** (on the linear propensity score): In addition to nearest neighbor matching on the linear propensity score, I also consider subclassification on the linear propensity score. The idea behind subclassification is to sort the estimated propensity score and then divide the observations

[3] Again, for now X enters into the propensity score equation linearly.

into M subclasses based on the ordered propensity scores.[4] Rosenbaum and Rubin (1984) show that subclassification on the propensity score with as few as five subclasses can substantially reduce bias in estimating treatment effects. Much like choosing the number of matches, choosing the number of subclasses is part of the choice of $\mathcal{M}$ in the algorithm. I consider both fixed and random M in my simulations. Within the algorithm, the linear propensity scores are estimated exactly as above, and the subclassification affects the choice of which observations contribute to the donor pool $\tilde{D}_i$. Observations in the same subclass as the observation to be matched are considered to be a part of the donor pool. I restrict the analyses to contain at least two treated and two control observations in every subclass. Because the linear propensity scores are estimated stochastically, within any specific iteration, it is possible to have subclasses that do not contain at least two treated and two control observations. In those rare instances, I decrease M by one for that iteration of the algorithm only until every subclass in that iteration meets the criteria.

The simulations presented compare the choice of one of these methods as well as the number of matches/subclasses and the set of variables to match on. All of these choices are captured in $\mathcal{M}$ in step 1 of the algorithm. As mentioned before, each simulation holds constant the choice of method and number of variables to match on. The number of matches/subclasses are either held constant or allowed to vary randomly within a range. Within a single iteration in a simulation, steps 1 and 2 produce a donor pool for every observation i, which is denoted $\tilde{D}_i$ in step 3. Using the donor pool, I then draw a value of $\tilde{\theta}_i^{mis}$ in step 4 by modeling the mean of the donor pool. For continuous outcome variables, I draw $\tilde{Y}_i^{mis}$ from the posterior predictive distribution $\mathcal{N}(\theta_i^{mis}, \sigma_i^{2\,mis})$.[5] For binary outcome variables, I draw $\tilde{Y}_i^{mis}$ from a Bern(θ_i^{mis}) distribution.

In addition to the four matching methods, I also consider two methods which do not use a matching procedure as a baseline.

1. (Bayesian) **regression imputation**: I take the simplest and most commonly used case where the imputations of the missing potential outcomes are generated from the coefficients of a Bayesian linear

[4] In the context of subclassification, I use M to refer to the number of subclasses rather than the number of matches. Increasing M actually decreases the number of subclasses holding sample size constant.

[5] $\sigma_i^{2\,mis}$ is also estimated from the donor pool with an $\mathcal{IG}\left(\frac{0.001}{2}, \frac{0.001}{2}\right)$ prior.

regression model. I fit a regression model of Y on W and X using the priors

$$\begin{aligned}\beta &\sim \text{improper uniform} \\ \sigma^2 &\sim \mathcal{IG}\left(\frac{0.001}{2}, \frac{0.001}{2}\right)\end{aligned}$$

The missing potential outcomes are then imputed from the coefficients $\tilde{\beta}$ such that

$$\tilde{Y}_i^{mis} = \tilde{\beta}_0 + \tilde{\beta}_1(1 - W_i) + \tilde{\beta}_X X_i$$

where β_0 is the intercept, β_1 is the coefficient on W, and β_X is the set of coefficients on X from the regression. Since I use fairly uninformative priors, the estimates from this Bayesian regression will be nearly identical to estimates from a non-Bayesian regression. I use a Bayesian regression simply to remain consistent with the other approaches I test. I use this model as a baseline since this is probably the simplest and most common regression model-based way to impute potential outcomes. Note that the imputations here come solely from an estimate of an average treatment effect.

2. **no matching (all)**: I consider the case where all of the j observations where $W_i \neq W_j$ are used as matches for i. In this specification, no matching algorithm is used since all observations of the other treatment group are used as matches. In the case where treatment assignment is randomized, one would expect that no matching would produce roughly the same quality of matches as other matching algorithms. The donor pool for this method is simply all observations with a different treatment status and the estimation of θ_i^{mis} and imputation of Y_i^{mis} follows the same process as the matching procedures above.

Within each method, I also test the sensitivity of the choice of the number of matches to use and the set of covariates to include where appropriate. Thus, for each specification of $\mathcal{M}$ that I test, I vary all three dimensions that the researcher can choose.

2.2 Setting Up the Simulations

In general, I will only discuss how I perform the simulations and the results for the case of a continuous dependent variable. I also repeat some of the simulations for a binary dependent variable, but the results are similar so I relegate those simulations to the appendix.

Data generating processes

To assess the performance of the different methods, I generate fake data from numerous linear and non-linear data generating processes to test how well the methods recover various causal estimands of interest. The data generating processes are borrowed from the ones used by Hainmueller (2012) and Frölich (2007) with a few changes tailored specifically to the framework used here. The best performing method(s) should ideally be fairly robust to deviations from non-linearity in the data generating process even though I only use linear specifications. I also consider three different sample sizes of 100 (small), 1000 (medium), and 5000 (large).

To begin, I generate ten covariates that completely determine the outcomes:

- $x_1 \sim \mathcal{N}(0, 2^2)$
- $x_2 \sim \mathcal{N}(0, 1)$
- $x_3 \sim \mathcal{N}(0, 1)$
- $x_4 \sim \mathcal{U}(-3, 3)$
- $x_5 \sim \chi^2_1$
- $x_6 \sim Bernoulli(.5)$
- $x_7 \sim \mathcal{N}(0, 1)$
- $x_8 \sim \mathcal{N}(0, 1)$
- $x_9 \sim \mathcal{N}(0, 1)$
- $x_{10} \sim \mathcal{N}(0, 1)$

Using these ten covariates, I generate the potential outcome $Y_i(0)$, the outcome without treatment, for each observation i. I consider three different outcome generating equations:

1. $Y(0) = x_1 + x_2 + x_3 - x_4 + x_5 + x_6 + x_7 - x_8 + x_9 - x_{10}$
2. $Y(0) = x_1 + x_2 + 0.2x_3x_4 - \sqrt{x_5} + x_7 + x_8 - x_9 + x_{10}$

3. $Y(0) = (x_1 + x_2 + x_5)^2 + x_7 - x_8 + x_9 - x_{10}$

The three equations vary in their degree of linearity, starting from a (1) linear relationship between Y and X and going to (2) a moderately non-linear and (3) very non-linear relationship. For each i, I then assign treatment in three different ways:

1. $p(W = 1) = 0.5$
2. $\eta = x_1 + 2x_2 - 2x_3 - x_4 - 0.5x_5 + x_6 + x_7$
   $W = 1$ if $\eta > 0$; otherwise $W = 0$
3. $\eta = 0.5x_1 + 2x_1x_2 + x_3^2 - x_4 - 0.5\sqrt{x5} - x_5x_6 + x_7$
   $W = 1$ if $\eta > 0$; otherwise $W = 0$

In the first case, treatment assignment is completely random with equal probability of being assigned treatment or control. In the second case, treatment assignment is linearly related to the first seven covariates. Since in my framework, there exists a set of covariates $X^{(p)}$ that completely explain the outcomes, I also allow a subset of the covariates (the first seven covariates) to be confounders that perfectly predict treatment assignment. In the third case, the first seven covariates are non-linearly related to treatment assignment. Note that in scenarios 2 and 3, conditioning on $x1$ through x_7 is sufficient to control for confounders. The three outcome equations and the three treatment assignment scenarios create nine different data combinations that range from unconfounded and linear in Y to (linear and non-linear) confounded treatment assignment and very non-linear in Y.

For most specifications, I draw each "true" τ_i independently from a $\mathcal{N}(2, \sqrt{3}^2)$ distribution. Drawing τ_i independently gives the most general situation in which each individual's τ_i gives no information about any other τ_i. If one considers the case where treatment effect heterogeneity is explained by some observed covariate, then matching on that covariate should improve the ability of the model to capture the different τ_i. Thus, drawing the true τ_i independently serves as a conservative test of the methods' ability to estimate the individual causal effects. In a few other specifications, I also vary the distribution from which τ_i is drawn. Specifically, I consider cases where the τ_i are drawn independently from:

1. $\mathcal{N}(2, \sqrt{3}^2)$

2. $\mathcal{N}(20, \sqrt{3}^2)$

3. $\mathcal{N}(2, \sqrt{100}^2)$

4. $\mathcal{N}(20, \sqrt{100}^2)$

5. mixture of $\mathcal{N}(2, \sqrt{3}^2)$ and $\mathcal{N}(20, \sqrt{3}^2)$ with equal probability on each

6. mixture of $\mathcal{N}(2, \sqrt{100}^2)$ and $\mathcal{N}(20, \sqrt{100}^2)$ with equal probability on each

By varying the mean of the τ_i distribution, I vary the size of the effects to see how well the methods perform as the effect sizes increase. I vary the standard deviation of the τ_i distribution to test how well the methods perform over a changing range of τ_i. I expect the methods to perform better with greater effect sizes (more power) and a smaller range over τ_i (less heterogeneity). I also consider the mixture distributions to simulate scenarios in which treatment effects are clustered such that treatment has a range of effects for one group and a different range of effects for another group. For example, treatment may hurt one group of individuals and help another group.

To complete the data generating processes, I generate $Y_i(1)$:

$$Y_i(1) = Y_i(0) + \tau_i$$

I then put together the "observed" dataset that the models use. To mirror the typical data analysis, I run the different model specifications that I test using the datasets containing the following variables:

- W
- $Y = W \times Y(1) + (1 - W) \times Y(0)$
- $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$

Causal estimands of interest to be recovered

Since I generate the individual causal effects τ_i in the simulations, τ_i and any other causal estimand is known. The goal of the simulations is to evaluate how well a method can recover the known true values of these estimands. I consider how well a method recovers the following causal estimands in the simulations:

- **Individual causal effects**: The most important estimands to recover are the N τ_i's themselves. For the case of binary dependent variables, τ_i can only take on values of -1, 0, and 1 so it is more difficult to actually evaluate how well the methods recover the τ_i since the posterior distribution is a mixture of two of the possible three values. Therefore, I only look at the aggregated estimands described below for the simulations with binary dependent variables.

- **Average treatment effect**: Another important quantity to recover is the ATE. Any method that can recover the ICEs should be able to recover the ATE correctly since the ATE is a simple linear function of the ICEs. Since the ATE is usually the easiest estimand to estimate, any method that performs poorly on recovering the ATE is probably not a very robust and useful method.

- **Average treatment effect on the treated**: The ATT is another average effect that calculates the average effect over a subset of the data. Since recovering the ICEs correctly implies recovering any aggregation of the ICEs, I should be able to randomly choose any subset and calculate the average effect and judge a method by its ability to recover this average effect.

- **Treatment effect quantiles** (0.5, 0.75, 0.95): Since I claim that estimating ICEs allows for unparalleled flexibility in recovering any other causal estimand, I put the method to a difficult test by attempting to recover the treatment effects at different quantiles. To calculate a quantile treatment effect, I sort the ICEs from lowest to highest and then take the desired quantile of these sorted effects. Even though my simulations have even numbered sample sizes, I take the quantiles without averaging, so the 0.75 quantile treatment effect for $N = 1000$ is the 750th ordered statistic for the sorted τ_i. As the quantiles become more extreme, I expect any method to perform worse so my model should recover the 0.5 quantile with more accuracy and precision than the 0.75 and 0.95 quantiles. The results available in the appendix confirm this to be true.

Performance metrics used to evaluate the methods

The typical simulation study uses performance metrics such as bias, mean squared error, confidence interval coverage, or power to evaluate a statistical method. All of these metrics stem from a frequentist perspective where the data is assumed to be sampled randomly many times and each time the method calculates a statistic that characterizes the sampled data. All the metrics used are concerned with how the method performs on average over repeated samples. These traditional metrics are inappropriate in the current context for two reasons. First, the method I propose is fundamentally a Bayesian method that does not rely on a repeated sampling framework. Instead, the data is assumed to be sampled once and a Bayesian method conditions on the actual observed dataset only, so using traditional metric to test the repeated sampling properties of a Bayesian method makes little sense. Second, the whole idea of individual causal effects as I present them here is incompatible with a repeated sampling framework. My framework assumes that the potential outcomes are fixed. Therefore, the estimand does not change regardless of how many times you sample. τ_i remains the same for individual i even if i was sampled repeatedly. Furthermore, since individual causal effects are specific to individual i, a repeated sampling framework would involve sampling i such that i appears in the dataset for some samples and not others. For samples that do not include i, τ_i is unestimable. Therefore, I cannot use traditional notions of repeated sampling to evaluate the methods proposed.

Instead, I develop and use Bayesian versions of bias, mean squared error, power, and coverage. Under the Bayesian version, I replace the repeated sampling framework by evaluating the methods over the N individuals in the dataset. For example, instead of evaluating how a method performs on average over repeated samples, I evaluate how a method performs by averaging over the N individuals observed. The Bayesian metrics that I use for ICEs and other causal estimands of interest include posterior mean bias, expected error loss, the proportion of the credible intervals not including 0, and calibration coverage.[6]

- **Posterior mean bias** ("bias"): Let θ be any estimand or parameter of interest. The traditional bias

[6] For the simulations with binary continuous variables, I only look at posterior mean bias and expected error loss because the latter two metrics are difficult to calculate when τ_i only takes on discrete values of -1, 0, and 1.

of an estimator $\hat{\theta}$ is

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

where the expectation is taken over $\hat{\theta}$ under repeated samples of the data. $\hat{\theta}$ is usually some "best" estimate of θ. Contrast this with the posterior mean bias metric that I use.

$$\text{posterior mean bias} \quad = \quad E(\theta|X) - \theta$$

where X represents the observed data and $\theta|X$ is the posterior distribution of θ conditional on the observed data. The expectation here is the expectation of the posterior distribution, or the posterior mean. Using decision theory and a quadratic loss function, it can be shown that the posterior mean is the Bayes estimator in that it minimizes the expected loss given θ.[7] Therefore, the posterior mean bias is a Bayesian analogue of bias in the frequentist sense. It represents a broad notion of how far off from the truth our "best" estimate is. For the aggregated estimands such as the ATE or ATT, posterior mean bias is calculated simply as the mean of the MCMC simulations from the posterior distribution minus the true value of the estimand calculated from the τ_i, which are generated from a known data generating process. For the ICEs themselves, I can look at the posterior mean bias for each of N τ_i's, but I choose to summarize them by the average[8] and standard deviation of the N posterior mean biases to make comparing the methods easier.

[7] In decision theory, one must take an action or make a decision a assuming that the true state of nature is θ. Using a quadratic loss function $L(\theta, a) = (\theta - a)^2$, the expected loss given our posterior is

$$\begin{aligned} E_{\theta|X}[L(\theta,a)] &= \int (\theta - a)^2 p(\theta|X) d\theta \\ &= \int \theta^2 p(\theta|X) d\theta - 2a \int \theta p(\theta|X) d\theta + a^2 \int p(\theta|X) d\theta \\ &= E(\theta^2|X) - 2aE(\theta|X) + a^2 \end{aligned}$$

One can minimize the loss by differentiating with respect to a and setting it equal to zero, giving us the posterior mean as the decision or estimate that minimizes expected loss.

$$\hat{a} = E(\theta|X)$$

[8] Averaging over the N posterior mean biases for the τ_i is actually equivalent to looking at the posterior mean bias for the ATE due to the linearity of expectations.

- **Expected error loss** ("root mse"): For any parameter θ and estimator $\hat{\theta}$, the typical root mean squared error calculation is

$$\sqrt{E[(\hat{\theta}-\theta)^2]} \quad = \quad \sqrt{\text{variance} + \text{bias}^2}$$

again with the expectation taken over repeated samples. The root mean squared error gives a rough estimate of how far off the estimator is from the truth, taking into account both bias and variance. The Bayesian analogue that I use is the expected error loss, which does not require expectations over repeated samples. For notational clarity, now let θ denote a random variable for the parameter of interest and let θ^* denote the true underlying value of the parameter.[9] Then

$$\text{expected error loss} \quad = \quad \sqrt{\int (\theta-\theta^*)^2 p(\theta|X)d\theta}$$

In contrast to the posterior mean bias metric, the expected error loss metric accounts for the deviations from θ^* for all possible values of θ rather than just the point estimate at the posterior mean. It is basically a weighted average of the squared error loss for the entire support of the posterior. In practice, the expected error loss is calculated by taking each draw $\tilde{\theta}$ from the posterior and calculating its squared error relative to θ^* and then taking the average across the draws. For aggregate estimands like the ATE, I look at the expected error loss whereas for the N τ_i's, I look at the average of the N expected error losses.

- **Proportion of the credible intervals[10] not including 0** ("power"): In hypothesis testing, the typical definition of the power of a statistical method is the probability of the method rejecting the null hypothesis given that the null hypothesis is false. In other words, it is the probability of detecting

[9] The notation used in this section may be confusing because I attempt to compare frequentist and Bayesian methods assuming a fixed underlying true parameter, which is usually reserved only for frequentists. Bayesians usually describe parameters probabilistically using random variables even though a true underlying parameter value may exist. Since I am comparing estimates of θ from Bayesian models to a true value of θ in my simulations, I assume a fixed parameter value. To clarify the notation, whenever I discuss frequentist methods, θ is the fixed parameter value. When discussing Bayesian methods, θ can refer to the random variable for the parameter or the true underlying value given by nature. I attempt to be more explicit by using θ^* to represent the true underlying value when discussing both the random variable and the true underlying value.

[10] I use 95% credible intervals here and throughout to refer to the central 95% region of the posterior to be consistent with the idea of a 95% confidence interval. The interpretation of a 95% credible interval is that the truth lies in the interval with probability 0.95. In practice, I calculate the 95% credible intervals by simply taking the 0.025 and 0.975 quantiles of the posterior draws.

an effect when one exists or the probability of not committing a Type 2 error. The statistical power of a method depends on the statistical significance criteria used (α level), the magnitude of the effect or the effect size, and the sample size. Using the typical $\alpha = 0.05$ criteria, one would usually test the statistical power with simulation by randomly drawing data with the same sample size and the same predefined effect size that matches the alternative hypothesis[11], calculating the statistic or test for each sample, and then determining the proportion of samples in which the test rejects the null hypothesis (e.g. the proportion of times that the test "gets it right"). One direct way is to calculate the proportion of 95% confidence intervals that do not contain the null hypothesis. This proportion is a calculation of the statistical power given the specified α, effect size, and sample size.

For the application of my Bayesian model to the estimation of ICEs, I cannot use the typical way to calculate power because as described earlier, there is no repeated sampling principle on which to rely. Instead, I rely on the N observations in the simulated dataset with N ICEs as N "repeated samples." I then calculate the proportion of 95% credible intervals for the N τ_i's that do not include 0 as a rough estimate of the "power" for a particular method. The estimate is rough and does not exactly satisfy the definition of power in the typical sense. Assuming that the null hypothesis is $\tau_i = 0$,[12] the data generating process for the case of continuous outcome variables always generates $\tau_i \neq 0$ for all i, which satisfies the condition of the null hypothesis being false.[13] However, unlike the case of the typical power calculation, τ_i is not constant for all i, so the proportion is calculated over varying effect sizes. Nevertheless, given my framework and goals, this calculation gives a rough estimate of power which will approach the more traditional power calculation as the standard deviation on the τ_i approaches 0.

- **Calibration coverage** ("coverage"): The way typical simulation studies assess the accuracy of confidence intervals generated by a method is by looking at its coverage probability, which is the proportion of the time that the interval contains or "covers" the true value of the parameter. Recall the cor-

[11] If the null hypothesis is that the effect size is zero and the alternative hypothesis is that the effect size is not equal to zero, then the effect size in the simulations is set to a value that is not equal to zero.

[12] The language here is not exactly correct since I am using hypothesis testing language in a Bayesian context. Nevertheless, I use this language of testing for power because I want to compare the performance of different methods in capturing the effects when they exist.

[13] This is due to the fact that τ_i is continuous and the probability of drawing any specific value is 0 for a continuous distribution.

rect definition of a confidence interval, say the (nominal) 95% confidence interval, is that in repeated samples, 95% of the calculated 95% confidence intervals should contain the truth. Ideally then, the actual coverage probability of the method equals the nominal probability of 0.95. Deviations from 0.95 would suggest that some assumptions of the model are not met. To derive the coverage probability in a simulation, one would simulate repeated samples from the data generating process, holding the parameter at a single "true" value, calculate the 95% confidence interval each time, and then calculate the proportion of the confidence intervals that contain the "true" value.

Under my Bayesian framework for estimating ICEs, repeated sampling once again does not make sense because of the Bayesian and the ICE aspects. Much like the "power" calculation, I once again leverage the N τ_i's as a substitute for repeated sampling. Here I appeal to the idea of Bayesian calibration with credible intervals. A Bayesian 95% credible interval has a much more intuitive definition as the interval in which the true value occurs with 0.95 probability. Probability here is subjective since it is a function of both the data and the subjective prior probability. However, the idea of calibration is that the Bayesian model should produce a 95% credible interval that is calibrated such that it can predict 95% of future observations correctly. Applying this logic to the simulation for ICEs, a method that performs well should have 95% credible intervals that contain the true values 95% of the time. In my simulations, I calculate the proportion of the N 95% credible intervals that contain the true ICEs. Note that as in the calculation of the rough "power" statistic above, each τ_i varies, which differs from the traditional coverage calculations. However, with Bayesian calibration, each ICE 95% credible interval should ideally contain its own τ_i with 0.95 probability, so I can look across all N τ_i and estimate the proportion that contain its own true τ_i as the calibration coverage probability. The best performing methods are the ones that have coverage probability closest to 0.95 using the 95% credible intervals in the calculation.

I assess the performance of the different matching methods and specifications using all four of these metrics when possible. Each metric conveys a different aspect of model performance and the methods that perform the best ideally perform well on all four metrics.

Different specifications

As alluded to above, I test the ability of the model and different matching methods in estimating the causal estimands of interest. For the first set of simulations, I run numerous simulations of the model, each time varying one aspect of the model specification or one aspect of the data generating process. A model specification includes

- **choice of method**: regression, all, mahalanobis, predictive mean, propensity score, or propensity score subclassification
- **number of matches** (for mahalanobis, predictive mean, and propensity score) **or number of subclasses** (for propensity score subclassification): small, medium, large, or random[14]
- **number of X variables to condition on**: 0 (for the method all only), 5, 7, or 10[15]

In addition to varying the model specifications, I also vary the data generating process for each specification. The data generating process specifications are

- **sample size**: 100, 1000, or 5000[16]
- **outcome generating equation**: linear, moderately non-linear, or very non-linear
- **treatment assignment**: unconfounded, confounded linearly, confounded non-linearly

[14] For the number of matches, small, medium, large, and random were defined as 2, 10, 25, and an integer uniformly drawn from the range 2 through 25 respectively. For the number of subclasses, small, medium, large, and random were defined differently depending on the sample size for each simulation. With sample size of 100, the number of subclasses used was 2,4,5, and an integer uniformly drawn from the range 2 through 5. With sample size of 1000, the number of subclasses used was 5,10,20, and an integer uniformly drawn from the range 5 through 20. With sample size of 1000, the number of subclasses used was 5,20,50, and an integer uniformly drawn from the range 5 through 50.

[15] The variables were conditioned on in order, so 5 X variables conditioned on means conditioning on x_1 through x_5 and so forth. Recall that for the confounded treatment assignments, the first 7 X variables were used in the confounding.

[16] When increasing the sample size, rather than regenerating a new dataset completely, I keep the previous sample and simply add on extra observations, so a dataset with sample size 1000 contains 100 observations from the previous simulation and adds 900 new observations. By adding on observations instead of regenerating completely new observations, I allow the datasets of different sizes to be comparable (conditional on the same generating equations) because the first 100 observations in the dataset are the same across the two sizes. I retain the condition that these "individuals" are the same, which is more coherent given the ICE framework.

For this first set of simulations, I hold the distribution of τ_i constant by generating them all from a $\mathcal{N}(2, \sqrt{3}^2)$ distribution. There are 52 combinations of model specifications and 27 combinations for the data generating processes, which lead to $52 \times 27 = 1404$ different simulations in the first set.

I then consider a second set of simulations that further tests the optimal number of matches to use. I hold the data generating sample size to 1000 with the nine different outcome/treatment assignment generating equations and only condition on 7 covariates. I only consider the case of predictive mean matching. The specification that varies is the number of matches, which I now specify as a percentage of the smaller treatment group. Given a simulated dataset, I take smaller of the treated or control groups and calculate the number of matches M as a percentage of this number (rounded up). The percentages I consider are

- every 1 percentage point between 1% and 9% inclusive
- every 10th percentage percentage point between 10% and 90% inclusive
- the case of 100%, which I then make equivalent to just the "all" matching method (so the 100% here is actually 100% of both treatment groups)

The different percentages produce 19 different specifications, combined with the 9 different data generating processes to produce $19 \times 9 = 171$ different simulations in the second set.

Finally I consider a third set of simulations to assess the sensitivity of the results to different ways of generating the true values of τ_i as I described above. I hold the sample size to 1000 again with the nine different outcome/treatment assignment generating equations, condition on only 7 covariates, and restrict the number of matches or subclasses to 25 (except for the case of the "all" method). For each of the six different ways of generating τ_i described previously, I vary the choice of method used. So for each of six different ways of generating τ_i, I have six different methods and nine different data generating processes, for a total of $6 \times 6 \times 9 = 324$ different simulations.

The three sets of simulations combined result in 1404+171+324=1899 different simulations. I then repeat for the case with a binary dependent variable. For each of the 1899 simulations, I derive the posterior from the algorithm described in the beginning. Due to computational and time issues, each MCMC is relatively short with a chain length of 2000. For the most part, my parameters are relatively independent so I am confident that my parameters are mixing well despite such a short chain length and no burn-in period.

2.3 Results from the Simulations

The simulations show that my model in general does a fairly good job at estimating ICEs, although with very high variance in both the estimates (posterior variance) and the quality of the estimates. Although the simulations produce many results and insights that are noteworthy, I only present a subset of the results that can guide researchers on the best practices and methods for estimating ICEs. The rest of the results from the simulations appear in the appendix. The general insights from the simulations are:

1. The model generally performs well in recovering ICEs and other causal estimands. Predictive mean matching generally outperforms all the other matching methods.

Figure 2.1 shows the results from the first set of simulations comparing the model using the different matching methods with different specifications and sample sizes. The metric here is the average ICE posterior mean bias, which is equivalent to the ATE posterior mean bias. A method or specification is judged by how close its posterior mean bias is to zero. In the top right quadrant with a linear outcome equation and unconfounded treatment assignment, most of the specifications are spot on in their estimate of the ATE.[17] As the outcome equations become more non-linear, the bias gets bigger, but the specifications on average have very little bias until the outcome equations become very non-linear. Looking across methods, the propensity score subclassification method is probably the most consistent in the sense that difference in bias across specifications is the smallest,[18] but the subclassification method is also the most easily biased. The propensity score matching method seems to be the most varied in performance across specifications and its bias seems to be somewhat larger as well. Although the differences are miniscule, it appears that

[17] For each method, the different points refer to different specifications of the number of matches, the number of conditioning variables, or the sample size (denoted by color). In all of these graphs, some points are not shown because they fall outside the general range of most of the specifications.

[18] Another way to put it is that the variance of the bias *across* specifications within the subclassification method is the smallest.

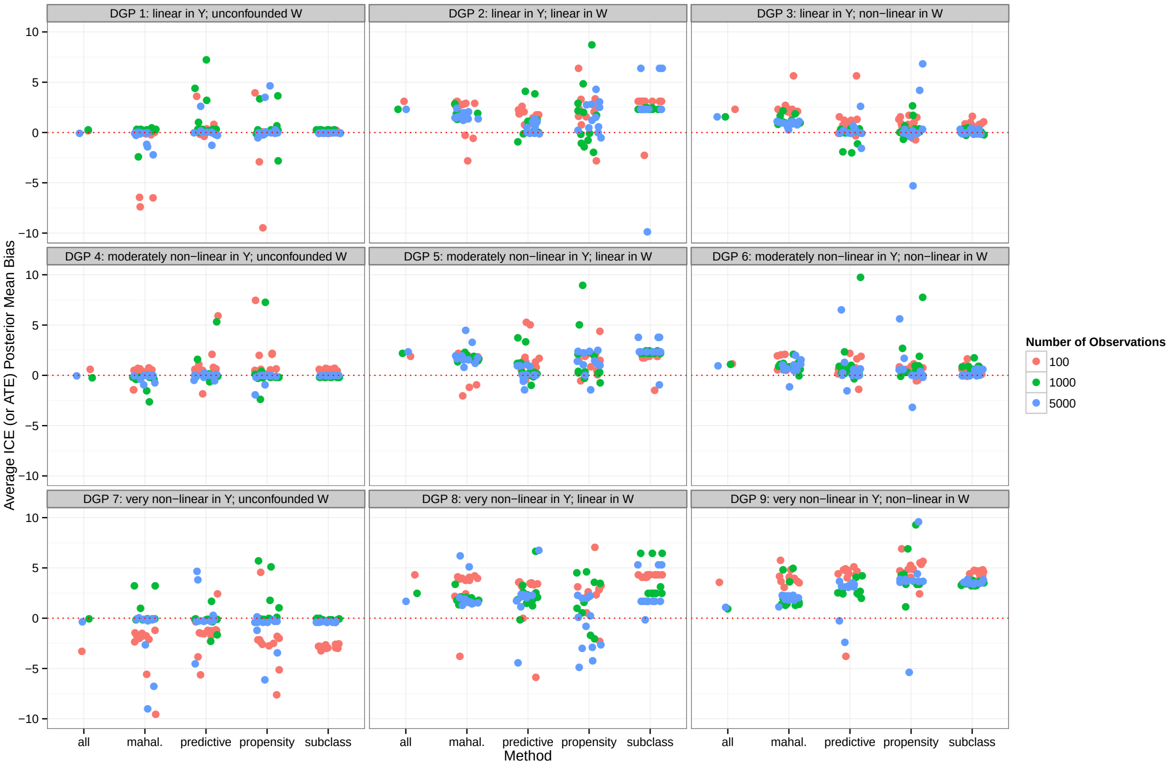

Figure 2.1: Comparing Average ICE (or ATE) Posterior Mean Bias for the Different Matching Methods (continuous outcome)

the predictive mean matching method is the method that is most consistent across specifications and has a relatively low bias.

Figure 2.2 shows the results using the average ICE expected error loss as the performance metric. Recall that this metric is analogous to the traditional root mean squared error and gives a sense of both the “bias” and the (posterior) variance of our estimates. A value closer to zero on this metric indicates a better performing method. One can see clearly that the mahalanobis and predictive mean matching methods almost always outperform the other matching methods. Given that the posterior mean bias was similar across the methods, this suggests that mahalanobis and predictive mean matching generally produce more precise estimates with smaller posterior variance. As expected, larger sample sizes also produce estimates with smaller expected error loss.

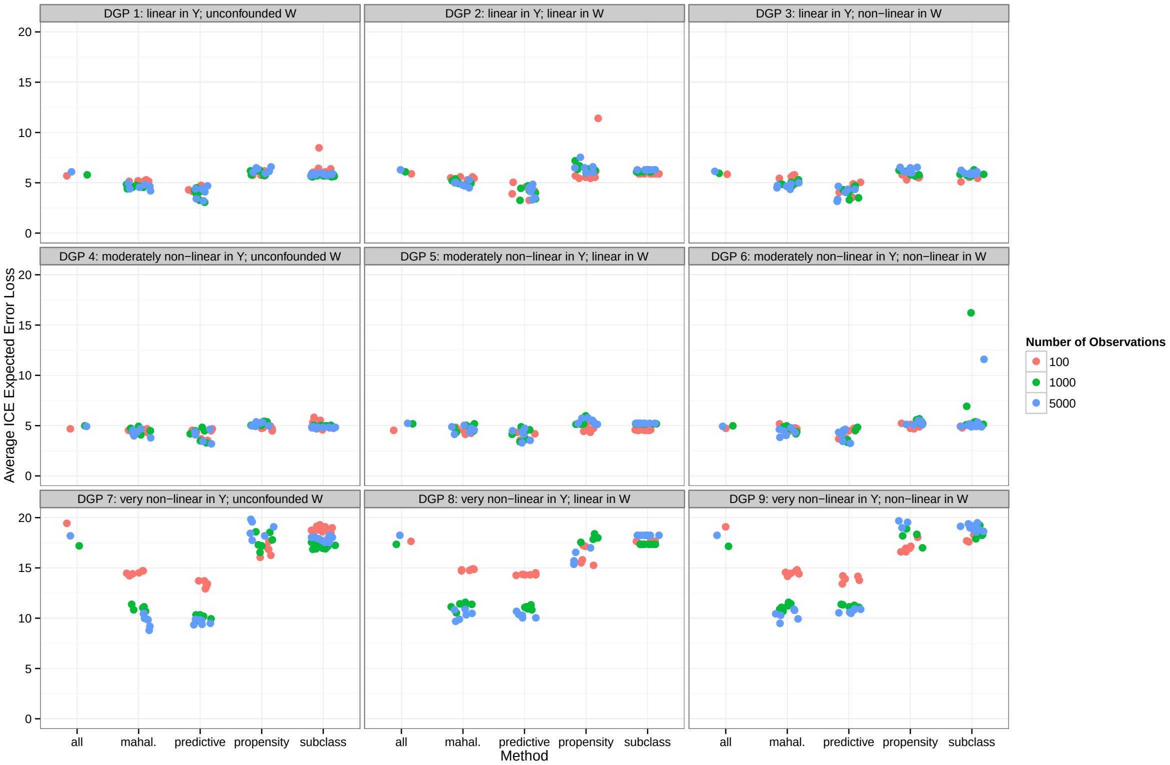

Figure 2.2: Comparing Average ICE Expected Error Loss for the Different Matching Methods (continuous outcome)

With a smaller posterior variance, one should also expect predictive mean matching to perform better on the “power” metric of the proportion of 95% credible intervals not including zero since the credible intervals should be smaller. Figure 2.3 confirms this result where values closer to one on this metric indicate better performance.

In almost all the different data generating processes, the predictive mean matching method performs just as well or better than the other methods. However, one thing to note is that for almost every method and specification, the performance on this metric is quite low. The proportion of credible intervals that does not include zero never exceeds 0.5, despite the fact that all the true τ_i are not equal to zero. This result, although undesirable, is expected since the matching methods use a finite and often small number of donor observations, so the posterior variance on the estimate of τ_i is quite high and the credible intervals are quite large. However, the “power” does improve as the actual τ_i get larger. Recall that in traditional methods, the

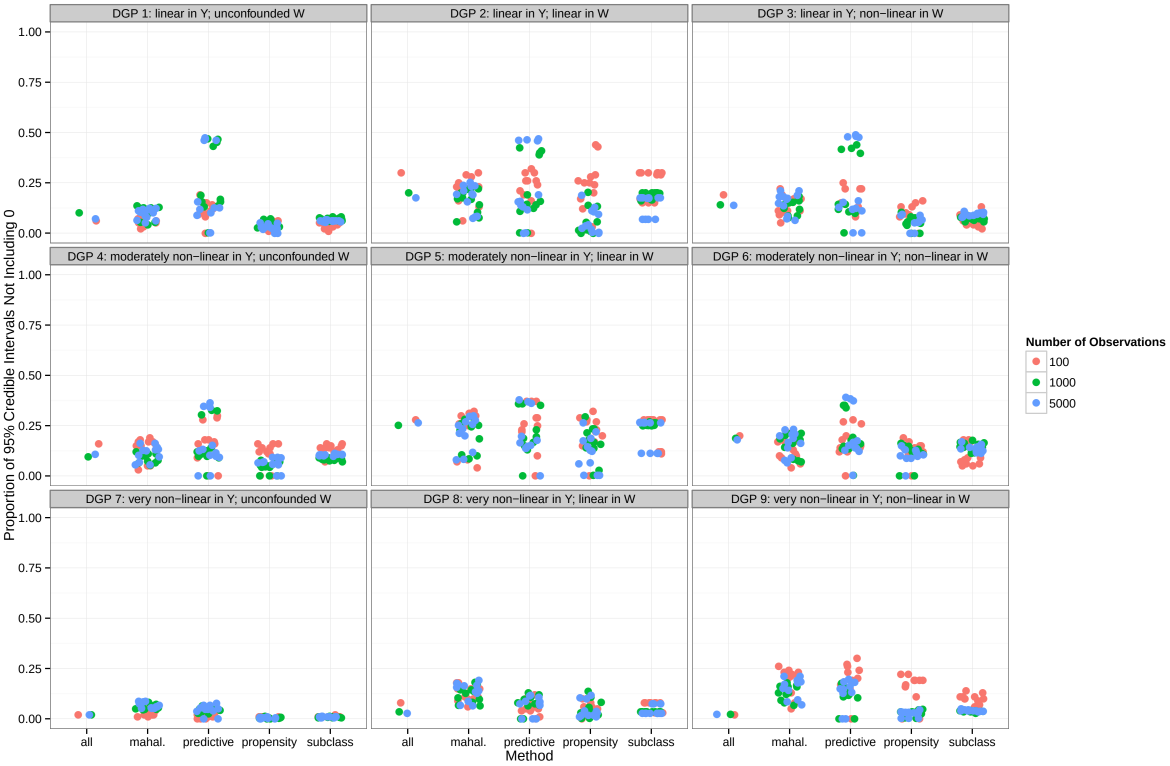

Figure 2.3: Comparing ICE "Power" for the Different Matching Methods (continuous outcome)

power of a method increases as the effect size gets larger. Figure 2.4 shows the proportion of 95% credible intervals including zero as a function of the different τ_i distributions.

When the mean of the τ_i distribution is high (at 20), then the "power" is actually quite high for many of the matching methods. Even with a low mean and a high standard deviation, some of the τ_i will be high and so the "power" increases. Drawing τ_i from the mixture distribution of both large and small effects can also increase power relative to only drawing from smaller effects. Thus, although the matching imputation method that I suggest frequently cannot detect small effects, it can do quite well with larger effects. Also, although I use "power" as one metric of judging the methods, the typical Bayesian model is not as concerned with "power" and hypothesis testing, but instead on whether the credible intervals are properly calibrated and whether the intervals accurately reflect our degree of uncertainty.

Even though my model's 95% credible intervals are quite large, they have the desirable property of being

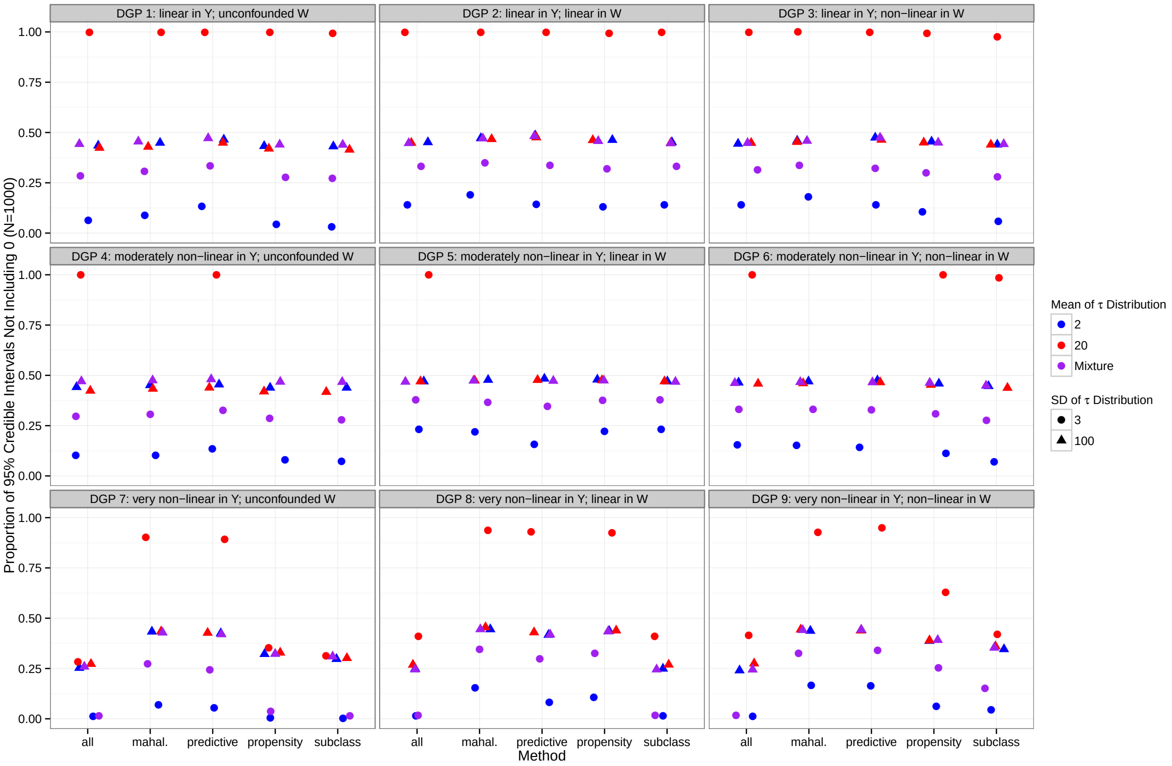

Figure 2.4: Comparing ICE "Power" with Different τ_i Distributions (continuous outcome)

very close to properly calibrated most of the time. Simply put, the large credible intervals have proper "coverage". Figure 2.5 shows this result. Since I am using 95% credible intervals, a method or specification is said to be properly calibrated if the calibration coverage is at 0.95.

Figure 2.5 that most of the specifications are around the 0.95 range. As the data generating process becomes more non-linear, the calibration coverage becomes worse, but it is still usually greater than 0.8. The calibration also improves with larger sample sizes. It does not appear that any particular method performs significantly better or worse on this metric. The results here suggest that the credible intervals from the matching methods give about the correct amount of estimation uncertainty.

The results from the first set of simulations confirm that my model performs as well as one might expected, although not perfectly. Predictive mean matching seems to perform as well or better than any other matching method in recovering the causal estimands of interest. Additional results in the appendix lead to a similar

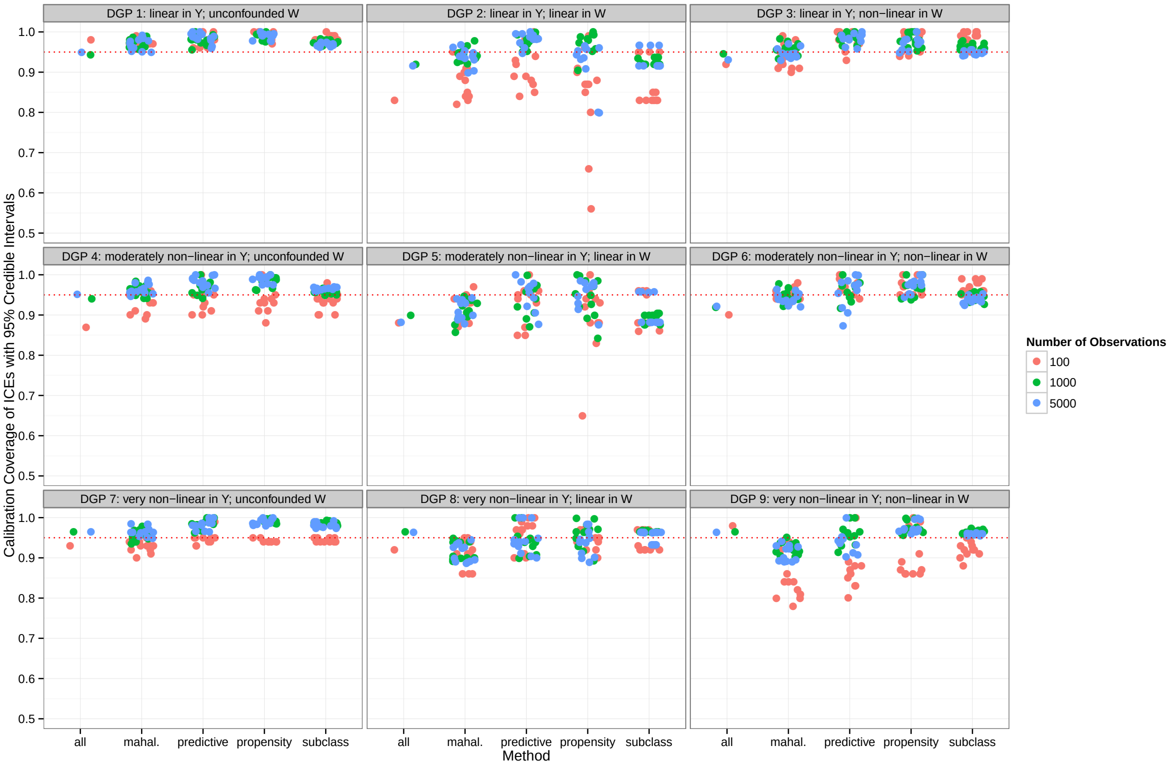

Figure 2.5: Comparing ICE Calibration Coverage for the Different Matching Methods (continuous outcome)

conclusion. While more research is needed into assessing why predictive matching performs better, I can offer at least one possible explanation. Recall that the point of the model and estimating ICEs is to impute the missing potential outcome for each observation. The basic idea of predictive mean matching is to first run a regression using all the data for one treatment group to predict the missing potential outcomes for the other treatment groups. The coefficients from that regression are used to match on the predicted means to form a donor pool for a missing potential outcome. This iterative process actually imputes twice; once to get a rough mean to determine the donor pool and again to actually impute from the donor pool. The objective of predictive mean matching most closely resembles the objective of estimating ICEs in imputing potential outcomes and the two-step iterative process allows for improvements in the imputations. In a sense, I propose a causal framework in my model but use a data mining/machine learning type algorithm in practice. This allows me to achieve optimal results while retaining the principle of modeling the causal process. This may explain why predictive mean matching in my model performs best in practice.

2. Regression imputation works well for estimating average effects. It offers more precise estimates for individual and average effects, but the uncertainty does not accurately reflect the correct uncertainty in estimating the ICEs. Compared to regression imputation, predictive mean matching in my model gives estimates that are almost as good and the uncertainty estimates are correct.

The simplest and most straightforward way to estimate ICEs is by imputing the missing potential outcomes with a regression model. Regression imputation takes the regression model of Y on W and X and imputes using the fitted values from the regression coefficients, simply changing the treatment assignment indicator to the missing one. I compare this simple method of (Bayesian) regression imputation with my Bayesian imputation model using predictive matching, which I showed was the best performing matching method.

Figure 2.6 shows the posterior mean bias of regression imputation versus predictive mean matching. Unsurprisingly, regression imputation performs very well in recovering the ATE, a quantity that it was designed to capture. It consistently gets it right over various specifications. Predictive mean matching performs almost as well, although it is less consistent across various specifications. As the data generating process becomes more non-linear, the functional form for both methods is incorrect and the estimates become less accurate. To confirm that both methods perform about as well in estimating ICEs, I look at "point estimation" for both methods in Figure 2.7. In Figure 2.7, I use the specification with sample size 100 and 7 conditioning variables for both methods and 25 matches for the predictive mean matching. For each method, I take the posterior means for each ICE and take the absolute differences between the posterior means and each true τ_i. This captures how far off each method is for each ICE. I then difference these differences to capture the relative performance of each method for each ICE. Each point on the graph represents the difference in difference for each ICE, so there should be 100 points for each data generating process. A point above the zero line indicates that the "point estimate" for predictive mean matching is closer to the true ICE for that specific ICE and a point below the zero line indicates that the "point estimate" for regression imputation is closer. The red points indicate the median on the difference-in-difference scale. It appears that there is no specific pattern to the distribution of differences-in-differences. Most of the points seem randomly distributed around the zero line, which indicates that for some observations, regression imputation does better and for others, predictive mean matching does better. For the very non-linear generating equations, the differences

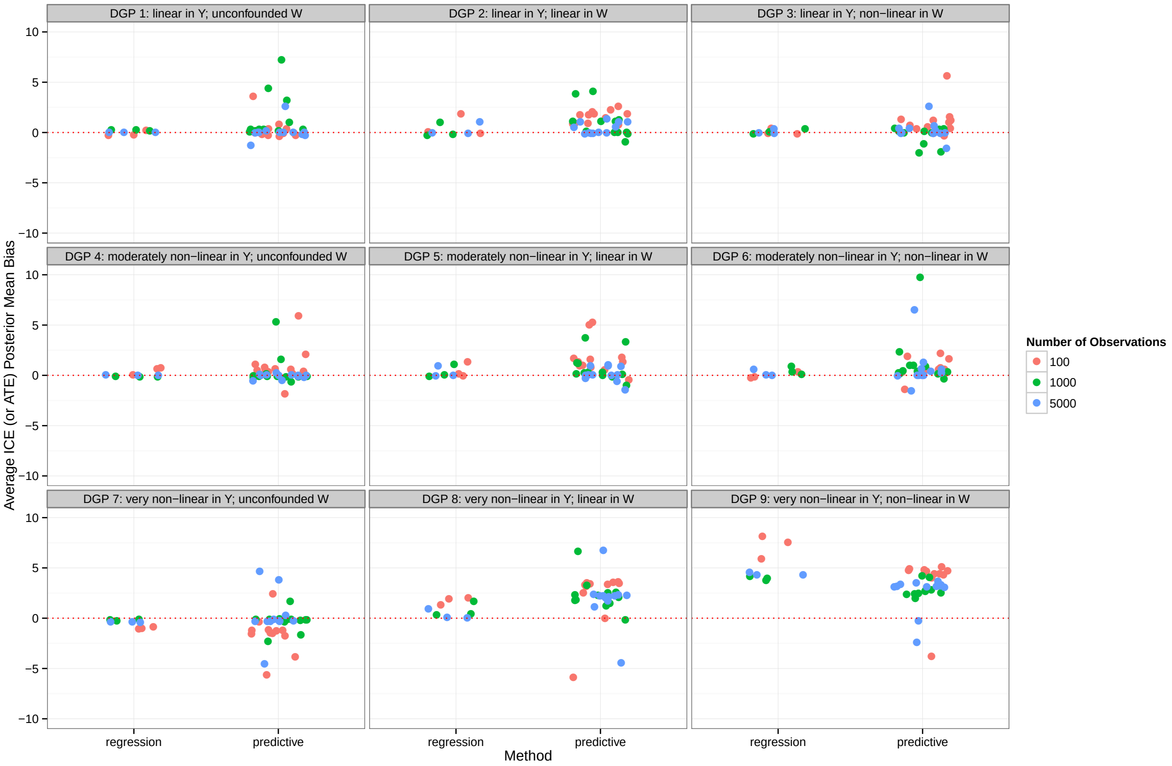

Figure 2.6: Comparing Average ICE (or ATE) Posterior Mean Bias for Regression Imputation versus Predictive Mean Matching Model (continuous outcome)

become more spread out and outliers occur more frequently. Nevertheless, it appears that both regression imputation and predictive mean matching perform similarly on “point estimation” of ICEs.

Although the performance on point estimation is similar for both regression and matching, regression imputation gives posteriors that have smaller variances, as shown in Figure 2.8, which plots the results of average ICE expected error loss. The expected error loss is generally smaller for regression imputation. This is also unsurprising since regression imputation makes an added assumption of modeling only the average. This added assumption allows for more precise estimates and subsequently more “power”, as Figure 2.9 demonstrates. Figure 2.9 shows that regression imputation is able to detect $\tau_i \neq 0$ at a much higher rate than predictive mean matching. By modeling only the average effect and imputing from the model, regression imputation results in much smaller posterior variance. Recall that in regression imputation, the model uses all the observations to model the ATE, which results in a relatively small posterior variance for the ATE.

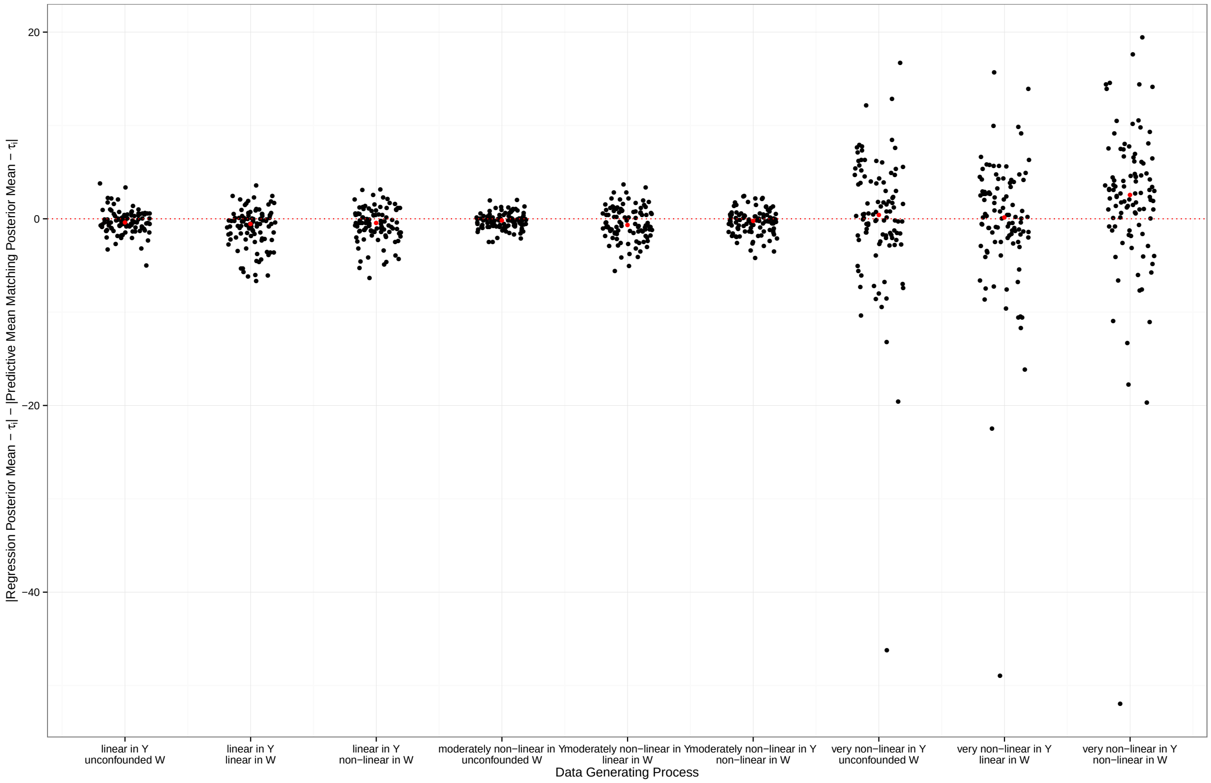

Figure 2.7: Comparing the Absolute Differences Between Posterior Means and the True ICE for Regression Imputation versus Predictive Mean Matching Model (continuous outcome)

This posterior is used directly in the imputation and posterior for each ICE, so the width of the posterior credible interval for the ICE is the same as the width of the credible interval for the ATE. Contrast this with my imputation model with predictive mean matching, where the width of the credible interval for an ICE is derived from matching on a smaller set of donor pool observations. It is straightforward to see that regression imputation results in smaller credible intervals, which in turn decreases the probability of zero appearing in the credible interval and thus more “power”.

Given that regression imputation produces estimates that are just as “correct” as my imputation model with predictive mean matching with smaller credible intervals and more power, why would one not use regression imputation for estimating ICEs? It turns out that the credible intervals are actually too small, which is unsurprising since they are credible intervals designed for the ATE rather than ICEs. Regression

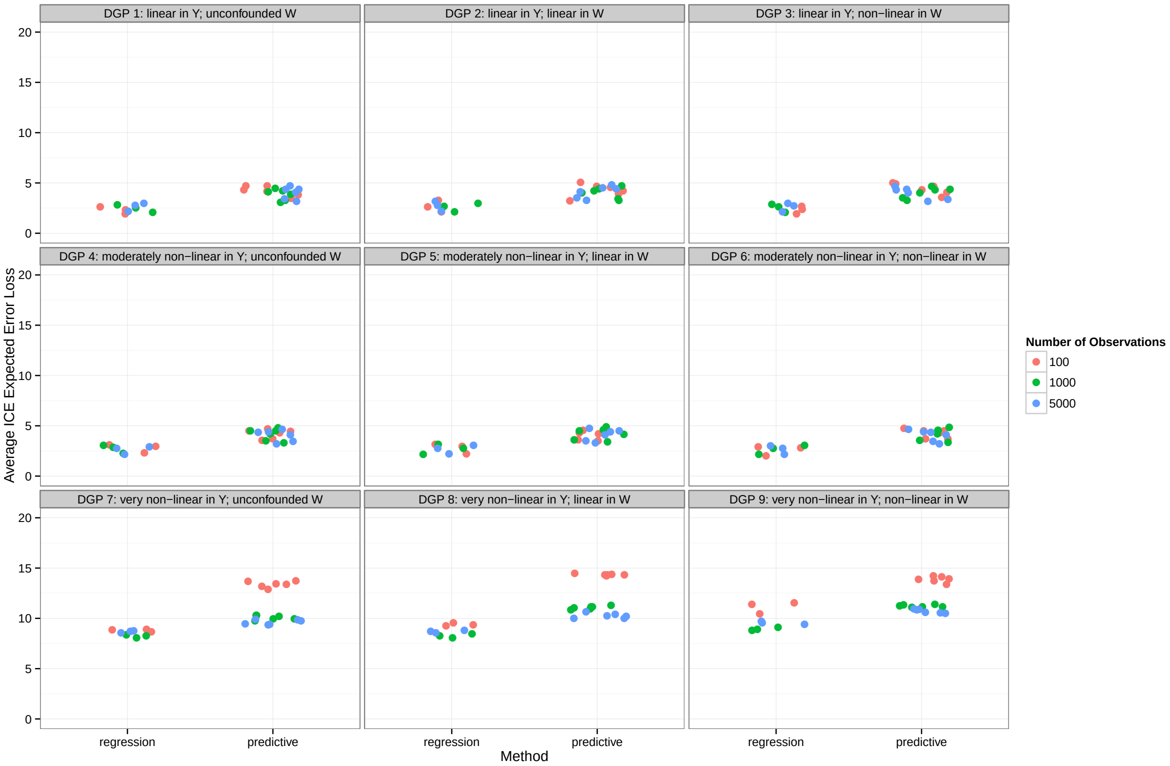

Figure 2.8: Comparing Average ICE Expected Error Loss for Regression Imputation versus Predictive Mean Matching Model (continuous outcome)

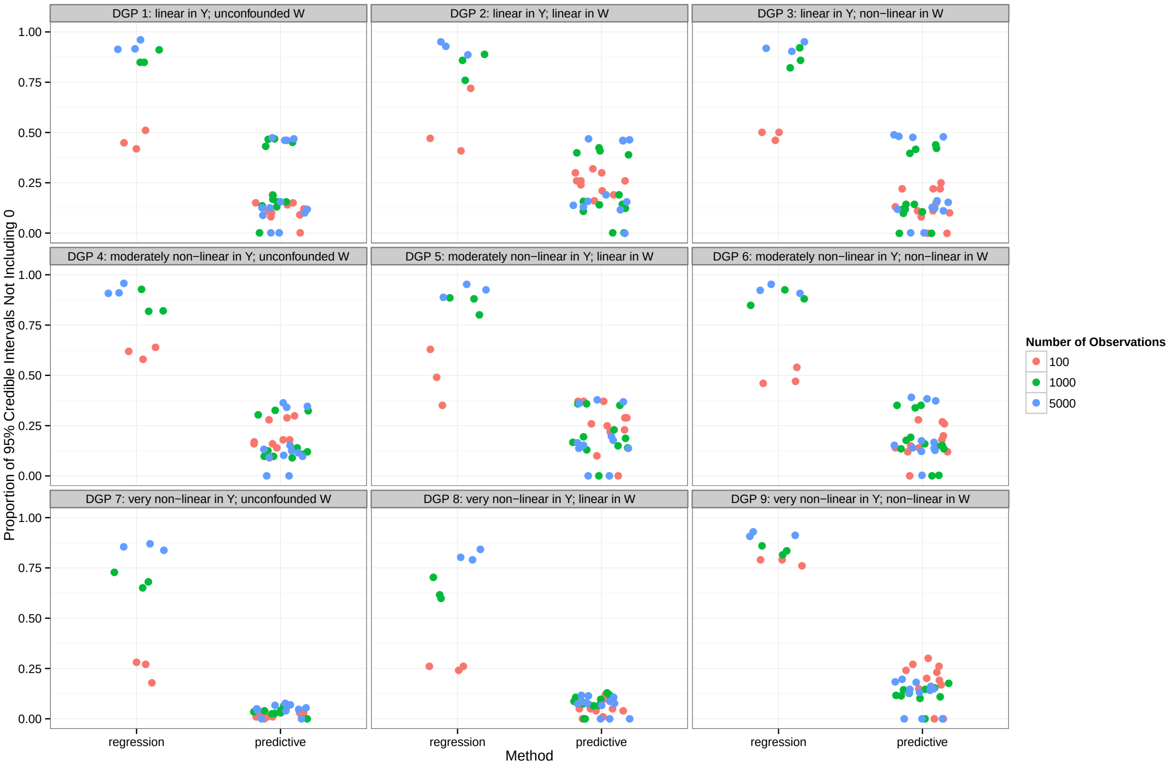

Figure 2.9: Comparing ICE "Power" for Regression Imputation versus Predictive Mean Matching Model (continuous outcome)

imputation is very poorly calibrated, and the uncertainty that is reflected by the posterior variance is incorrect for ICEs. Figure 2.10 shows the results of the calibration coverage for the 95% credible intervals. While approximately 95% of the 95% credible intervals cover the true τ_i for predictive mean matching, most of the time less than 50% of the 95% credible intervals do so for regression imputation. The smaller credible intervals lead to incorrect inferences more than half the time.

Figure 2.11 shows the different calibration coverages when drawing τ_i from different distributions. While predictive mean matching is accurately calibrated regardless of the distribution of the true τ_i, regression imputation is also very poorly calibrated regardless of the distribution of the true τ_i. There appears to be a general pattern that the calibration for regression imputation is better when the τ_i are more spread out (higher standard deviation of the τ_i distribution). One possible explanation is that when the τ_i are more spread, the posterior variance for the ATE is larger and so the credible intervals for the ICE are also

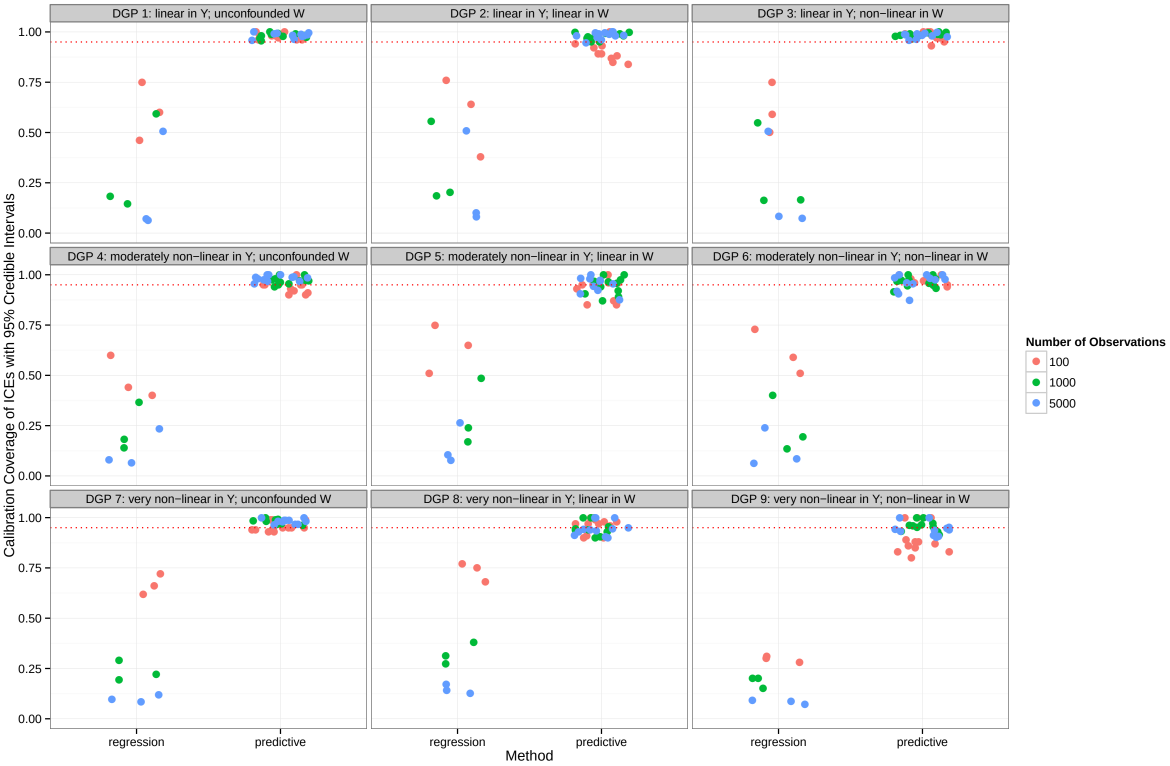

Figure 2.10: Comparing ICE Calibration Coverage for Regression Imputation versus Predictive Mean Matching Model (continuous outcome)

larger and thus will include the true τ_i a greater proportion of the time. Nevertheless, it is clear that while regression imputation does just as well as my matching imputation model in point estimation of the ICEs, it is a poor technique for estimating the uncertainty of the ICEs and should be used only for modeling averages rather than individual effects. The typical method of imputing from a regression model is incorrect when looking at individuals.

3. There appears to be no discernible difference in the number of X variables to condition on as long you condition on all (or almost all) confounders.

Although more simulations are needed to fully test the effect of omitting or including conditioning variables, it appears that as long as one conditions on all or close to all of the confounders, adding extra prognostic

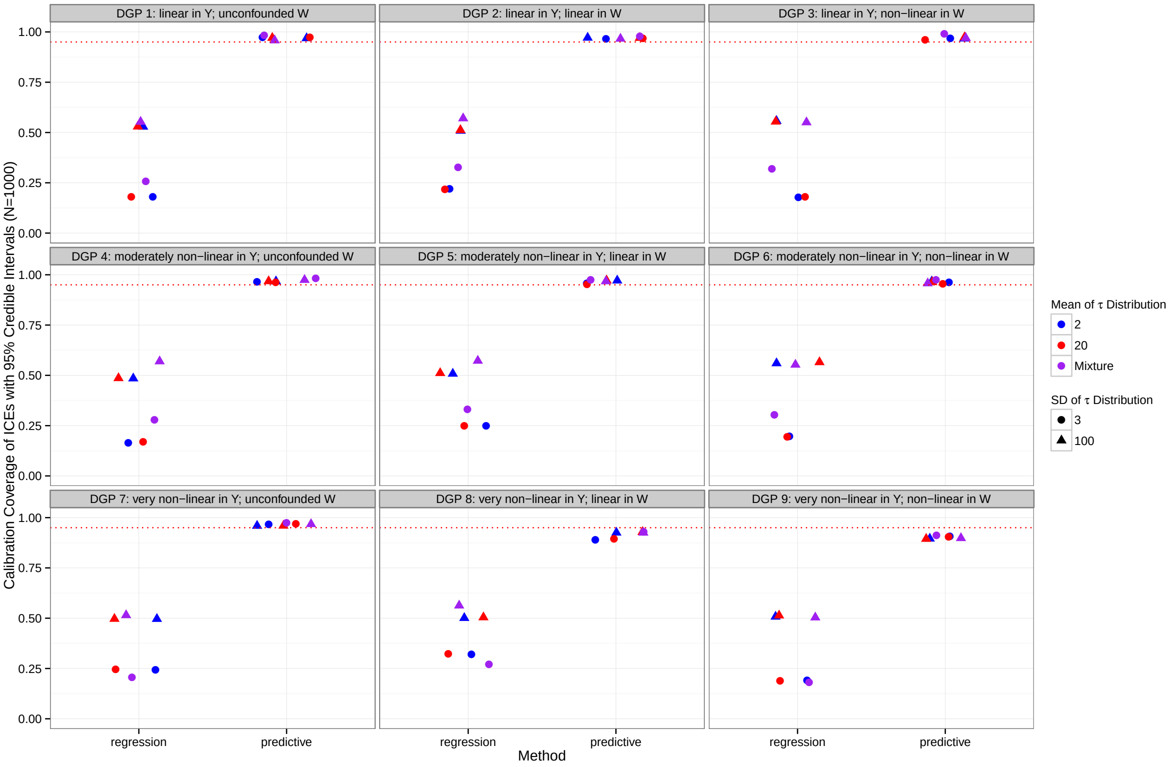

Figure 2.11: Comparing ICE Calibration Coverage with Different τ_i Distributions (continuous outcome)

variables to the conditioning set does not result in drastic improvements. Figure 2.12 shows the results of average ICE expected error loss across all the matching methods with 0, 5, 7, and ten conditioning variables. Recall that for the specifications with confounded treatment assignment, 7 is the correct number of confounders. Conditioning on ten X variables means conditioning on all the confounders and all the prognostic variables. The results suggest that there are no discernible differences in performance when conditioning on 5, 7, or 10 confounders across all the different data generating processes. This suggests that as long as one controls for approximately the correct confounders, the results should be quite stable. However, I do not test the effect of omitting very important versus less important confounders or including or excluding very important prognostic variables. Future research should look into these questions in more detail.

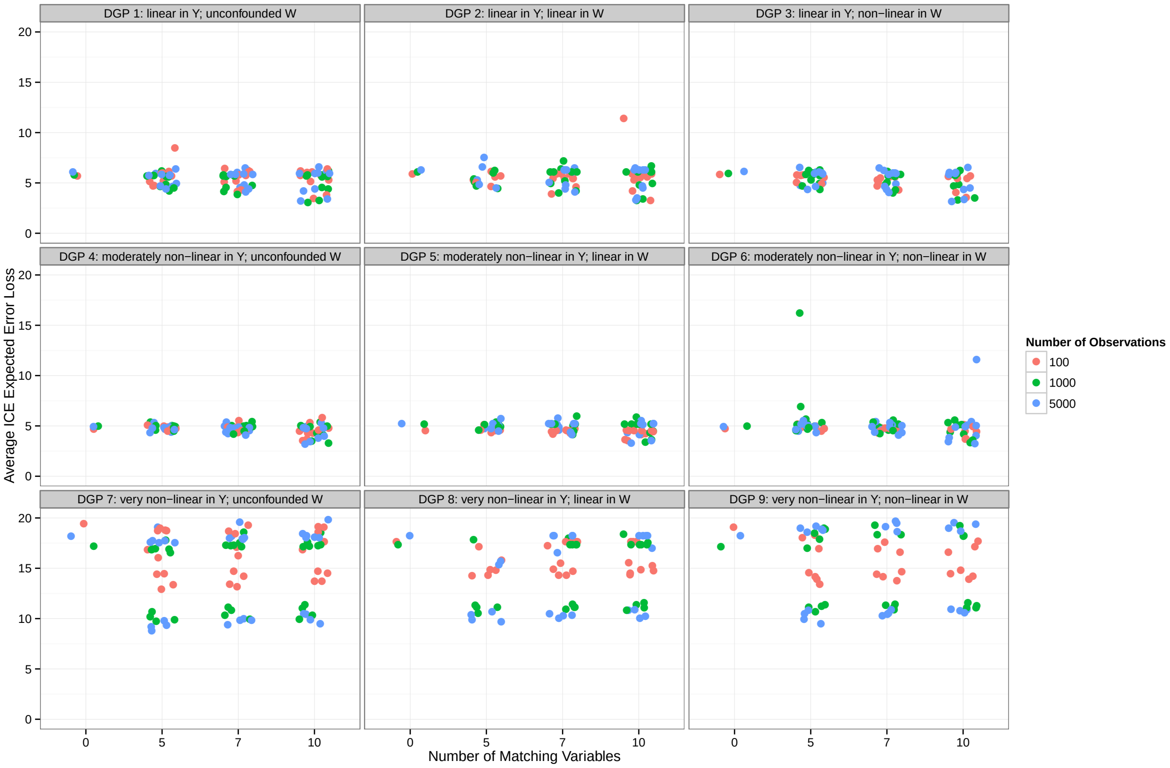

Figure 2.12: Comparing Average ICE Expected Error Loss for Different Conditioning Sets (continuous outcome)

4. The optimal number of matches to use is dependent on the data generating process, although one should not use a very small number or a very large number of matches. A random number of matches does not seem to provide a huge improvement compared to a fixed number of matches.

In typical matching analyses, there is a bias-variance tradeoff between using too few versus too many matches. When using a small number of matches, bias is small since only high quality matches are used, but variance is large with such a small donor pool. When using a large number of matches, variance is smaller but lower quality matches are included in the donor pool, which may increase bias. There is a slightly different story when using matching to estimate ICEs in my model. Figure 2.13 shows the posterior mean bias from using different sizes of donor pools across the different matching specifications.[19] Recall that in my specifications,

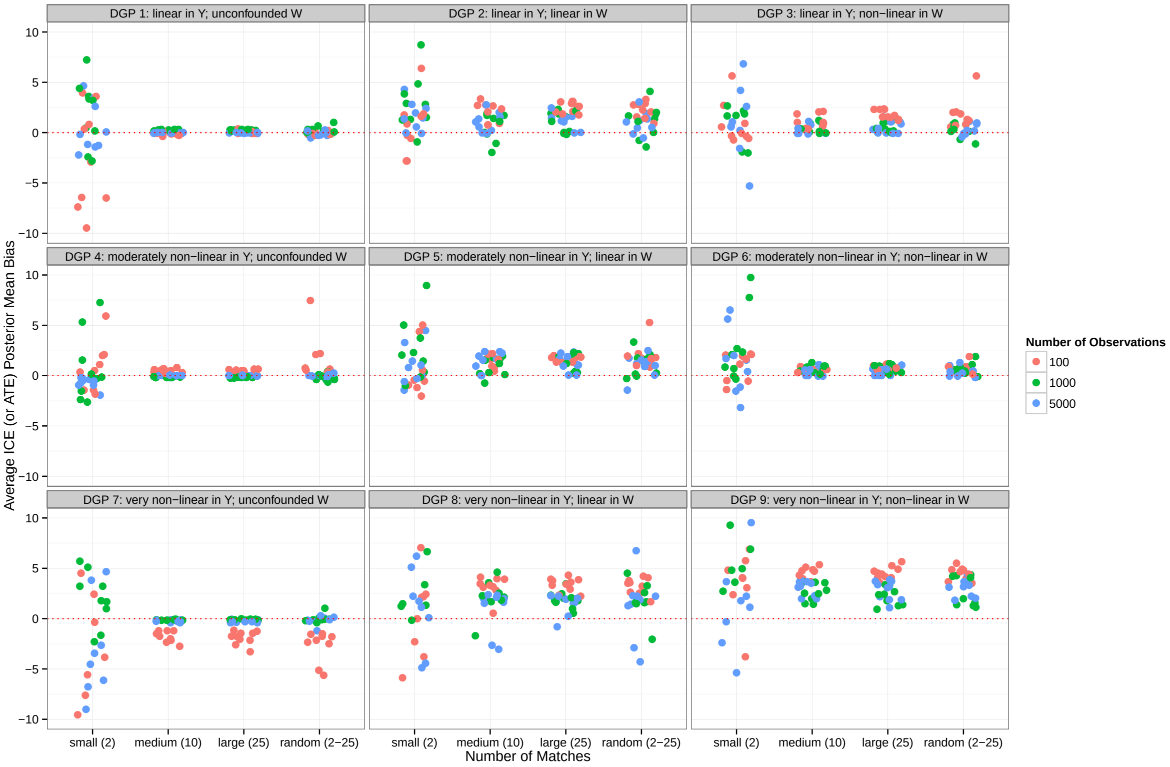

Figure 2.13: Comparing Average ICE (or ATE) Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

a small number of matches is 2, medium is 10, large is 25, and random is a randomly drawn integer between 2 and 25 for each iteration of the algorithm. For a very small number of matches, the posterior mean bias is quite unstable across various specifications. Using a medium or large number of matches seems to give better and more consistent results. There seems to be no benefit to using a random versus fixed number of matches. Figure 2.14 shows the results of average ICE expected error loss, which takes into account posterior variance. When using only two matches, there is a large error, which represents both poor "point estimates" and large posterior variance. Using a slightly larger number of matches shrinks the expected error significantly. Also, using a random number of matches increases the variance of the results without a large increase in posterior mean bias.

[19] Each point represents a different specification of matching method and number of conditioning variables. The subclassification method is not included in these results.

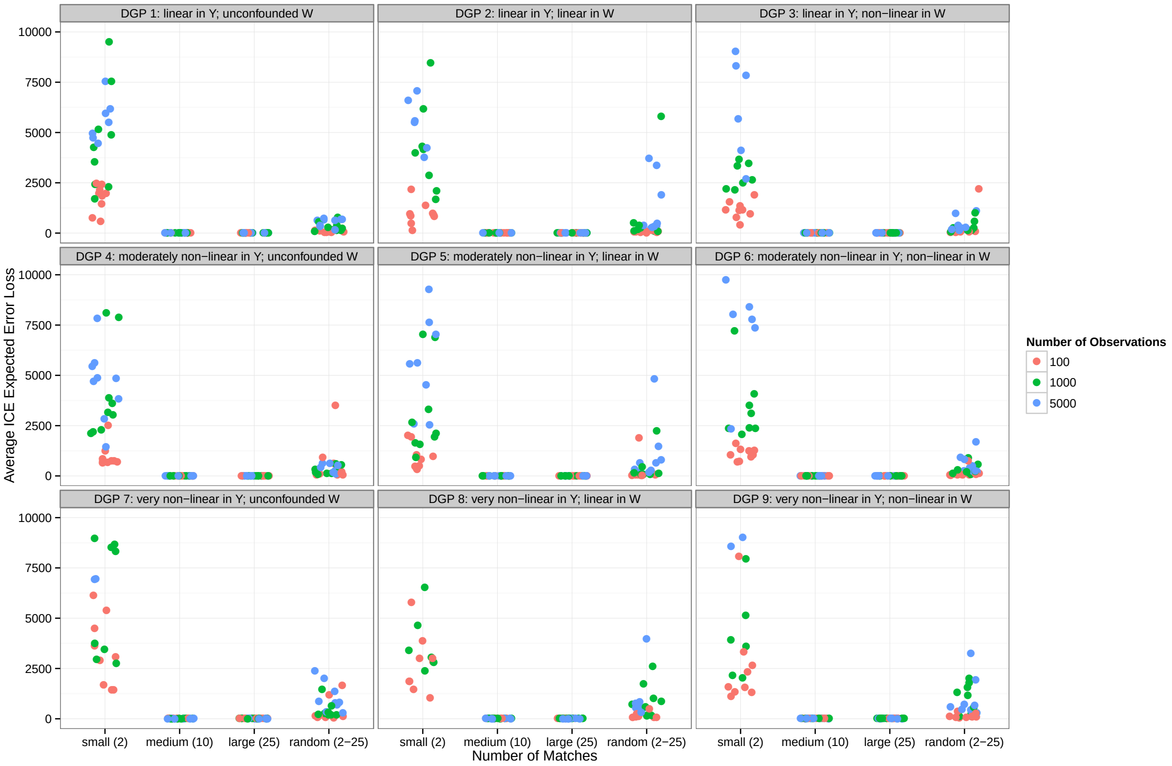

Figure 2.14: Comparing Average ICE Expected Error Loss for Different Numbers of Matches (continuous outcome)

While one can look at the previous results and conclude that larger numbers of matches are better, using 25 matches is large for a sample size of 100 but quite small for a sample size of 5000. I further test the idea of optimal number of matches by looking at the number of matches as a percentage of the number of observations in the smaller treatment group. Figure 2.15 shows the posterior mean bias for the different match percentages using a specification with predictive mean matching on 7 confounders with sample size of 1000. As the match percentage (or equivalently the number of matches) increases, the posterior mean bias also tends to increase, which suggests that larger donor pools are incorporating poorer quality matches and inducing “bias”. There does not appear to be an optimal match percentage for all data generating processes, although I suggest that 10% of the smaller treatment arm seems to be a good number to use that consistently gives decent results. The results on other metrics (in the appendix) also confirm that there is no optimal number and 10% seems to work well.

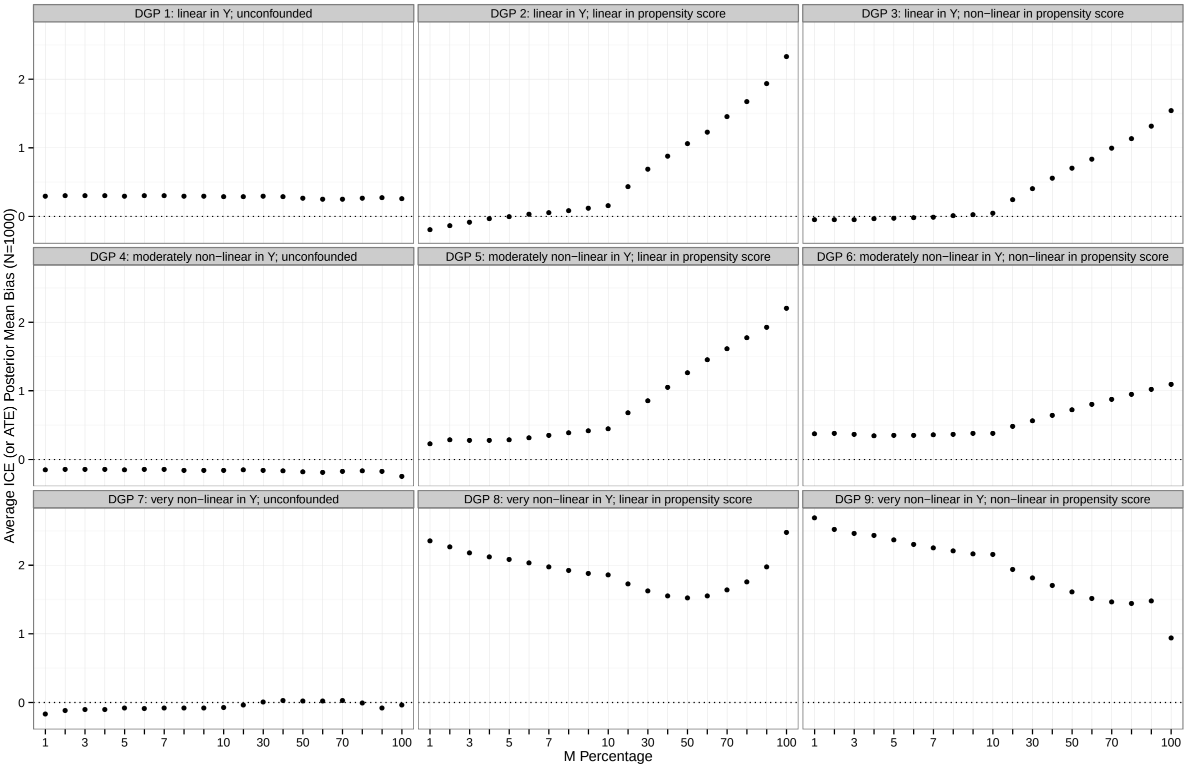

Figure 2.15: Comparing Average ICE (or ATE) Posterior Mean Bias for Different Match Percentages (continuous outcome)

2.4 Conclusion

The simulation results I have presented here and in the appendix are only the tip of the iceberg for testing my model and the different specifications. I have tried to test my model and compared it to imputation from regression, which is the simplest and most widely used way to estimate and predict individual effects. I conclude that predictive mean matching performs the best out of the matching methods I propose. I also show that both regression imputation and predictive mean matching do fairly well in “point estimation” of the ICEs, but regression imputation gives uncertainty estimates that are wildly incorrect whereas my model is properly calibrated. For practical use, I suggest using predictive mean matching with a fixed donor pool size of approximately 10% of the smaller treatment arm, conditioning on all observed confounders.

Chapter 3

Estimating ICEs in Two Applications

I now apply the estimation framework and estimate ICEs in two applications from political science and economics. The first application revisits a field experiment from Olken (2007) on the effects of different forms of corruption monitoring on actual corruption. The second application looks at the effects of a national job training program known as JobCorps, using data from a randomized study known as the National Job Corps Study conducted by Mathematica Policy Research, Inc. I follow a similar approach and use the same dataset found in Frumento et al. (2012). The two applications are interesting for estimating ICEs for various reasons. Both are very important substantively and address issues of interest to many scholars. The corruption monitoring study is a unique and interesting field experiment that has made a substantial contribution to the study of corruption. The question of the effect of job training on employment outcomes is perhaps the most widely studied area by economists and statisticians interested in causal inference and program evaluation. In addition, the data available for both applications provide an opportunity to demonstrate the flexibility of the estimation framework and the different ways in which estimating ICEs can increase knowledge and discovery. They incorporate ICE estimation with both binary and continuous dependent variables, binary and continuous treatment variables, single-stage and two-stage estimation, and somewhat randomized and non-randomized treatment assignment settings.

3.1 Estimating ICEs: A Review

3.1.1 Framework

Recall that the main idea for estimating the individual causal effects is to estimate or impute the missing potential outcome for each observation. Knowing the missing potential outcome allows us to directly calculate the ICE and any other causal estimand. I use the combination of matching and a Bayesian model to get point estimates and uncertainty intervals for the missing potential outcome.

Under the treatment assignment ignorability and SUTVA assumptions, the distribution of potential outcomes is identical for observations with the exact same values on the observed covariates. This implies that the distribution of the missing potential outcome for observation i can be approximated with the observed potential outcomes for a set of donor observations with the opposite treatment assignment. Since exact matching is only possible in large samples with discrete covariates, I use predictive mean matching (as described previously) to find donor pools of matches that are similar on the covariate values. To derive the posterior for the ICEs, I incorporate the matching step in a Bayesian model. The Bayesian model captures the uncertainty in the matching process, the donor pool, the parameters of the distributions of missing potential outcomes, and the imputations themselves through the joint posterior.

3.1.2 Estimation

The general algorithm for estimating ICEs is as follows:

For a binary treatment and continuous outcome variable, I simulate from the posterior through the following steps. Let observation i be a treated (control) observation. Choosing m-to-1 predictive mean matching with m approximately equal to 10% of the smaller treatment arm, I first estimate the parameters of the predictive mean matching $\theta_{\mathcal{M}}$ with a draw $\tilde{\beta}_c$ ($\tilde{\beta}_t$) from the posterior of a Bayesian linear regression of Y_c on X_c (Y_t on X_t). I then calculate a predictive mean score for observation i as $X_i\tilde{\beta}_c$ ($X_i\tilde{\beta}_t$) and also calculate a predictive mean score for all control (treated) observations j as $X_j\tilde{\beta}_c$ ($X_j\tilde{\beta}_t$). I then find the m control (treated) observations with the closest predictive mean score to i and designate them as the donor

MCMC Algorithm for the Posterior of τ_i

```
Repeat the following $n_{sim}$ times:

Gibbs Sampler:
    1. Draw a matching procedure $\tilde{\mathcal{M}}$ from $p(\mathcal{M})$.
    2. Draw $\tilde{\theta}_{\mathcal{M}}$ from $p(\theta_{\mathcal{M}}|Y, X, W, D, \theta^{mis}, \mathcal{M})$.

for ($i$ in 1:$N$){
    3. Determine $\tilde{D}^{(i)}$ from matching procedure. (matching step)
    4. Draw $\tilde{\theta}_i^{mis}$ to estimate $\theta_i^{mis}$.
}

Draw from PPD and Calculate $\tau_i$:
for ($i$ in 1:$N$){
    5. Draw $\tilde{Y}_i^{mis}$ from $f(\cdot|\tilde{\theta}_i^{mis})$. (imputation step)
    6. Calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$.
}
```

observations. I then draw $\tilde{\theta}_i^{mis}$ by modeling the donor pool with a Normal likelihood and Normal prior for a model with mean and variance unknown. Using $\tilde{\theta}_i^{mis}$, I draw an imputation of the missing potential outcome $\tilde{Y}_i^{mis}$ from a Normal distribution and then calculate $\tilde{\tau}_i = W_i(Y_i - \tilde{Y}_i^{mis}) + (1 - W_i)(\tilde{Y}_i^{mis} - Y_i)$. I repeat this process for all observations i for $n_{sim} = 2000$ iterations with a burn-in length of 100.

3.2 Application 1: Monitoring Corruption

3.2.1 The Setup and Data

The first application of estimating ICEs comes from a study conducted in Olken (2007) on the effectiveness of corruption monitoring.[1] Corruption is an important topic in both the economics and political science literature, and various ways to combat corruption have been suggested. The Olken study is unique in that it is a randomized field experiment that tested the effectiveness of two types of corruption monitoring in Indonesian villages: top-down monitoring and grassroots bottom-up monitoring. Olken concluded that top-

[1] I obtained the data from the study from Olken's website at `http://economics.mit.edu/faculty/bolken/data`

down monitoring is effective in reducing corruption while bottom-up monitoring had little impact. The study is a good example to demonstrate the use of my model because it is a relatively straightforward study that may have heterogenous treatment effects and it also collected data on multiple levels, which I use to demonstrate the flexibility of using the ICE framework.

The setting of the project is 608 villages in the Indonesian provinces of East Java and Central Java between September 2003 and August 2004.[2] Through a national Indonesian government program (Kecamatan Development Project) funded from the World Bank, each village proposes a usually infrastructure related project and is usually given some money for it. The most common type of infrastructure project is a project to surface an existing dirt road with a surface made of sand, rocks, and gravel. The study is limited to villages with such projects.

In order to ensure the proper use of funds, there are various monitoring mechanisms. Each project is associated with a series of approximately three village-level accountability meetings. In the beginning, only 40 percent of the funds are released to the implementation team. At the first village accountability meeting, the implementation team must present an accountability report explaining how the funds were used. Only after the meeting has approved the report would the other 60 percent of the funds be released. These meetings are open to the public but are typically attended by only 30-50 people, most of whom are members of the village elite.

A second accountability mechanism is the threat of an audit by an independent government development audit agency known as the BPKP. Each project has approximately a 4 percent baseline chance of an audit from the BPKP. The audit process involves auditors checking all financial records and inspecting physical infrastructure. Corruption findings from the audit can lead to officials forcibly returning the money publicly or even criminal action.

In the experimental design for the study, Olken was able to randomize the two types of corruption monitoring. Broadly speaking, the experiment consisted of four treatment conditions: audit, participation either with invitations only or invitations plus comment form, or control. The audit and participation treatments were randomized independently, so a village can possibly receive both an audit treatment and a participation

[2] The following description of the study is mostly taken from Olken (2007).

treatment.

- **audit treatment**: The audit treatment is a "top-down" mechanism in which an outside entity (in this case the BPKP) monitors the project for signs of corruption. For the audit treatment, villages were cluster randomized at the subdistrict level to ameliorate spillover effects (all villages within a subdistrict either received an audit treatment or not). The randomization was also stratified or blocked by district and number of years the subdistrict had participated in the program. The audit treatment consisted of increasing the probability of an audit by BPKP from 4 percent to 100 percent. Villages were informed before planning for construction that they would be audited with probability 1 either during or after construction. They were also told that the results of the audit would be presented at a village meeting, so village officials faced a possibility of punishment by the villagers, possible cutoff of funding from future KDP projects, or even criminal action. Of the 608 villages in the study, 283 received the audit treatment and 325 did not.

- **participation treatments**: The participation treatments are intended to be grassroots mechanisms in which local villagers themselves are an integral part of the corruption monitoring. The idea of the participation treatments is to increase village attendance at the village-level accountability meetings, which are open to the public but usually dominated by the village elite. Randomization of the participation treatments was done at the village level, and each village either got the intervention of invitations, invitations and comments, or control. In the invitations intervention, either 300 or 500 invitations were distributed throughout the village prior to each of the three accountability meetings. The invitations were distributed either by sending them home with school children or by asking the heads of hamlets and neighborhood associations to distribute them. The distribution method and number of invitations were also randomized by village. In the invitations and comments intervention, villages received the invitations exactly as the invitations intervention, but in addition to the invitations, there was a comment form asking for villagers' opinions of the road project. The comment forms are anonymous and summarized by a project enumerator at each accountability meeting. Thus, the comment form produced an additional anonymous avenue through which villagers can monitor corruption without fear of retribution from village leaders. Of the 608 villages in the study, 105 received the invitations intervention, 106 received the invitations and comments intervention, and 114 did not receive a participation intervention.[3]

The corruption and misuse of funds for the projects usually came in the form of either collusion with suppliers to inflate prices or quantities of supplies used or inflated labor costs. Olken and his team measured corruption by doing an independent assessment of the "correct" costs of the project through sampling the materials used in the roads and surveys with suppliers and workers. The difference between this independent assessment and the actual costs of the project is an unbiased measure (with high error) of the corruption. For each village, Olken defined the dependent variable as the log of the reported amount minus the log of the independent assessment amount, which is approximately the *percent expenditure missing*.[4] He reports several different measures of the percent missing variable:

- Percent missing for major items in road project: sand, rocks, gravel, and unskilled labor
- Percent missing for major items in roads and ancillary projects
- Percent missing for materials in road project
- Percent missing for unskilled labor in road project

I consider all four of these continuous measures of corruption in my analyses.

Due to circumstances such as missing data, attrition, or audit treatment randomization at the subdistrict level, the treatment assignment in the complete dataset may not be as clean as one would like. Fortunately, Olken also collected a few background covariates at the village level to allow for possible covariate adjustment. The covariates measured include

- Distance to subdistrict
- Education of village head

[3] From here on out, I refer to the invitations treatment as "invites" and the invitations and comments treatment simply as "comments".

[4] Due to the noisiness of both the reported amount spent and the independent assessments, the estimates of percent missing are sometimes negative or greater than 1. Such values do not make sense in the context of percent missing so I consider the variable as simply a continuous measure of corruption.

- Age of village head
- Salary of village head
- Percent of households that are poor
- Village population
- Mosques per 1,000 population
- Mountainous village dummy
- Total village budget
- Number of subprojects

In addition, Olken also collected data on the village-level accountability meetings including attendance levels. I first replicate the results from Olken's initial analyses using ICEs. I then demonstrate the flexibility of the model in estimating other quantities of interest and with different treatment variables and outcome variables.

3.2.2 The Effect of Monitoring Treatments on Corruption (binary treatments and continuous outcomes)

Olken's main result in the paper is that the audit treatments on average reduce corruption by about 8 or 9 percentage points while the two participation treatments have no consistent statistically significant effect on corruption. The main specification that he uses is a linear regression of the following form:

$$\begin{aligned}\text{PercentMissing}_{ijk} &= \alpha_1 + \alpha_2\; I(\text{Audit})_{jk} + \alpha_3\; I(\text{Invites})_{ijk} \\ &\quad + \alpha_4\; I(\text{Comments})_{ijk} + \epsilon_{ijk}\end{aligned}$$

where i indexes a village, j is a subdistrict, k is a stratum for the audits, and $I(\cdot)$ are indicator variables for whether a village got a specific treatment. The coefficients α_2, α_3, and α_4 are the average treatment effects for the three treatments respectively. Due to the form of the linear regression, Olken's estimated effect for

Table 3.1: Treatment and Control Groups for Average Treatment Effects using ICEs and the Corresponding Regression Parameters from Olken

Treatment Group	Control Group	Olken Parameter
audit; no participation	no audit; no participation	α_2
audit; invites	no audit; invites	α_2
audit; comments	no audit; comments	α_2
invites; no audit	no invites; no audit	α_3
invites; audit	no invites; audit	α_3
comments; no audit	no comments; no audit	α_4
comments; audit	no comments; audit	α_4

one treatment averages over the distribution of the other treatments in the sample. Specifically, α_1 assumes that the treatment effect of getting the audit treatment versus no audit treatment is the same regardless of whether the village got a participation treatment or not. This assumption may be violated for example, if the effectiveness of an audit is smaller with the presence of a participation treatment as well.

I first demonstrate the flexibility and comparability of estimating ICEs by comparing aggregated average effects from ICEs versus the specification found in Olken. Using the same linear specification above, I run a Bayesian linear regression with improper uniform priors to get the same results as Olken. I then run the ICE algorithm using predictive mean matching on the 10 covariates to get ICEs. To get the various average treatment effects, I simply aggregate the ICEs. There are two important differences between my approach and the original Olken approach. First, I use the covariate adjustment to deal with the less than perfect randomization, which Olken does not include in his specification. Second, I carefully define treatment and control groups to estimate the treatment effects and I allow for treatment effects to differ depending on the presence or absence of other treatment conditions.[5]

Table 3.1 shows the different treatment and control groups for the seven average treatment effects estimated using ICEs and the corresponding parameters from the linear regression. The rows represent the seven possible interactions between the different treatments. Since Olken did not include interaction terms in his initial model, he constrains the seven possible treatment effect interactions to three treatment effect parameters.

[5] This is equivalent to a linear regression specification with interaction terms between the treatments.

Figure 3.1 compares the results of both the average treatment effects calculated from the estimated ICEs and the average treatment effects estimated from the regression model for the four different measures of corruption. The red lines indicate the point estimates and 95% credible intervals from the regression method. Note that for each graph, the regression method only produces three distinct estimates corresponding to α_2, α_3, and α_4. The results shown in the graph suggest a few conclusions.

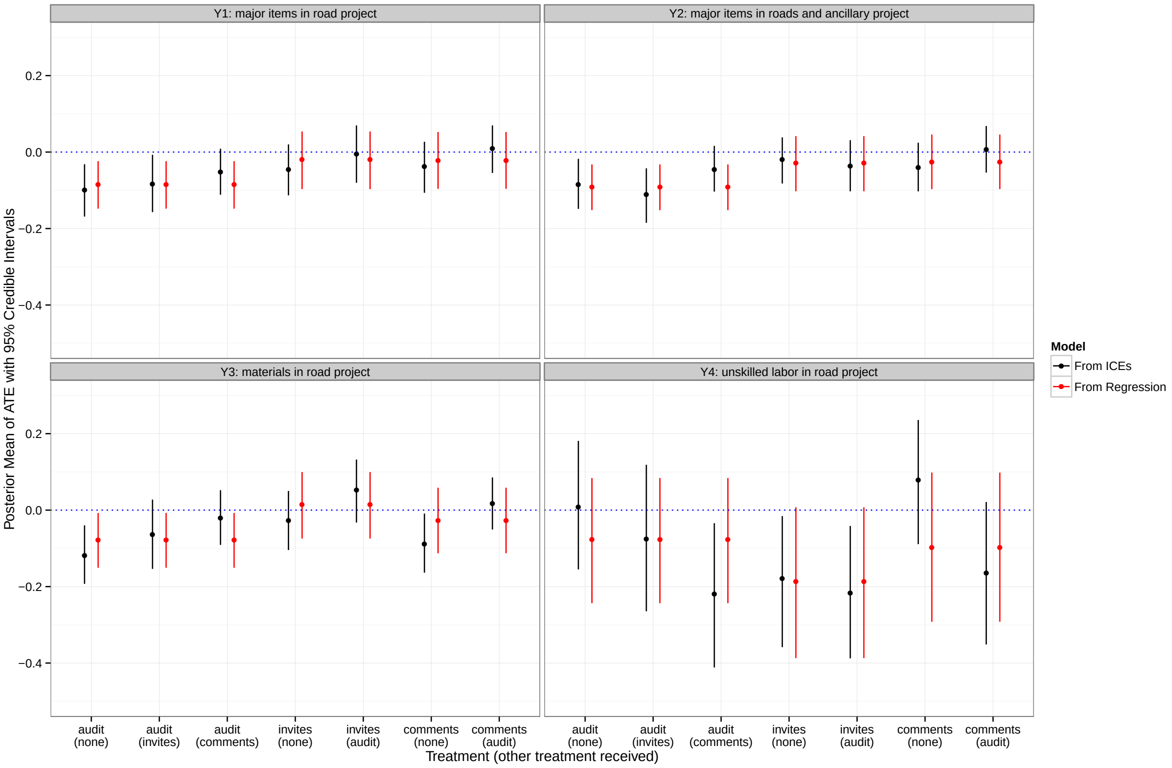

Figure 3.1: Comparing ICE Average Treatment Effects to Regression

- The treatment effects estimated from the ICEs are relatively close to the ones estimated from the regression method. This likely suggests that the ICE method of aggregating for average effects can recover the same estimates as the regression method, which is known to have good properties given certain assumptions. What this suggests is that the ICE model is giving reasonable answers that are similar to other tried and true methods. The slight differences between the two models are likely due to conditioning and matching on covariates and treatment effect heterogeneity given the presence or absence of other treatment conditions.

- The magnitude of the interactions between the treatments is relatively small, indicating that the presence of one treatment does not dramatically affect the effectiveness of another treatment. From the graphs, it appears that for the same treatment, the red estimates from regression are usually averages of the different black estimates from the ICE method. For example, the audit treatment effect from regression looks to be an average of the three different audit treatment effects from the ICE model. This is not surprising given how the problem was set up. There appears to be weak evidence that the treatments can crowd out one another. For example, looking at the ICE models (black lines) for $Y3$ in the bottom right panel, it is clear that the significant audit treatment in the first column is no longer significant when an invites or comments treatment is added, as made clear in the second and third columns. Although the differences are themselves small and likely insignificant, this does confirm intuition that multiple monitoring treatments are not necessarily additive.

- The results do also seem to confirm the substantive conclusion that Olken reaches that the audit treatment leads to an approximately 8-9 percentage point decrease in corruption while the participation treatments do not have a consistent effect. However, the results also suggest that the "statistical significance" of the effects from a hypothesis testing standpoint is very tedious, and the presence of multiple treatments can render the results insignificant.

This first result demonstrates the ability of the ICE estimation method to recover various causal quantities accurately when benchmarked against more traditional methods. The estimation process also forces the researcher to think very clearly about what constitutes the treatment and control groups, which leads to a more clear exposition of what the treatment effect represents. Finally, the results presented also show a simple example of how the ICE method can estimate treatment effect heterogeneity in a straightforward manner that mirrors the use of interaction terms in regression. In this case, the treatment effects were estimated separately in the presence and absence of other treatment effects.

3.2.3 Audit Treatment Effect Heterogeneity

Since the audit treatment seemingly has a significant positive effect, I explore this effect further by looking at treatment heterogeneity and other types of treatment effects using ICEs. I condition on the presence of the other treatments by only comparing observations with the same status on the participation treatments within the matching step. For example, for observations that received the audit treatment and the invites treatment, I only match to observations that do not receive the audit treatment but do receive the invites treatment to estimate the ICEs. I do the same for observations receiving the comments treatment and for those that do not receive a participation treatment. The estimated ICEs are then used to calculate other quantities of interest.

One of the main benefits of the ICE approach is the ability to estimate any treatment effect by simple aggregation. Figure 3.2 shows the results of average treatment effects for the audit treatment within different subgroups of the data for each of the four measures of corruption.

The first three columns of each panel represent the posterior of the average treatment effect (ATE), average treatment effect for the treated (ATT), and average treatment effect for the controls (ATC) using the ICE estimates. The posteriors are derived simply by averaging the posteriors for all observations, treated observations only, and control observations only respectively. Typically, in observational studies, the ATT and ATE may be different if treatment assignment depended on some covariate that was also correlated with the treatment effects. Since treatment assignment was more or less randomized in this case, it is unsurprising that the ATE, ATT, and ATC are very similar.

The next four columns of each panel show average treatment effects for various subgroups of the data defined by specific covariate values. “Populous” subsets the ATE to villages with population greater than the dataset average. One theory may be that larger villages may be prone to more corruption because it may be harder for citizens to monitor officials due to collective action problems, so an outside audit may be more helpful. “Poor” indicates the ATE for villages with greater percent of households that are poor than the dataset average. One might expect that villages with more poor households may be more susceptible to corruption and thus an outside monitoring mechanism such as an audit may have a greater effect than in wealthy villages. “Mountainous” denotes the ATE for villages that are located in a mountainous region. One

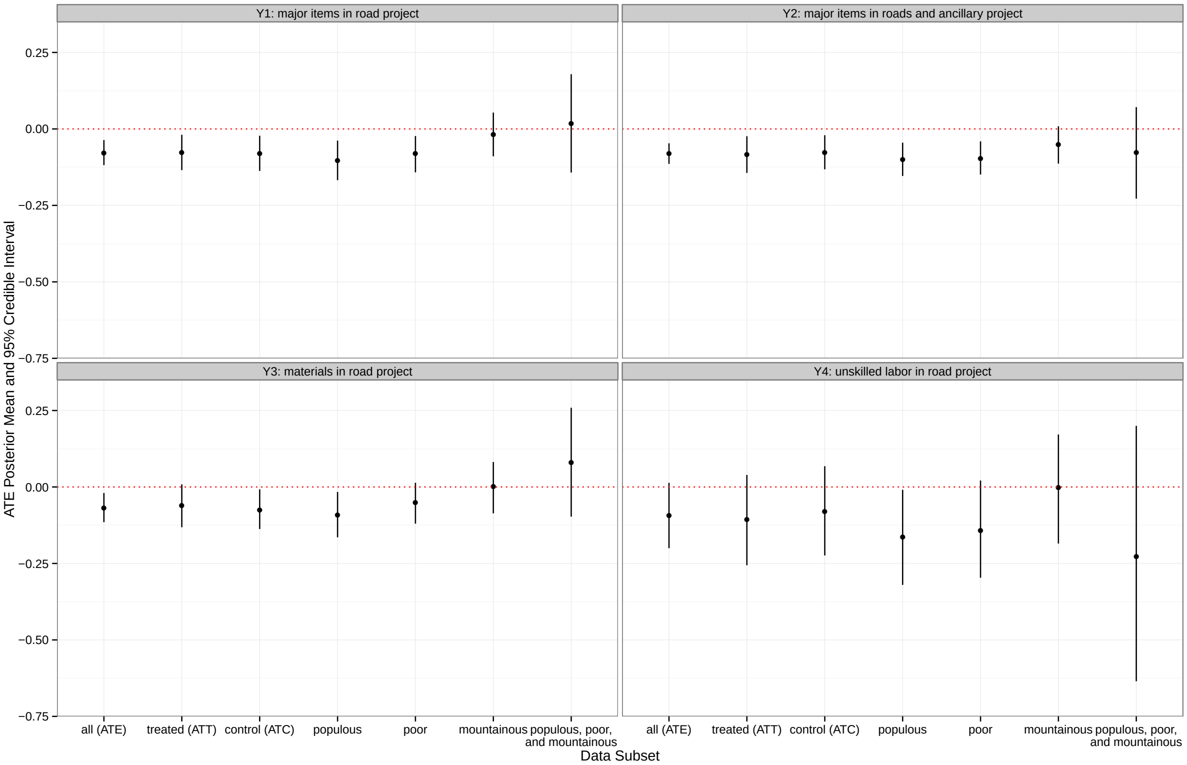

Figure 3.2: Audit Average Treatment Effects within Subgroups

can argue that geographically isolated villages have a stronger social bond, which allows for more monitoring within the village, so outside audits may be less helpful. And finally, “populous, poor, and mountainous” denotes villages that are large, poor, and within a mountainous region. The results show that the ATEs for populous and poor regions is not significantly different from the overall average, but audits seem to have a smaller and insignificant effect in mountainous villages. Subsetting the dataset by all three criteria together renders the sample size too small and the uncertainty intervals become quite wide. The results from Figure 3.2 suggest that treatment heterogeneity by subgroup may not be a huge problem. It also shows the flexibility of examining treatment effect heterogeneity by simply combining ICEs for various subgroups of observations.

Detecting treatment effect heterogeneity by finding average treatment effects within subgroups is very similar to existing methods and practices. However, estimating ICEs also allows researchers to look at the individuals themselves and look for treatment effect heterogeneity through various graphical methods. As an example, suppose the researcher would like to know whether the audit treatment would have a large effect on specific villages and how that effect differs across villages. One benefit of the Bayesian approach is that it allows the researcher to make probability statements about parameters in a coherent manner. Suppose a large effect for the audit treatment is defined as decreasing the percent missing by 20 percentage points. Then the probability of a large effect for any village is simply the probability of an individual causal effect of less than or equal to -0.2. With simulations from the posterior, this becomes simply the proportion of draws less than or equal to -0.2 for a specific τ_i.

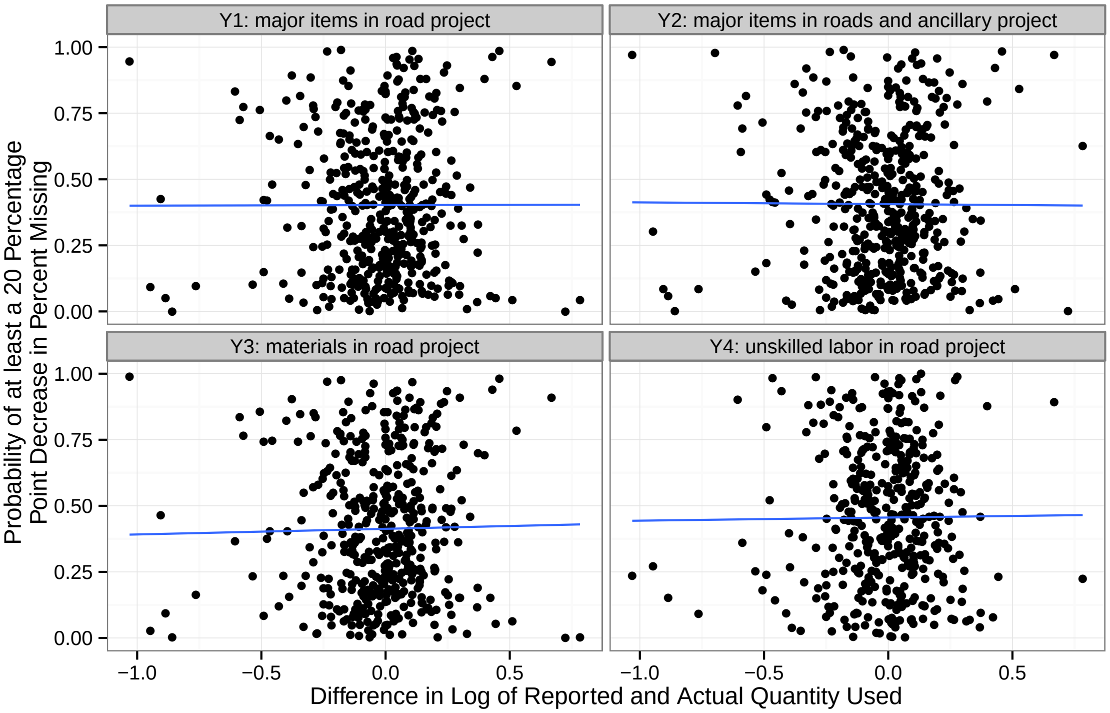

Figure 3.3: Probability of a Large Audit Treatment Effect by Quantity Overreporting

Figure 3.3 plots the probability of $\tau_i \leq -0.2$ on the y-axis and the difference in log of reported versus actual quantity of materials or labor used on the x-axis with a best-fit line drawn. Each point on the plot is a single village and each of the four panels on the graph represents one of the four different corruption variables. The y-axis is simply the probability that the audit treatment has a large effect. The x-axis represents how much a village over-reports its materials and labor usage. Recall that corruption can occur through over-reporting of quantity and/or inflating of prices. The results suggest that there is no relationship between how well the audit works and how much quantity over-reporting there is.

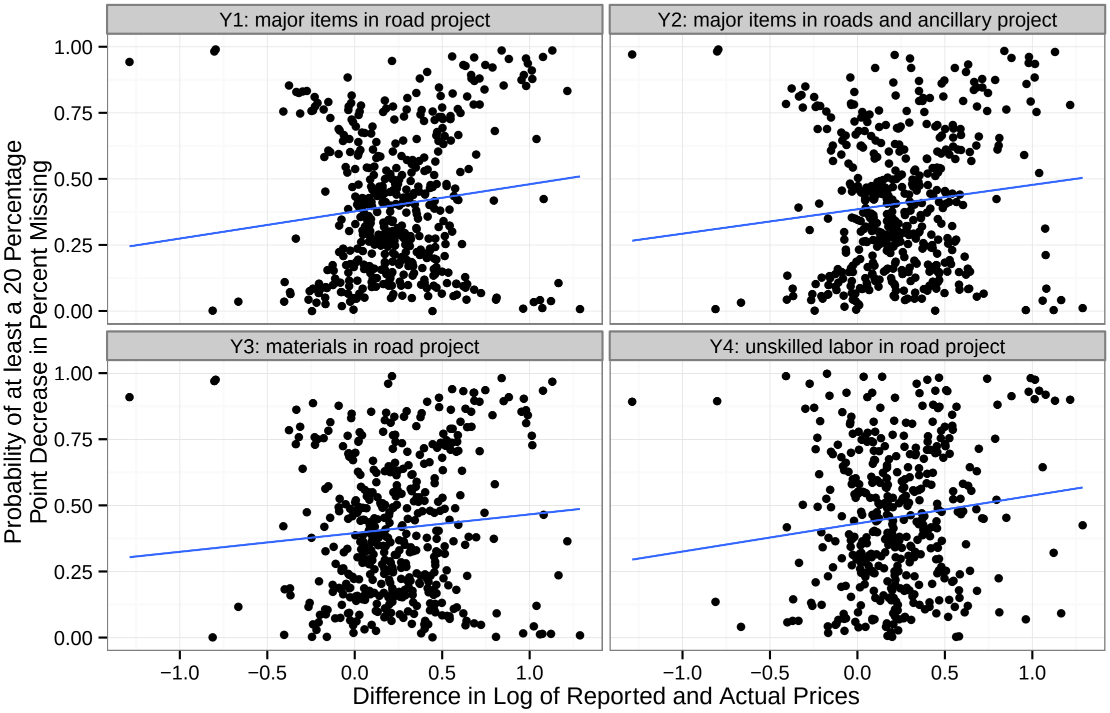

Figure 3.4: Probability of a Large Audit Treatment Effect by Price Overreporting

However, Figure 3.4 suggests there may be a relationship between audit treatment effectiveness and price inflation, which is plotted on the x-axis. It seems there is a slightly positive relationship where the probability of a strong audit effect increases with an increase in price inflation. The effect may be even stronger after discarding outliers in the top left of the graphs. The positive relationship suggests that audit treatments may be more effective in villages that over-report their prices. One explanation may be that prices are probably easier to check in an audit by comparing various outside sources, while quantity used may be harder to check in an audit. Therefore, audits work much more effectively in catching price inflation than quantity inflation. Figures 3.3 and 3.4 demonstrate one simple graphical way of detecting treatment effect heterogeneity. Given the posteriors of all the individual causal effects, treatment effect heterogeneity is straightforward to examine and researchers can make simple probability statements about the heterogeneity without resorting to hypothesis testing and the many issues that associated with it.

3.2.4 Treatment Effect Quantiles

The ICEs from the entire sample form a distribution of causal effects, which researchers may also be interested in. As mentioned before, the ICEs are only in-sample quantities, so any extrapolation from sample quantities to population quantities requires assumptions about how representative the sample is to the population. Nevertheless, the entire distribution of ICEs allows researchers to see what the entire range of effects are and to also look at treatment effect quantiles. However, an important distinction must be made between treatment effect quantiles and quantile treatment effects, the latter of which researchers have tried to develop methods for. A treatment effect quantile refers to the quantiles of the treatment effects whereas a quantile treatment effect refers to the difference of potential outcomes at a specific quantile for each of the two potential outcome distributions. Let $q(\cdot)$ be a quantile function for any quantile. Then

$$\begin{aligned}\text{treatment effect quantile} &= q(Y(1) - Y(0)) \\ \text{quantile treatment effect} &= q(Y(1)) - q(Y(0))\end{aligned}$$

In the case of average effects, the average treatment effect is equal to the difference in the average of the potential outcome distributions because of the linearity in expectations property. However, in the case of quantiles, the two quantities are different unless strong assumptions about rank order are made. Existing methods such as quantile regression try to estimate the quantile treatment effects, but I argue that treatment effect quantiles are the actual quantities researchers are interested in. Previous methods were unable to estimate treatment effect quantiles due to identification problems.

Figure 3.5 plots the treatment effect quantiles for the three treatments at the 25th, 50th, and 75th quantiles. The results suggest that the range of individual treatment is quite large and can vary from -0.5 to 0.5. Intuitively, this does not make sense as one would not expect corruption monitoring to increase corruption. There are several possible explanations for this result. The first is that the dependent variables are measured with such noise, with quite a few observations receiving nonsensical values of greater than 1 or less than -1, that the treatment effect quantiles results are driven by such measurement errors. The second explanation may be that quantiles on the extremes of the distribution are estimated with less accuracy as my simulations showed. Therefore, one should consider the treatment effect median to be more accurate than the other quantiles. Finally, one may consider that there may actually be some cases where monitoring inadvertently leads to more corruption. For example, in the audit treatment, the auditors themselves may be corrupt, and there exists possible collusion or bribery opportunities between the auditor and the project managers, especially since the audits were announced ahead of time. This possible collusion may inadvertently lead to more corruption. Nevertheless, Figure 3.5 shows that it is possible to get estimates of treatment effect quantiles by looking at the distribution of ICEs.

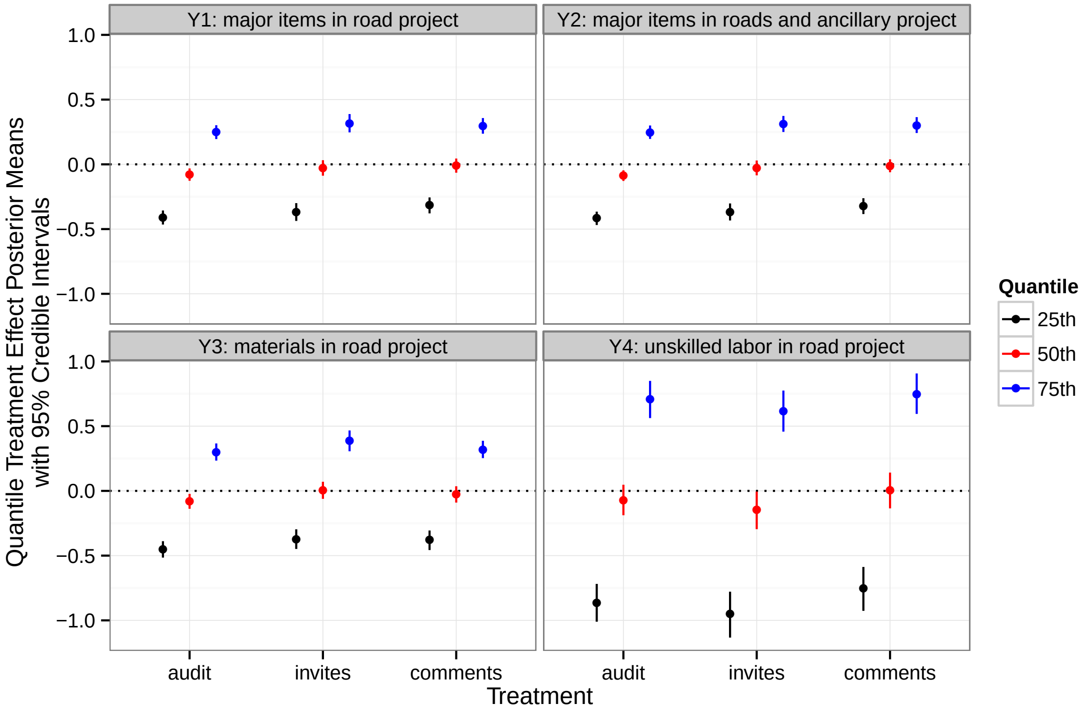

Figure 3.5: Three Treatment Effects Quantiles

3.2.5 The Effect of Participation Treatments on Outsider Village Meeting Attendance

Despite the results from above suggesting that only the audit treatment has a significant effect on corruption, I look more closely at the participation treatments and its mechanisms. The participation treatments also provide for an opportunity to demonstrate the flexibility of estimating ICEs because they can be thought of as part of a two-stage data structure. Recall that the participation treatments were theorized to be effective through the grassroots mechanism of increasing non-elite village turnout at village accountability meetings (first stage) and the increased attendance of outsiders should decrease the likelihood of corruption (second state). It is important to note that this is the only channel through which participation should reduce corruption. Olken was able to record actual attendance data at the three accountability meetings in each village, so I can use this data to estimate the effect of the first stage of treatment on non-elite (outsider) village attendance.

Figure 3.6 shows the results of the participation treatments on the raw outsider meeting attendance numbers and outsider meeting attendance as a percent of total attendance for each village averaged across three meetings. "Invites" refers to the treatment of sending invitations only, whereas "comments" refers to both an invitation and anonymous comment form, and "participation" lumps the two treatments together into a broad category. Recall also that the treatments were distributed randomly either by sending them home with children at schools or through neighborhood heads. The red and blue lines separate out the two delivery mechanisms. I use the same method to estimate the ICEs as before and the dependent variable is treated as a continuous variable. Since there are a variety of treatments and delivery mechanisms, I focus here only on the average treatment effects for the treated (ATTs) rather than the ATEs. These two quantities of interest should be equal given random assignment of treatment.

The results from Figure 3.6 lead to several conclusions. First, it appears that the participation treatments generally do lead to a significant increase in outsider attendance at the accountability meetings. Receiving a participation treatment in general increases outsider attendance by an average of around 7.5 people or around a 5 percentage point increase of outsiders as a percentage of the audience. Second, it appears that the invitations alone are more effective at increasing outsider attendance than an invitation and an anonymous comment form. This makes sense since the comment forms are a way for villagers to express opinions about the projects without fear or identification and retribution, so they act as a substitute for actually attending the meeting. And finally, it appears that the treatments are slightly more effective when distributed through schools as opposed to through neighborhood heads. This also makes sense since it may be the case that the neighborhood heads are more likely to be corrupt and less likely to have an incentive to

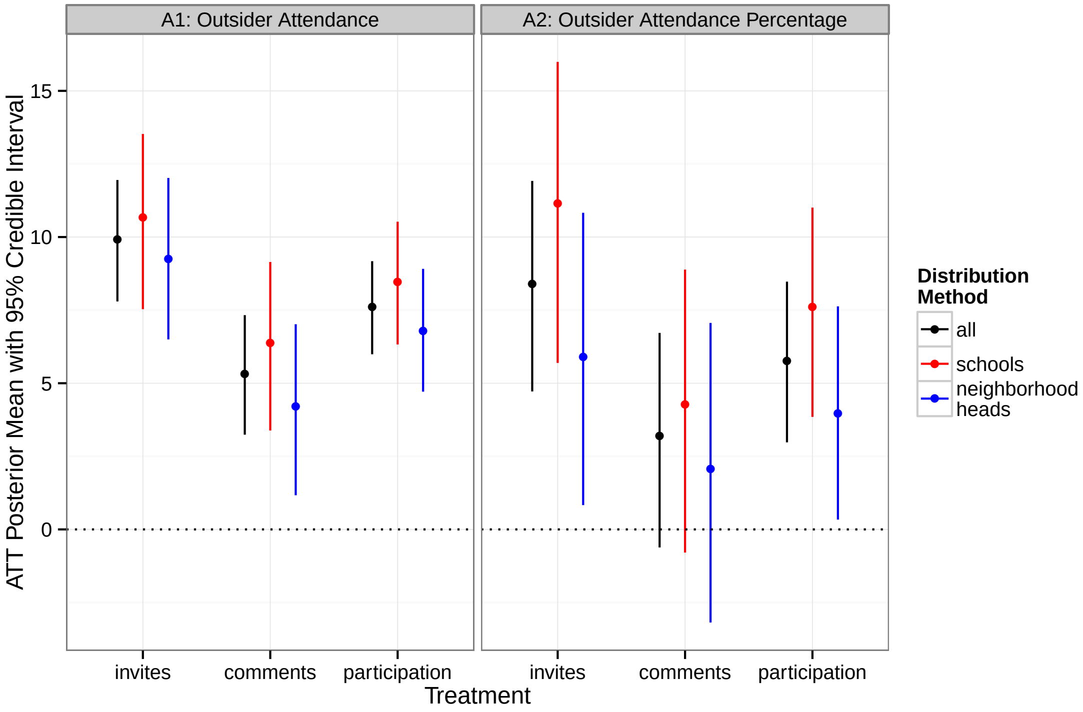

Figure 3.6: Participation ATTs on Outsider Meeting Attendance

increase outsider attendance at the meetings. Overall, the results suggest that the participation treatments actually work as intended in increasing outsider attendance to the accountability meetings.

3.2.6 The Effect of Outsider Village Meeting Attendance on Corruption (continuous treatments and outcomes)

The previous subsection showed that the participation treatments have a positive and significant average effect on outsider attendance at the village accountability meetings. In this subsection, I look at whether increasing outsider village meeting attendance has the effect of reducing corruption. Here I look at this second stage independently of the first stage. The next subsection will incorporate the two stages together into one model.

This second stage also provides an opportunity to demonstrate how the ICE model I use can be adapted to accommodate non-binary treatments. In this case, both outsider meeting attendance and outsider meeting attendance percentage are considered continuous "treatment" variables, denoted A.[6] The key assumption required for continuous treatments is a linearity assumption, where the effect of continuous treatment A is assumed to be linearly related to the outcome Y. The linear ICE is then simply the effect of increasing A by one unit. The way to conceptualize this is that there are an infinite number of potential outcomes $Y(A)$ since there are an infinite number of possible values for A. The linearity assumption imposes a structure where the ICE is

$$\tau_i = Y_i(A+1) - Y_i(A);\ \forall A$$

Note that this is equivalent to the previous definition of τ_i for binary treatments if $A = 0$.

To simulate from the posterior for τ_i with continuous treatments, only a few minor adjustments are necessary to the original algorithm.

- Previously, the set of possible donor observations for observation i was all observations with the opposite treatment status. For continuous treatments, the set of possible donor observations for observations i is any observation with a different value on the treatment variable A. Since A is continuous, the set of possible donors is likely to be nearly every other observation in the dataset.
- Denote the counterfactual treatment status[7] for observation i as $A_i + 1$. Then $Y_i^{mis} = Y_i + \tau_i$.
- Once the donor pool has been determined from the matching step, to draw the equivalent of $\tilde{\theta}_i^{mis}$, simply run a linear regression step of Y on A with the donor pool. Let $\tilde{\lambda}_{0i}$ and $\tilde{\lambda}_{1i}$ be the intercept and slope draws from this regression. Then $\tilde{\theta}_i^{mis} = \tilde{\lambda}_{0i} + \tilde{\lambda}_{1i}(A_i + 1)$.
- To draw $\tilde{Y}_i^{mis}$, simply draw from a Normal distribution (for continuous outcome variables) with mean $\tilde{\theta}_i^{mis}$ and the standard deviation equal to $\tilde{\sigma}$ from the regression step above. Then $\tilde{\tau}_i = \tilde{Y}_i^{mis} - Y_i$

[6] I refer to the attendance variable as treatment variables here when looking at this second stage independently. They are treatments in the sense that I am interested in their effects on corruption looking only at the second stage. However, in the overall scheme of the study, the treatments are still the participation and audit interventions. I denote these second stage "treatments" with A to avoid confusion.

[7] With the linearity assumption, one can really define any counterfactual to estimate the ICE. I use $A_i + 1$ here for simplicity.

The main differences between this and the previous algorithm is simply modeling the donor pool with a linear regression rather than with a Normal model and then specifically defining a counterfactual treatment status. The counterfactual treatment status in the case of binary treatments is already strictly defined as the opposite treatment status whereas in this case, there are an infinite number of potential counterfactual treatment statuses.

Using this algorithm and setup for continuous treatments, Figure 3.7 shows the results of the effects of outsider attendance and outsider attendance percentage on the four corruption measures.[8] Note that since outsider attendance was not randomly assigned, this second stage analysis resembles an observational study. I calculate the (linear) ATE, which is simply the average of all the ICEs in the data. In the case of continuous treatments, the "treatment" and "control" groups are not well defined, so ATT and ATC are also not well-defined. The black lines represent the ATE using all observations while the red and blue lines indicate the ATEs for the subgroups of observations that received or did not receive the audit treatment respectively.

The results from Figure 3.7 suggests that increasing outsider attendance by one person or increasing outsider attendance percentage by one percentage point does not really have a significant effect on decreasing corruption. In fact, the point estimates seem to suggest that increasing outsider attendance may actually increase corruption, although the credible intervals often cover zero. With the same caveats about the corruption variables measured with high error, it seems that the grassroots approach to corruption monitoring is ineffective. Although the participation treatments do increase participation, this increase does not appear to lead to a similar increase in accountability.

3.2.7 Two-Stage Analyses of the Effect of Outsider Meeting Attendance on Corruption

A proper analysis of the effect of increasing outsider meeting attendance should take into account both stages of data. The previous subsection only looked at the effect in the second stage without taking advantage of

[8] In the matching specification for these models, I also include the treatment statuses for the invites, comments, and audit treatments, whether the participation treatments were distributed through schools or neighborhood heads, and total meeting attendance as control variables in addition to the original ten covariates. None of these are post-treatment since the treatment in this case is outsider attendance.

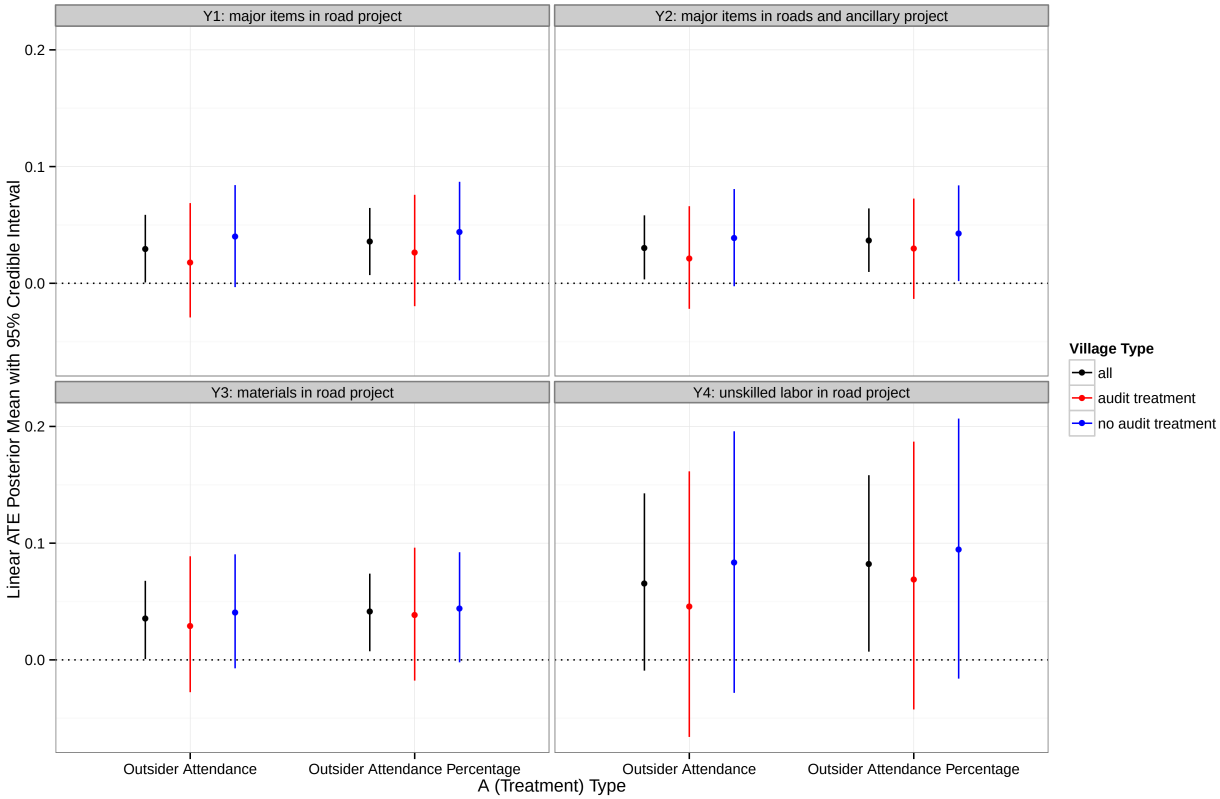

Figure 3.7: Linear ATE of Outsider Meeting Attendance on Corruption

the randomization of the participation treatments. In this subsection, I demonstrate how to estimate ICEs in a two-stage framework that mirrors existing methods. I consider the participation interventions here to be one intervention without differentiating between invites and comments. There are two ways to conceptualize the two-stage analysis, both based on broad sets of existing methods. The first and more common way to think about the problem is to look at it through the lens of instrumental variables. The second way is to think about it as a problem of identifying causal mechanisms. I use the instrumental approach here, although the framework can be used to identify causal mechanisms as well.

The hypothesized causal pathway is as follows. Villages get assigned to either receive a participation intervention or not. Villages that receive a participation intervention should experience an increase in outsider meeting attendance because of the treatment. The increase in outsider meeting attendance should then result in more corruption monitoring, which should then lead to lower levels of corruption. So far, I have shown that participation interventions do increase outsider meeting attendance on average, but increasing outsider meeting attendance on average does not reduce corruption. However, since both estimates were averages, I have yet to show the effect of outsider meeting attendance on corruption *in those villages where participation increased outsider meeting attendance*. The ICE framework allows me to examine this problem further by specifically linking the two stages together on an individual village level.

In the typical instrumental variables setup, there is a treatment variable of interest where treatment assignment is not ignorable. However, there exists an instrument that has ignorable assignment and is correlated with the treatment variable. The analysis then leverages the ignorable assignment in the instrument to identify the effect for the treatment variable. In this case, the participation treatment would be the instrument and the outsider meeting attendance would be the treatment variable of interest.[9] Under certain assumptions, the instrumental variables analysis can estimate and identify a local average treatment effect (LATE), which is the average treatment effect for compliers. Compliers here are defined as the subgroup of individuals for whom the instrument affects the treatment variable in the hypothesized direction when given the instrument and has no effect when not given the instrument. In our example, a village is classified as a complier if outsider meeting attendance increases when receiving the participation intervention and stays the same when not receiving the participation intervention. The LATE is then the effect of outsider meeting attendance on corruption for complier villages.

To identify the LATE in this example (and generally speaking for instrumental variables), the following assumptions must hold:

- **Stable treatment value assumption (SUTVA)**: assumed to hold, although slightly violated by the differing treatments of invites and comments.
- **Ignorable assignment of the instrument**: assumed to hold because of random assignment of participation.
- **Exclusion restriction**: assumes that the participation interventions affect corruption only through the channel of outsider meeting attendance; assumed to hold.

[9] For this subsection, I only consider the raw outsider meeting attendance number rather than outsider meeting attendance percentage.

- **Non-zero average causal effect of participation intervention on outsider meeting attendance**: shown to hold in previous sections.
- **Monotonicity**: participation interventions only affect outsider meeting attendance in one direction; assumed to hold although I relax this assumption later.

The key to identifying LATE is to identify which villages are compliers and which are not. If compliance status is known, then LATE would be easy to estimate. However, compliance status is not known, but I can estimate compliance status in the first stage using the ICE framework and then use the ICE framework in the second stage as well to estimate LATE given compliance status.

Consider the following way to use ICEs in an instrumental framework setting. In the first stage, estimate the ICEs for all observations to get the individual effects of the participation intervention on outsider meeting attendance. The posterior of the ICEs represent the uncertainty over compliance status. For each draw from the posterior, consider a village to be a complier village if the ICE is positive and not a complier if the ICE is not positive. For each iteration, classify every village as either a complier or non-complier based on the first stage ICE. The draws from the entire posterior of this first stage characterize the uncertainty over whether or not a village is a complier. The probability of village i being a complier village is simply the proportion of posterior draws greater than 0 in this first stage.

Next, denote the missing potential outcomes from the first stage as A^{mis}. Then, in the second stage, implement the ICE algorithm a second time with the corruption measure as the outcome and outsider meeting attendance as the treatment. This is the same algorithm as above for continuous outcomes and continuous treatments. However, one key difference is that the counterfactual treatment here is the A^{mis} from the first stage, whereas before, the counterfactual was arbitrarily chosen to be $A - 1$. The idea behind this is that A_i^{mis} is the imputed outsider meeting attendance for observation i if it had received the opposite participation intervention. Then Y_i^{mis} is the potential outcome for corruption given a hypothetical outsider attendance value of A_i^{mis}. A second key difference is that in the potential donor pool at the second stage, donors must be of the same compliance type. So if observation i is drawn as a complier in the iteration, then the donor observations must also be drawn as compliers in that iteration. The ICE algorithm simply imputes two missing potential outcomes for the opposite participation treatment. For each draw of the algorithm, I draw a set of compliers and then draw an estimate of LATE.

The specifications of this two-stage model can vary in several ways. For example, one can include control variables to match either in the first stage or the second stage or both. The assignment for the instrument must be ignorable, so it must be randomly assigned or ignorable after controlling for covariates. In the second stage, including control variables in the matching is optional and may or may not increase the precision of the estimates. One can also choose not to include matching variables, in which case the donor pools in the first and second stages would simply be all observations with a different instrument and treatment statuses respectively.

Another way to alter the specification is to impose the monotonicity assumption. In the specification I initially described, the monotonicity assumption is not strictly necessary and not imposed. It allows the participation intervention to actually decrease outsider meeting attendance. However, if a monotonicity assumption makes sense substantively, imposing it in the algorithm will improve estimates and reduce noise. Let i be an observation that receives the participation intervention. To impose the monotonicity assumption in this example, I must constrain A_i^{mis} produced from the first stage ICE to be less than or equal to the observed A_i. If $A_i^{mis} > A_i$ for any draws of A_i^{mis}, then I simply change the imputation of A_i^{mis} such that $A_i^{mis} = A_i$

Figure 3.8 presents the results of various specifications of this two-stage model of the raw average outsider meeting attendance number on corruption using the participation intervention as an instrument. I consider four different specifications: two models with the monotonicity assumption, with and without second stage matching, and two models without the monotonicity assumption. I consider two quantities of interest for the four dependent variables: the LATE and the non-complier average treatment effect (NCATE). The LATE considers only compliers whereas the NCATE considers only non-compliers.

The results from Figure 3.8 lead to several conclusions. First, consider the NCATE estimates. The NCATE is a way to test the validity of the exclusion restriction. Recall that the exclusion restriction states that the instrument only affects the outcome through the treatment. If the exclusion restriction holds, then the NCATE should be zero since the instrument should not be affecting the outcome for non-compliers. The blue lines in Figure 3.8 confirm that the NCATE is likely zero, suggesting that the exclusion restriction is

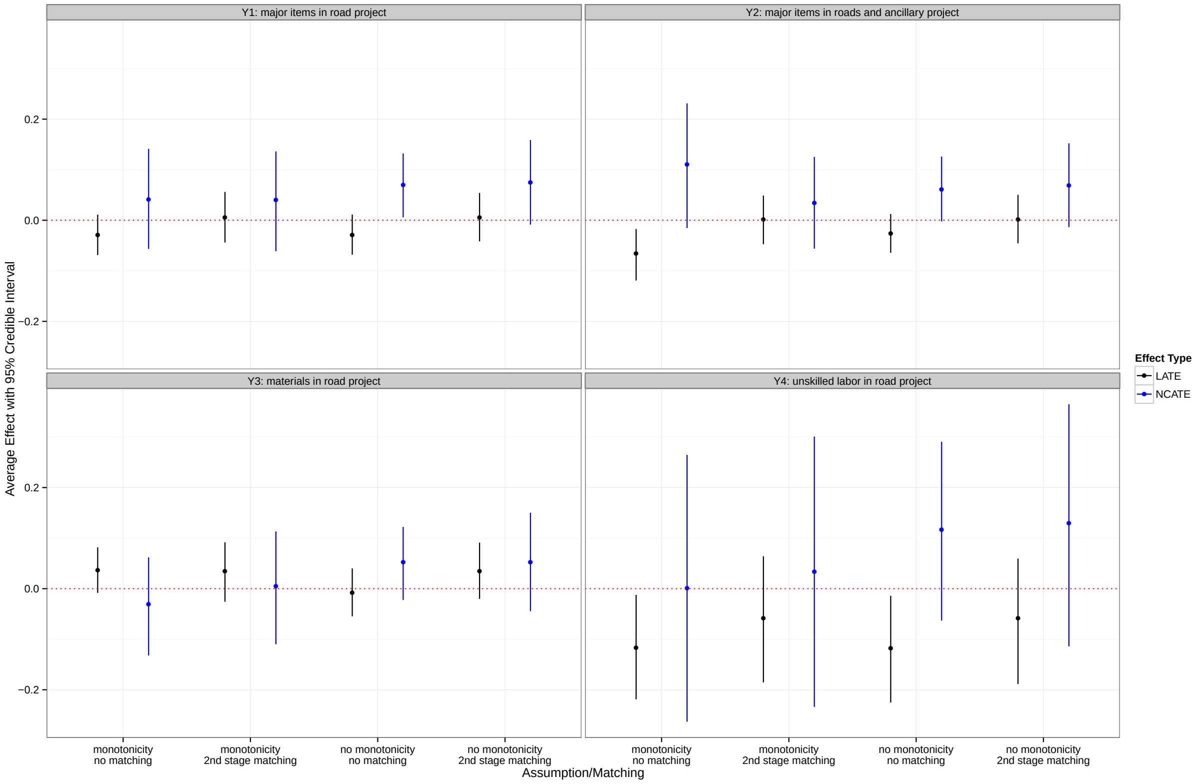

Figure 3.8: Two-Stage ATEs of Outsider Meeting Attendance on Corruption

a valid assumption. The LATE estimates across the specifications and corruption measures suggest that outsider meeting attendance does not have a significant effect on corruption. This confirms the result from before that grassroots monitoring is not very effective in reducing corruption.

3.3 Application 2: The National Job Corps Study

The second application implements the ICE algorithm on a randomized study of a job training program in the US. The question of whether or not job training programs are effective is one of the most widely evaluated questions in the fields of economics and causal inference. The specific data used here comes from the National Job Corps Study conducted by Mathematica Policy Research, Inc. The job training program, known as Job Corps, offers job training for disadvantaged youths between the ages of 16 and 24. The study here involved a random sample of all eligible applicants for the program in late 1994 and 1995. I obtained the dataset from Frumento et al. (2012) and closely mirrored the analyses in their paper.

In the original study, 15,386 individuals were sampled and assigned either a treatment (9,409) or control (5,977) intervention. The treatment group was offered the opportunity to enroll in the program while the control group was denied access to the program for three years. Interviews were then conducted with the entire experimental population at baseline and then at 52, 130, and 208 weeks after the random assignment. Due to problems with incomplete baseline interviews, individuals who died during the follow-up, and people who were admitted to the program even though they were assigned to control, the resulting experimental population consisted of 13,987 individuals. Of the individuals that were in the treatment group, not all of them chose to enroll in the program. The treatment group compliance rate (those who were assigned to treatment and then enrolled) was about 68%. The following background covariates were collected on all individuals at baseline:

- Gender
- Age
- Has children
- Years of education
- Mother's years of education
- Father's years of education
- Has job
- Months employed in previous year
- Had job in previous year
- Earnings in previous year
- White or non-white
- With or without a partner

- Ever arrested
- Whether household income > \$6000
- Whether personal income > \$6000

I deal with missingness in the covariates by using only one imputation from a set of multiple imputations, following the same method as Frumento et al. (2012). They justify using only a single imputation by stating that there was very small variability in the results across multiple imputations. At the follow-up interviews, two outcomes are measured: employment and wages. For the purposes of this application, I only look at the binary employment outcome (employed or not), although future extensions should also look at wages.

Frumento et al. (2012) address three issues with the study in their paper: treatment assignment noncompliance, partially defined wages due to nonemployment, and unintended missing outcomes. Because my focus is on estimating ICEs and showing the flexibility of the model in examining treatment effect heterogeneity, I only address the first problem of noncompliance. I exclude the second problem by looking only at employment rather than wages, and I ignore the third problem by dropping observations with missing outcomes. The latter may induce bias when looking at population estimands, but theoretically poses no problems when limiting the analysis to the sample or individual estimands. I deal with the problem of noncompliance by using principal stratification in a formal two-stage model. The principal strata are defined by estimating ICEs in the first stage. While the first application of monitoring corruption also included a two-stage model, I more formally define the model in the second stage. This application is also different from the first in that all the outcomes and treatments in the two stages are binary variables, which allows for easier notation.

3.3.1 A Two-Stage Model for the Effect of Job Training on Employment with ICEs

The outcomes I am interested in are the employment statuses of individuals in the experiment at 52 weeks, 130 weeks, and 208 weeks after randomization. Assignment to being offered the choice of enrolling into the job training program is randomized, but actual enrollment in the program is not. Let Z denote the binary treatment assignment and let W denote the binary enrollment in the program indicator. Y denotes any one of the three binary outcome variables. The two-stage model here incorporates the first stage of the effect of Z on W and the second stage effect of W on Y. The setup is a typical instrumental variables study where Z is the instrument. Since W was not randomized, I rely on the randomization of Z to identify treatment effects. All the typical IV assumptions of SUTVA, monotonicity, exclusion restriction, non-zero average effect of Z on W, and ignorable assignment of Z are assumed here.

Researchers are generally interested in two types of average treatment effects in this setup: the intention-to-treat effect (ITT) of Z on Y and the local average treatment effect (LATE), which is the effect of W on Y for compliers. Compliance here is defined as enrolling in the program if offered and not enrolling if not offered. Due to the nature of the program, I assume that there is only one-sided noncompliance in that individuals can choose not to enroll if offered treatment but they cannot choose to enroll if not offered treatment (monotonicity assumption). Notationally, I define compliance with the potential outcomes framework where W(Z) is the enrollment status given treatment status Z. Then, for compliers,

$$\begin{aligned} W(1) &= 1 \\ W(0) &= 0 \end{aligned}$$

and for non-compliers,

$$\begin{aligned} W(1) &= 0 \\ W(0) &= 0 \end{aligned}$$

Let G be a binary indicator for whether an individual is a complier or not and let $Y(Z)$ denote the potential outcome for Y given treatment assignment Z. Then, the typical treatment effects estimated under this setup are

$$\begin{aligned} \text{ITT} &= E[Y(1)] - E[Y(0)] \\ \text{LATE} &= E[Y(1)|G = 1] - E[Y(0)|G = 1] \end{aligned}$$

The unbiased estimate of the ITT is simply a difference in means of Y given randomization of Z. Estimating

LATE requires first estimating group membership for each individual.

I use principal stratification (Frangakis and Rubin 2002) and stratify observations given their Z and W indicators. Let $S(Z, W)$ denote a strata of observations with observed values of Z and W. Due to the assumption of only one-sided non-compliance, there are three strata in the data: $S(1,1)$, $S(1,0)$, and $S(0,0)$. The compliance statuses of individuals in $S(1,1)$ and $S(1,0)$ are known. Since individuals not assigned treatment cannot enroll in the program, it must be the case that everybody in $S(1,1)$ are compliers and everybody in $S(1,0)$ are non-compliers. The only uncertainty in compliance status is with the 5,299 individuals in $S(0,0)$.

I can estimate group membership status using the ICE algorithm in the first stage. Since estimating group membership for $S(0,0)$ is equivalent to estimating $W(1)$, the problem can be considered as one of estimating the ICEs of Z on "outcome" W. This first stage estimation gives the posterior probability of any individual belonging to the compliers group. Using the draws from the first stage, I can then implement a second stage where I estimate the ICEs of W on Y conditional on individuals being drawn as compliers to find complier treatment effects. Simply put, the algorithm is very similar to before. For each iteration of the MCMC, draw a value for W_i^{mis} for $i \in S(0,0)$. Determine compliance status using W_i and W_i^{mis}. For complier treatment effects then, take all individuals labeled as compliers and estimate ICEs with W as treatment and Y as the outcome. In both the first and second stages, matching on covariates is not strictly necessary since Z is randomized. However, using matching can improve estimates by subsetting the potential donor observations to a smaller set of more similar observations rather than using the entire set of observations in the other treatment group. At worst, matching poorly will simply produce a random draw from the potential donor pools. Given the large number of observations in this study, matching done correctly will almost certainly reduce the variance of the estimates.

Recall the original setup for estimating ICEs. Let Y_i^{mis} be the missing potential outcome to be imputed and let θ_i^{mis} be the mean of the distribution for Y_i^{mis}. Let $D_j^{(i)}$ be an indicator for whether the jth observation is a match for observation i. The random component in the model is the outcome when observation j is a match to observation i when i is the individual of interest. Y_i is fixed and therefore not a quantity of interest for modeling. The simplified[10] version of the original posterior was

$$p(\theta|Y,X,W) \propto p(D|\theta)p(Y|D,\theta)p(\theta)$$

where the posterior was augmented with D. In the two-stage model here, the posterior is augmented again with compliance status G.

Let π_j denote the probability of observation j being a complier. For the simple case where compliance status is estimated without matching, the empirical complier proportion can be used:

$$\hat{\pi}_j = \frac{\sum_{i=1}^{N} I(i \in S(1,1))}{\sum_{i=1}^{N}[I(i \in S(1,1)) + I(i \in S(1,0))]}$$

where $I(\cdot)$ is an indicator variable. Recall the previous complete likelihood[11] for the one-stage ICE model:

$$\begin{aligned} \mathcal{L}_{comp}(\theta^{mis}|Y,D) &= p(Y,D|\theta) \\ &= p(Y|D,\theta^{mis})p(D|\theta) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N}\left[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}},\mathcal{M})\right]^{D_j^{(i)}} \end{aligned}$$

With the two-stage model, there is a second data augmentation using compliance status G. If D and G were observed, the complete data likelihood would be

$$\begin{aligned} \mathcal{L}_{comp} &= p(Y|D,G,\theta)p(D|\theta)p(G|\theta) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{N}\left(\left[p(Y_j|\theta_i^{mis})p(D_j^{(i)}|\theta_{\mathcal{M}}^{(G)},\mathcal{M})\right]^{D_j^{(i)}} \pi_j^{G_j}(1-\pi_j)^{(1-G_j)}\right)^{I(G_j=G_i)} \end{aligned}$$

The likelihood here differs from before in that only donor observations within the same compliance status as i contribute information when i is of interest. The likelihood terms for any observation not in the same compliance group as i provide no information for θ_i^{mis} and are dropped. The matching parameters are also estimated separately for each compliance group, as denoted by $\theta_{\mathcal{M}}^{(G)}$. The product over all i's denotes the

[10] I suppress the notation for the matching to keep things simple.

[11] Like before, assume the matching parameters are estimated separately and given.

complete set of ICEs for all observations in the data. Integrating out G in the likelihood involves piecing the likelihood together from the three principal strata. However, like before, the researcher can simply approximate the integrals using Bayesian simulation.

One can complicate the model further by estimating π_j using an ICE step in the first stage, matching[12] and imputing W^{mis}.[13] Let ω^{mis} denote the mean of the distribution W^{mis} drawn by modeling the donor pool in the first stage.[14] The full MCMC algorithm for the two-stage ICE model that I implement contains the following steps.

Two-Stage MCMC Algorithm[a] for the Posterior of τ_i

```
Repeat the following $n_{sim}$ times:

    1. Draw a matching procedure $\tilde{\mathcal{M}}_1$ where the subscript denotes the
       first stage matching.
    2. Draw $\tilde{\theta}_{\mathcal{M}_1}$.

for ($i$ in 1:$N$){
    3. Determine $\tilde{D}_1^{(i)}$ from matching procedure.
    4. Draw $\tilde{\omega}_i^{mis}$ to estimate $\omega_i^{mis}$.
    5. Draw $\tilde{W}_i^{mis}$ from Bern($\tilde{\omega}_i^{mis}$).
    6. Calculate $\tilde{G}_i = Z_iW_i + (1 - Z_i)(W_i^{mis} - W_i)$
    7. Draw a matching procedure $\tilde{\mathcal{M}}_2$.
    8. Draw $\tilde{\theta}_{\mathcal{M}_2}^{(G)}$ separately for the two compliance groups.
    9. Determine $\tilde{D}_2^{(i)}$ from second stage matching conditional on $\tilde{G}$.
    10. Draw $\tilde{\theta}_i^{mis}$ to estimate $\theta_i^{mis}$.
    11. Draw $\tilde{Y}_i^{mis}$ from Bern($\tilde{\theta}_i^{mis}$).
    12. Calculate $\tilde{\tau}_i = Z_i(Y_i - \tilde{Y}_i^{mis}) + (1 - Z_i)(\tilde{Y}_i^{mis} - Y_i)$.
}
```

[a] Steps 3-5 may be skipped for observations in $S(1,1)$ and $S(1,0)$ since their compliance status is known. The equation in step 6 accounts for this.

The parameter θ_i in the two-stage model is the individual intention-to-treat effect. The algorithm also outputs the draws from the distribution of compliance status G. Using the draws of θ_i and G_i, the researcher

[12] If the covariates are uninformative about compliance status, then the first stage ICE would simply be an approximation of the empirical estimate $\hat{\pi}_j$ from above.

[13] Note that W^{mis} is imputed with certainty for individuals in $S(1,1)$ and $S(1,0)$.

[14] The parameters ω^{mis} are the first stage equivalent of θ^{mis} in the one-stage ICE algorithm.

can calculate any causal effect of interest including the ITT, LATE, and NCATE (non-complier average treatment effect) and explore any treatment heterogeneity. I implement this algorithm on the job training data with predictive mean matching on the 15 covariates with $M = 20$ in both the first and second stages of the algorithm with an MCMC of length 2000.[15]

3.3.2 Treatment Effects and Treatment Effect Heterogeneity from a Two-Stage Model

I first estimate three average treatment effects (ITT, LATE, NCATE) across the three survey timepoints of 52, 130, and 208 weeks after randomization. The dependent variable is whether the individual is employed at each timepoint. The average Job Corps participant stays in the training program for 1-2 years, so some of the participants in the program may still be enrolled at the first timepoint of 52 weeks. The ITT measures the average effect of treatment assignment on employment regardless of whether an individual enrolls in the program. The LATE measures the average effect of treatment assignment amongst compliers and the NCATE measures the average effect of treatment assignment amongst non-compliers. To calculate the LATE (NCATE), I take the draws of τ_i for each iteration of the algorithm and average the ones for individuals that were drawn as compliers (non-compliers) within that iteration. This vector consists of draws from the posterior for the LATE (NCATE). I then take the mean and the quantiles of the vector for the point estimate and credible interval.

Figure 3.9 presents the results from the posteriors using the two-stage ICE algorithm. At 52 weeks, all the effects are negative, which indicates that job training actually decreases the probability of being employed at 52 weeks. At 130 and 208 weeks, the average effects become positive, suggesting that job training does actually increase employment prospects. There are a few things to note from these results. First, the fact that the effects are negative at 52 weeks is unsurprising. There are at least two possible explanations. The first is that participants in the Job Corps program are likely to still be enrolled in the program and thus have not had an opportunity to search for jobs. Their counterparts that did not enroll probably have higher

[15] The donor pool size $M = 20$ is significantly lower than the 10% of smallest treatment arm number that I used before. Because the dataset is quite large, 10% of the smallest treatment arm would result in $M > 500$. My simulations thus far have not covered such a large dataset so I chose a much smaller number to allow for sufficient variation in the composition of the donor pools.

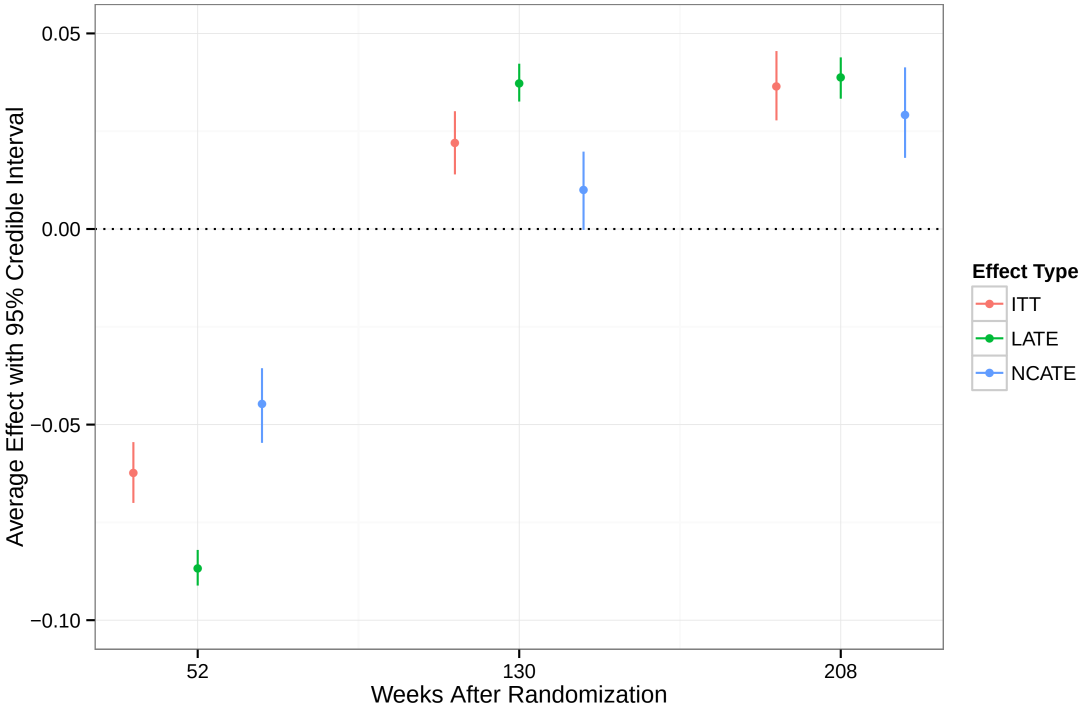

Figure 3.9: Three Average Treatment Effects at Three Timepoints

employment rates since they have spent the 52 weeks looking for jobs. The second explanation is that even if participants have already finished the program, the resulting skills they have acquired lead them to search for higher income jobs, which may take longer to find. The idea is that participants now have a higher "reservation wage", the lowest wage at which they are willing to work. Because I only look at employment outcomes, it can be misleading since those that take job training may demand a higher wage whereas those that did not take job training may be willing to settle for lower-paying jobs. If one imagines the ease of finding a job is inversely related to the wage paid by the job, then lower-paying jobs are easier to obtain and individuals with a higher reservation wage are likely to be unemployed longer. I explore this idea further through exploring treatment effect heterogeneity.

A second finding to note is that the LATE is always stronger in magnitude than the NCATE. This is to be expected as the effect of treatment assignment should be much stronger for those that actually take the treatment than those that do not. However, with the exception of possibly the result in week 130, the

NCATE effects, though weaker than LATE, do not seem to be zero. It appears that simply being assigned treatment does actually have an effect on individuals independently of actually enrolling in the job training program. From a methodological perspective, this seems to call into question the validity of the exclusion restriction, which requires that treatment assignment only affects the outcome through actually enrolling in the program. There may be a couple explanations for this. First, it may be the case that individuals who are assigned treatment are given a boost of confidence from simply being offered acceptance into the program. The offer itself may spur the individual to think about the future and to look harder for employment even without enrolling in the program. Second, it may also be the case that individuals who are offered a spot in the program may decide to decline the invitation in favor of another competing job training program or opportunity. Being offered the treatment may simply open their eyes to the opportunities available to them, and they may decide to pursue other opportunities that lead to employment. Nevertheless, a non-zero NCATE may indicate a violation of the exclusion restriction, which likely causes an upward bias in the estimate of the LATE.

The LATE results so far suggest that actual enrollment into the job training program decreases the probability of employment in the beginning and while still in the program, but has a positive effect on employment after completing the program. I now explore treatment effect heterogeneity further using the posterior of the ICEs. The first avenue I explore is whether the LATE is stronger for certain types of individuals characterized by the covariates. I take each of the nine binary covariates in the data and I estimate the LATE for $X = 1$ and $X = 0$.[16] I then take the difference in the LATE for $X = 1$ and $X = 0$. Figure 3.10 shows the results for the difference in LATEs between the two binary groups for four of the binary covariates.

The way to interpret the lines in Figure 3.10 is that a positive value on the y-axis indicates that the LATE for $X = 1$ is greater than the LATE for $X = 0$. For example, in the top left corner, at week 52, the LATE for individuals with children is about 5 percentage points greater than the LATE for individuals without children. This suggests that the effect of job training on employment at week 52 is greater on individuals with children. This result may conflate two mechanisms. First, it may be the case that individuals with

[16] The process to calculate the LATEs here is similar to before. For each iteration of the MCMC, I identify those individuals drawn to be compliers with $X = 1$ and those draw to be compliers with $X = 0$. I repeat this process for the entire length of the MCMC to get draws from the posteriors for the LATEs for $X = 1$ and $X = 0$.

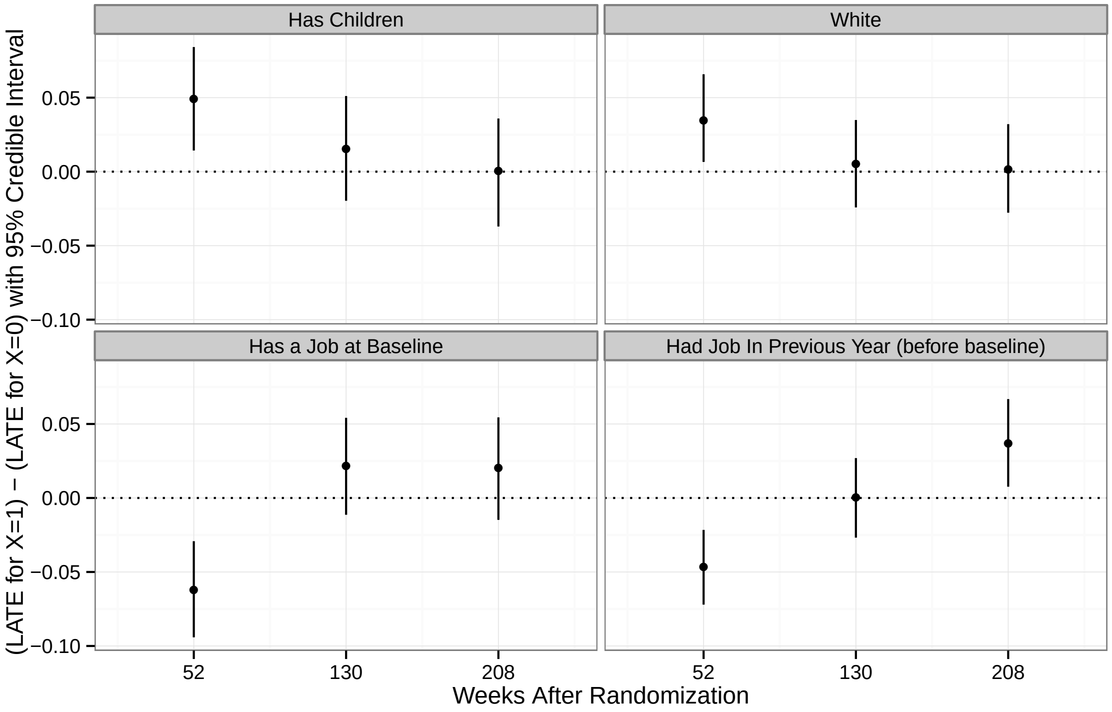

Figure 3.10: Difference in LATEs for Four Binary Covariates

children are more likely to finish the job training program sooner and thus be employed sooner due to the need for a steady job to raise children. Second, it may also be the case that individuals with children have a lower reservation wage because they cannot afford to hold out for a higher-paying job and may settle for lower-paying jobs to support their children.

In the top right panel, it appears that the job training works slightly better for whites than for non-whites. There may be numerous explanations for this. Race may be correlated with a large number of other factors which may result in the appearance that whites finish the program faster and/or have an easier and faster time to employment after the program. I control for education and household income in the matching specification, but that may not account entirely for the heterogeneity in effects across races.

The two bottom panels look at the differences in LATE between individuals who were employed before the program and individuals who were not. Employment was measured as having a job when the baseline survey was taken (left) or having a job within the previous year before the baseline survey (right). The two variables are undoubtedly quite highly correlated. In both cases, it appears that having a job before the program significantly decreases the effect of the program on employment at week 52. It may be the case that those already holding jobs beforehand are opting into training for jobs that require more skills, so they must stay in the program longer than others. It may also be the case that their reservation wage after the program is much higher than individuals who had not had a job prior to baseline. Their baseline jobs may have been of the lower-paying variety, so after the program, they expect an upgrade in their employment whereas those who had not previously held a job may opt to take lower-paying jobs after the program and become employed more quickly. Through evaluating treatment effect heterogeneity by aggregating ICEs, I find some heterogeneity that confirms the two theories of longer duration in the program and higher reservation wages leading to higher initial unemployment.

In Figure 3.10, I explored treatment effect heterogeneity by looking at LATE for for different groups of individuals based on covariates. The ICE framework also allows for exploring treatment effect heterogeneity in the reverse way by first dividing individuals into groups based on their ICEs and then comparing covariate information for the different groups. The two approaches are slightly different in that the first asks the question “What is the effect of job training for people that look a certain way (based on covariates)?” This second approach asks the question “What do people who benefited/were hurt from job training (in terms of employment) look like?”

In the study, the treatment effects can take on three possible values: 1, 0, and -1.[17] I first classify individuals into one of three effect categories: helped (1), no effect (0), or hurt (-1). I limit the analysis to compliers so that the effects are from the job training program itself. I also limit the analysis here to look only at employment in week 52. I then compare the mean value of the covariates for people in each effect category and see how they differ. To account for the uncertainty in the classifications and in compliance status, I repeat this process for each iteration of the MCMC. Specifically, for each iteration, $\tilde{\tau}_i$ classifies the individual i into an effect category. I then subset to compliers given the drawn compliance status $\tilde{G}$ and

[17] Note that given treatment assignment and the outcomes in the data, the possible values each ICE can take is constrained to two values out of 1, 0, and -1. For example, an individual that was assigned treatment and is employed can only have an ICE of either 0 or 1 since treatment assignment could not have hurt employment given that the individual got treatment and is employed.

record the mean value of the covariates for each of the three effect categories. I repeat this process for all iterations and the result is a series of vectors of covariate means, one vector for each covariate-effect category. The vectors represent the posteriors of the covariate means.

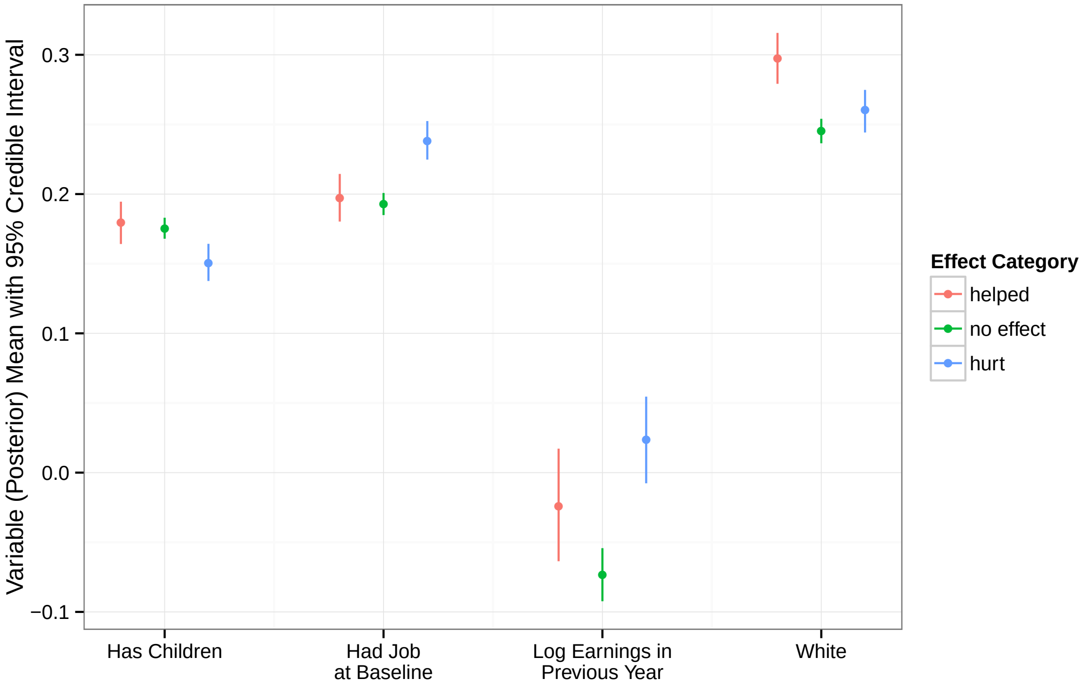

Figure 3.11: Comparing Covariates by Effect Category for Employment at Week 52

Figure 3.11 displays the the covariate means by effect category for four of the covariates. Recall that “helped” implies a positive effect of job training on employment at week 52 and “hurt” implies a negative effect. Around 18 percent of those that were helped by job training had children compared to 15 percent for those who were hurt. Those that were hurt by job training also were more likely to have held jobs at baseline and more likely to have had higher earnings in the year before baseline. Finally, those that were helped by job training were also more likely to be white. These results are all consistent with the previous hypotheses that individuals with children are more likely to benefit more quickly from job training in terms of finding employment while those with previous jobs are more likely to take longer to become employed, possibly due to staying longer in the program or holding a higher reservation wage.

The results presented here are largely consistent with the idea that the job training program actually works well in the long run in getting people employed. However, there is a short-term cost in terms of immediate employment. The ICE framework allows me to explore this result and find evidence that some individuals are more willing to bear this short-term cost whereas others are more likely to seek immediate employment after finishing the program. Further work can extend the ICE framework to address the issue of wages in conjunction with employment outcomes.

3.4 Conclusion

I have presented the results from two separate experimental studies related to monitoring corruption and job training. In both cases, the studies originally made a huge contribution to their respective fields. I use the ICE framework that I propose to mostly confirm those results, but I also demonstrate how the framework allows for a more flexible way to approach the problems. I show how to use ICEs to explore treatment effect heterogeneity and I contribute some interesting results that were not addressed in the original studies. The two applications allowed me to show how to adapt the ICE framework to different types of outcome and treatment variables and to embed it within the framework of instrumental variables and two-stage analyses. The ICE framework can be extended even further to account for all types of data structures and patterns. In the final chapter, I address some further extensions using ICEs and discuss some other remaining issues for future work.

Chapter 4

Concluding Remarks and Extensions

In the first three chapters, I have argued the case for a new approach to causal inference through the direct estimation of individual causal effects, laid out a framework to do so, presented a model to estimate the ICEs through a Bayesian approach with matching, showed that such a model can recover ICEs and other causal estimands through simulation, and demonstrated how the model works in two different applications. I now address some pertinent issues relating to the approach and also suggest some extensions and further applications of the model for future research.

4.1 Issues Relating to Matching

One of the crucial aspects of the model is the matching process, which chooses the donor observations that ultimately informs the imputation of the missing potential outcomes. The simulations I show suggest that predictive mean matching seems to generally work well across the simulated datasets. However, in any specific application, any number of other matching methods may perform even better. There is likely no single specification that dominates across all datasets. In the causal inference literature, the choice of matching specifications is often ad-hoc and ultimately can cause problems. Researchers can choose from the method to use, the number of matches, the weight of matching variables, etc. The practical solution is often to use a variety of specifications and to compare them on some balance metric to choose the best method. However, this process also has its pitfalls. The choice of balance metric is usually another specification itself. Also, the idea of choosing the best specification may be problematic since it assumes that the specification is the correct one and all others provide no additional information. Short of exact matching, it is unlikely that one specification is the correct one to use. In the case of estimating ICEs, the problems are magnified because it can be the case that one specification works well for certain individuals whereas another specification works well for other individuals. I have also not presented any methods for checking balance in estimating ICEs.

The framework I have presented allows for averaging of specifications by allowing the researcher to choose a different specification for each draw of the algorithm and even possibly for each individual within a draw. I believe this is a more appropriate way to do matching since it does not put all weight on a single specification and leverages the power of model averaging. However, as of now, the pool of specifications and the relative likelihood of choosing any one specification is completely up to the researcher. The prior for the matching specifications completely determines which specifications get used. Ideally, one would be able to calibrate the probabilities of matching specifications through information from the data. For example, for any individual, if some balance metric could be derived such that the specifications are used in proportion to how well they perform on the balance metrics, then the matching specifications would actually be “informed” by the data. This would likely improve the accuracy of the imputations. However, the process for developing a method to incorporate balance metrics within the current ICE algorithm is very computationally intensive and left for future research.

A second way to test the various matching specifications is to see how well they predict observed outcomes. Instead of using the matching to form donor pools to impute Y^{mis}, one can use the same process to predict the observed Y instead and see how well each of the specifications perform. This becomes analogous to a machine learning problem. In fact, the matching method that performs the best does not even have to be a causal inference matching algorithm at all. One can imagine using a myriad of the existing machine learning algorithms to estimate θ^{mis} using Y as the training and test outcomes.

4.2 Prediction and Population Extrapolation

Testing the performance of the matching algorithms brings up another point about predicting causal effects for individuals outside of the data. As I have alluded to before, ICEs are fundamentally in-sample estimands since there is data only about individuals in our dataset. However, the goal of statistics and causal inference is almost always to predict and generalize to out-of-sample individuals and datasets. Suppose the researcher is presented with the covariate vector for an out-of-sample individual and is asked to predict the individual causal effect of some treatment for this individual. How can the researcher adapt the ICE framework to make a prediction?

The simplest solution for the researcher is to think about the problem as needing to impute two missing potential outcomes, one for the hypothetical treatment and one for the hypothetical control. This involves matching to both in-sample control and treated units, creating two separate donor pools, modeling two separate means, and then drawing two separate Y^{mis}. The resulting ICE would be a predicted ICE for the out-of-sample individual based on the two imputed potential outcomes.

Another important issue related to prediction is also the issue of generalizing the results to some larger population. In most empirical work, the goal is to use the data to make inferences about some population. Usually, these population inferences rely on some assumptions that may or may not be explicit. For the purposes of the ICE framework, generalizing to a population would theoretically imply knowing covariate information for every individual in the population and then predicting each of their ICEs. However, since the ICE framework allows us to aggregate to calculate average effects in the sample, one can also use these estimates to generalize to the population given certain assumptions. The major assumption that is needed to generalize aggregated ICEs is a random sampling assumption. The sample that one estimates the ICEs on must be a representative sample of the population that one wants to generalize to. One can use population weights or other corrections in the data to meet these assumptions. Given the correct sampling assumption, one can say that the individuals in the data are similar to individuals in the population. And while one cannot say anything about ICEs simply based on this fact, one can say that the aggregated ICE average effects are good estimates of population average effects.

The main way in which estimates of population estimands differ from estimates of sample estimands is in the uncertainty estimates. The variance of population estimates is usually higher to account for the sampling uncertainty. In the ICE framework, one way to simulate this uncertainty to get more accurate population uncertainty estimates is through bootstrapping. There are two possible ways to do the bootstrapping. In the first, one can bootstrap the data first, run the ICE algorithm on the bootstrapped dataset, and then calculate the aggregated effects and repeat. The second way is to calculate the ICEs in the full dataset first, then bootstrap the ICEs themselves and aggregate and repeat. The second way uses all of the information in the dataset to do the matching and imputations while the first way only uses observations in the bootstrapped datasets for each bootstrap iteration. Future research should consider which of the two ways is a better choice for getting uncertainty estimates of population estimands.

4.3 Non-parametric Imputation

The Bayesian model for the imputation of the missing potential outcomes requires the researcher to specify a distribution to draw from. Additionally, modeling the mean and variance of the donor pools in the matching step requires at least two observations in the donor pool. If either of these requirements are not met, the researcher can still impute via a non-parametric approach. Instead of modeling the mean of the donor pool and then drawing from a specified distribution, the researcher can simply impute by drawing one of the observed outcomes in the donor pool as the imputation. This is analogous to multiple "hot-deck" imputation (Cranmer and Gill 2013). The assumption is that the empirical distribution of the donor pool is the discrete distribution that is used in the posterior predictive step. This also allows for 1-to-1 matching where the imputation is simply the outcome of the donor observation. A non-parametric approach may be more desirable if the researcher does not want to make any distributional assumptions. However, the tradeoff is that the researcher assumes the the outcome values of the donor pools are sufficient to characterize the distributions of the missing potential outcomes. If there are not enough distinct values for the donor pool outcomes (in the continuous case), then the posterior of the ICEs become very discrete.

4.4 Convergence

As with any MCMC simulation, convergence to the stationary distribution is necessary and must be checked. The algorithm I propose really only contains dependence among parameters at the matching step (the imputation is only dependent on D, which is dependent on the matching parameters), so non-convergence may be less of an issue than typical MCMC simulations with high dependence among parameters. Nevertheless, convergence should be checked. Unfortunately, the number of parameters in the model is greater than or equal to the number of observations in the data, so checking convergence on each one is tedious at best. However, it is also necessary to check each parameter as non-convergence on even one parameter may be problematic for all the results (Gill 2008). Further research should be done on ways to test convergence on a large number of parameters. I defer to the vast literature on convergence diagnostics for this. However, one suggestion is that researchers can check convergence on the aggregations of the ICEs. For example, if checking convergence on all N ICEs proves to be too tedious, one can check convergence on the aggregated draws of the ATE or the ATT. If the ATE draws do not seem to converge, then this indicates that one or more of the ICEs have not converged. Unfortunately, the inverse is not true. Convergence on the ATE does not necessarily imply convergence on all the ICEs.

4.5 Extensions

4.5.1 Incorporating ICEs into (almost) any possible (causal) model

One benefit of the idea of estimating ICEs is that researchers can incorporate ICEs and potential outcome imputations into nearly any type of causal model that one can run. The potential outcomes framework is a powerful framework that clearly specifies the research design and the problem at hand. The ICE framework simply builds off the potential outcomes framework. Then any regression model, no matter how sophisticated, is really a means to estimate parameters in the potential outcomes framework. Often, including ICEs in a more sophisticated regression model simply boils down to choosing the right relevant set of donor observations.

Consider the fixed effects regression model often used in economics and political science.

$$Y_{ik} = \alpha_k + X_{ik}\beta_k + \epsilon_i$$

where k denotes a certain cluster (for example, countries). This fixed effects model estimates different intercept and (possibly) slope terms for each cluster. Another way to conceptualize the goal of fixed effects models is simply to match observations only within clusters (Imai and Kim 2013). Within the ICE framework, this simply boils down to limiting the potential donor pool within each iteration to observations in the same cluster and then aggregating by cluster to get the cluster-specific intercepts and slopes. In more complicated multilevel models, one can simply impute the missing potential outcomes and then aggregate either on a first or second level variable to get specific causal effects.

Now consider complicated regression models that attempt to model time components. Often, such models boil down to including a lagged dependent variable or other terms on the right-hand side of the regression equation. In the matching framework, this simply means adding a variable to the matching specification. To be more precise, the researcher can set the matching algorithm to exact match on certain variables, which is again simply an adjustment on the potential donor pool. More complicated time-dependent models may include certain parametric specifications, such as a spline to account for time in binary dependent variable models (Beck, Katz and Tucker 1998). Researchers can include such specifications either during or after matching to adjust the imputations. One way would be to run a regression within the donor pool using only the spline variables to estimate θ^{mis}.

The general idea is that including matching and reframing causal inference at the level of ICEs is compatible with almost any existing method. Furthermore, I argue that it has the added benefit of forcing researchers to seriously consider the causal quantities they are estimating by being explicit about modeling individuals. Adding a spline may be simple to implement in any statistical package, but forcing the researcher to understand that the spline simply models how other observations in different time periods contribute to the missing potential outcome of a certain observation of interest is valuable in promoting the understanding of the role of regression models in causal inference.

4.5.2 ICEs and Causal Inference Assumptions

Another benefit of working with ICEs and a possible avenue for extending the framework is through the testing and relaxing of typical causal inference assumptions. As I alluded to in the examples using a two-stage model model with instrumental variables, the typical exclusion restriction can be tested using the ICE framework by estimating the non-complier average treatment effect (NCATE). If the assumption of the exclusion restriction were correct, then the NCATE should be zero. In the job training example, using the ICE framework to estimate NCATE, I found that there may be some reason to doubt the exclusion restriction that is typically assumed.

Consider also the conventional SUTVA assumption that is required in almost all causal inference studies. The SUTVA assumption has two parts:

1. Treatment assignment on one observation does not affect the potential outcomes of another observation.
2. No varying treatment intensity.

The second part of the assumption may be violated, for example, if individuals assigned to a drug can take either a regular strength or extra strength version. In the corruption monitoring example, the participation treatment actually violated the second part of the assumption since villages received either invitations only or invitations and comment forms. For some parts of the analyses, I assumed that the two were the same. However, since the ICE framework results in a Bayesian posterior, I can actually make probability statements about how true the assumption actually is. Let village i be assigned control (neither invites nor invites and comments). Suppose I then impute the potential outcome for i being assigned invites (by matching to villages that received invites only) and then I also impute the potential outcome for i being assigned invites and comments (by matching to villages that received both invites and comments). Then I would have two posteriors for the two potential outcomes of receiving the two different versions of the participation treatment. I can then compare the two posteriors and calculate the probability that $\tau_i^{(inv)} = \tau_i^{(inv\&com)}$. This would be an estimate of the probability that the second part of SUTVA holds for village i. I can do the same calculation for all i and have a sense of how likely SUTVA is violated. More research must be done into how much to trust the results of such an analysis, but it at leasts suggests potential for testing the sensitivity of certain studies to certain assumptions.

Another key assumption that is often made in causal inference with instrumental variables is the monotonicity assumption. Let Z be an instrument for W with outcome Y. The monotonicity assumption (also called the no defiers assumption) states that $W_i(1) \geq W_i(0)$. In simple terms, the assumption is that there are no individuals who would take the treatment when assigned control but not take the treatment when assigned treatment. In many situations, this assumption makes sense. However, for certain studies, this assumption is very important and may not be fully satisfied. Consider the now famous paper on the effect of institutions on economic performance by Acemoglu, Johnson and Robinson (2001). In that paper, the authors use settler mortality as an instrument for extractive institutions. The theory states that in countries where settler mortality was high, settlers built extractive institutions since they did not settle there themselves. In countries with low settler mortality, the settlers actually installed less extractive and "better" institutions for economic growth. The author use this design to conclude that institutions matter in economic growth.

To simplify the analysis, let settler mortality (Z) and non-extractive institutions (W) both be measured with binary variables. The monotonicity assumption states that the relationship between settler mortality and institutions can only go one way for all countries. No country exists that would establish extractive institutions with low settler mortality but non-extractive institutions with high settler mortality. However, this assumption is fundamentally untested and given the myriad of variables that interact with both Z and W, it is conceivable that the monotonicity assumption could be violated. For example, one can make the case that the relationship would depend on which groups of settlers were affected the most by mortality. Suppose the settlers can be partitioned into "royalists" who supported the Crown and "colonialists" who supported more independent institutions. If mortality affected the royalist camp disproportionately, then it could be the case that increasing mortality actually increases the odds of less extractive institutions while absent high mortality, the royalists have enough political power to enact extractive institutions. This is but one possible scenario in which the monotonicity assumption is violated. Using the ICE framework, one can actually relax the monotonicity assumption and estimate the probability that a country is a "defier" country. By allowing for defiers and jointly estimating the compliance group memberships, one can get a better estimate for LATE and also estimate a defier average treatment effect.

By thinking about causal inference at the individual level and estimating ICEs, researchers are given tools to think about the assumptions they make and relax some of the assumptions or test the sensitivity of their results.

4.6 Final Words

In this dissertation, I have presented an argument for why researchers should shift their focus from estimating average effects to estimating individual effects. I am in no way arguing that existing methods for average effects should no longer be used. I believe that estimating ICEs in conjunction with existing methods can produce great results. The contribution of the dissertation is in opening up new avenues and helping scholars rethink how existing methods fit into the causal inference framework using potential outcomes. The algorithm and the models themselves are very much works in progress. There are other areas that my work touch on, such as the merging of matching and Bayesian methods, the use of model averaging with matching, or the idea that complicated regression models can be integrated with a matching approach. All of these areas deserve much more future research. I simply hope that my dissertation will spur more interest in how to deal with treatment effect heterogeneity and how to reconcile small n research with large n studies as well as provide a unified and straightforward framework for thinking about causal inference.

Bibliography

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105(490):493–505.

Abadie, Alberto and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1):235–267.

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review* 91(5):1369–1401.

An, Weihua. 2010. "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference." *Sociological Methodology* 40(1):151–189.

Arceneaux, Kevin and David W. Nickerson. 2009. "Who Is Mobilized to Vote? A Re-Analysis of 11 Field Experiments." *American Journal of Political Science* 53(1):1–16.

Beck, Nathaniel, Jonathan N. Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42(4):1260–1288.

Cranmer, Skyler J. and Jeff Gill. 2013. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science* 43(2):425–449.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens and Oscar A. Mitnik. 2008. "Nonparametric Tests for Treatment Effect Heterogeneity." *The Review of Economics and Statistics* 90(3):389–405.

Diamond, Alexis and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* . Forthcoming.

Dorresteijn, Johannes A., Frank L. Visseren, Paul M. Ridker, Annemarie M. Wassink, Nina P. Paynter, Ewout W. Steyerberg, Yolanda van der Graaf and Nancy R. Cook. 2011. "Estimating Treatment Effects for Individual Patients Based on the Results of Randomised Clinical Trials." *British Medical Journal* 343:d5888.

Feller, Avi and Chris C. Holmes. 2009. "Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments." Unpublished.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1):21–29.

Frölich, Markus. 2007. "Propensity Score Matching without Conditional Independence Assumption with an Application to the Gender Wage Gap in the United Kingdom." *Econometrics Journal* 10:359–407.

Frumento, Paolo, Fabrizia Mealli, Barbara Pacini and Donald B. Rubin. 2012. "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data." *Journal of the American Statistical Association* 107(498):450–466.

Gadbury, Gary L., Hari K. Iyer and Jeffrey M. Albert. 2004. "Individual Treatment Effects in Randomized Trials with Binary Outcomes." *Journal of Statistical Planning and Inference* 121(1):163–174.

Gaines, Brian J. and James H. Kuklinski. 2011. "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection." *American Journal of Political Science* 55(3):724–736.

Gill, Jeff. 2008. "Is Partial-Dimension Convergence a Problem for Inferences from MCMC Algorithms?" *Political Analysis* 16(2):153–178.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Gutman, Roee and Donald B. Rubin. 2012. "Robust Estimation of Causal Effects of Binary Treatments in Unconfounded Studies with Dichotomous Outcomes." *Statistics in Medicine* .

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hansen, Ben B. 2008. "The Prognostic Analogue of the Propensity Score." *Biometrika* 95(2):481–488.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition ed. New York: Springer.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Iacus, Stefano M., Gary King and Giuseppe Porro. 2012. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20(1):1–24.

Imai, Kosuke and Aaron Strauss. 2011. "Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-out-the-vote Campaign." *Political Analysis* 19(1):1–19.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstanding Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society Series A* 171(2):481–502.

Imai, Kosuke and In Song Kim. 2013. "On the Use of Linear Fixed Effects Regression Models for Causal Inference." Unpublished.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* . Forthcoming.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87(3):706–710.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86(1):4–29.

Imbens, Guido W. and Donald B. Rubin. 1997. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *The Annals of Statistics* 25(1):305–327.

Jin, Hui and Donald B. Rubin. 2008. "Principal Stratification for Causal Inference with Extended Partial Compliance." *Journal of the American Statistical Association* 103(481):101–111.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

Kravitz, Richard L., Naihua Duan and Joel Braslow. 2004. "Evidence-Based Medicine, Heterogeneity of Treatment Effects and the Trouble with Averages." *The Millbank Quarterly* 82(4):661–687.

Lagakos, Stephen W. 2006. "The Challenge of Subgroup Analyses - Reporting without Distorting." *The New England Journal of Medicine* 354(16):1667–1669.

Little, Roderick J. A. and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: Wiley.

McCullagh, Peter and John A. Nelder. 1989. *Generalized Linear Models.* Second edition ed. New York: Chapman and Hall.

Ming, Kewei and Paul R. Rosenbaum. 2000. "Substantial Gains in Bias Reduction from Matching with a Variable Number of Controls." *Biometrics* 56(1):118–124.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2):200–249.

Pattanayak, Cassandra W., Donald B. Rubin and Elizabeth R. Zell. 2012. "A Potential Outcomes, and Typically More Powerful, Alternative to "Cochran-Mantel-Haenszel"." Working Paper.

Pocock, Stuart J., Susan E. Assmann, Laura E. Enos and Linda E. Kasten. 2002. "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems." *Statistics in Medicine* 21(19):2917–2930.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111–163.

Rosenbaum, Paul R. 1984. "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society Series A* 147(5):656–666.

Rosenbaum, Paul R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society Series B* 53(3):597–610.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387):516–524.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias Due to Incomplete Matching." *Biometrics* 41(1):103–116.

Rothwell, Peter M. 2005. "Subgroup Analysis in Randomised Controlled Trials: Importance, Indications and Interpretation." *The Lancet* 365(9454):176–186.

Rubin, Donald B. 1973*a*. "Matching to Remove Bias in Observational Studies." *Biometrics* 29(1):159–183.

Rubin, Donald B. 1973*b*. "The Use of Matching Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29(1):185–203.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5):688–701.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6(1):34–58.

Rubin, Donald B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74(366):318–328.

Rubin, Donald B. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36(2):293–298.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, Donald B. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2:169–188.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100(469):322–331.

Rubin, Donald B. 2008. Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics. In *Handbook of Statistics*, ed. C.R. Rao, J. Philip Miller and D.C. Rao. Vol. 27 Elsevier pp. 28–63.

Rubin, Donald B. and Neal Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95(450):573–585.

Rubin, Donald B. and Richard P. Waterman. 2006. "Estimating the Causal Effects of Marketing Interventions Using Propensity Score." *Statistical Science* 21(2):206–222.

Steyer, Rolf. 2005. "Analyzing Individual and Average Causal Effects via Structural Equation Models." *Methodology* 1(1):39–54.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1):1–21.

Tanner, Martin A. and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82(398):528–540.

Appendix A

Extra Simulation Results

A.1 Comparing Methods for Continuous Outcomes

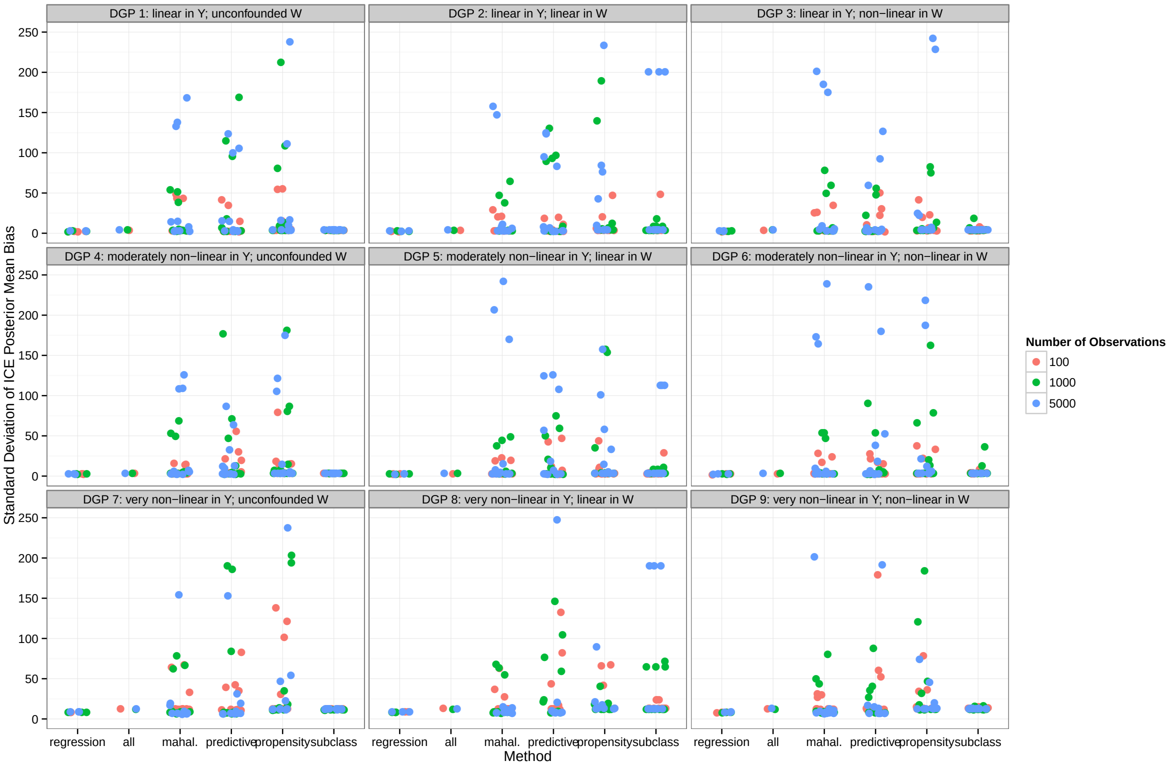

Figure A.1: Comparing Standard Deviations of ICE Posterior Mean Bias for the Different Matching Methods (continuous outcome)

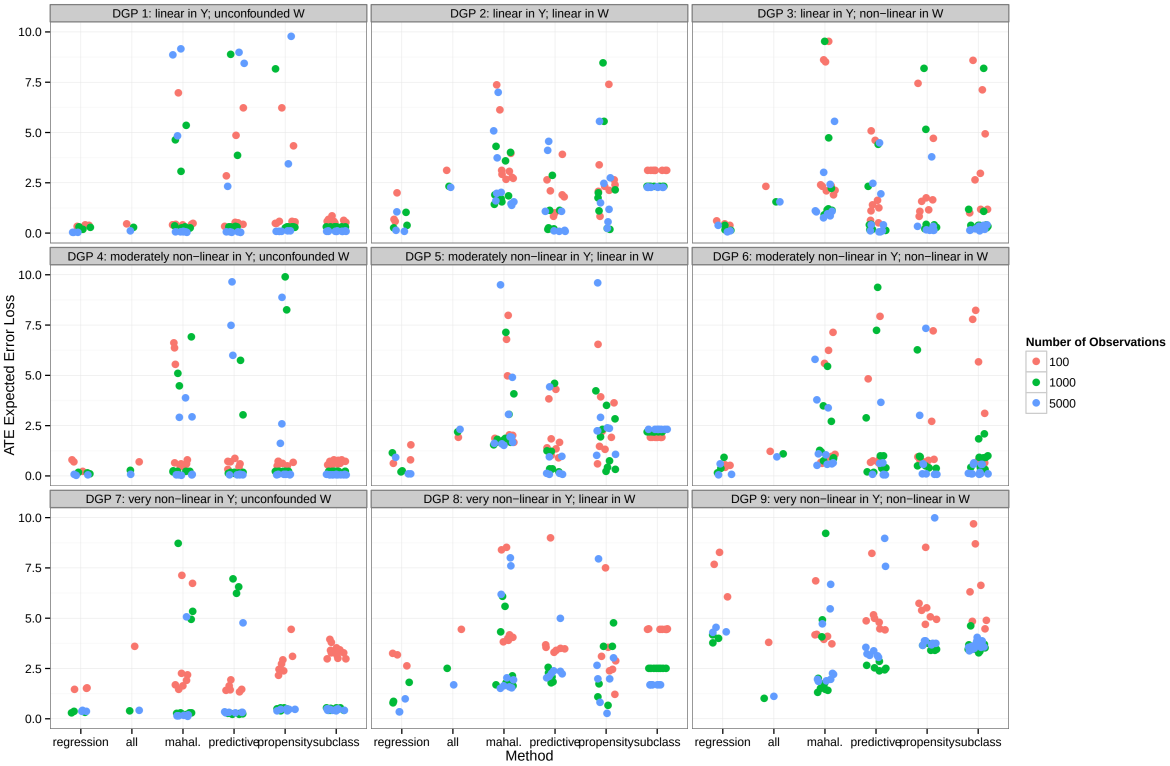

Figure A.2: Comparing ATE Expected Error Loss for the Different Matching Methods (continuous outcome)

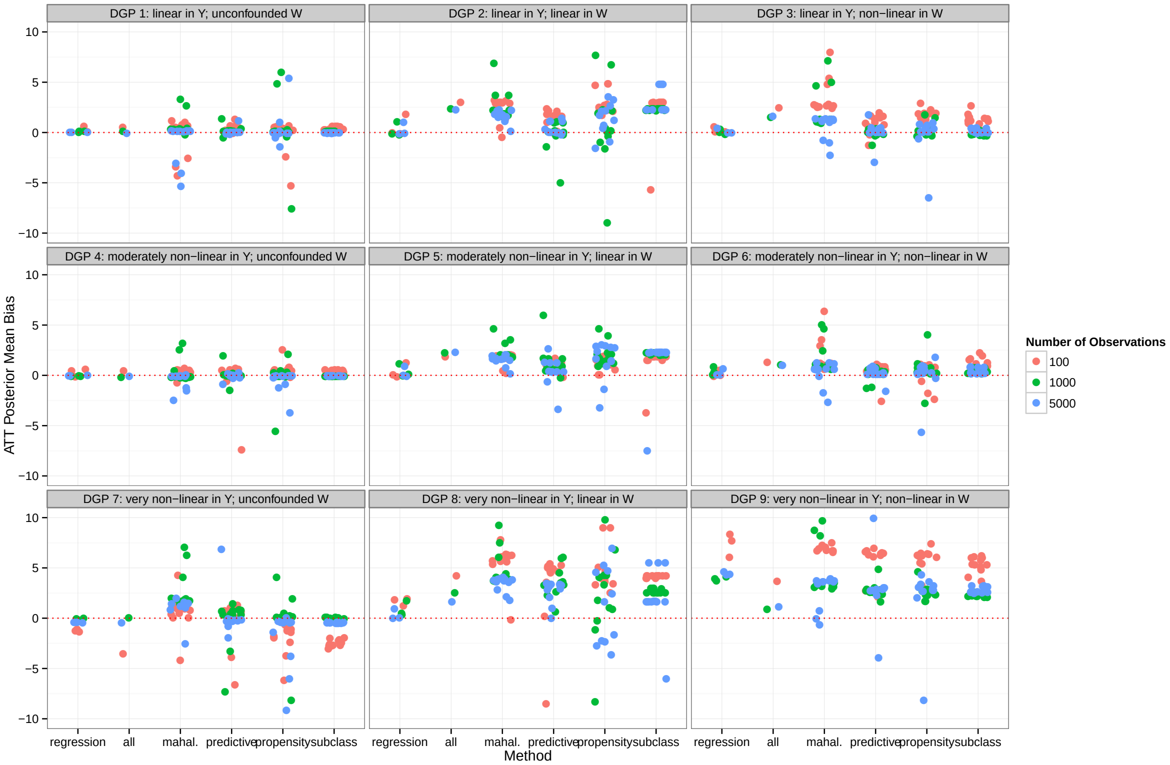

Figure A.3: Comparing ATT Posterior Mean Bias for the Different Matching Methods (continuous outcome)

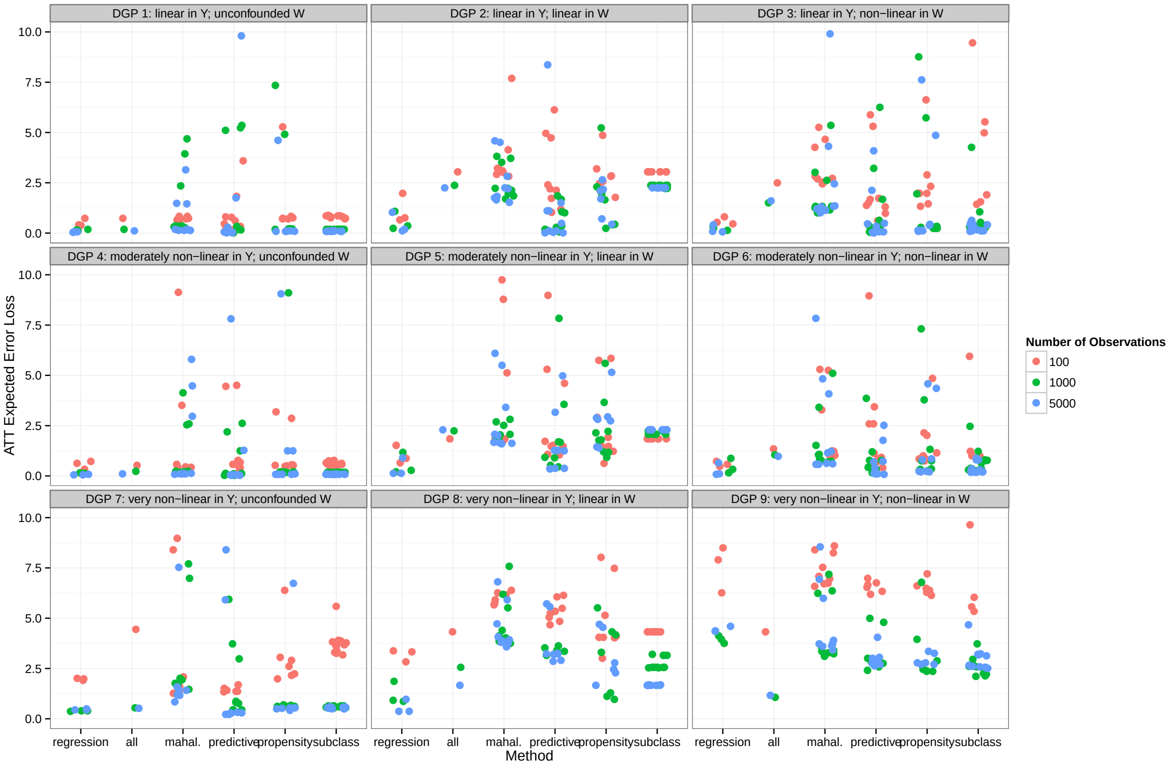

Figure A.4: Comparing ATT Expected Error Loss for the Different Matching Methods (continuous outcome)

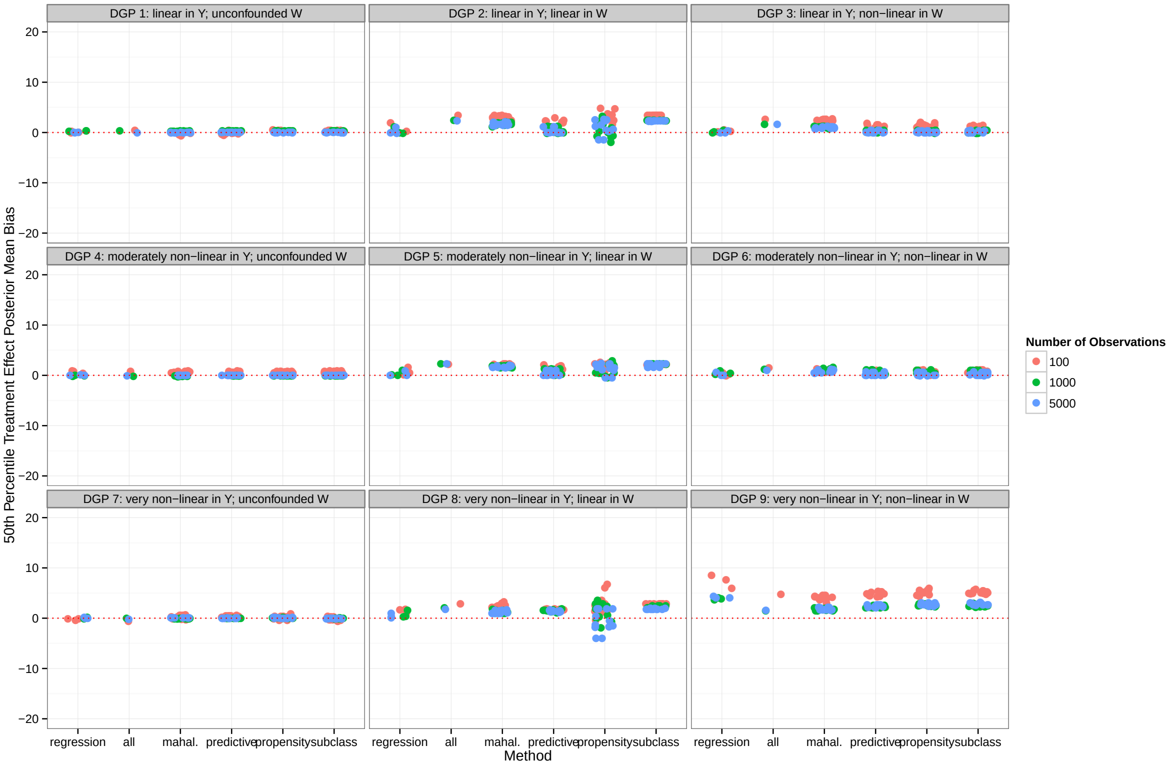

Figure A.5: Comparing 50th Percentile Treatment Effect Posterior Mean Bias for the Different Matching Methods (continuous outcome)

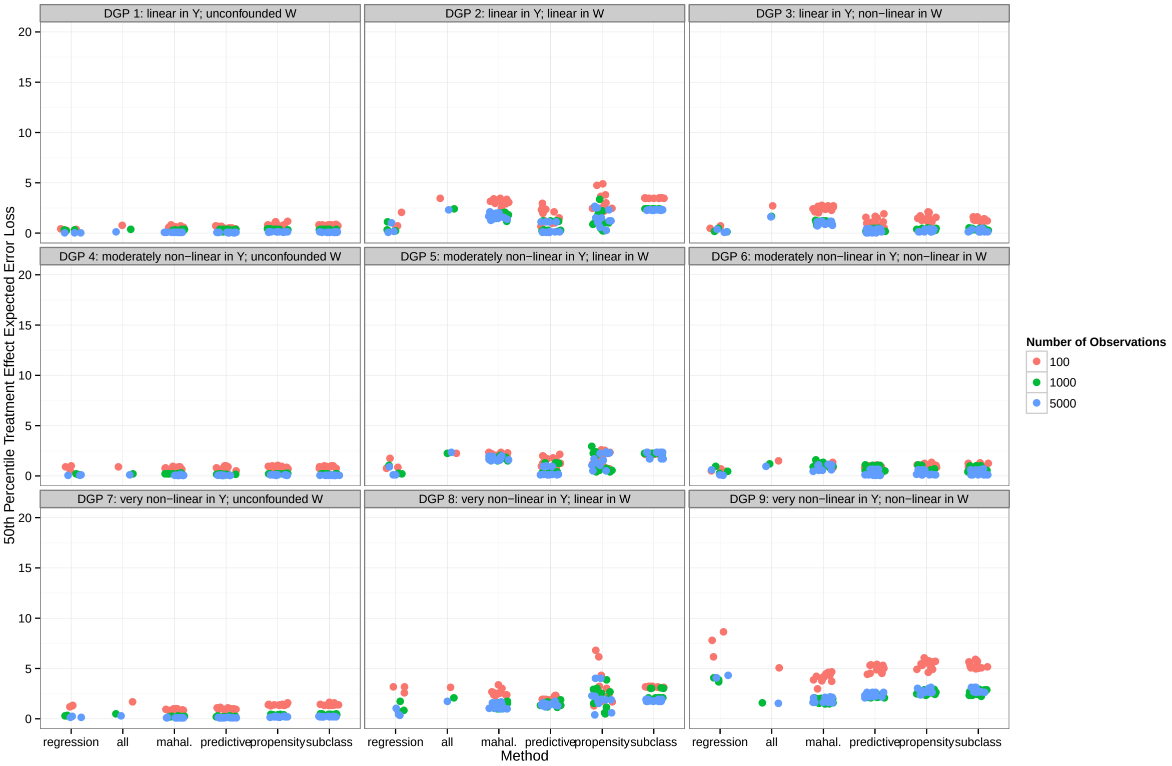

Figure A.6: Comparing 50th Percentile Treatment Effect Expected Error Loss for the Different Matching Methods (continuous outcome)

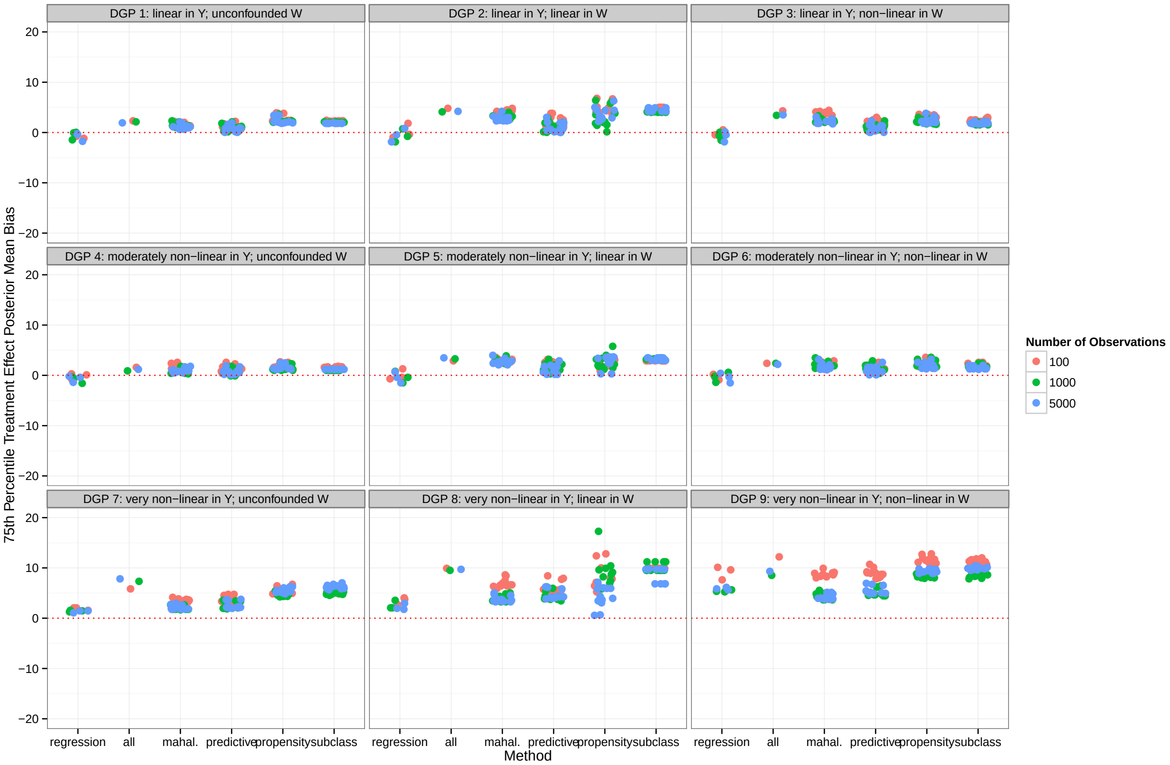

Figure A.7: Comparing 75th Percentile Treatment Effect Posterior Mean Bias for the Different Matching Methods (continuous outcome)

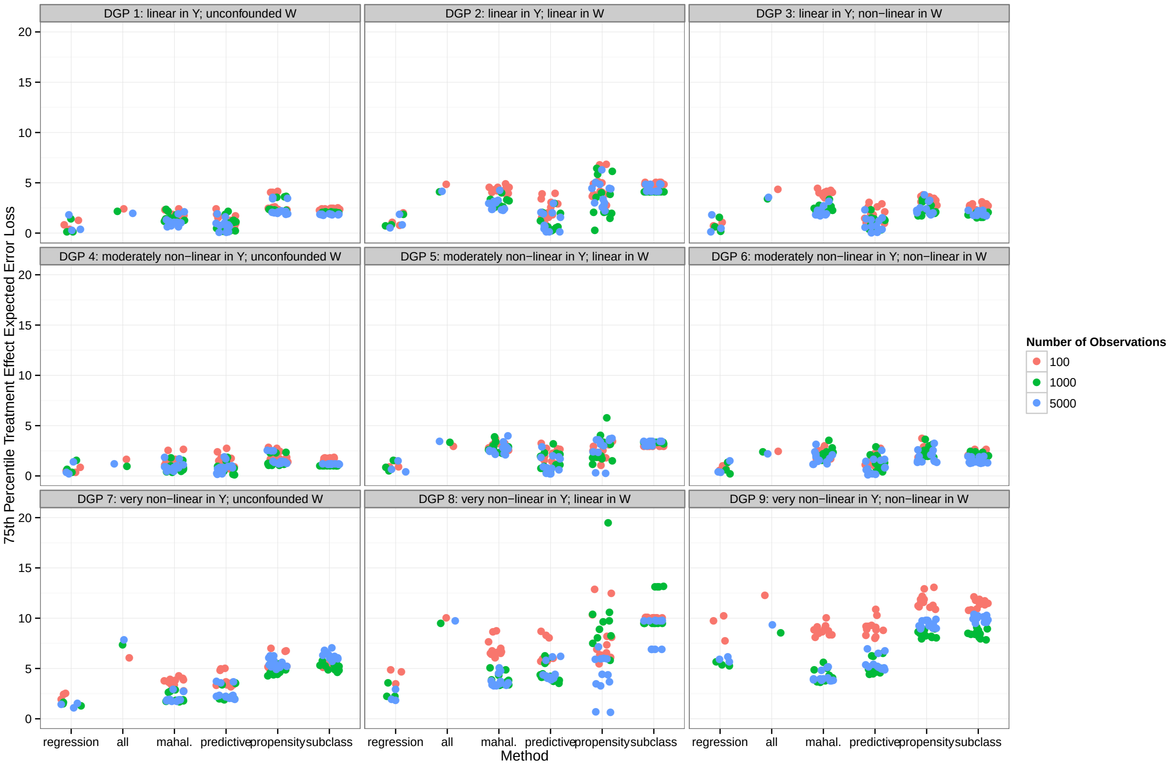

Figure A.8: Comparing 75th Percentile Treatment Effect Expected Error Loss for the Different Matching Methods (continuous outcome)

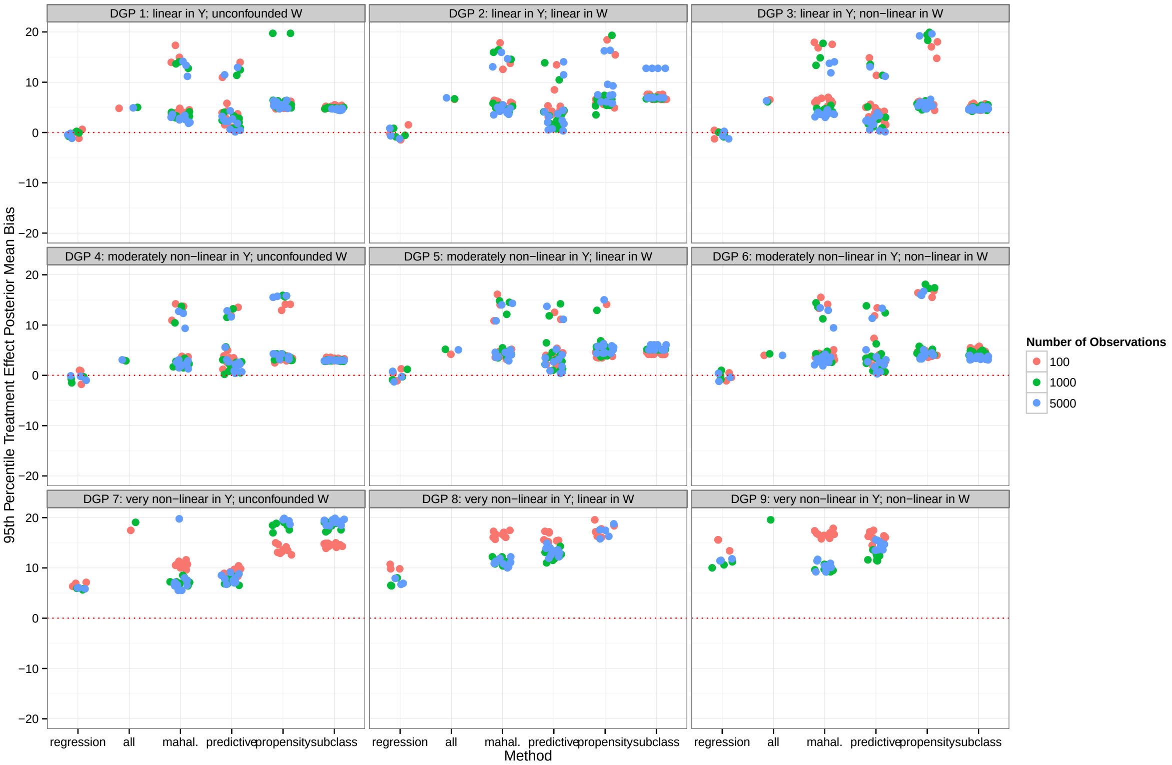

Figure A.9: Comparing 95th Percentile Treatment Effect Posterior Mean Bias for the Different Matching Methods (continuous outcome)

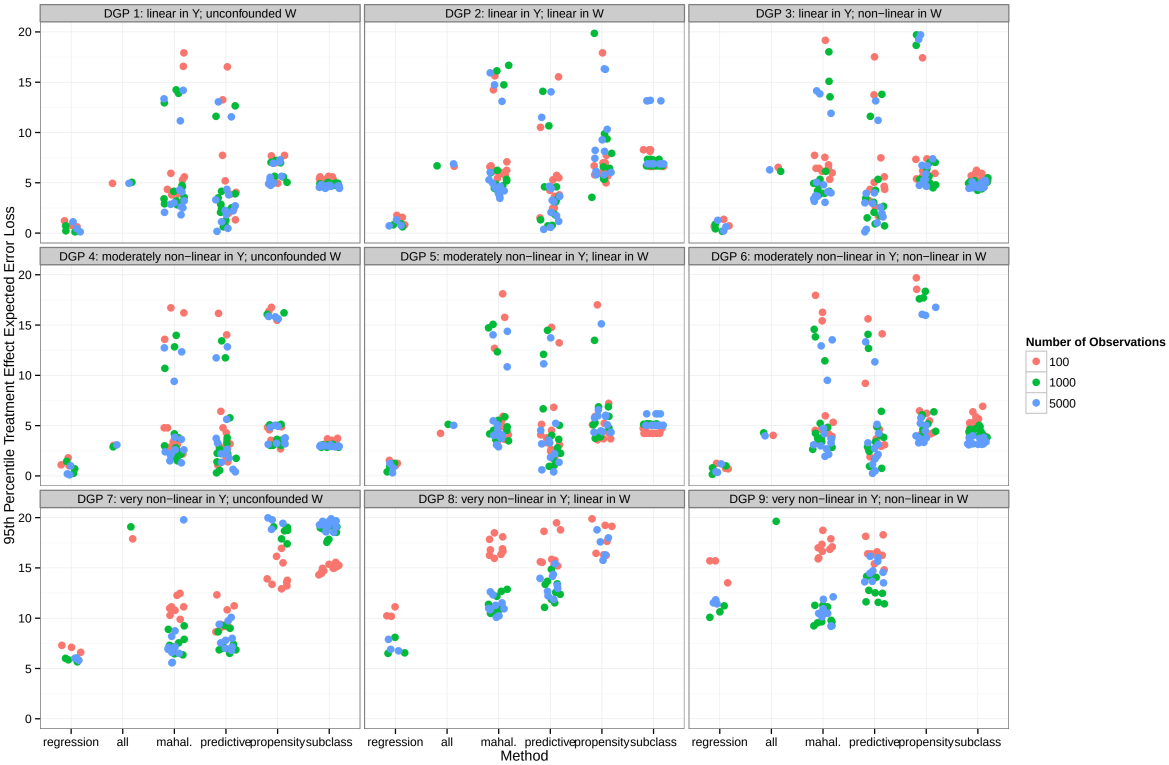

Figure A.10: Comparing 95th Percentile Treatment Effect Expected Error Loss for the Different Matching Methods (continuous outcome)

A.2 Comparing Number of Conditioning Variables for Continuous Outcomes

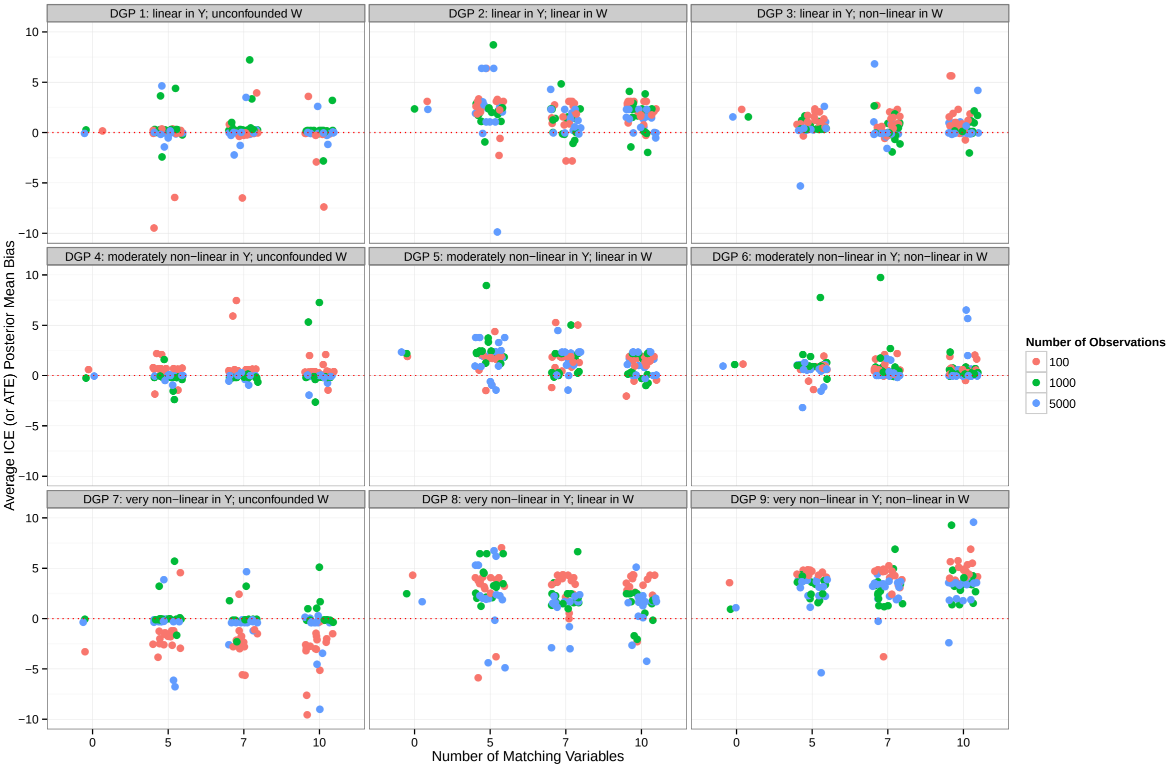

Figure A.11: Comparing Average ICE (or ATE) Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

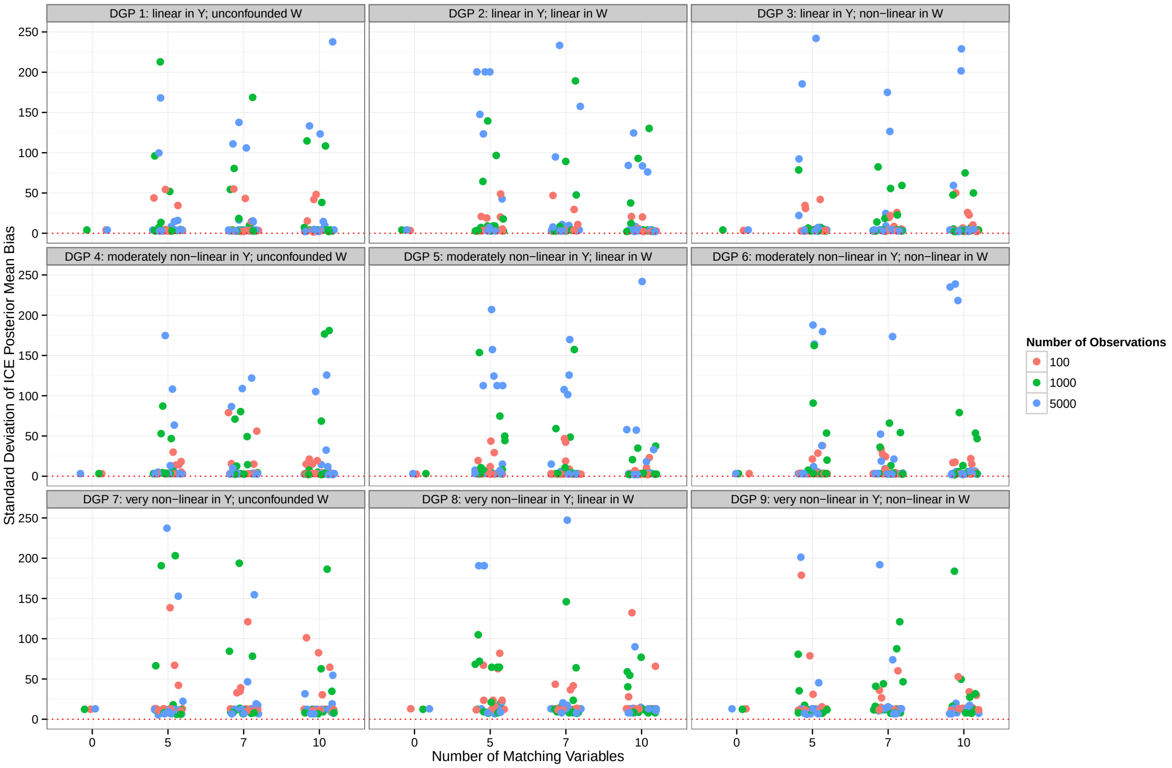

Figure A.12: Comparing Standard Deviations of ICE Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

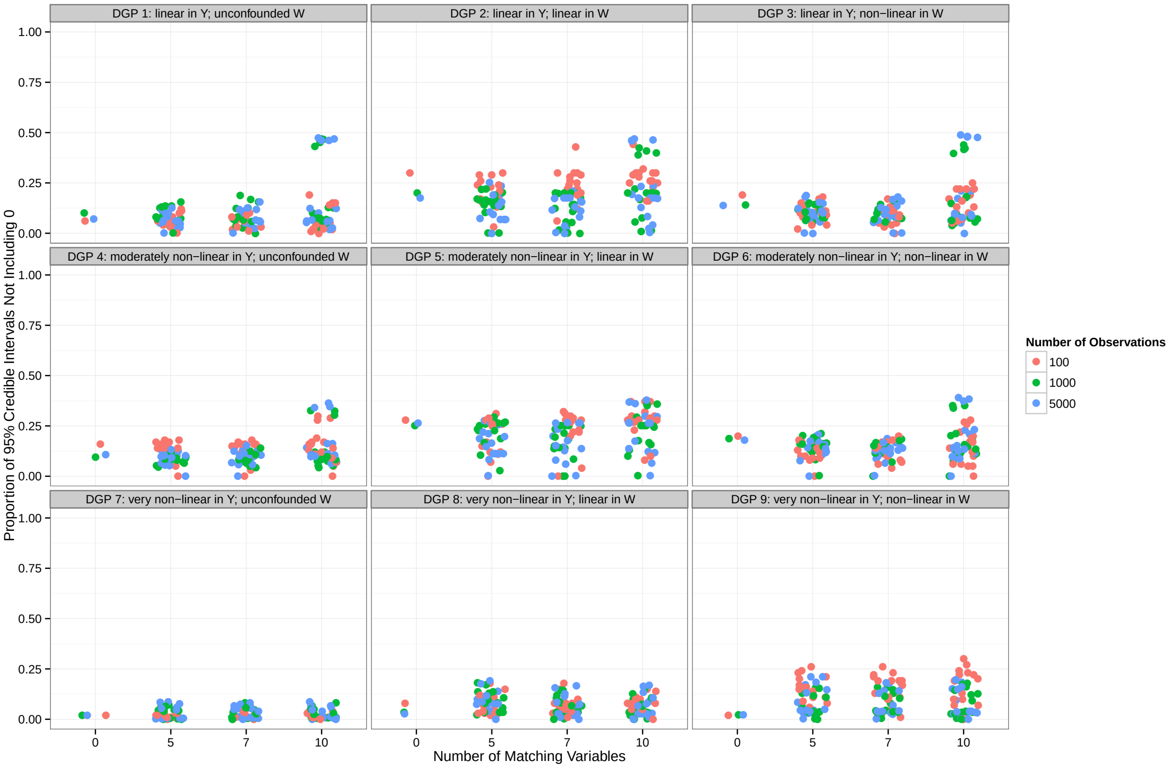

Figure A.13: Comparing ICE "Power" for Different Conditioning Sets (continuous outcome)

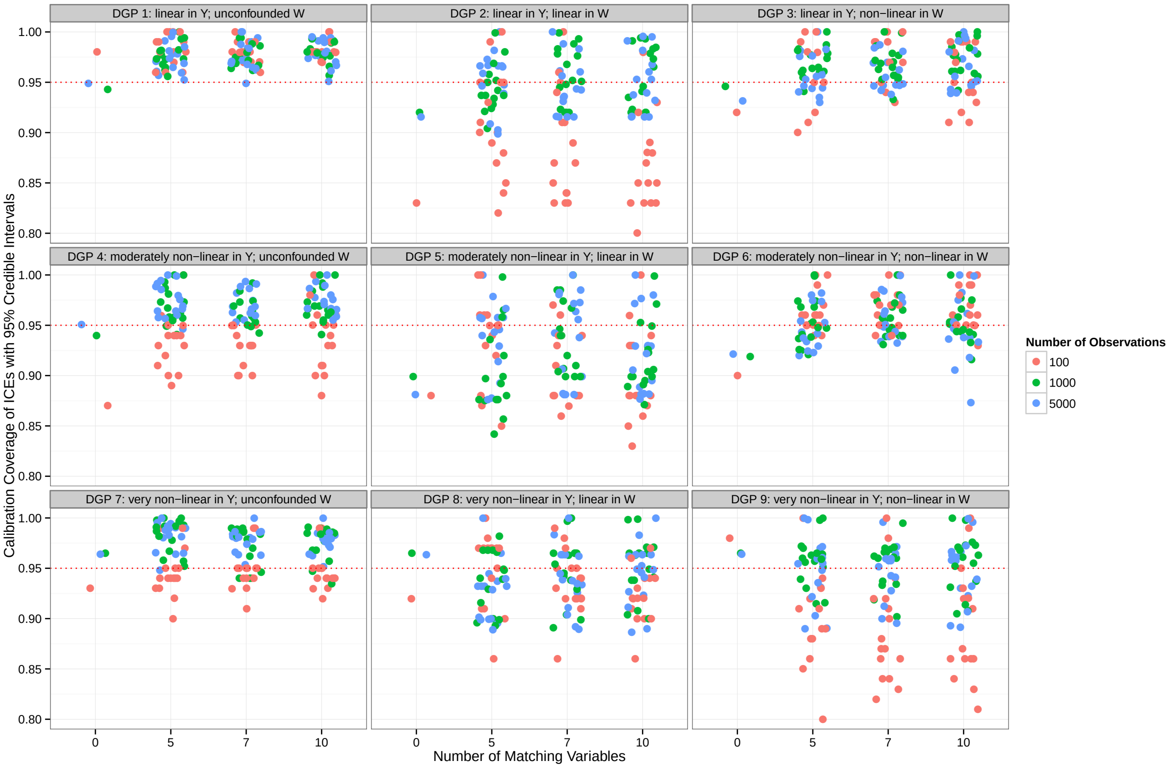

Figure A.14: Comparing ICE Calibration Coverage for Different Conditioning Sets (continuous outcome)

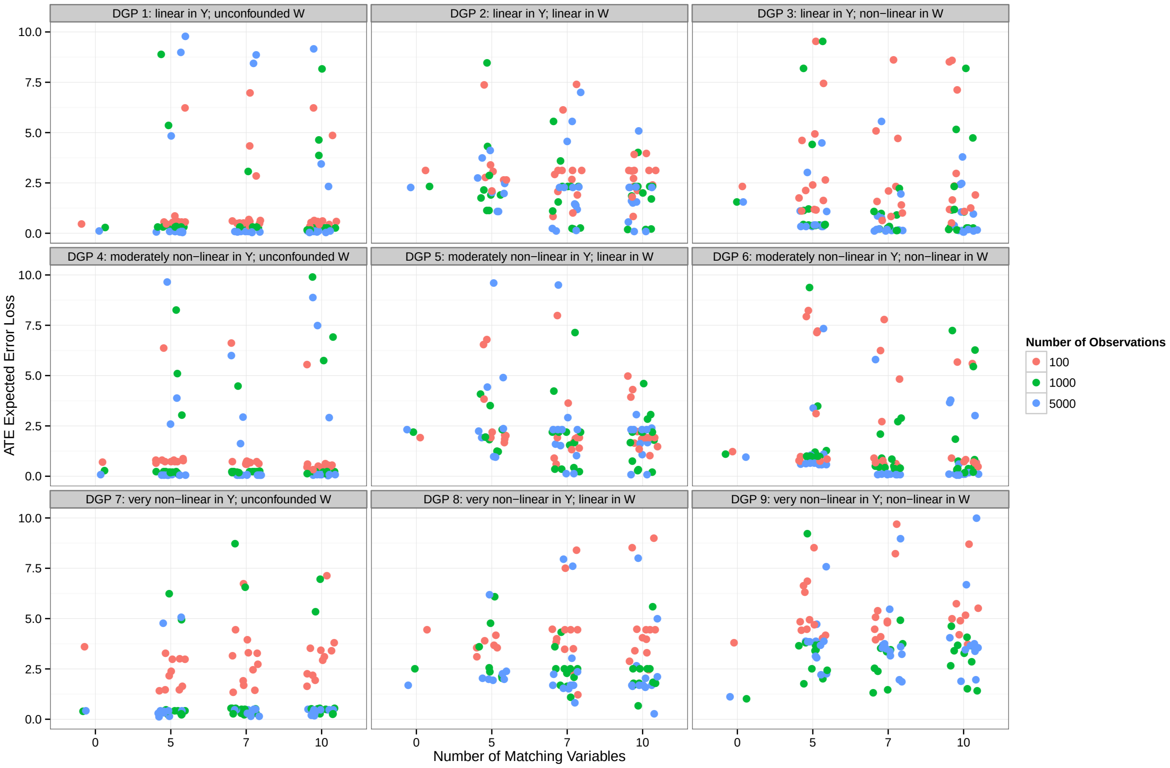

Figure A.15: Comparing ATE Expected Error Loss for Different Conditioning Sets (continuous outcome)

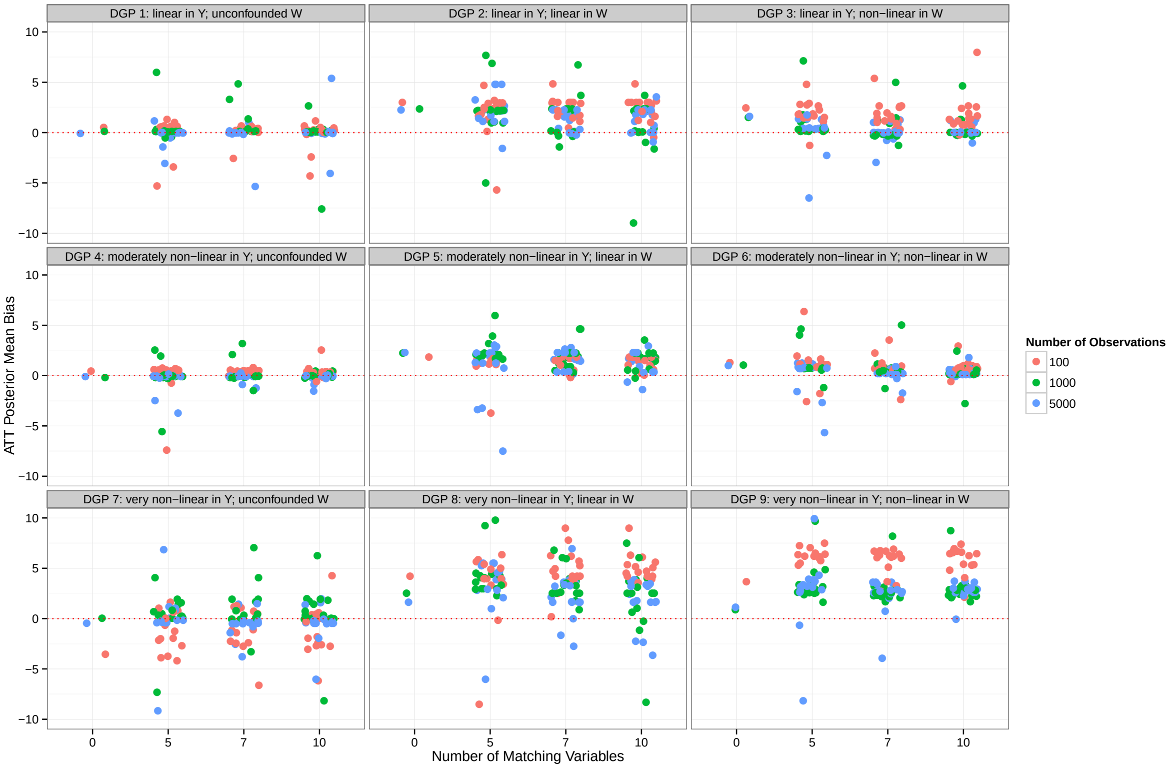

Figure A.16: Comparing ATT Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

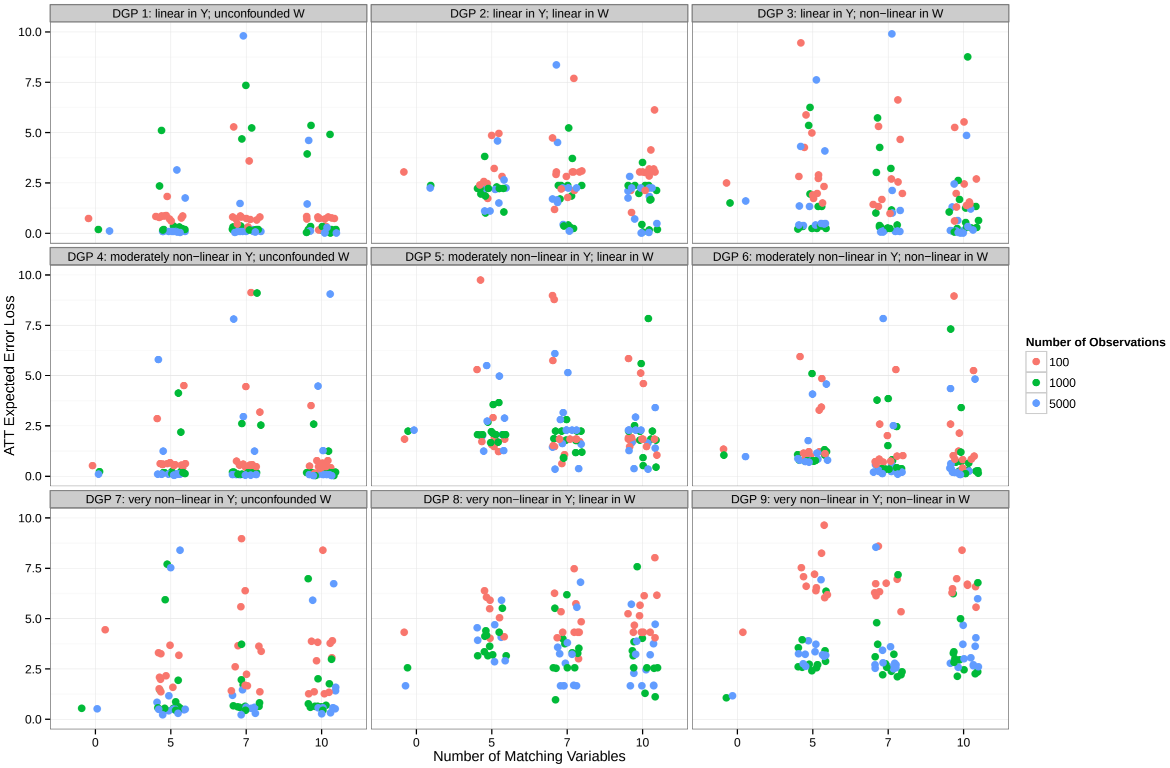

Figure A.17: Comparing ATT Expected Error Loss for Different Conditioning Sets (continuous outcome)

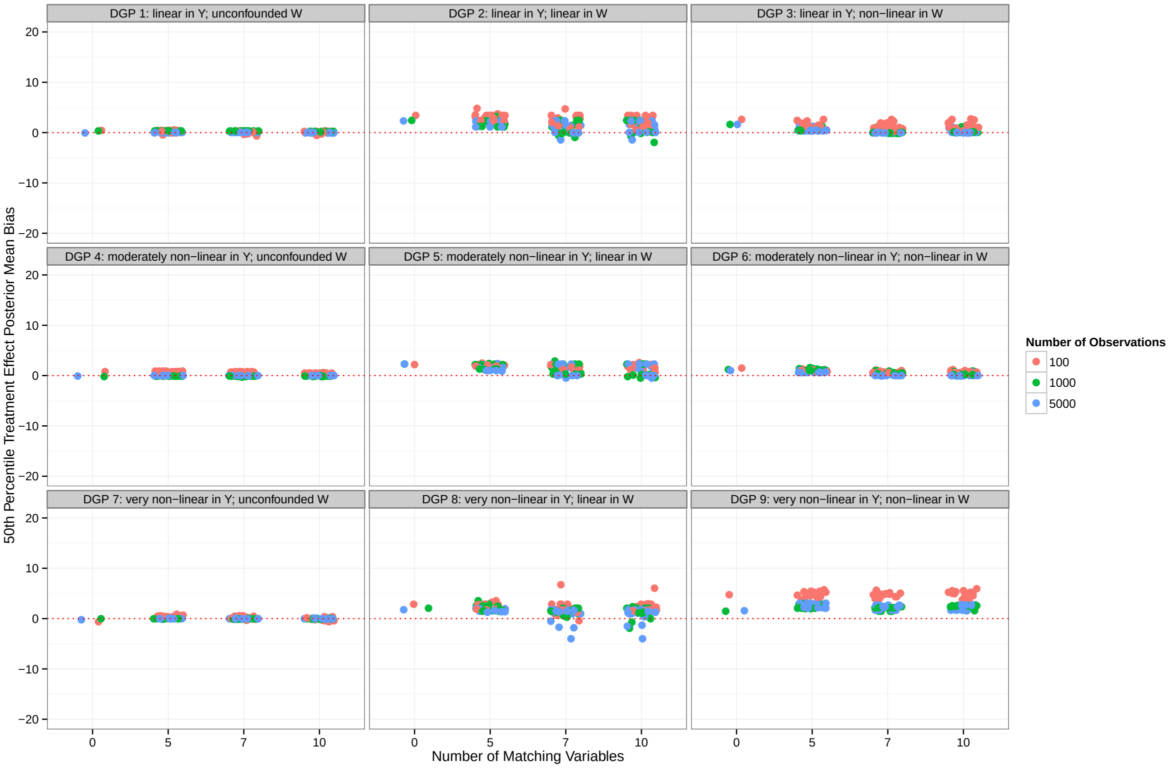

Figure A.18: Comparing 50th Percentile Treatment Effect Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

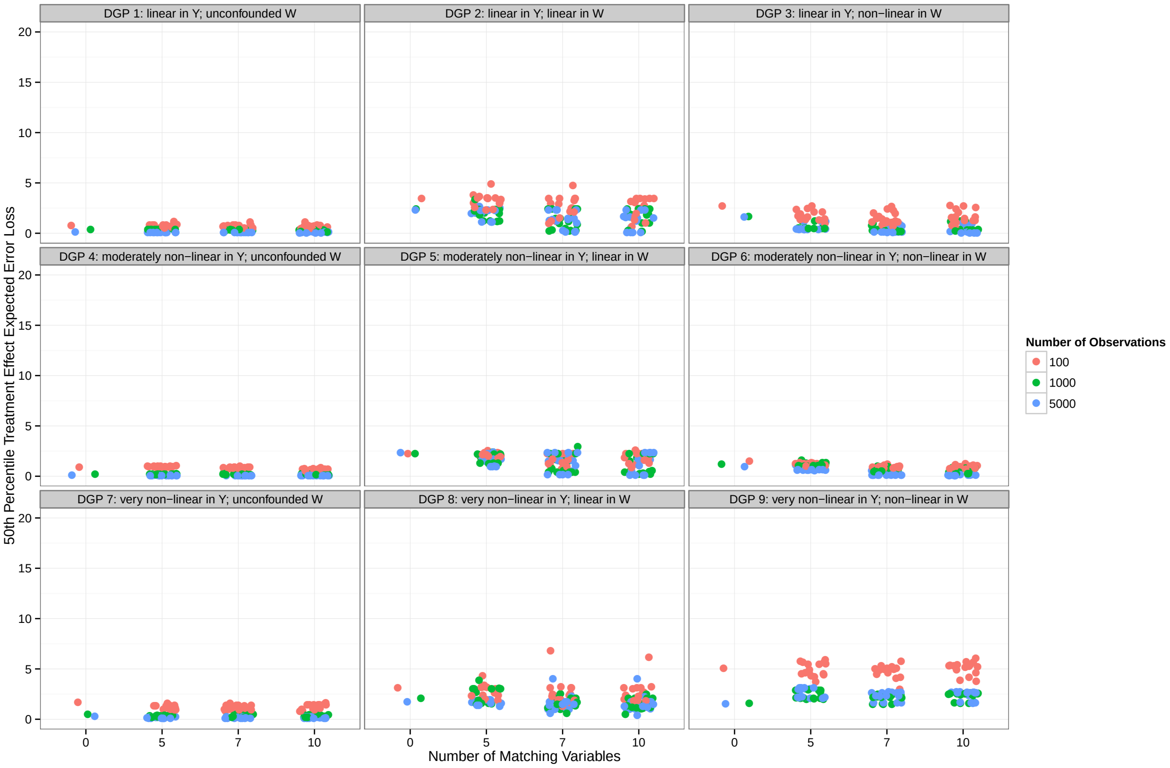

Figure A.19: Comparing 50th Percentile Treatment Effect Expected Error Loss for Different Conditioning Sets (continuous outcome)

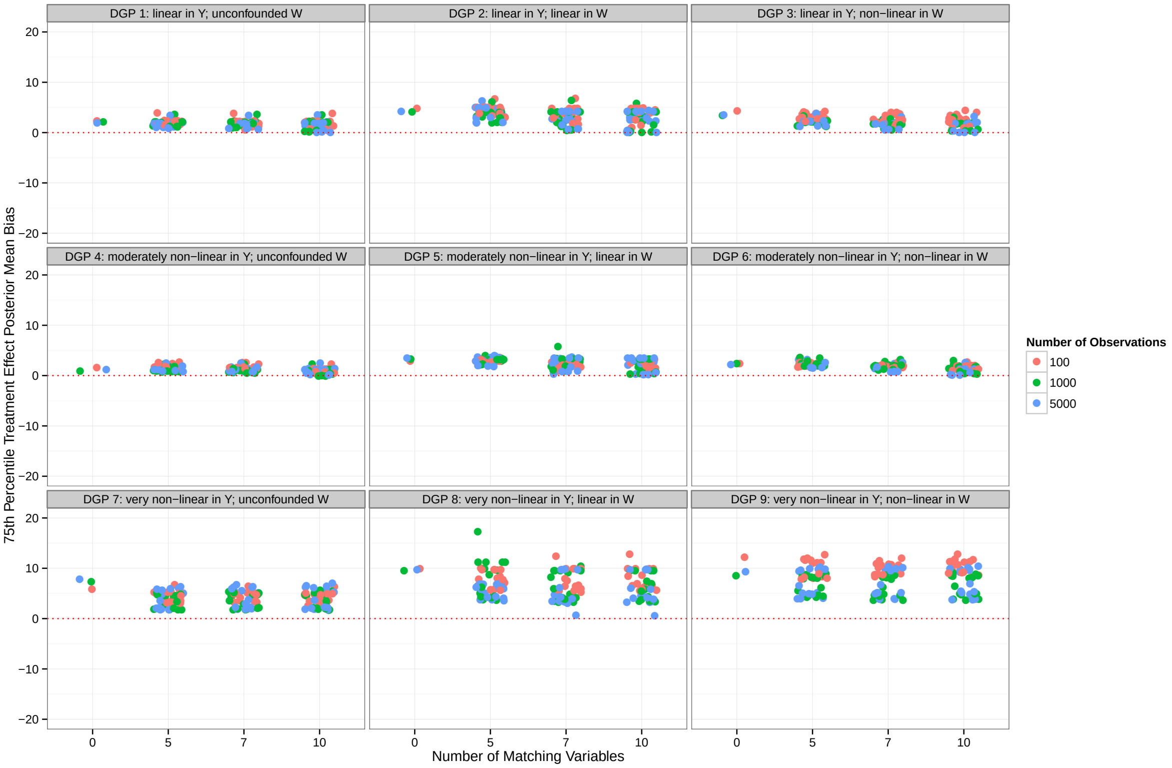

Figure A.20: Comparing 75th Percentile Treatment Effect Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

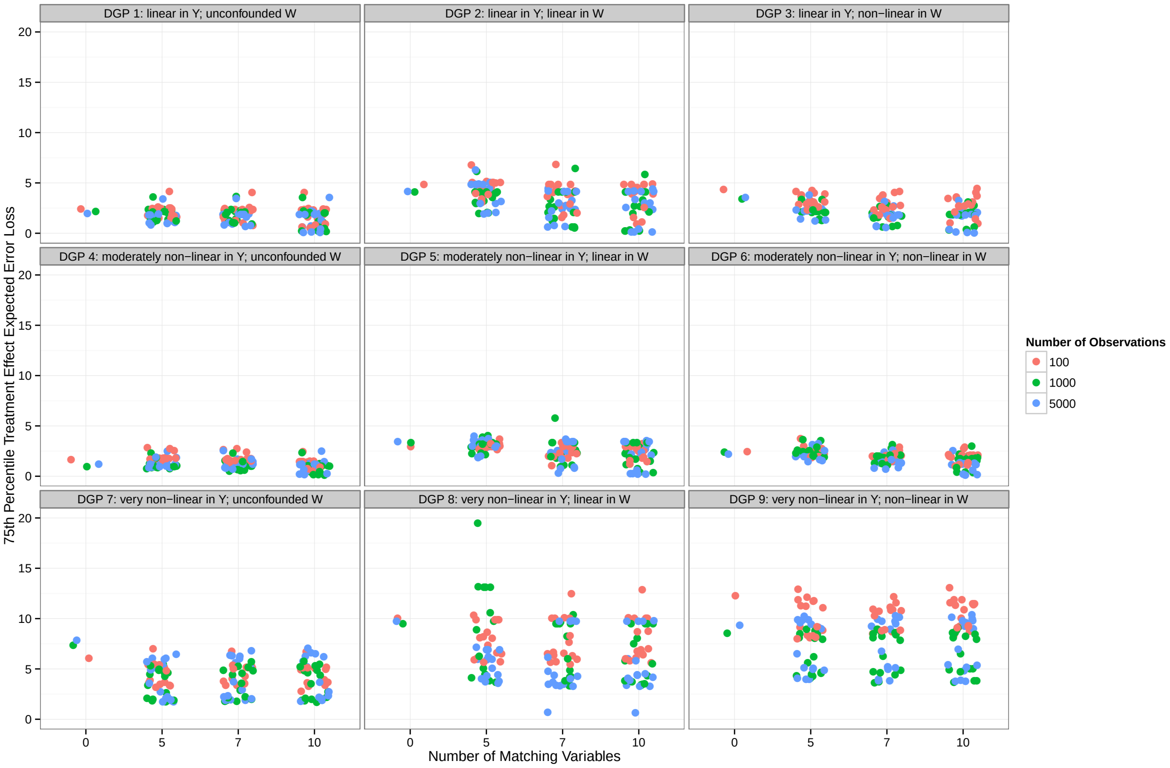

Figure A.21: Comparing 75th Percentile Treatment Effect Expected Error Loss for Different Conditioning Sets (continuous outcome)

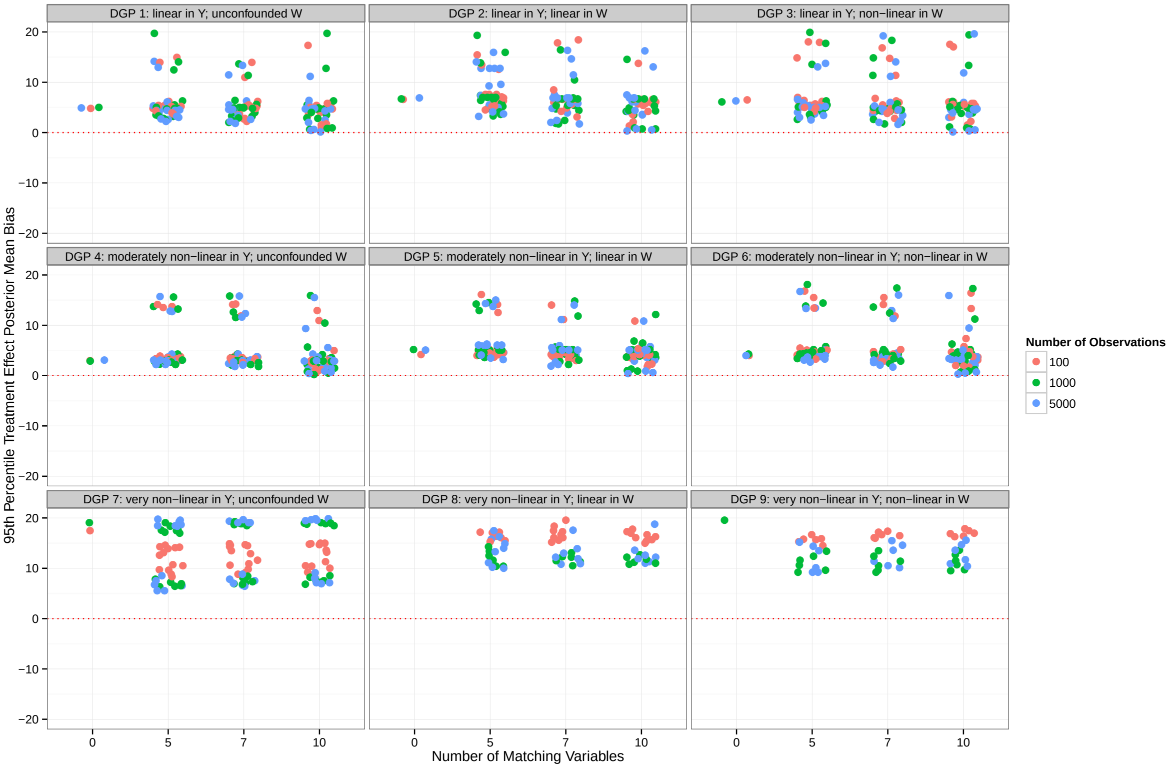

Figure A.22: Comparing 95th Percentile Treatment Effect Posterior Mean Bias for Different Conditioning Sets (continuous outcome)

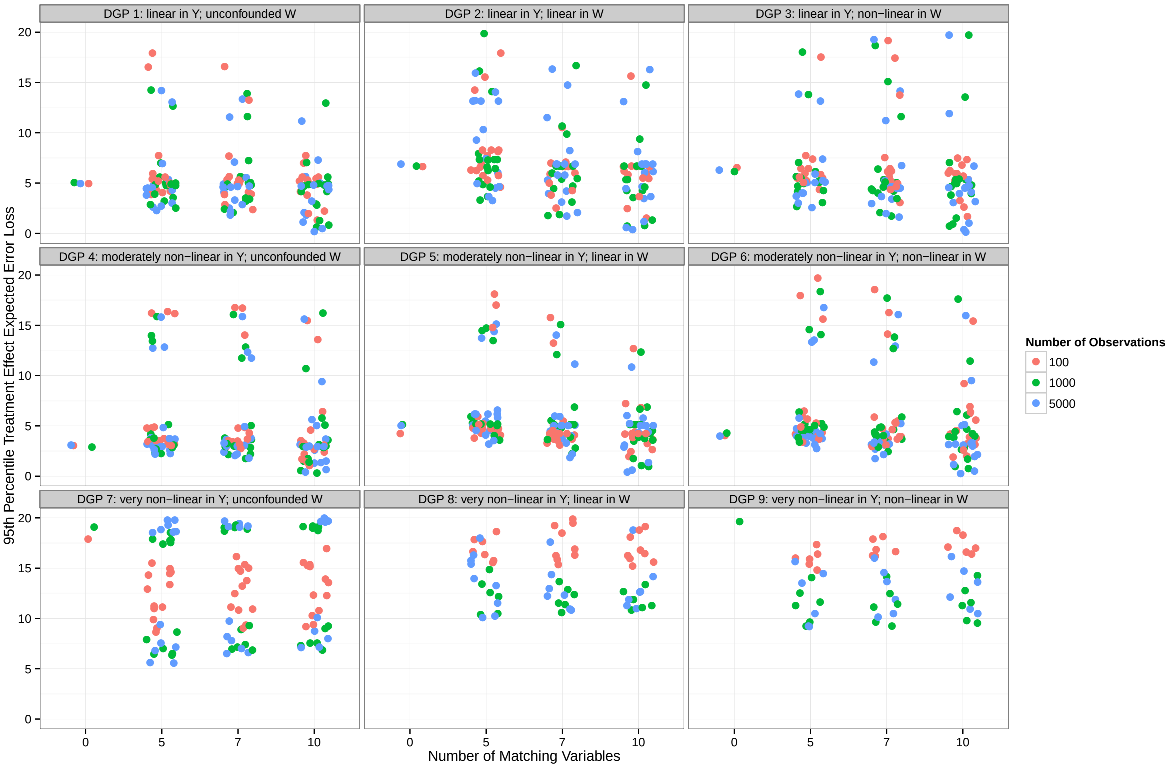

Figure A.23: Comparing 95th Percentile Treatment Effect Expected Error Loss for Different Conditioning Sets (continuous outcome)

A.3 Comparing Number of Matches for Continuous Outcomes

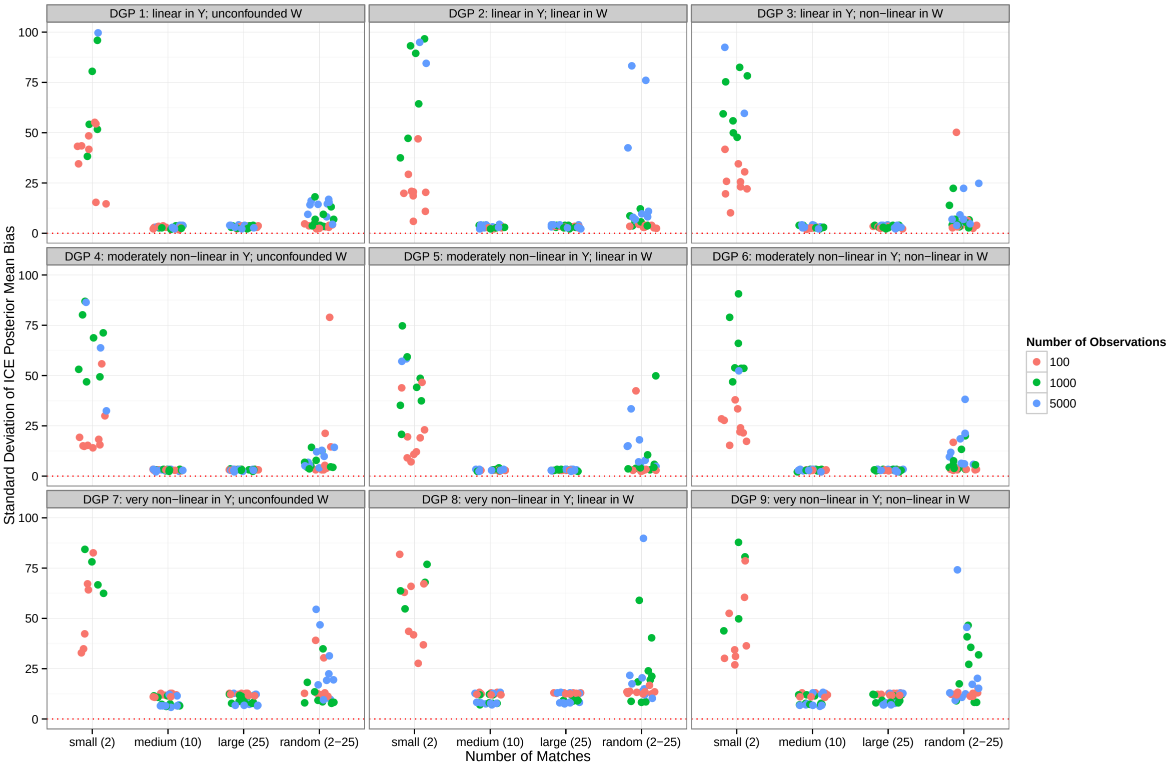

Figure A.24: Comparing Standard Deviations of ICE Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

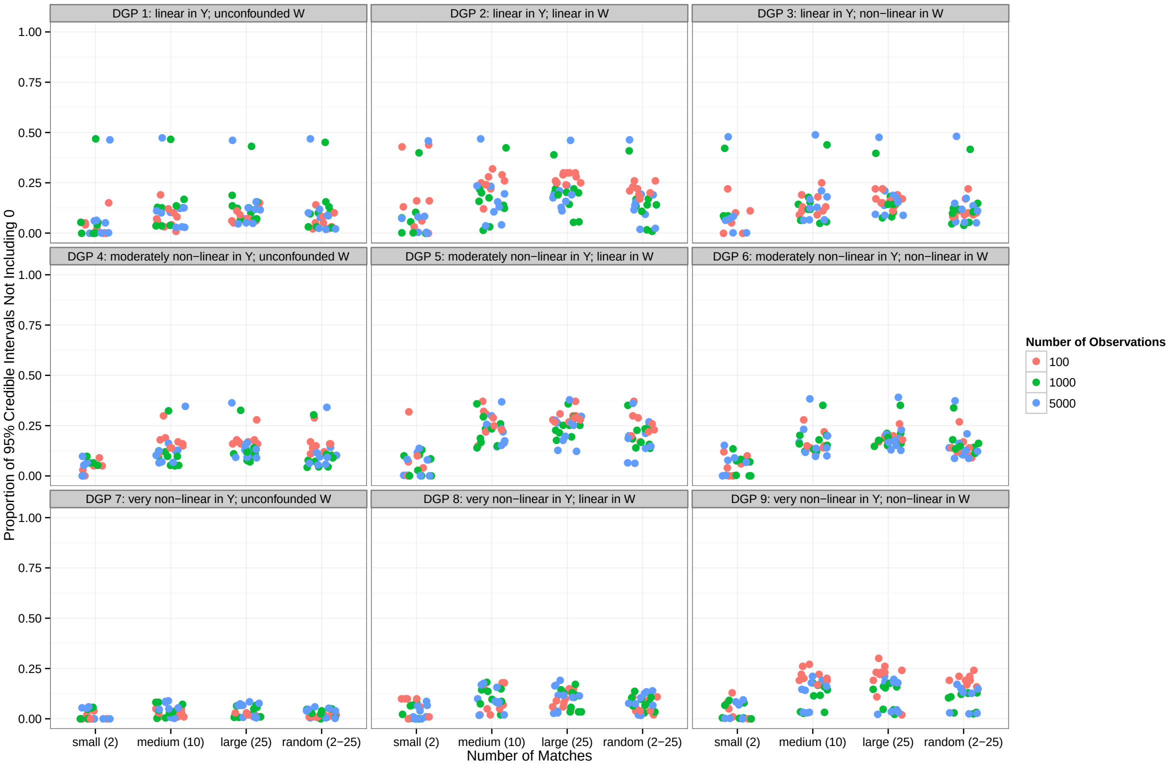

Figure A.25: Comparing ICE “Power” for Different Numbers of Matches (continuous outcome)

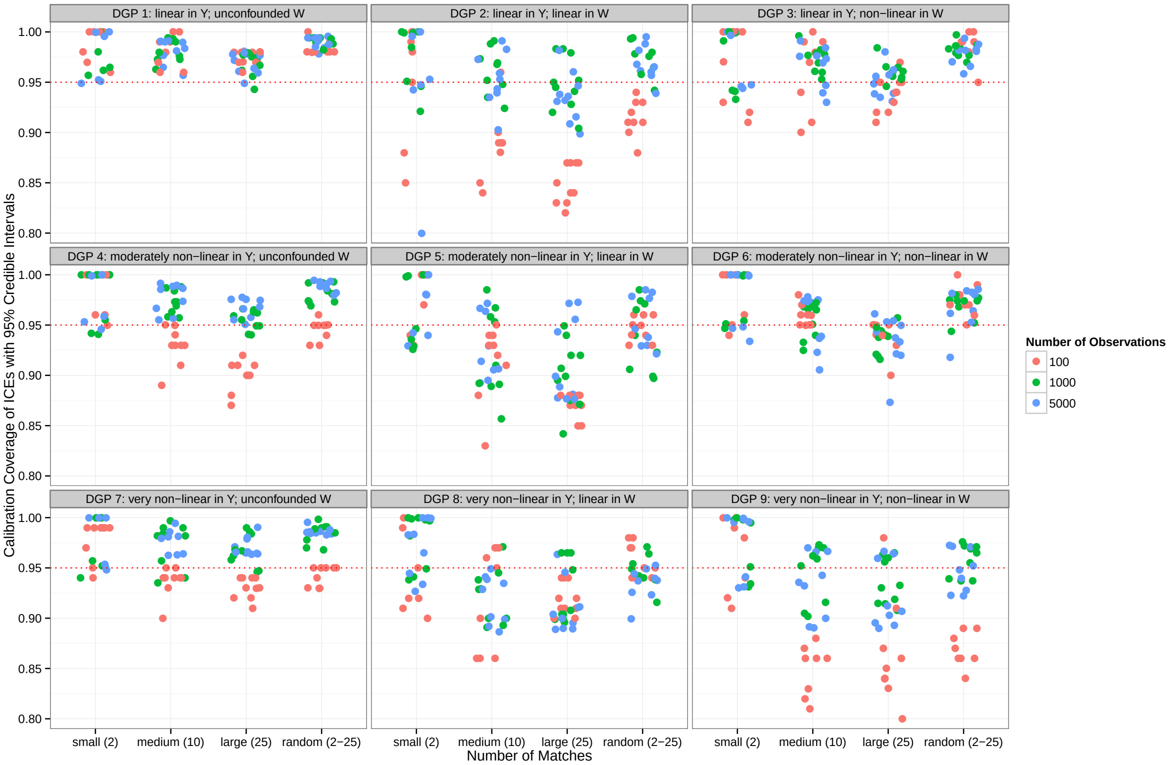

Figure A.26: Comparing ICE Calibration Coverage for Different Numbers of Matches (continuous outcome)

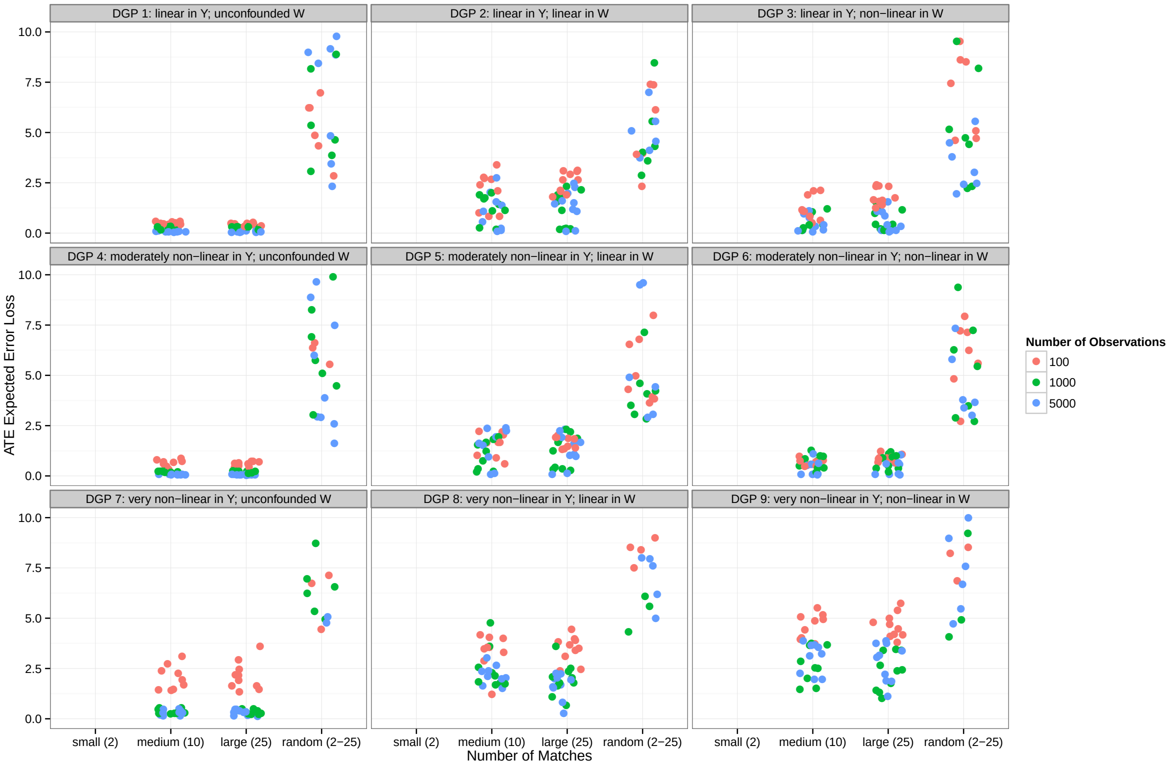

Figure A.27: Comparing ATE Expected Error Loss for Different Numbers of Matches (continuous outcome)

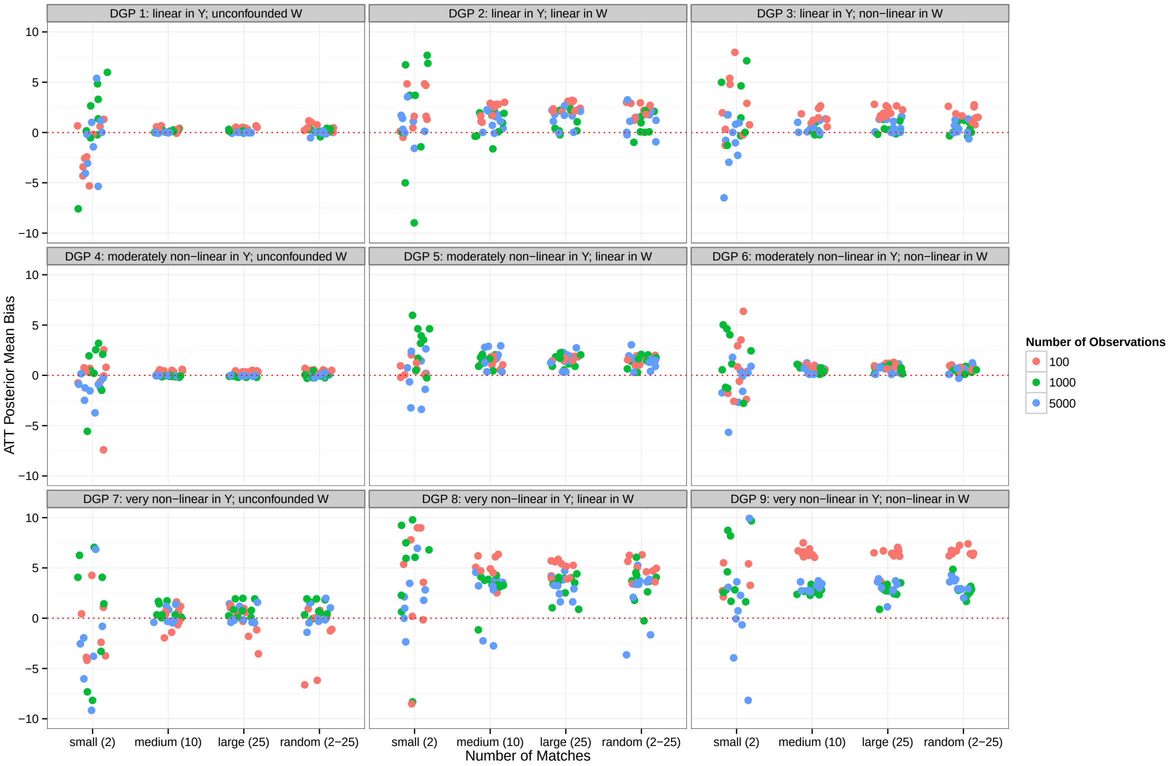

Figure A.28: Comparing ATT Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

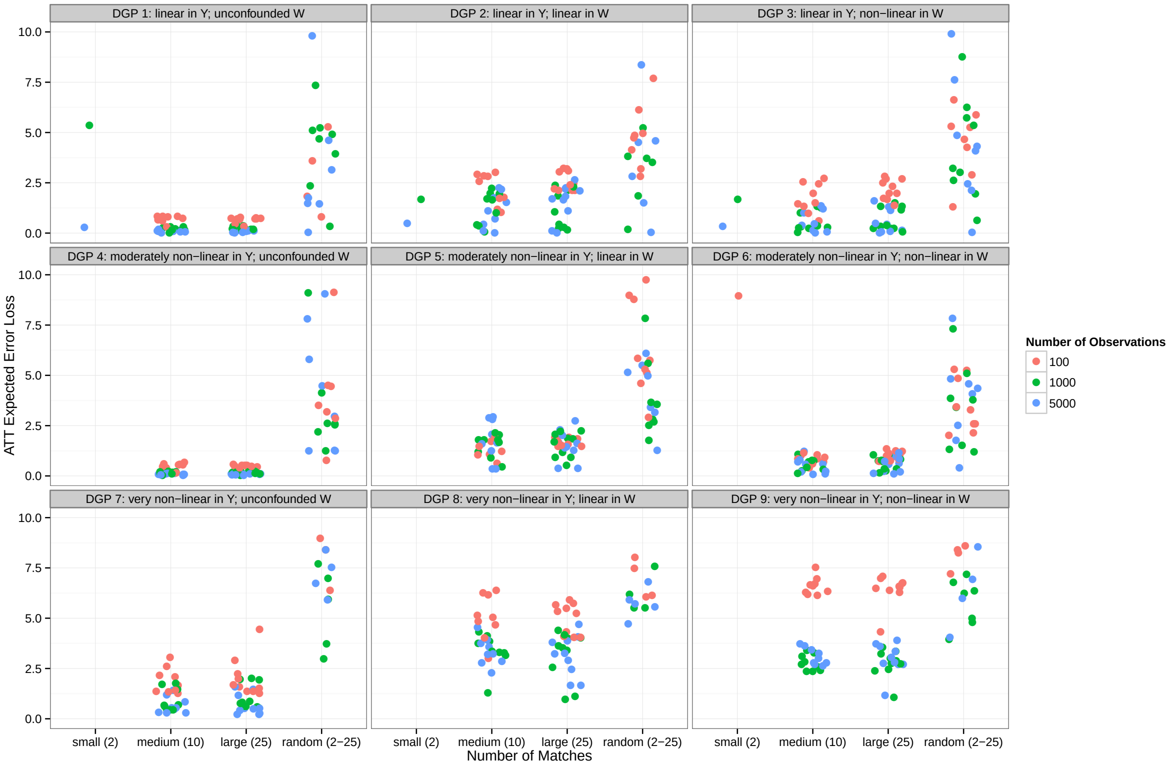

Figure A.29: Comparing ATT Expected Error Loss for Different Numbers of Matches (continuous outcome)

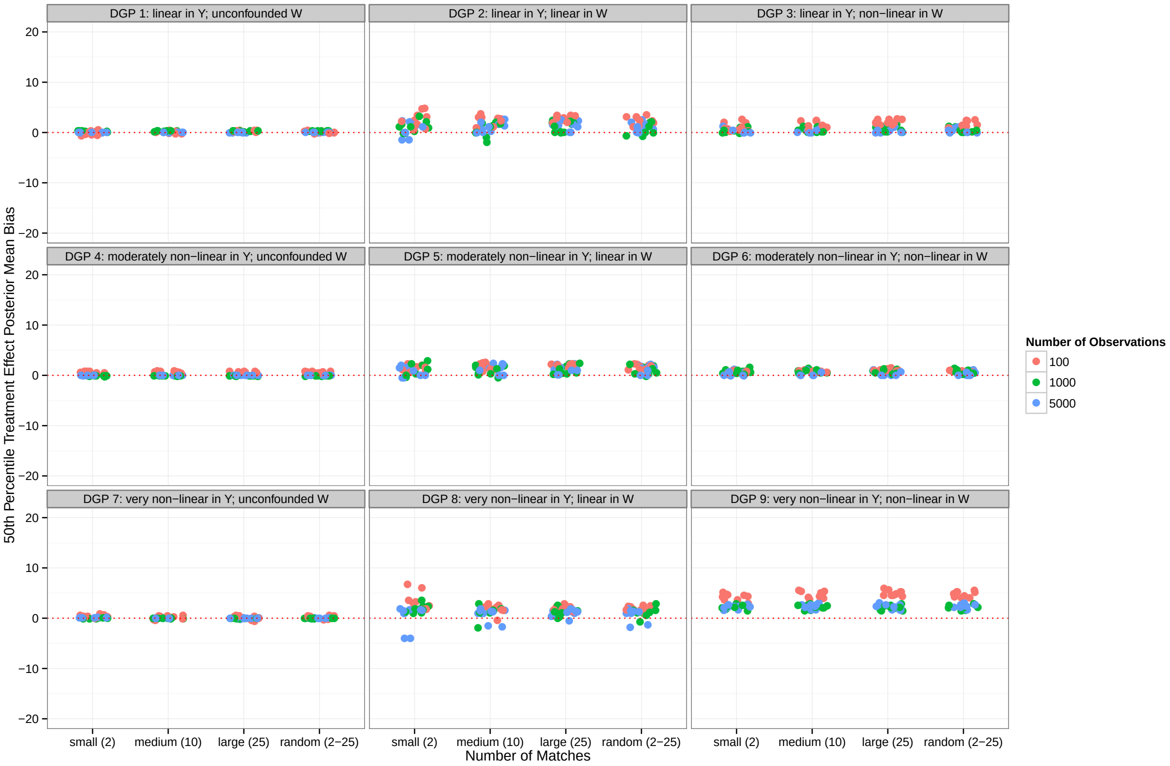

Figure A.30: Comparing 50th Percentile Treatment Effect Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

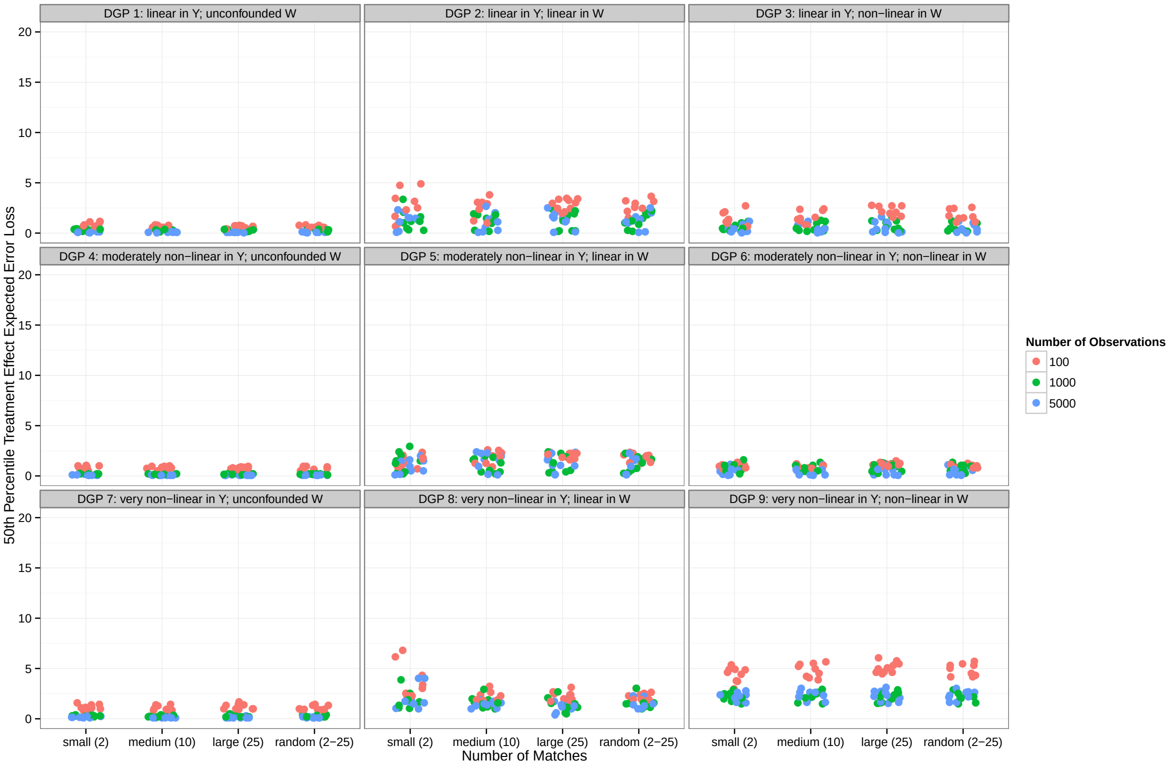

Figure A.31: Comparing 50th Percentile Treatment Effect Expected Error Loss for Different Numbers of Matches (continuous outcome)

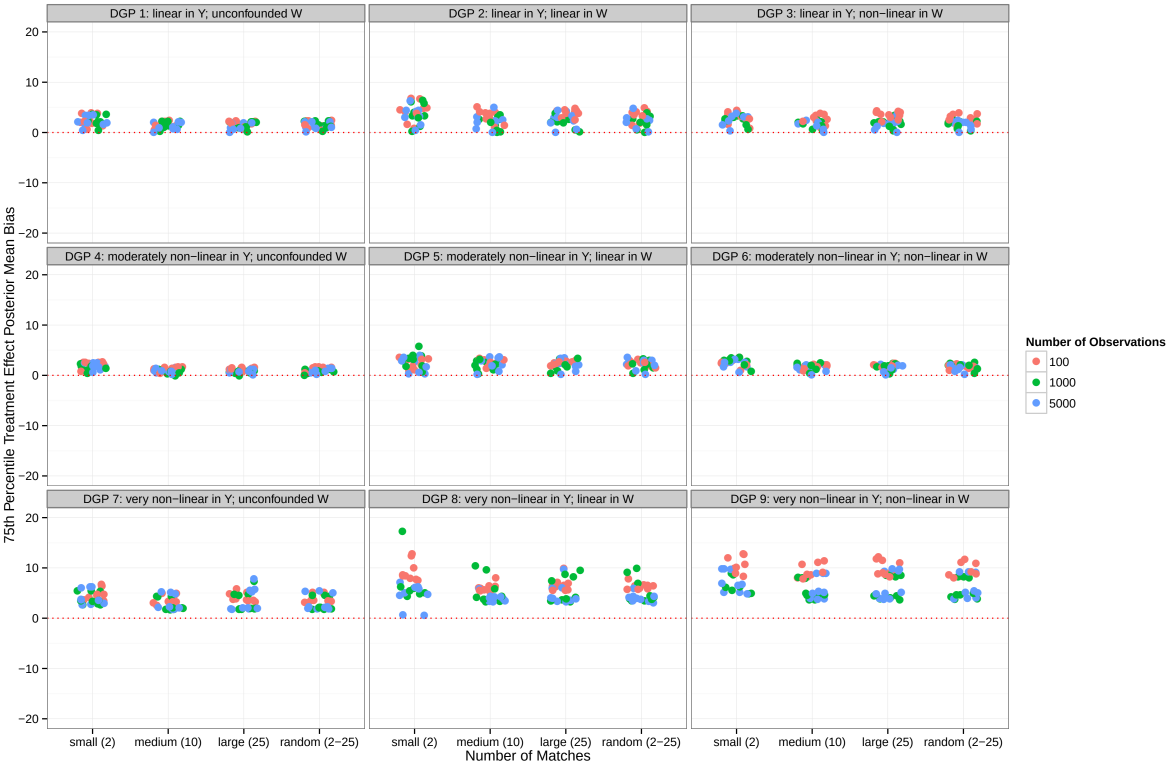

Figure A.32: Comparing 75th Percentile Treatment Effect Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

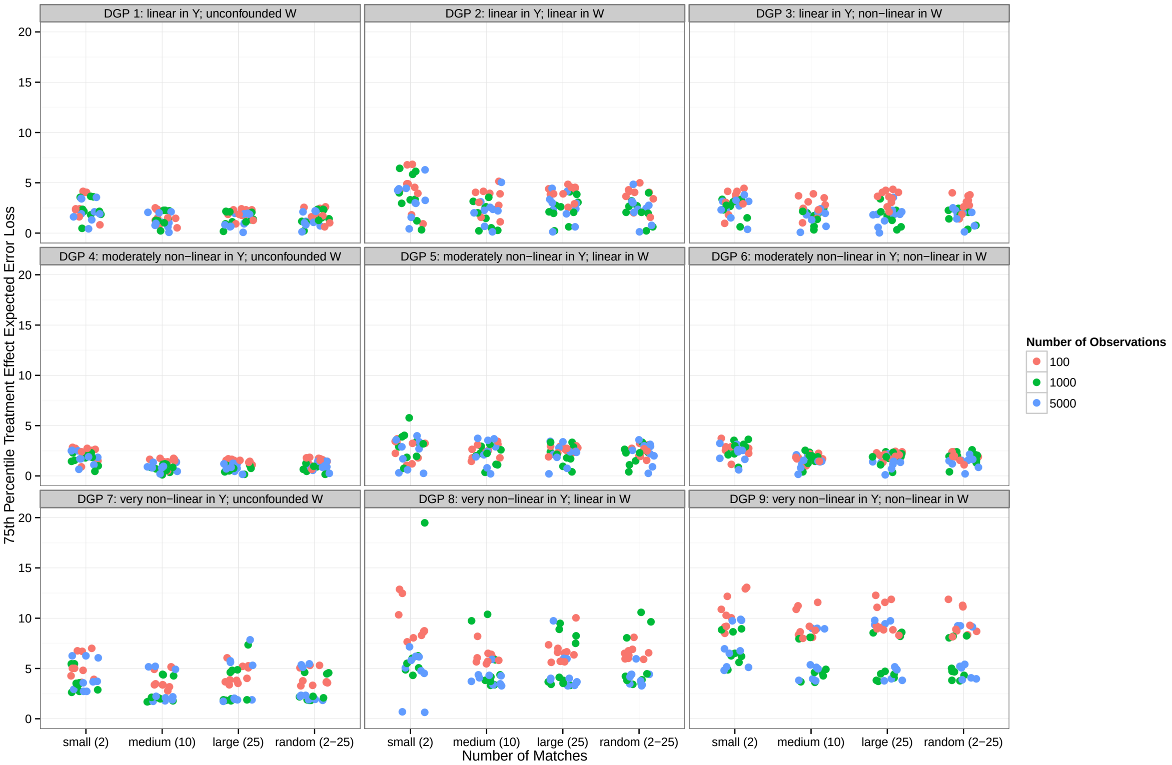

Figure A.33: Comparing 75th Percentile Treatment Effect Expected Error Loss for Different Numbers of Matches (continuous outcome)

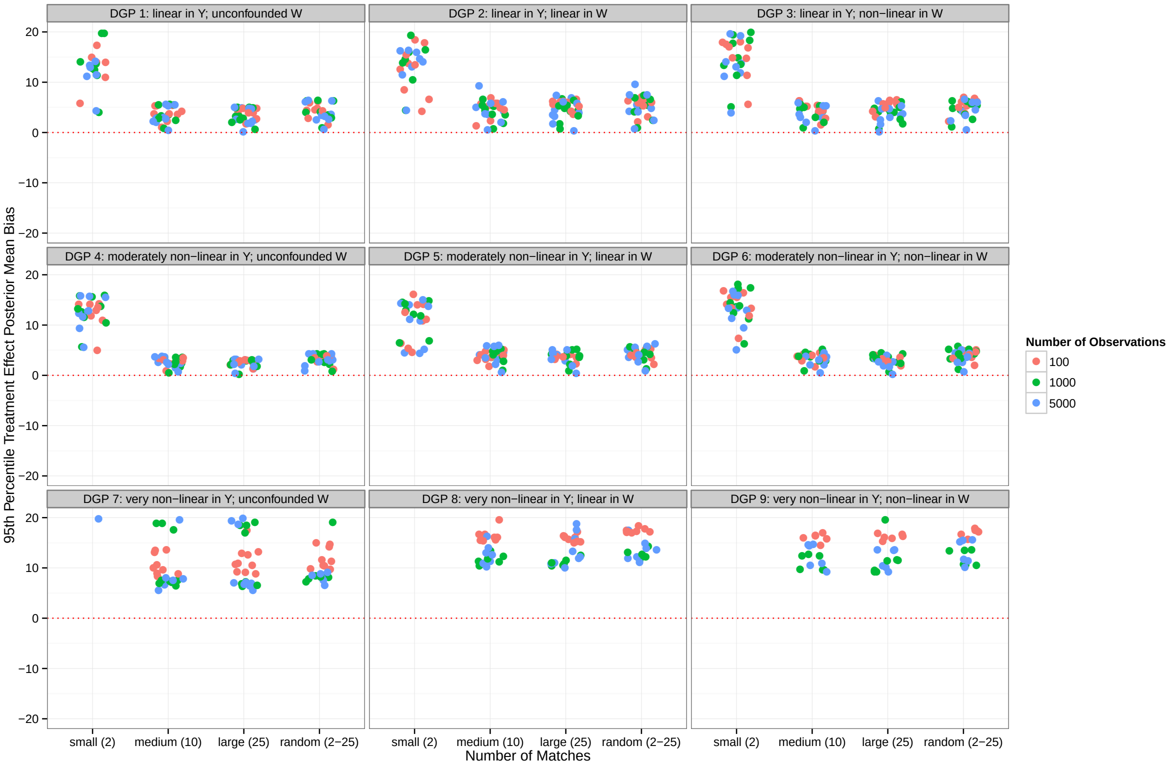

Figure A.34: Comparing 95th Percentile Treatment Effect Posterior Mean Bias for Different Numbers of Matches (continuous outcome)

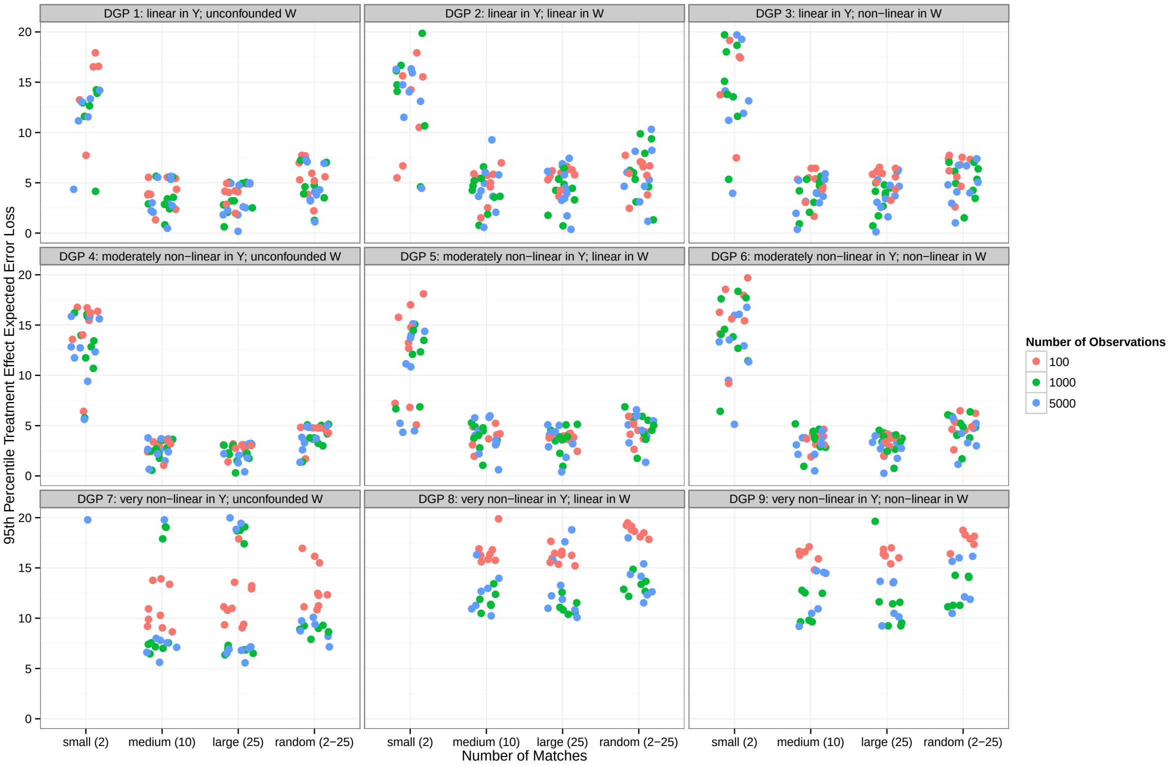

Figure A.35: Comparing 95th Percentile Treatment Effect Expected Error Loss for Different Numbers of Matches (continuous outcome)

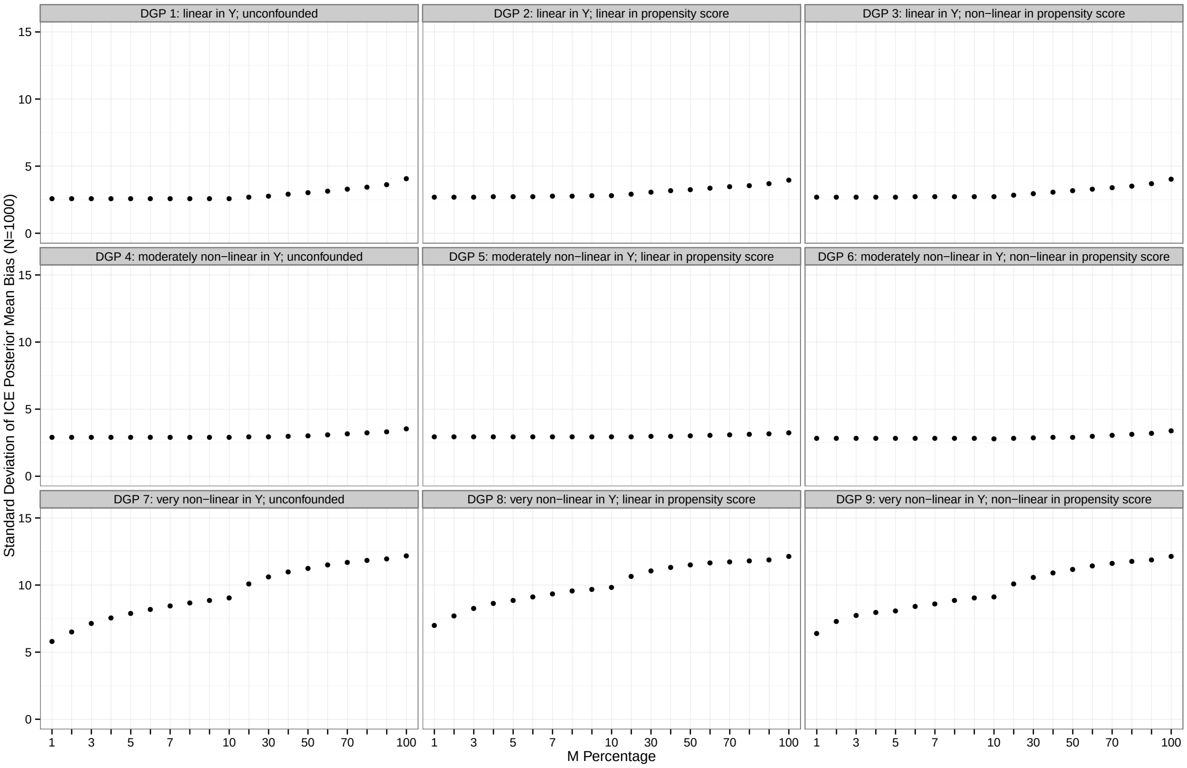

Figure A.36: Comparing Standard Deviations of ICE Posterior Mean Bias for Different Match Percentages (continuous outcome)

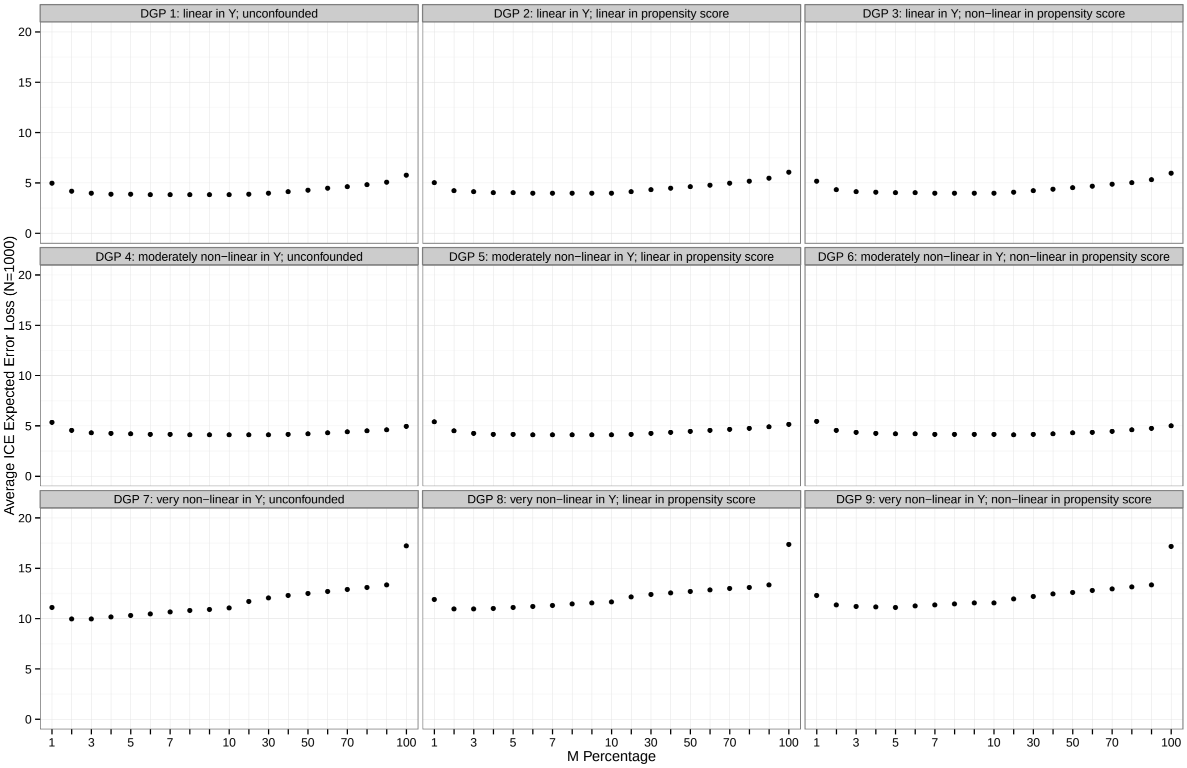

Figure A.37: Comparing Average ICE Expected Error Loss for Different Match Percentages (continuous outcome)

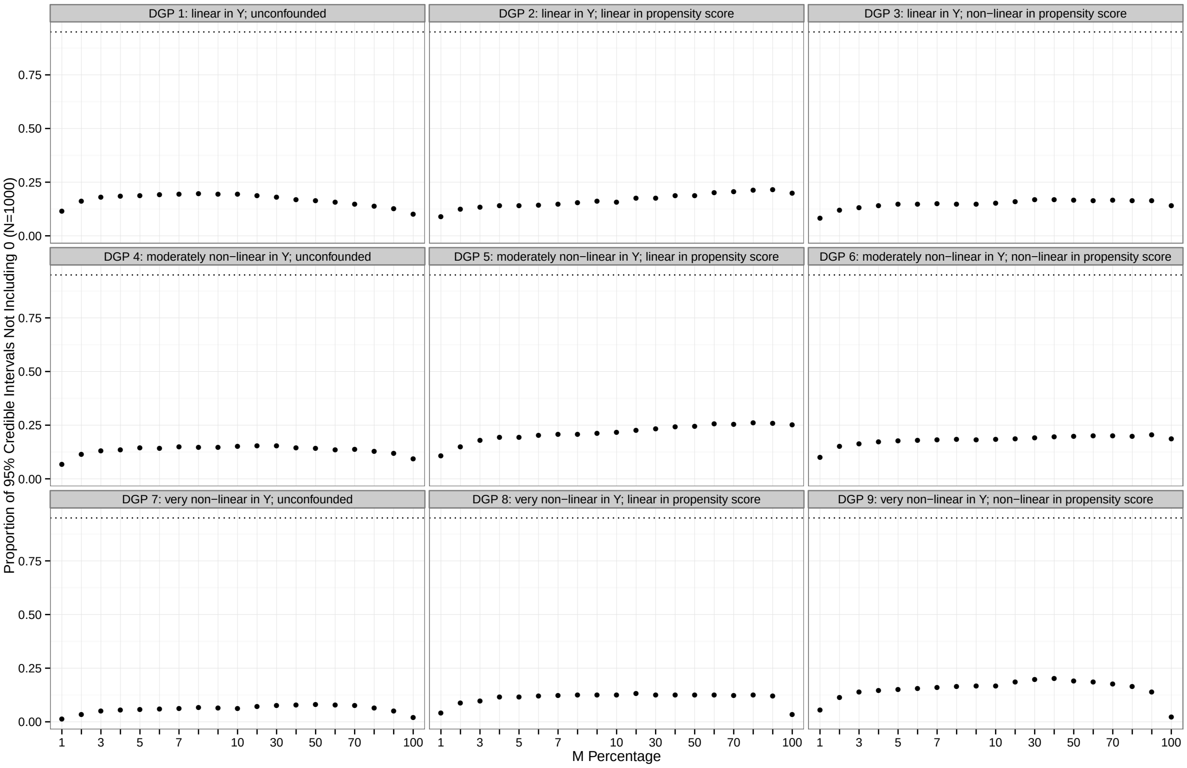

Figure A.38: Comparing ICE "Power" for Different Match Percentages (continuous outcome)

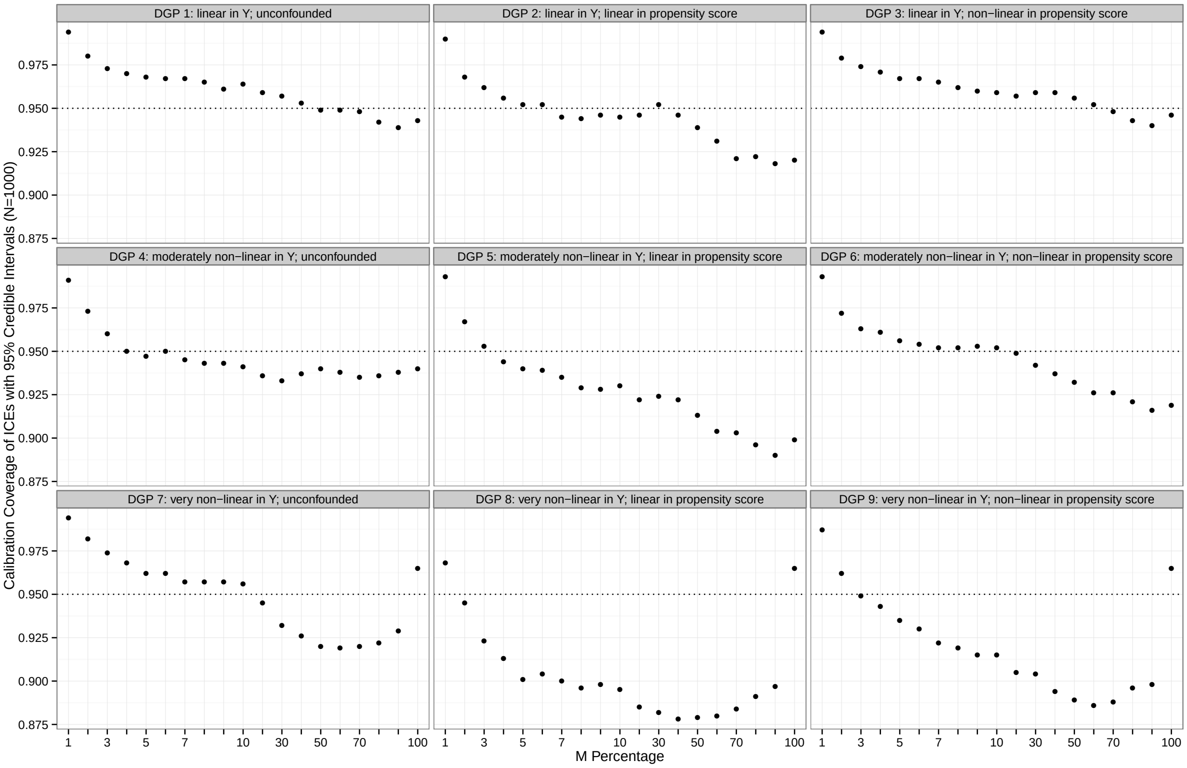

Figure A.39: Comparing ICE Calibration Coverage for Different Match Percentages (continuous outcome)

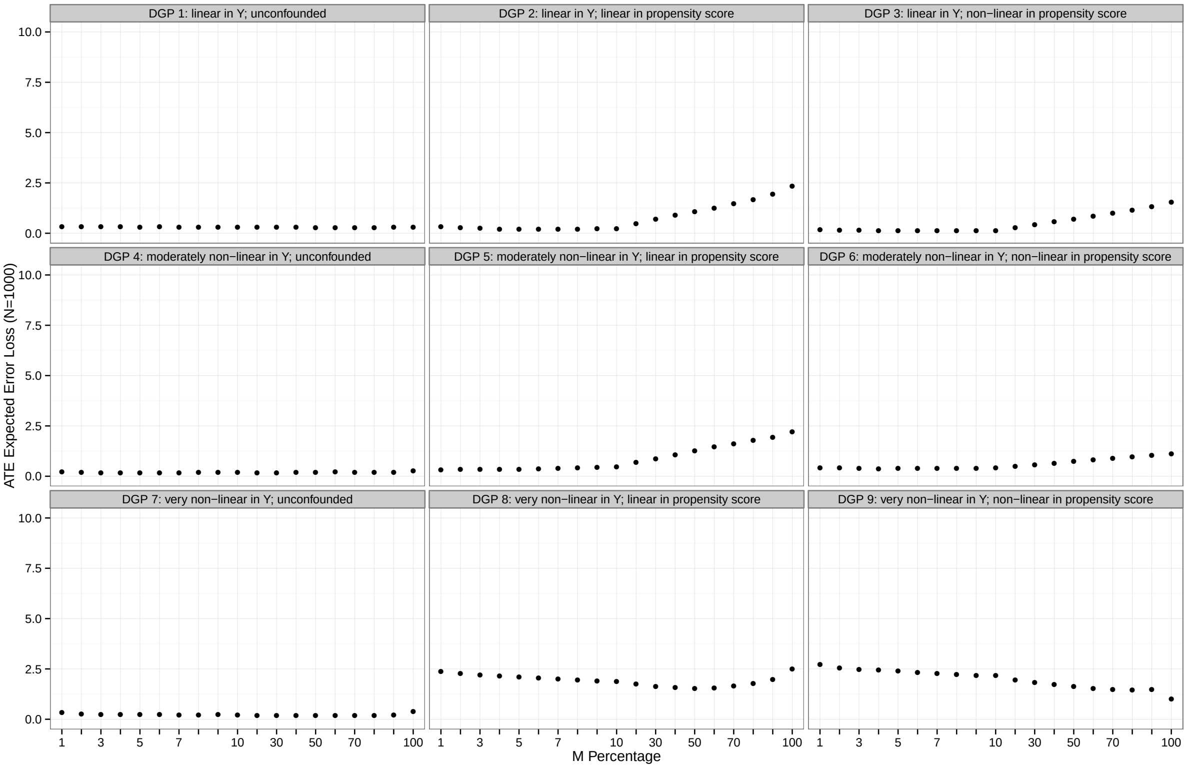

Figure A.40: Comparing ATE Expected Error Loss for Different Match Percentages (continuous outcome)

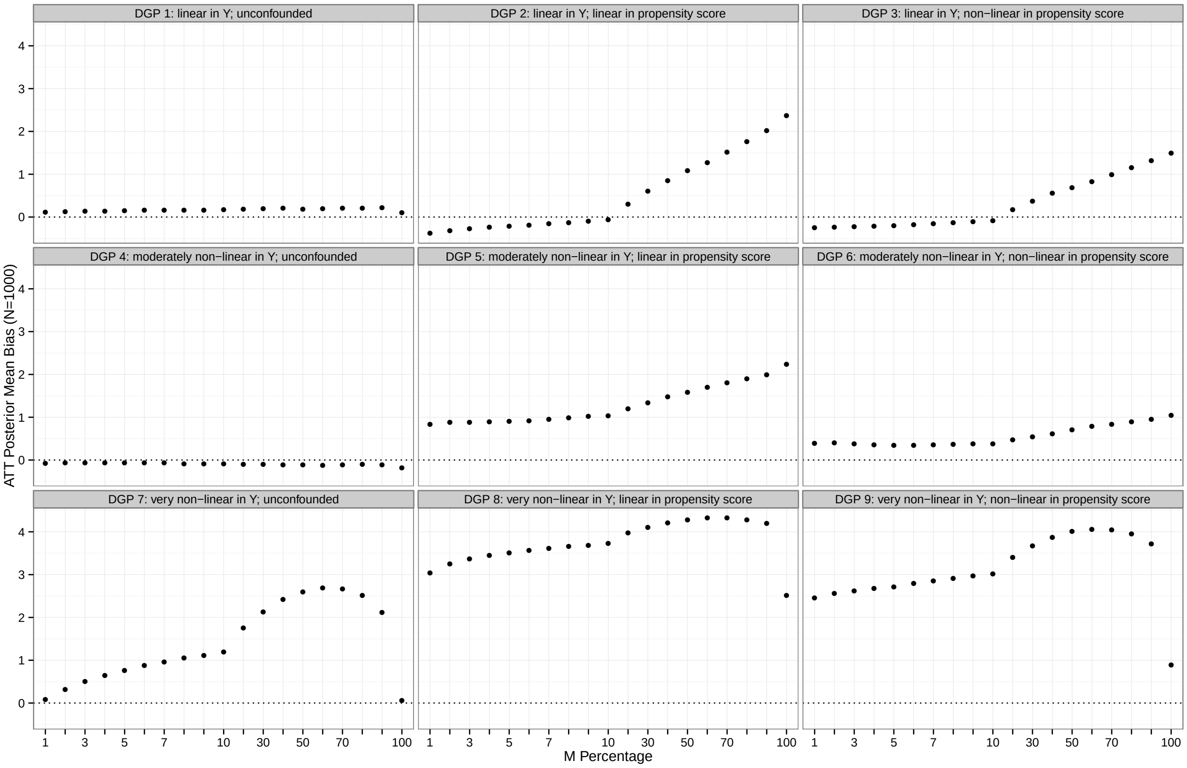

Figure A.41: Comparing ATT Posterior Mean Bias for Different Match Percentages (continuous outcome)

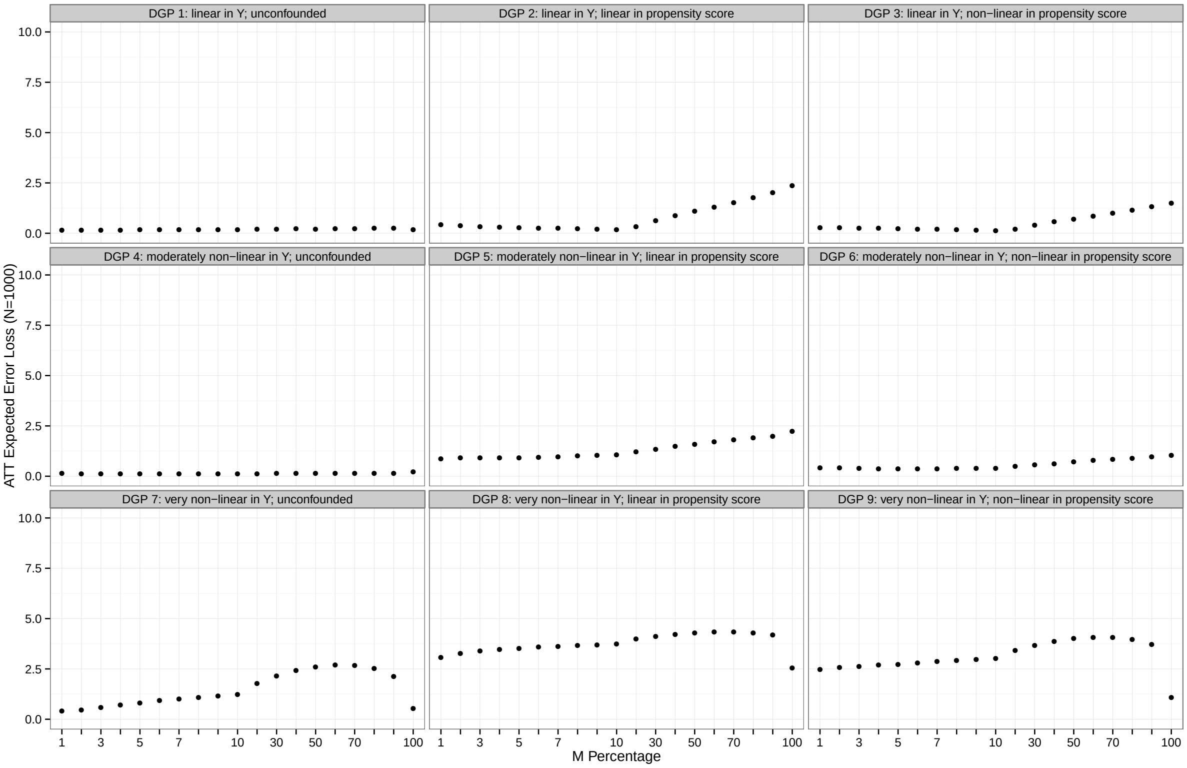

Figure A.42: Comparing ATT Expected Error Loss for Different Match Percentages (continuous outcome)

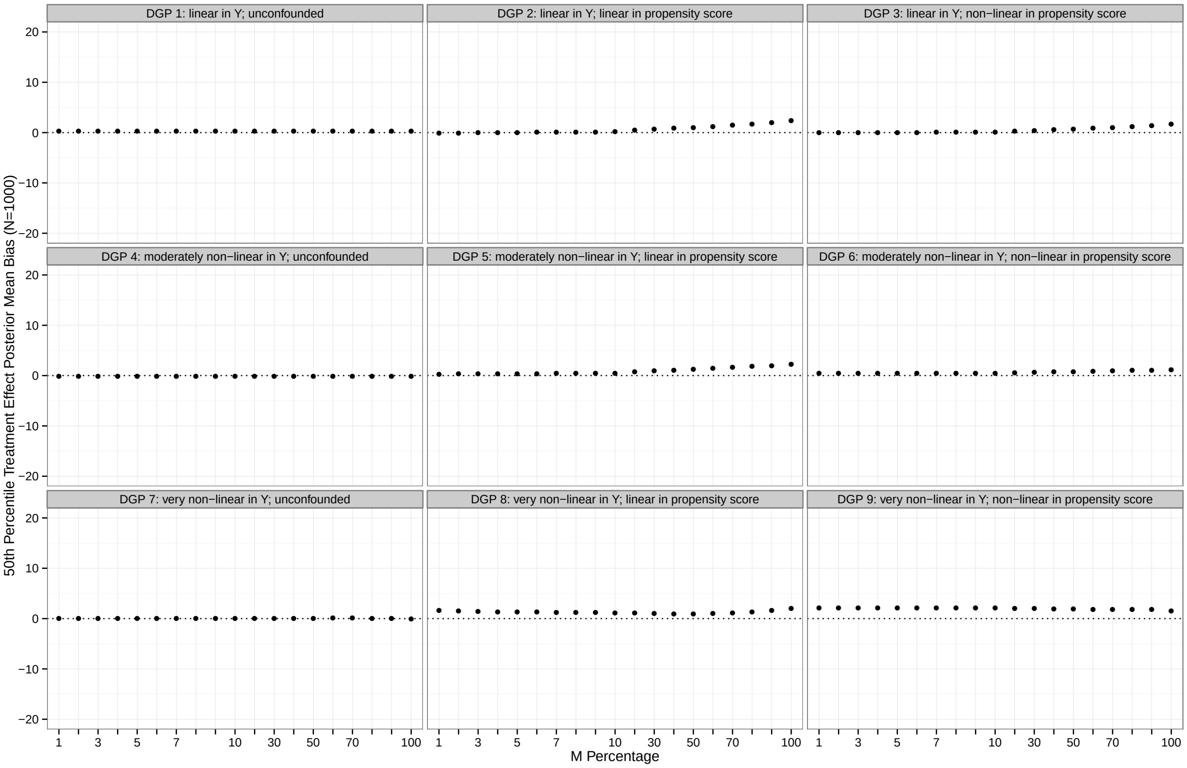

Figure A.43: Comparing 50th Percentile Treatment Effect Posterior Mean Bias for Different Match Percentages (continuous outcome)

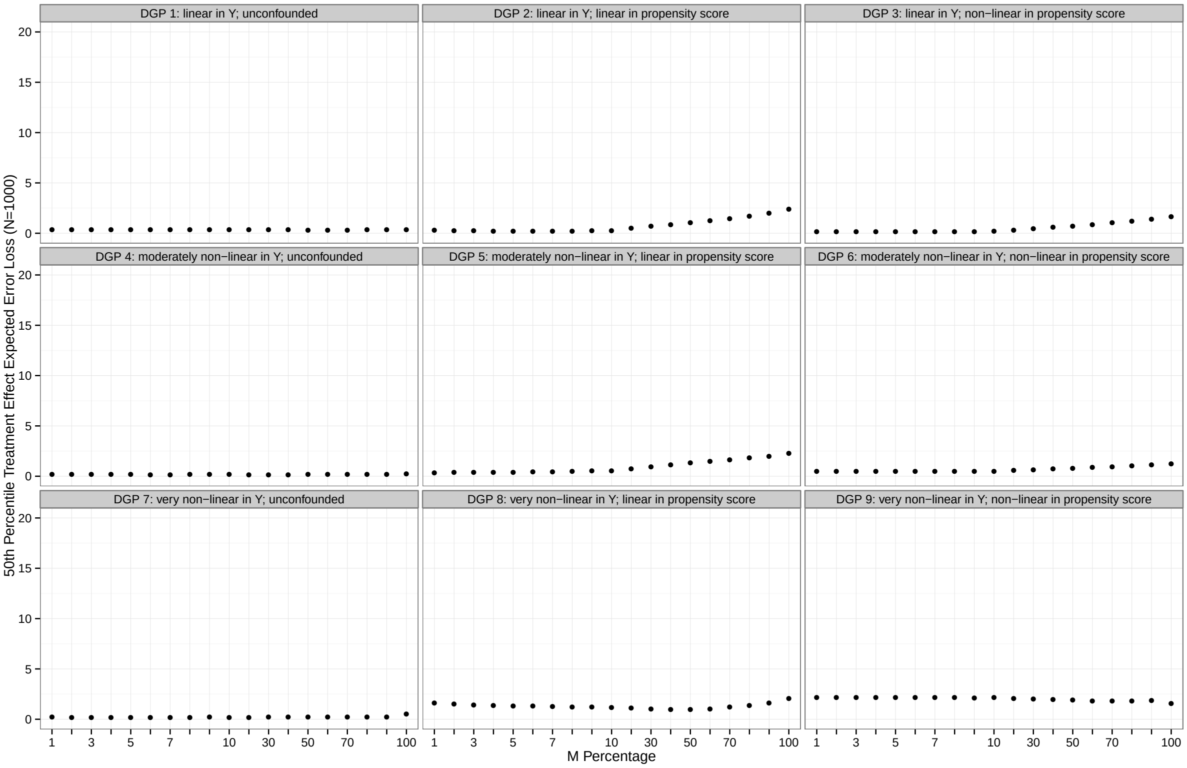

Figure A.44: Comparing 50th Percentile Treatment Effect Expected Error Loss for Different Match Percentages (continuous outcome)

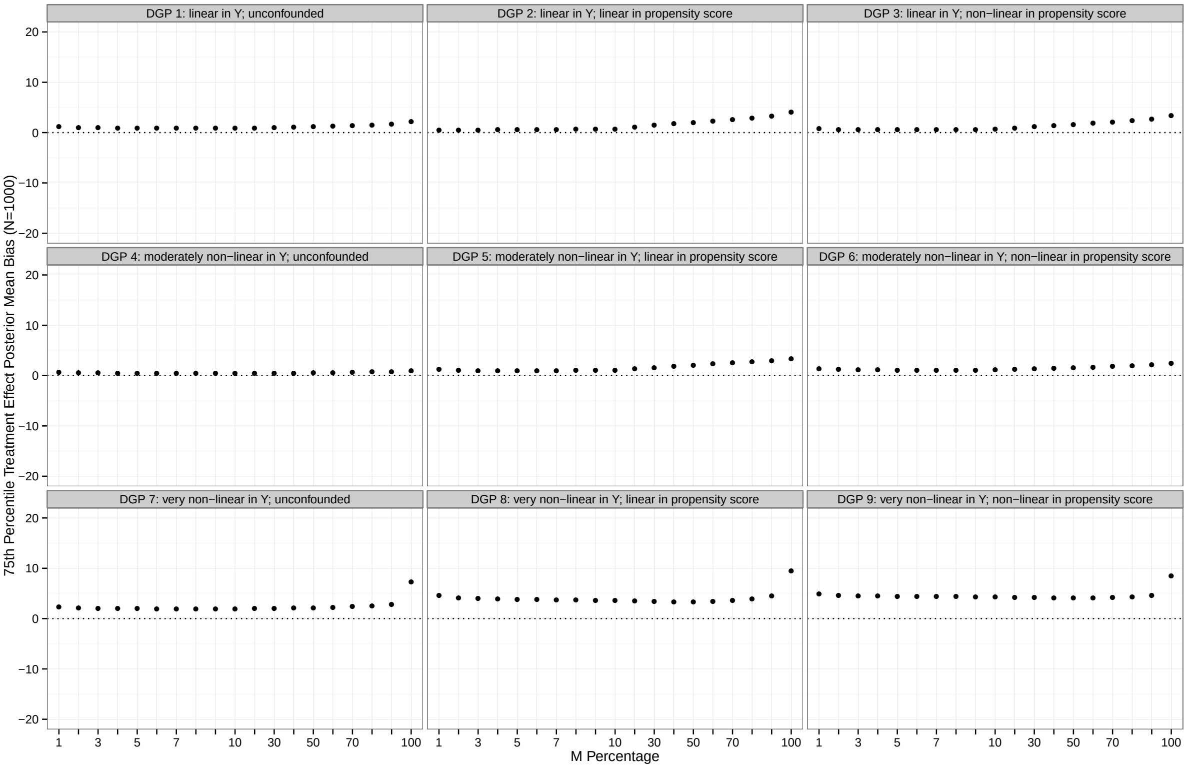

Figure A.45: Comparing 75th Percentile Treatment Effect Posterior Mean Bias for Different Match Percentages (continuous outcome)

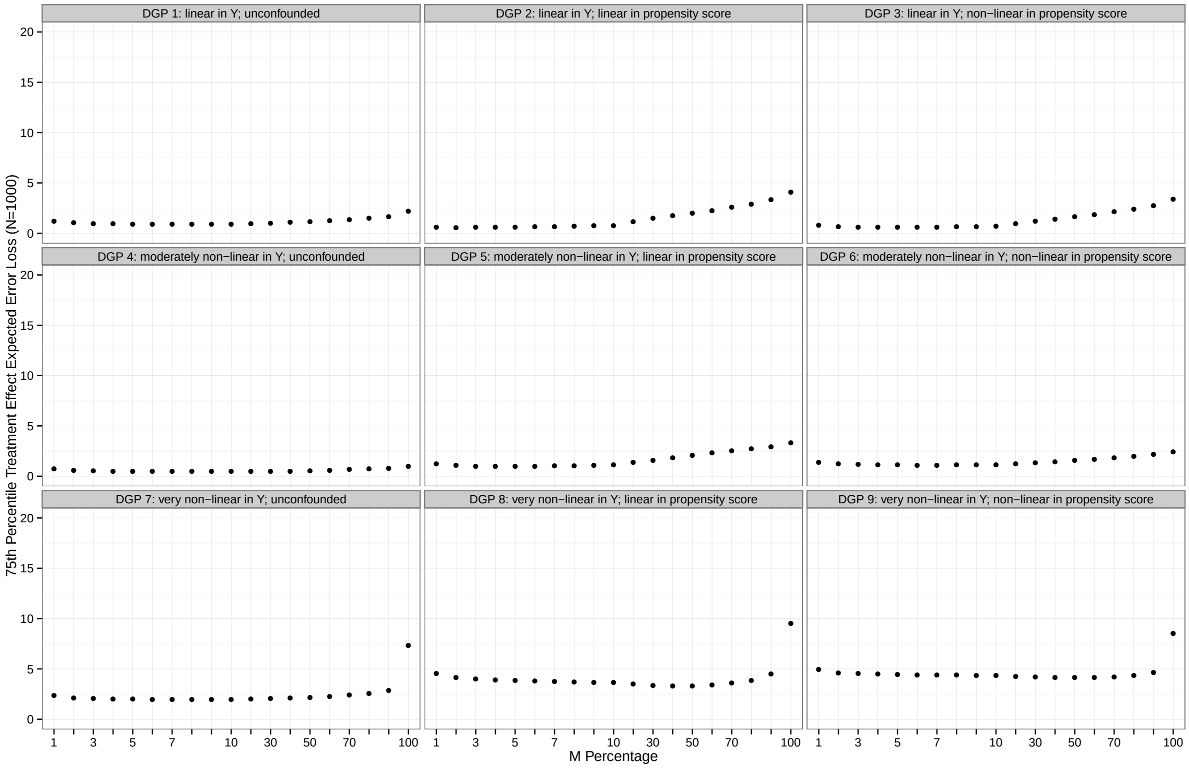

Figure A.46: Comparing 75th Percentile Treatment Effect Expected Error Loss for Different Match Percentages (continuous outcome)

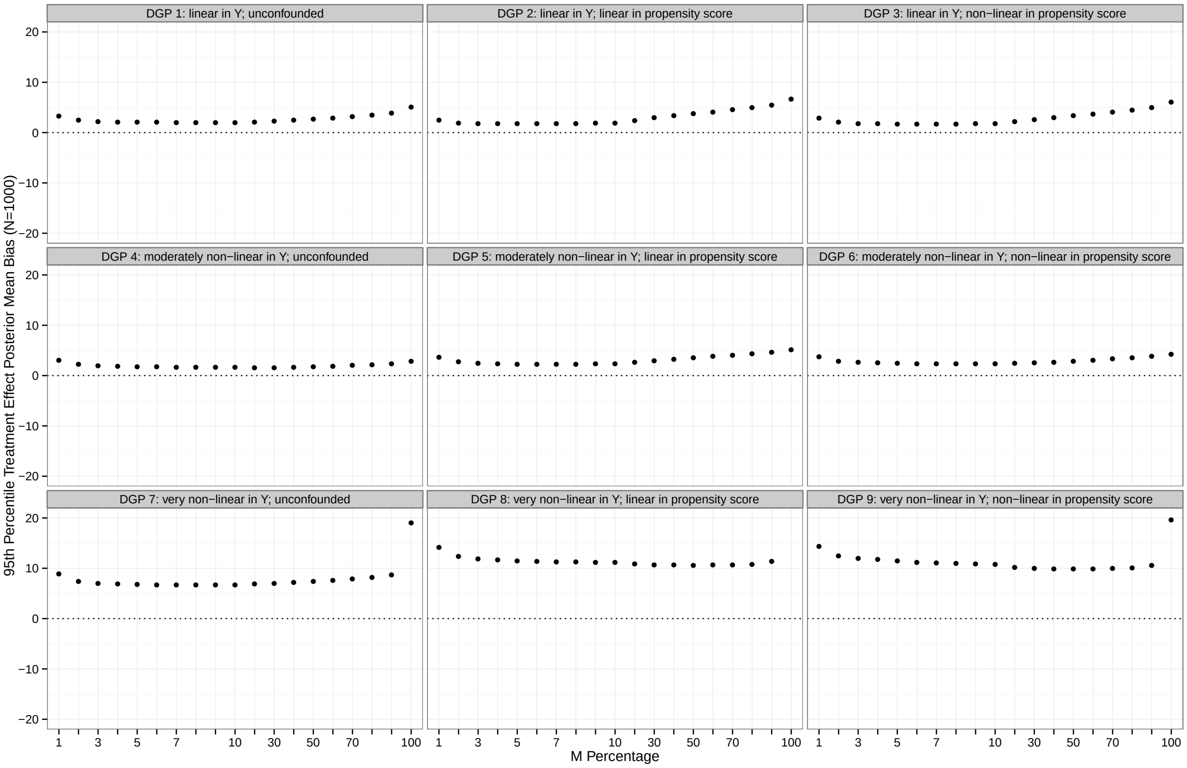

Figure A.47: Comparing 95th Percentile Treatment Effect Posterior Mean Bias for Different Match Percentages (continuous outcome)

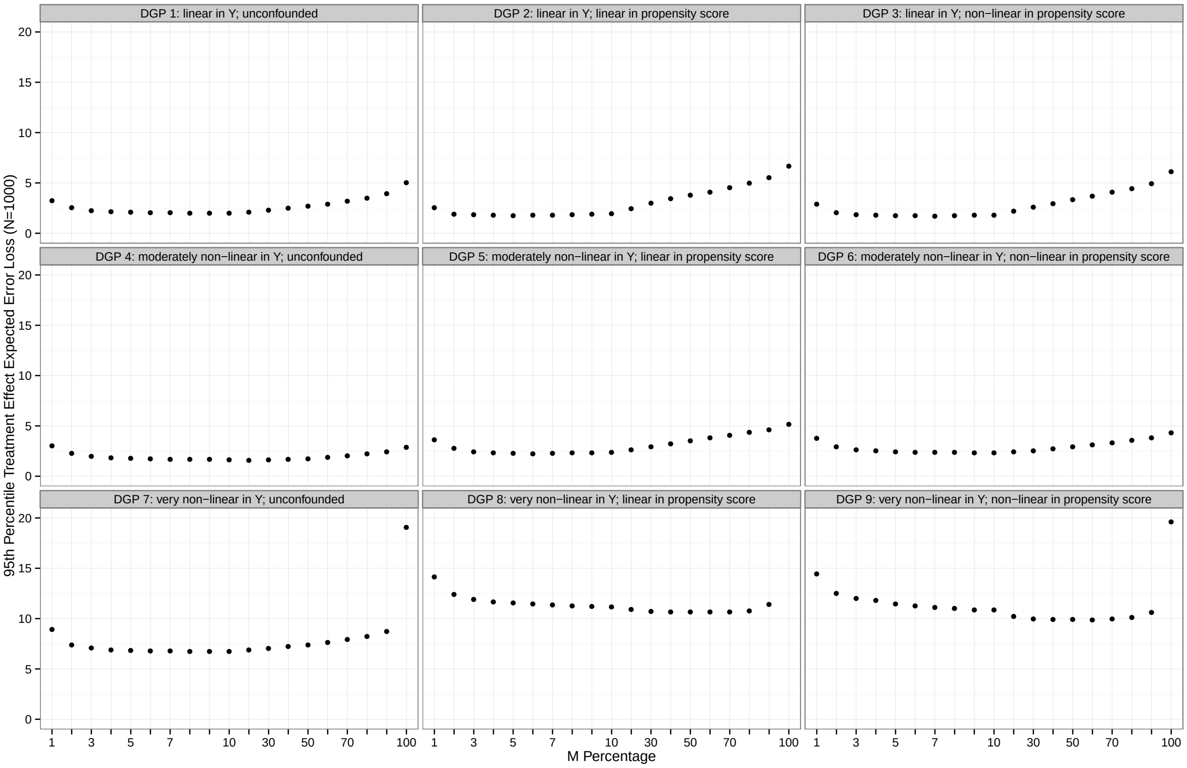

Figure A.48: Comparing 95th Percentile Treatment Effect Expected Error Loss for Different Match Percentages (continuous outcome)

A.4 Comparing Different τ_i Distributions for Continuous Outcomes

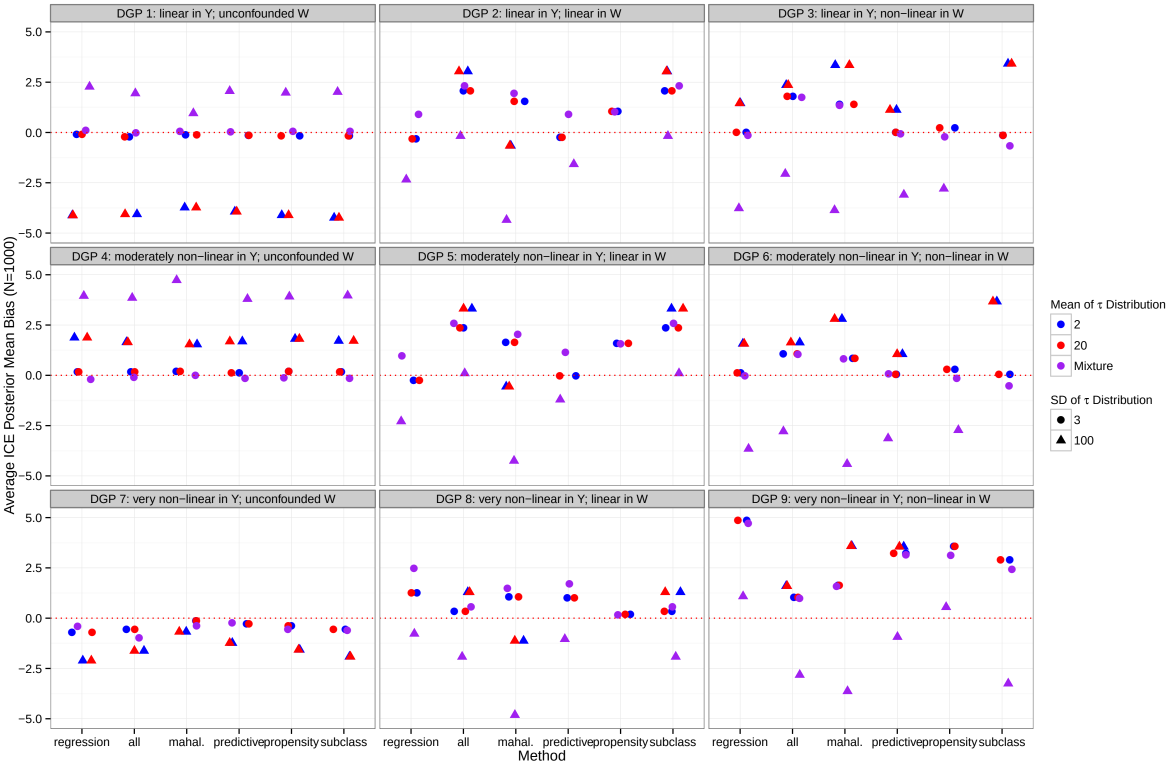

Figure A.49: Comparing Average ICE (or ATE) Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

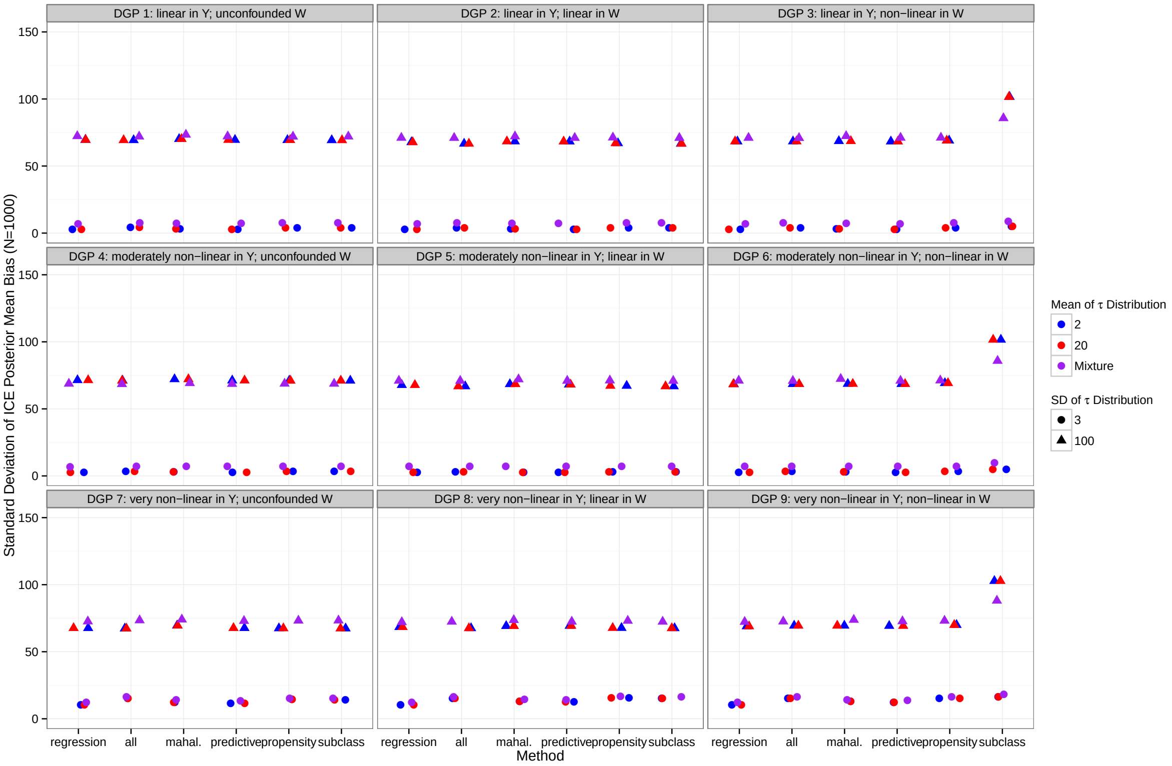

Figure A.50: Comparing Standard Deviations of ICE Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

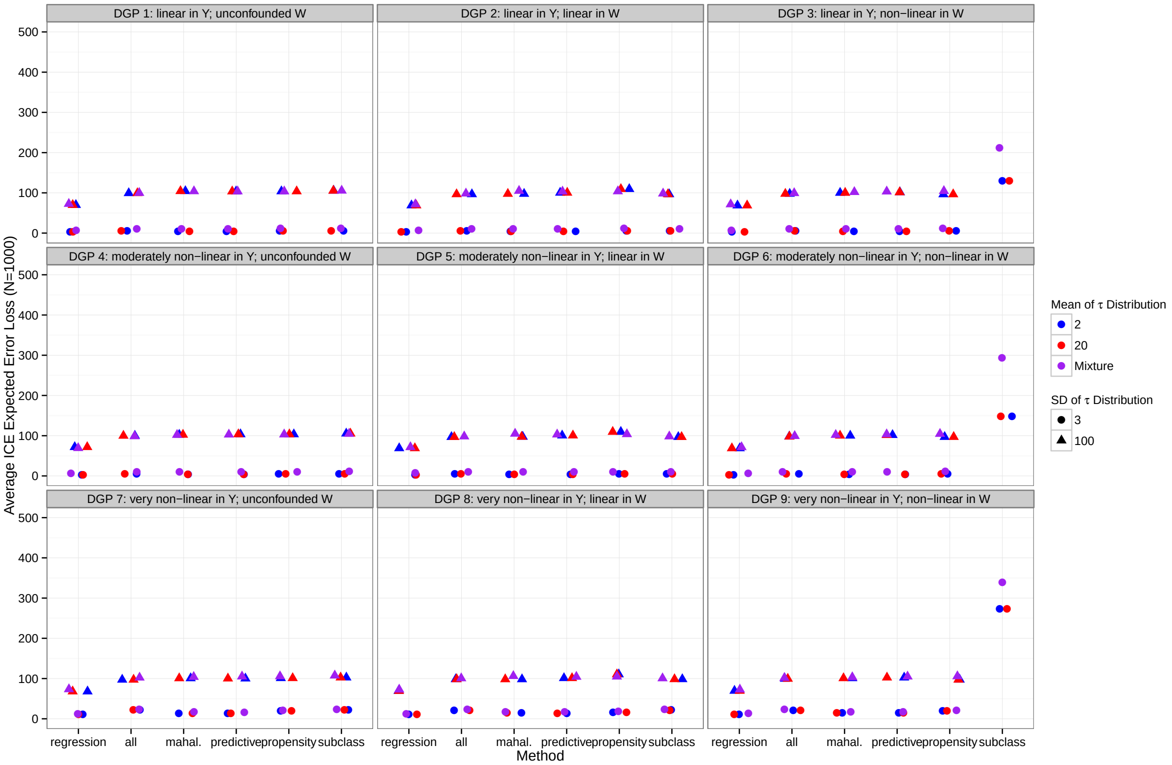

Figure A.51: Comparing Average ICE Expected Error Loss with Different τ_i Distributions (continuous outcome)

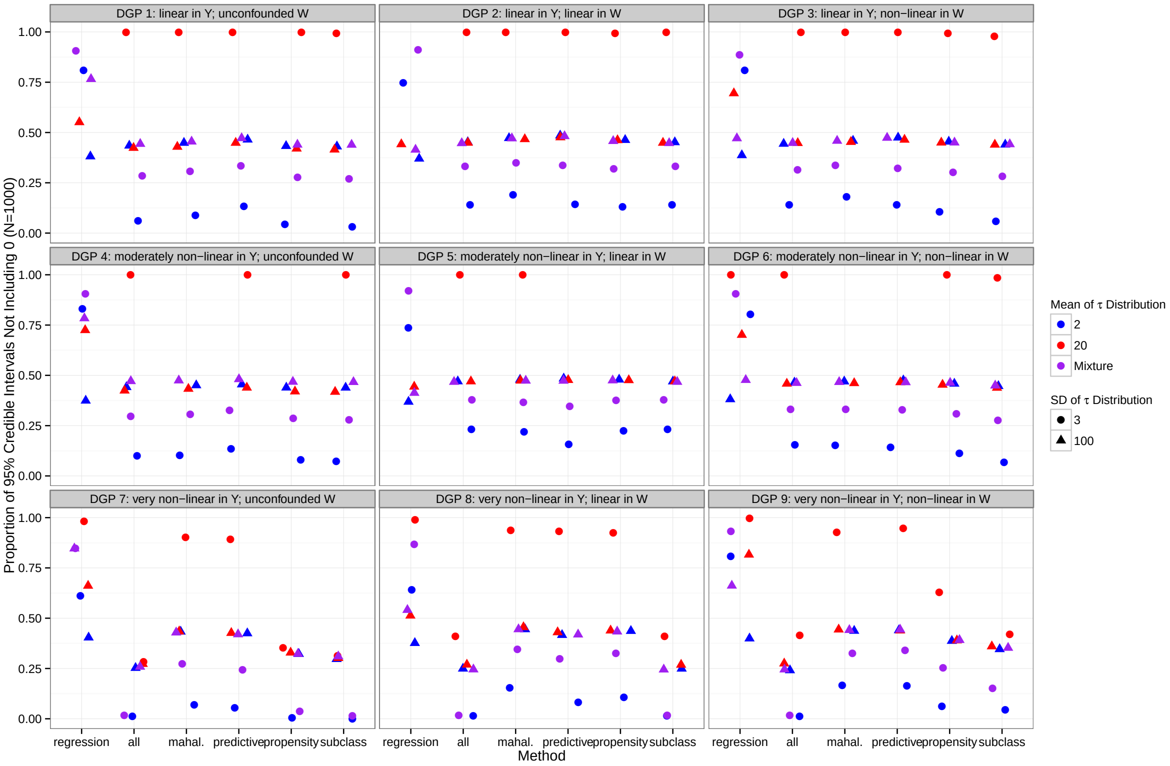

Figure A.52: Comparing ICE "Power" with Different τ_i Distributions (continuous outcome)

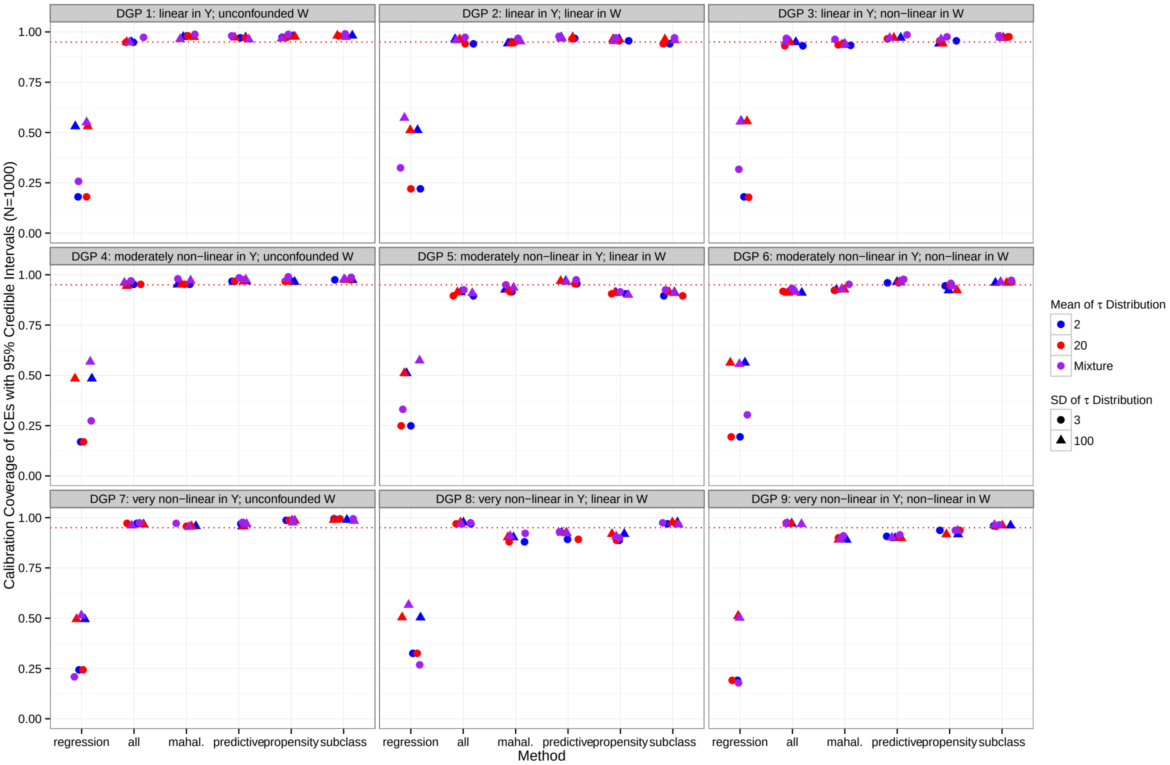

Figure A.53: Comparing ICE Calibration Coverage with Different τ_i Distributions (continuous outcome)

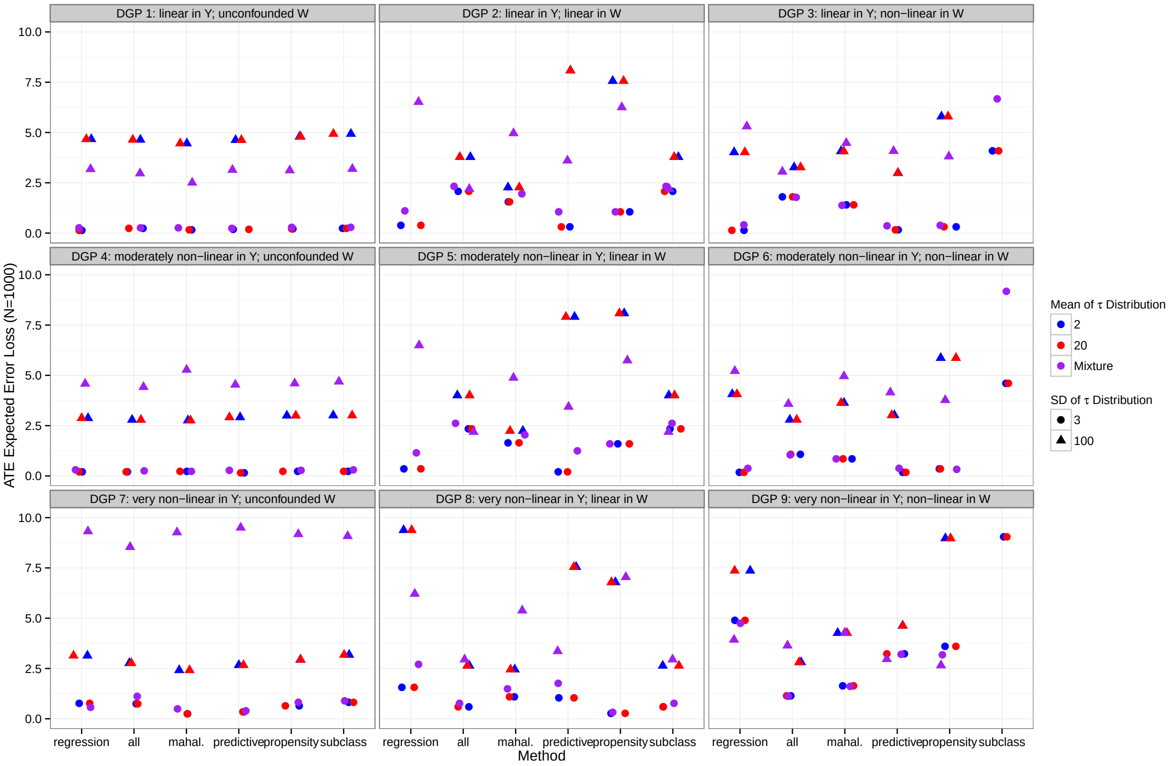

Figure A.54: Comparing ATE Expected Error Loss with Different τ_i Distributions (continuous outcome)

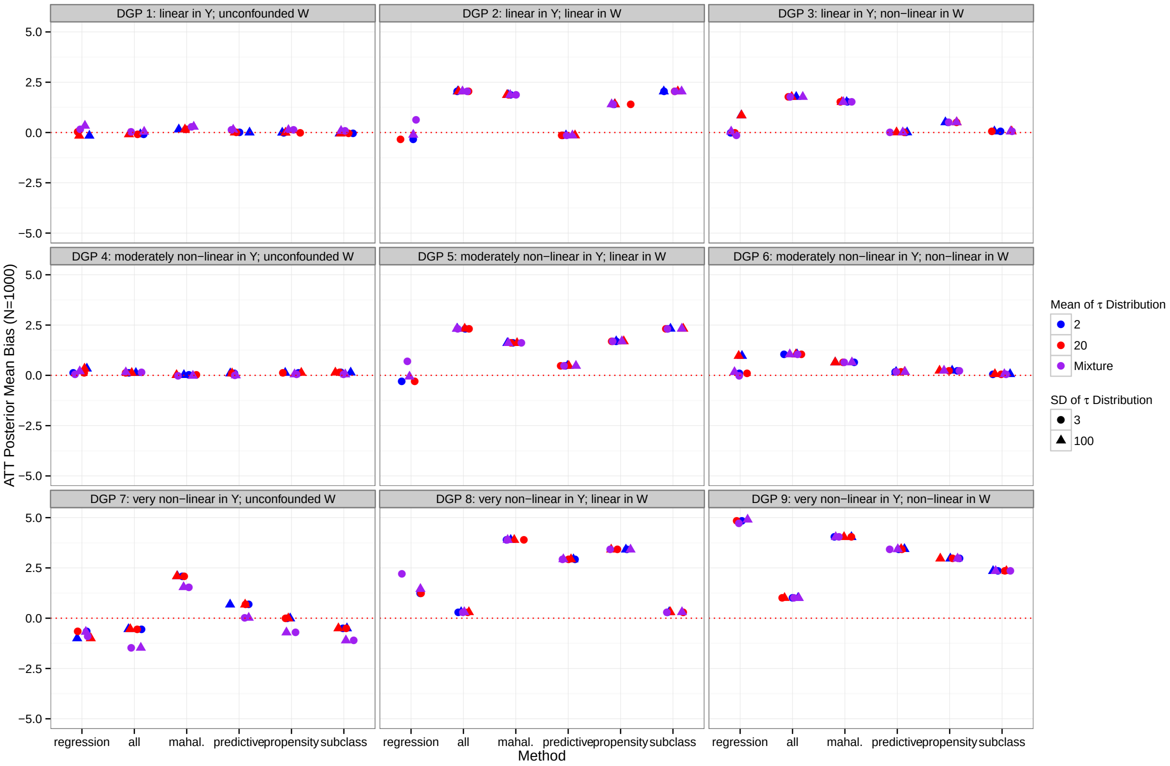

Figure A.55: Comparing ATT Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

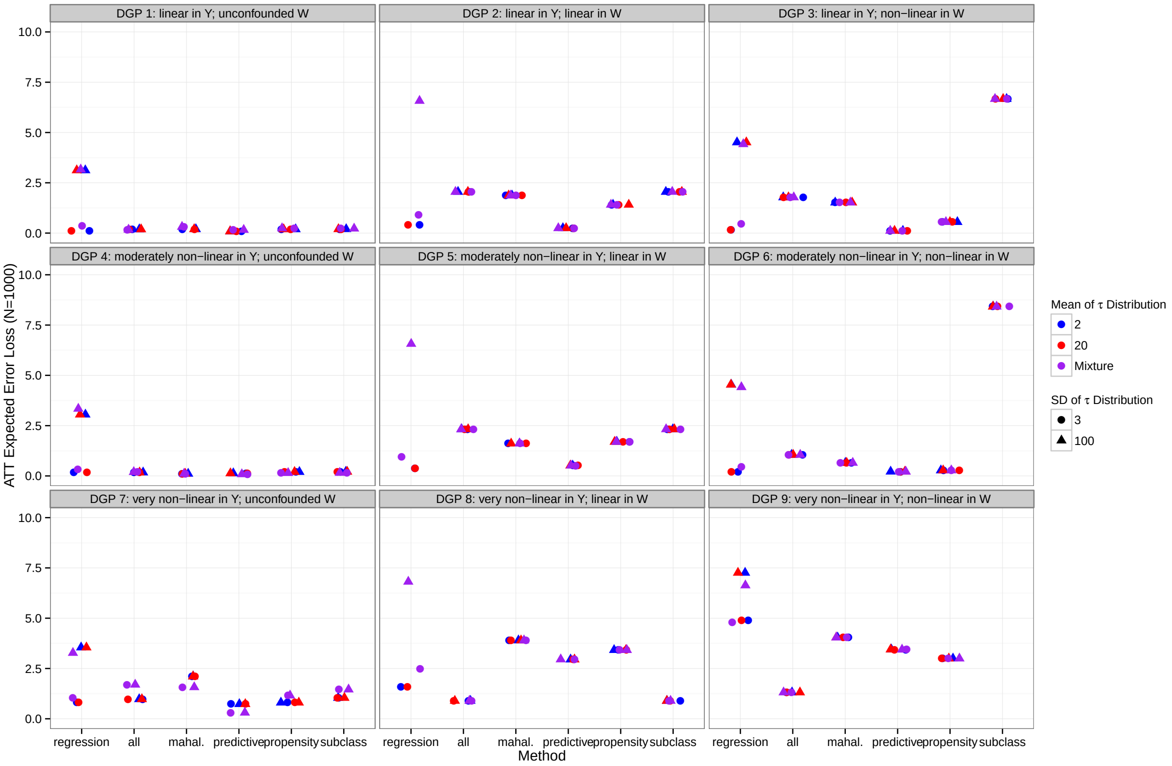

Figure A.56: Comparing ATT Expected Error Loss with Different τ_i Distributions (continuous outcome)

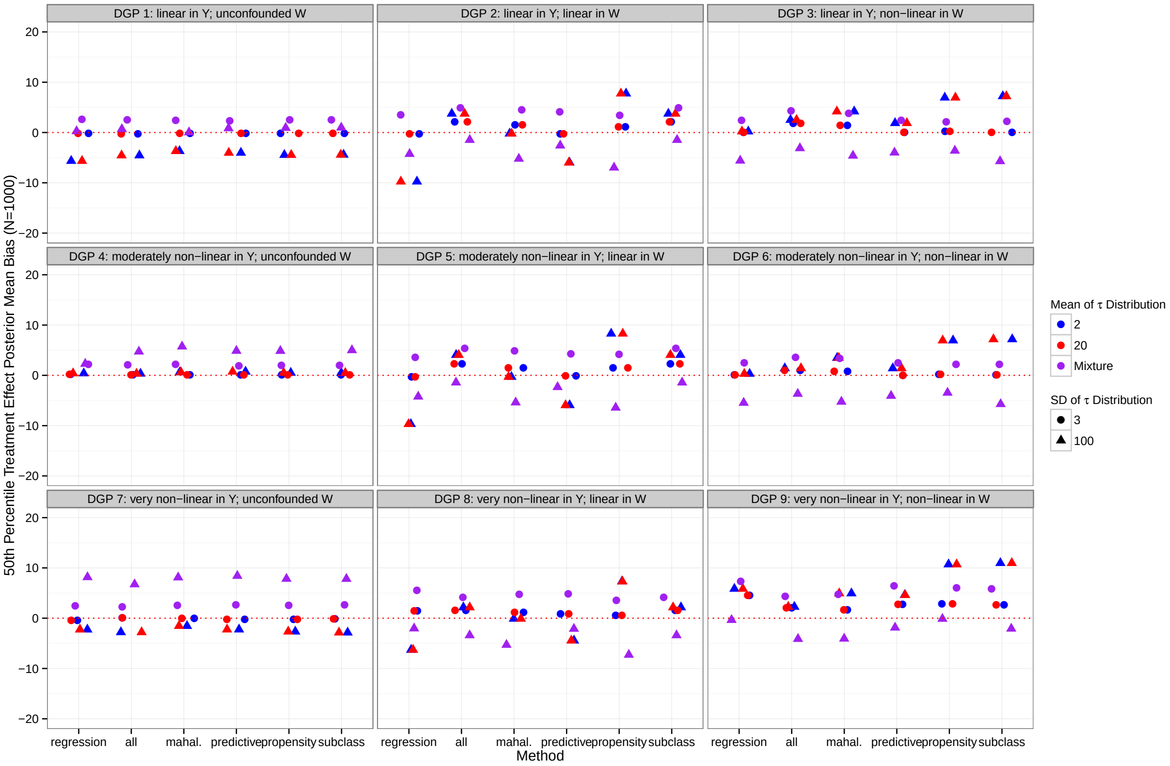

Figure A.57: Comparing 50th Percentile Treatment Effect Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

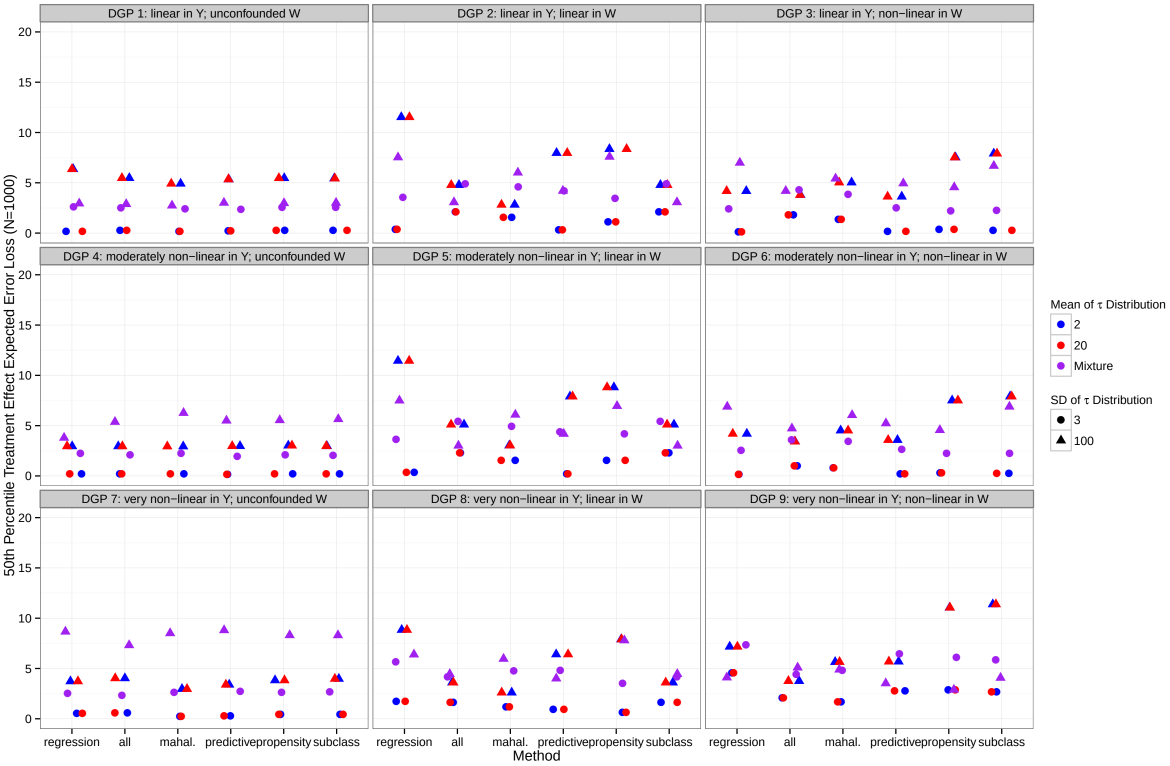

Figure A.58: Comparing 50th Percentile Treatment Effect Expected Error Loss with Different τ_i Distributions (continuous outcome)

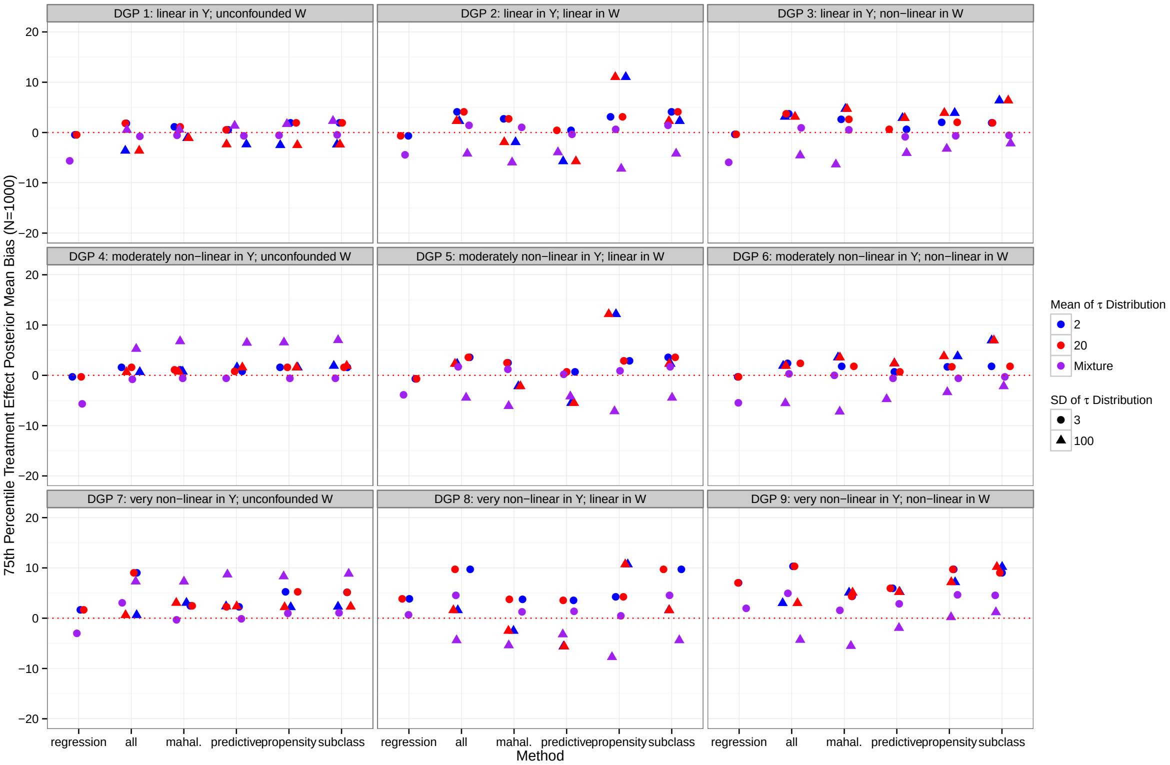

Figure A.59: Comparing 75th Percentile Treatment Effect Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

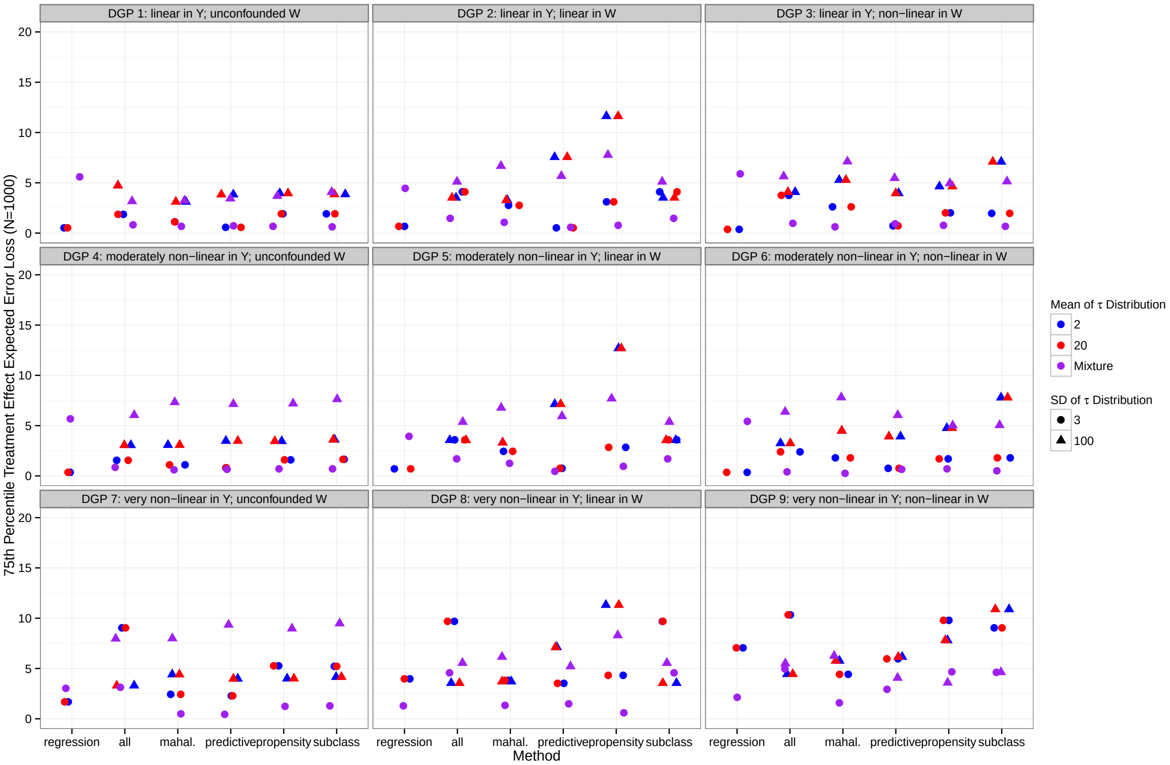

Figure A.60: Comparing 75th Percentile Treatment Effect Expected Error Loss with Different τ_i Distributions (continuous outcome)

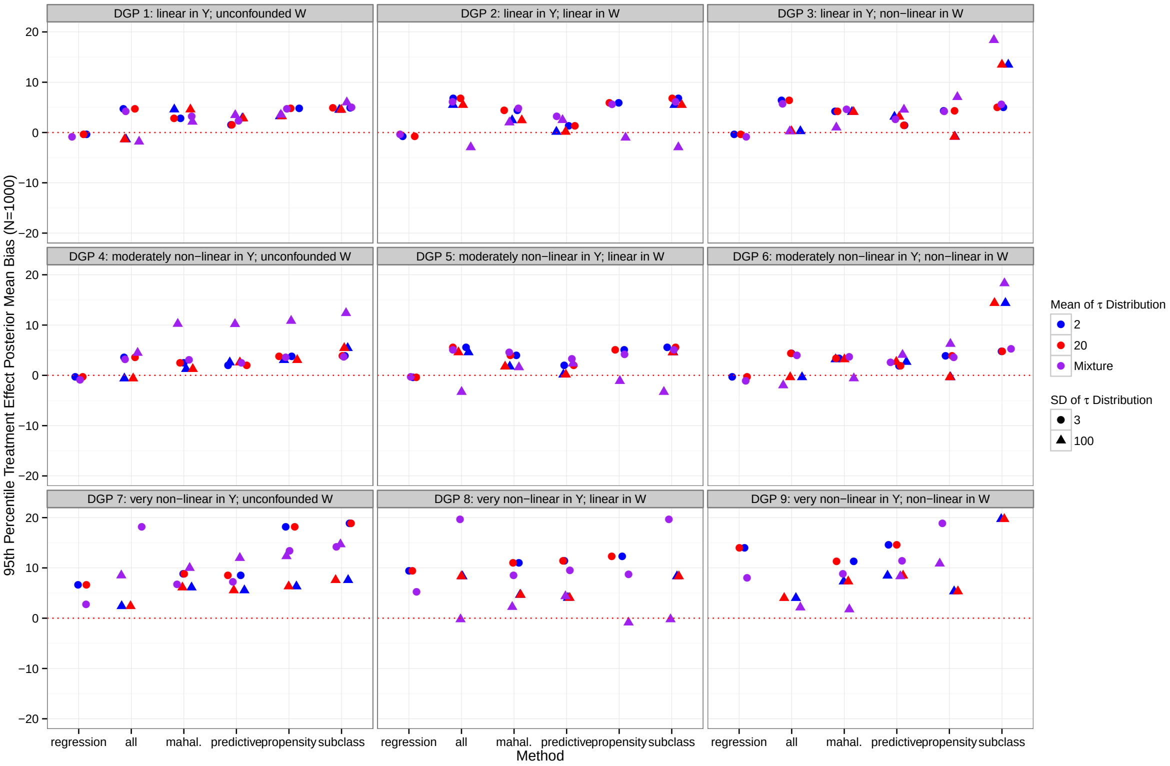

Figure A.61: Comparing 95th Percentile Treatment Effect Posterior Mean Bias with Different τ_i Distributions (continuous outcome)

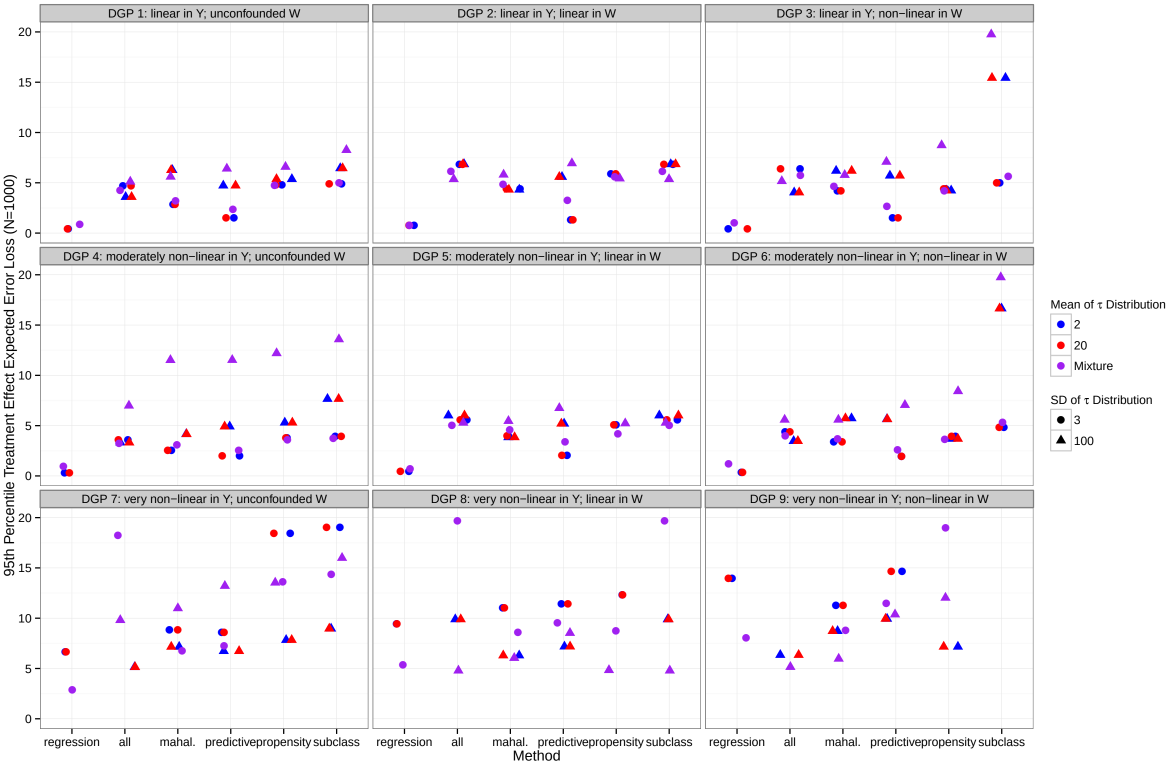

Figure A.62: Comparing 95th Percentile Treatment Effect Expected Error Loss with Different τ_i Distributions (continuous outcome)

A.5 Simulations for Binary Outcomes

The simulations for binary outcomes test the same methods and ideas as the simulations for the continuous outcomes. Because of the nature of ICEs for binary outcomes with binary treatment, where each τ_i can only take on a value of -1, 0, 1, it is tough to develop the ICE proportion of 95% credible intervals including 0 ("power") or the ICE calibration coverage ("coverage") metrics. The posterior draws for each ICE consist of values of 0 and 1 or -1, depending on the treatment assignment and observed outcomes. Since it is unlikely that 95% or more of the posterior draws are of the same value, the 95% credible intervals almost certainly contain both possible values, so the two metrics are meaningless. Therefore, I only present the metrics of posterior mean bias ("bias") and expected error loss ("root mse"). For similar reasons, I only calculate and present results for two causal estimands, the ATE and the ATT. The simulation testing capabilities for binary outcomes are much more limited, so I rely mostly on the simulations for continuous outcomes to reach my conclusions. However, the results that I do calculate for the simulations for binary outcome variables are very similar to the results for continuous outcomes.

The data generating process for binary outcomes is also very similar to that of continuous outcomes. I use the same covariates generated before. Once again, there are nine different data generating processes with three different sample sizes. I first take the continuous outcomes for $Y(0)$ from before:

1. $Y(0) = x_1 + x_2 + x_3 - x_4 + x_5 + x_6 + x_7 - x_8 + x_9 - x_{10}$
2. $Y(0) = x_1 + x_2 + 0.2x_3x_4 - \sqrt{x_5} + x_7 + x_8 - x_9 + x_{10}$
3. $Y(0) = (x_1 + x_2 + x_5)^2 + x_7 - x_8 + x_9 - x_{10}$

To generate a binary outcome, I simply assign $Y_i(0) = 1$ if the continuous outcome is greater than the mean of all the $Y(0)$ and $Y_i(0) = 0$ if the continuous outcome is less than the mean. The treatment assignment generating process stays the same as before.

1. $p(W = 1) = 0.5$
2. $\eta = x_1 + 2x_2 - 2x_3 - x_4 - 0.5x_5 + x_6 + x_7$

$W = 1$ if $\eta > 0$; otherwise $W = 0$

3. $\eta = 0.5x_1 + 2x_1x_2 + x_3^2 - x_4 - 0.5\sqrt{x5} - x_5x_6 + x_7$

   $W = 1$ if $\eta > 0$; otherwise $W = 0$

To generate τ_i, I use the following formula:

If $Y_i(0) = 0$,

$$\begin{aligned} P(\tau_i = 1) &= 0.75 \\ P(\tau_i = 0) &= 0.25 \end{aligned}$$

If $Y_i(0) = 1$,

$$\begin{aligned} P(\tau_i = -1) &= 0.4 \\ P(\tau_i = 0) &= 0.6 \end{aligned}$$

Given $Y_i(0)$, W_i, and τ_i, then

$$\begin{aligned} Y_i(1) &= Y_i(0) + \tau_i \\ Y_i &= W_iY_i(1) + (1 - W_i)Y_i(0) \end{aligned}$$

A.6 Comparing Methods for Binary Outcomes

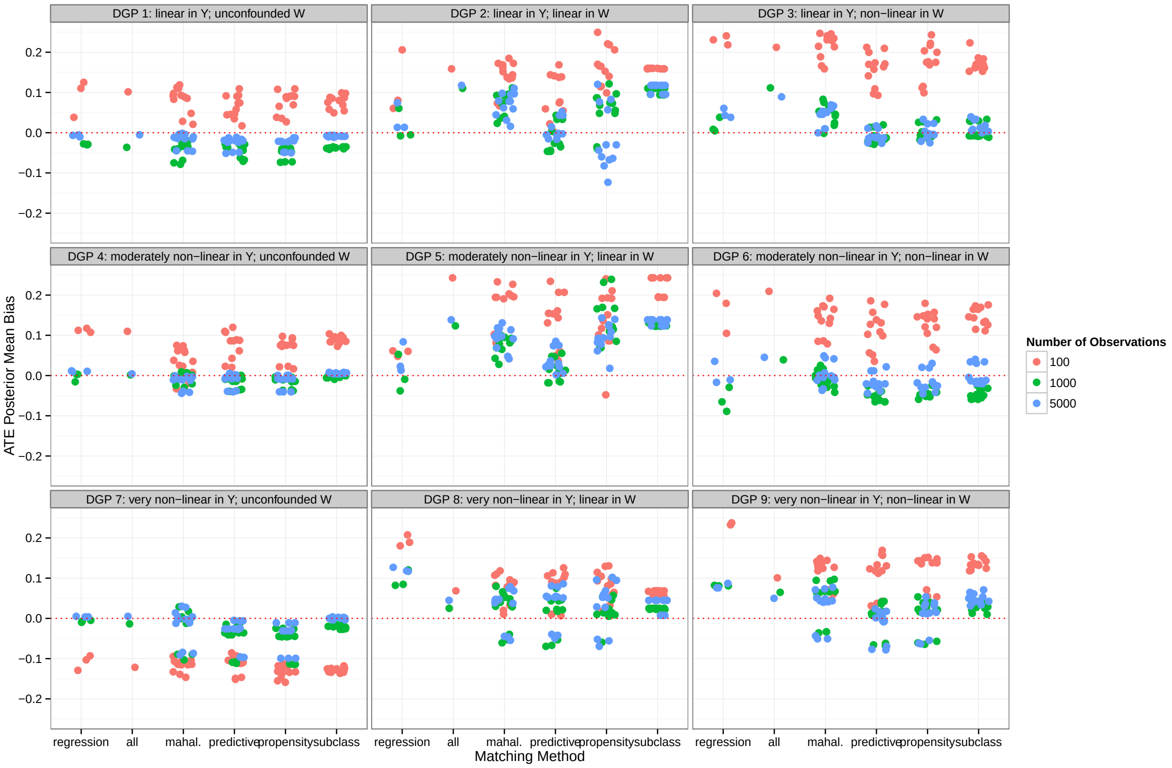

Figure A.63: Comparing ATE Posterior Mean Bias for Different Matching Methods (binary outcome)

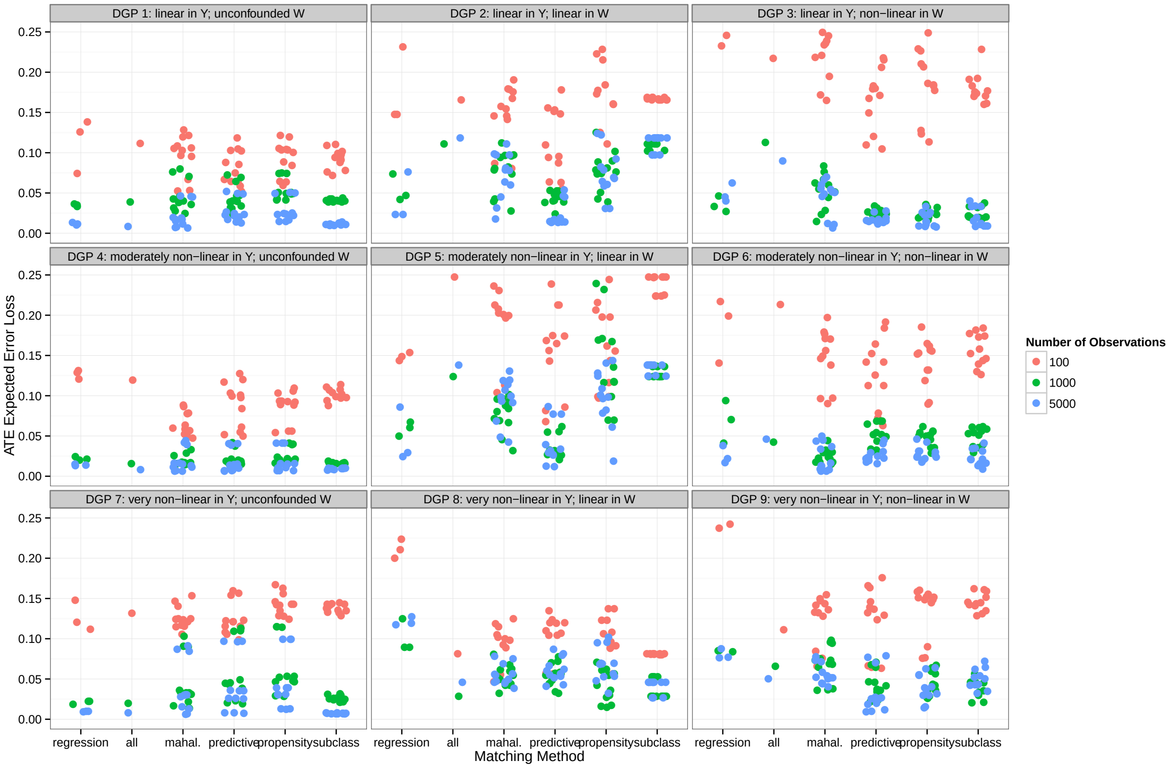

Figure A.64: Comparing ATE Expected Error Loss for Different Matching Methods (binary outcome)

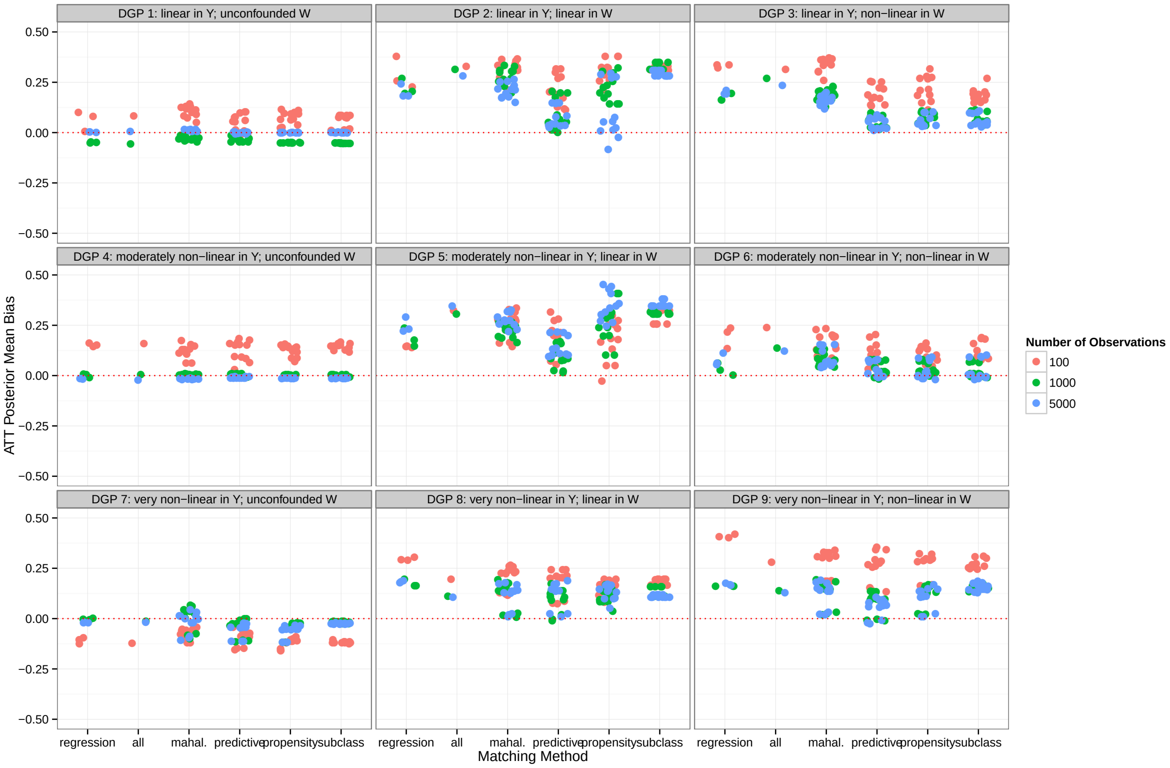

Figure A.65: Comparing ATT Posterior Mean Bias for Different Matching Methods (binary outcome)

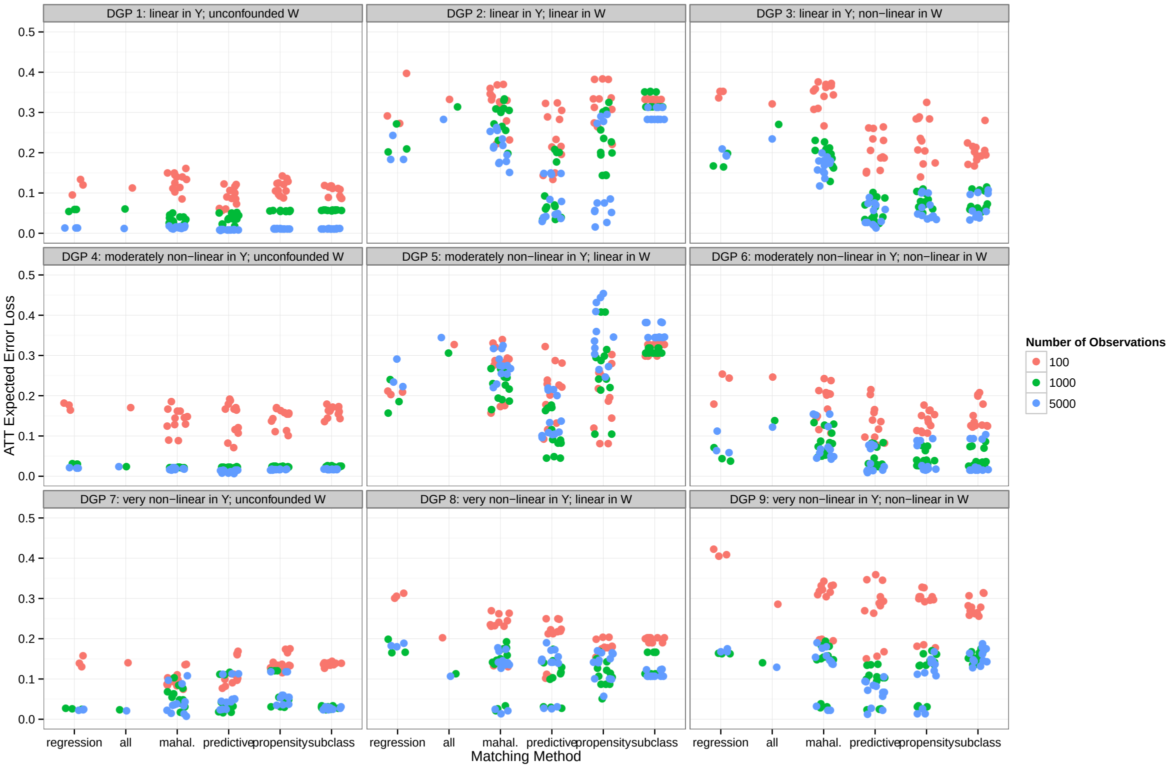

Figure A.66: Comparing ATT Expected Error Loss for Different Matching Methods (binary outcome)

A.7 Comparing Number of Conditioning Variables for Binary Outcomes

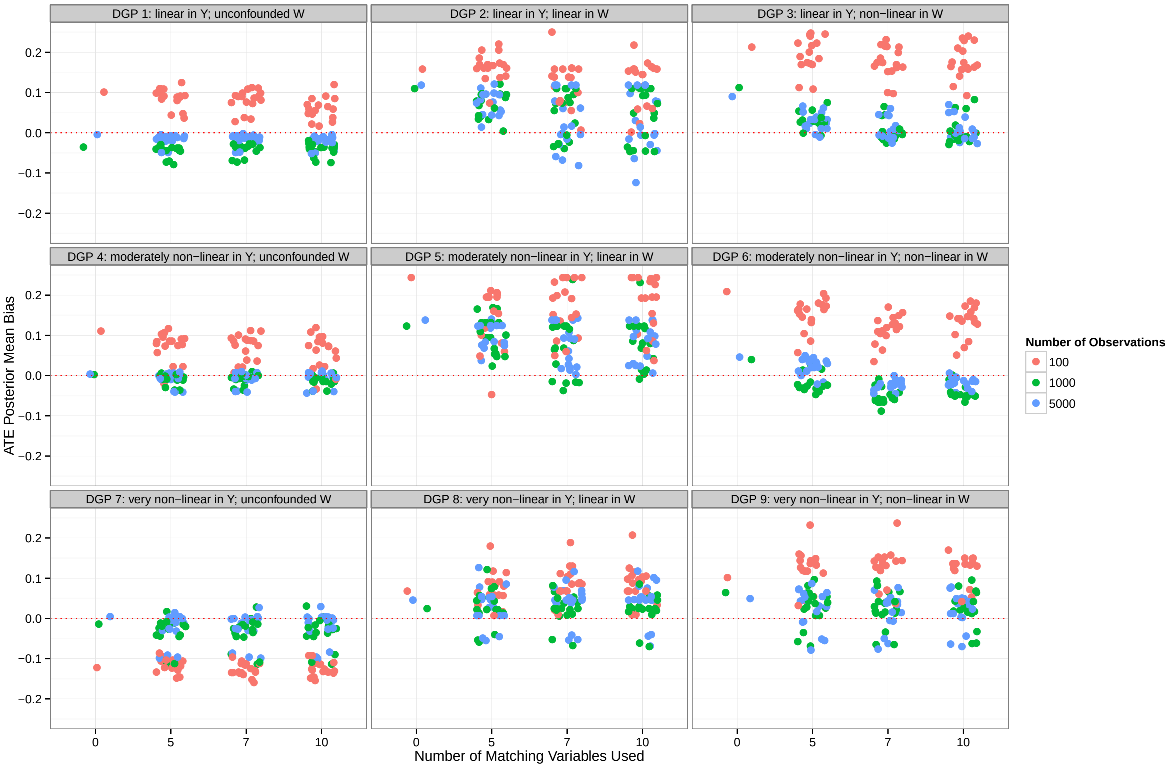

Figure A.67: Comparing ATE Posterior Mean Bias for Different Conditioning Sets (binary outcome)

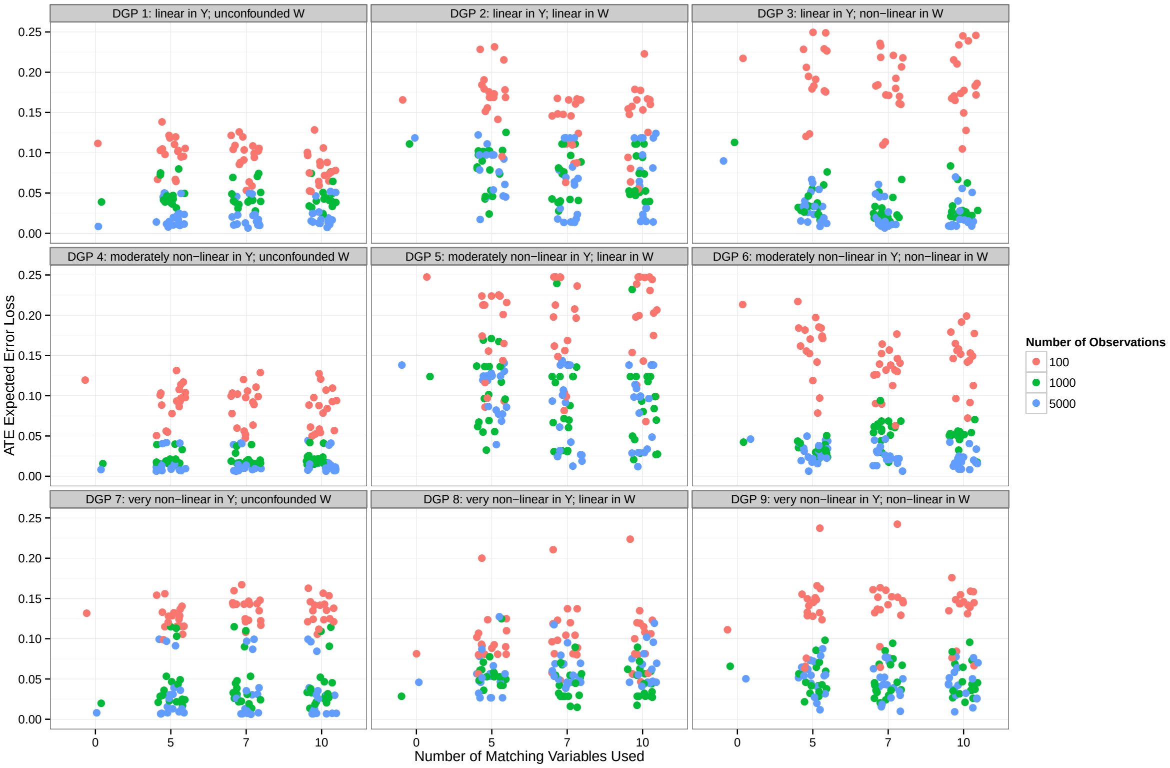

Figure A.68: Comparing ATE Expected Error Loss for Different Conditioning Sets (binary outcome)

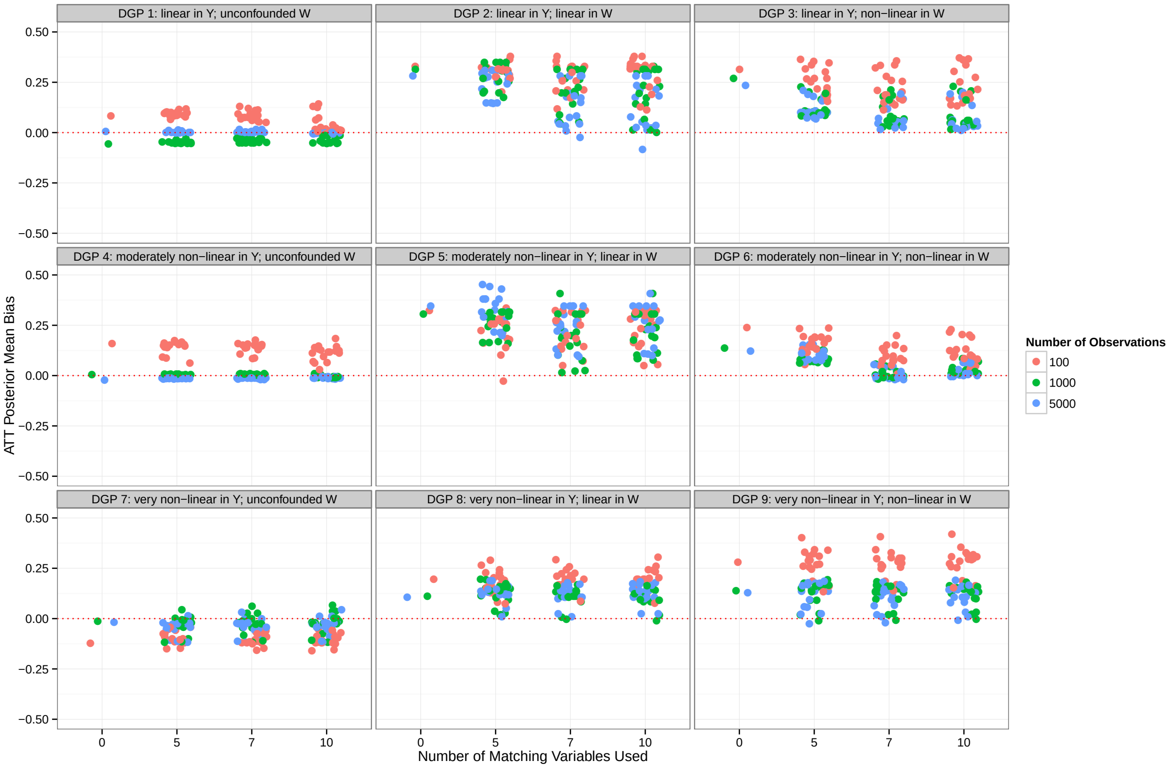

Figure A.69: Comparing ATT Posterior Mean Bias for Different Conditioning Sets (binary outcome)

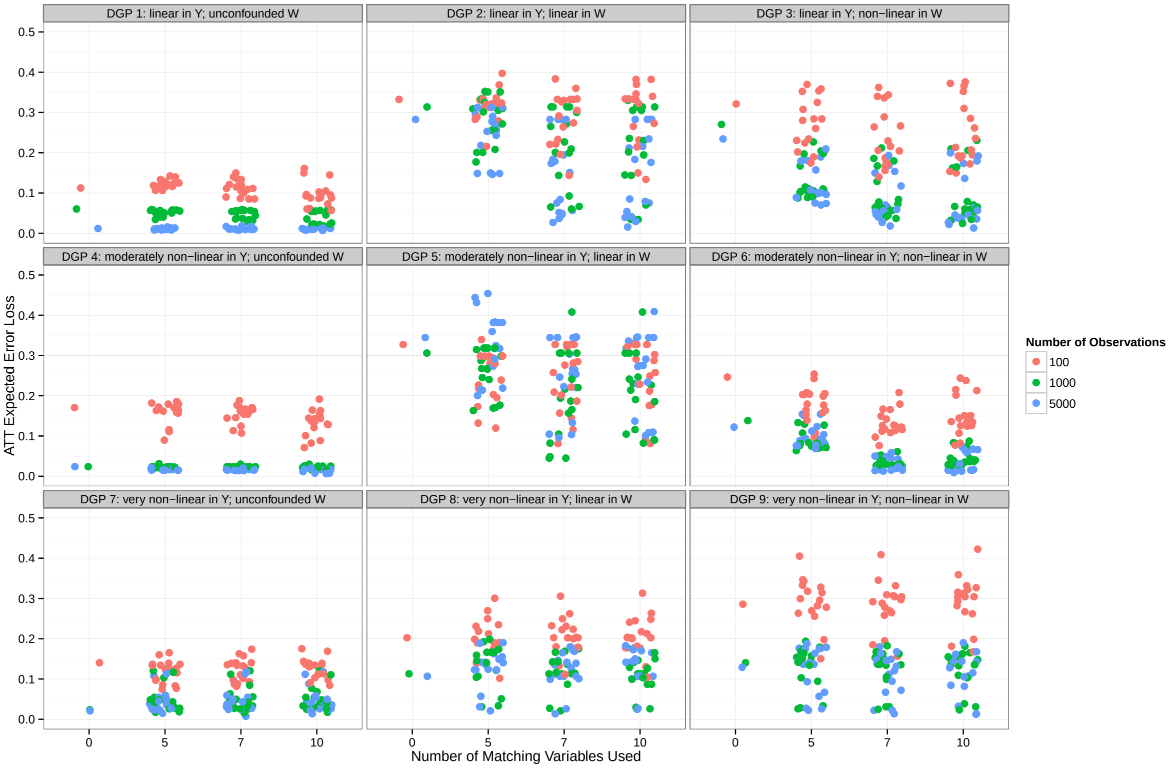

Figure A.70: Comparing ATT Expected Error Loss for Different Conditioning Sets (binary outcome)

A.8 Comparing Number of Matches for Binary Outcomes

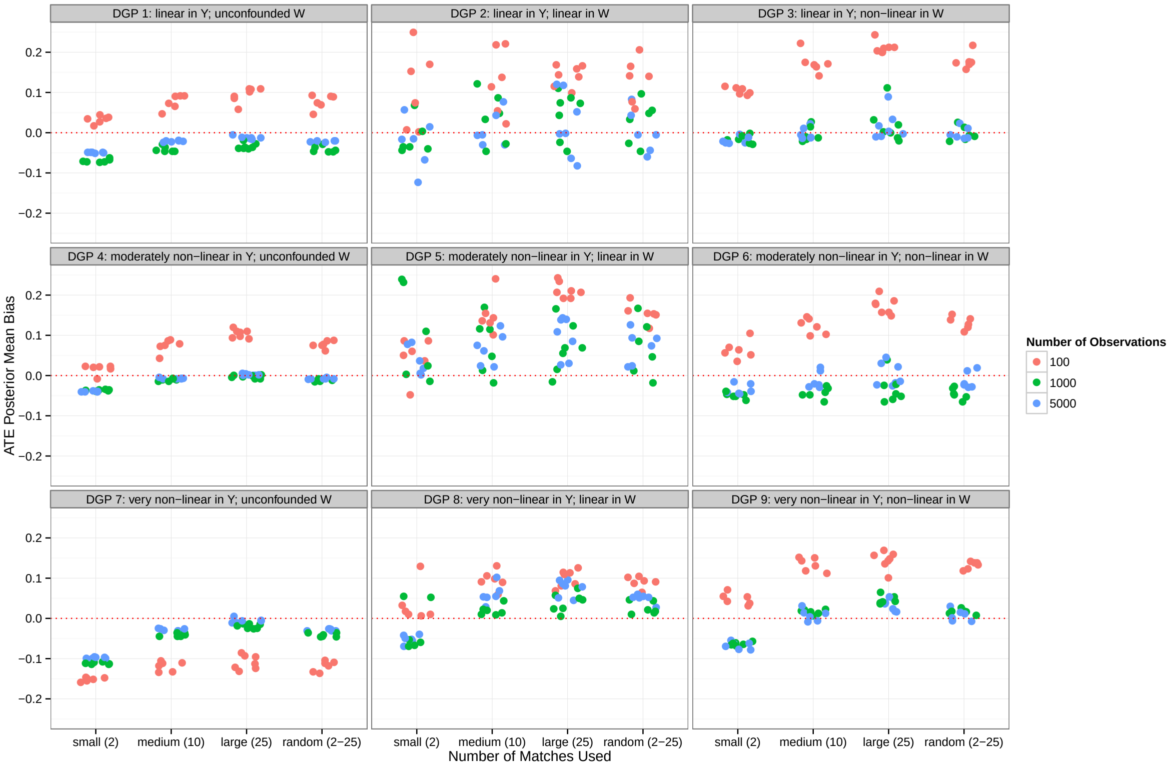

Figure A.71: Comparing ATE Posterior Mean Bias for Different Numbers of Matches (binary outcome)

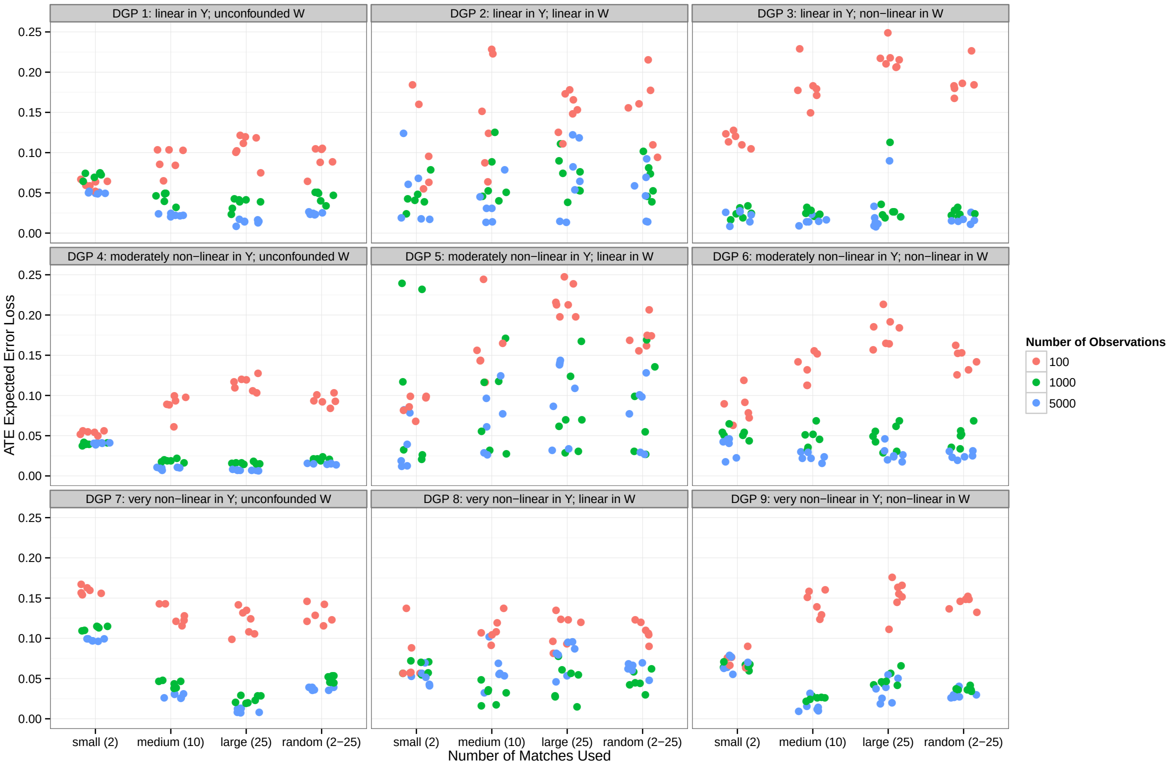

Figure A.72: Comparing ATE Expected Error Loss for Different Numbers of Matches (binary outcome)

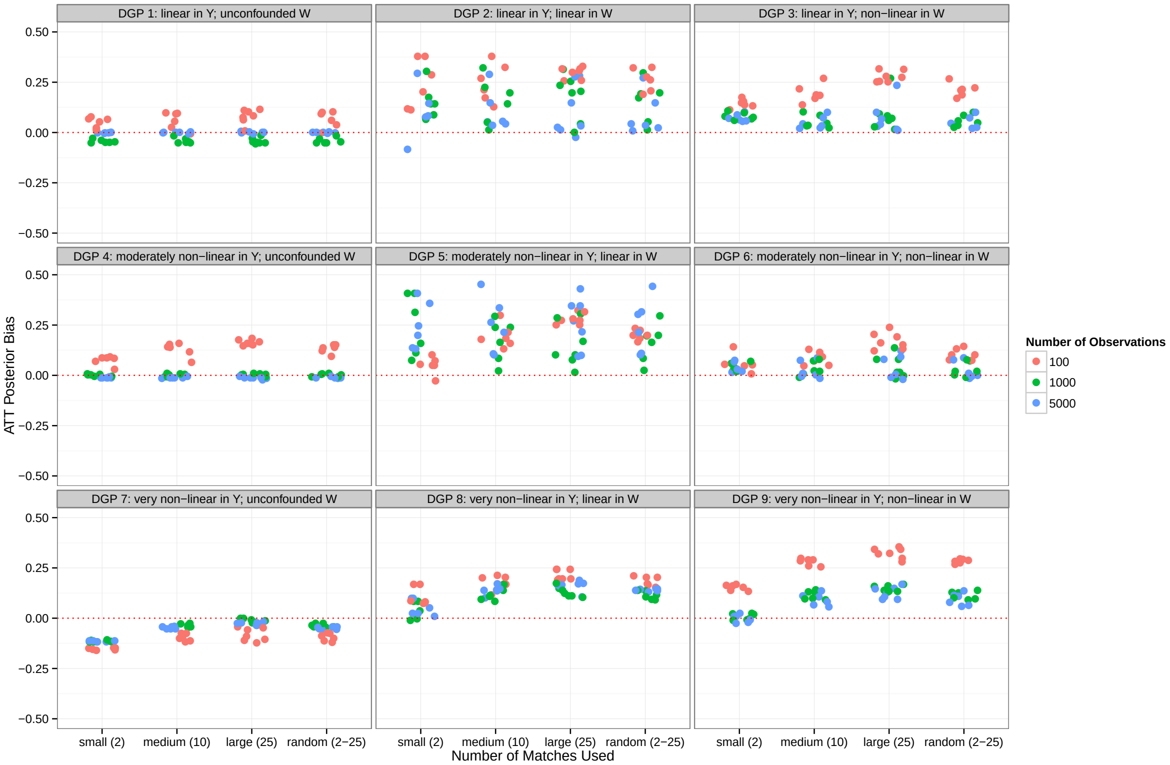

Figure A.73: Comparing ATT Posterior Mean Bias for Different Numbers of Matches (binary outcome)

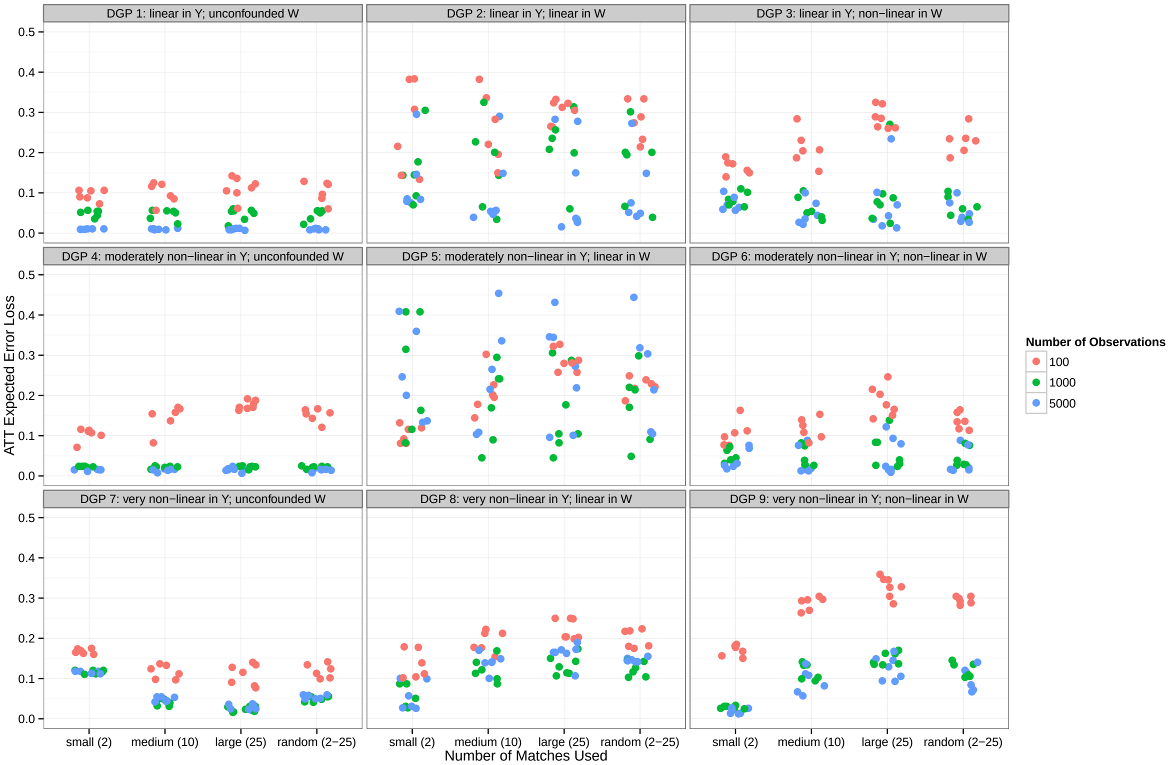

Figure A.74: Comparing ATT Expected Error Loss for Different Numbers of Matches (binary outcome)