DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Topics in Causal Inference

Citation

Kolokotrones, Thomas. 2020. Topics in Causal Inference. Doctoral dissertation, Harvard T.H. Chan School of Public Health.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:42676021

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

TOPICS IN CAUSAL INFERENCE

THOMAS MICHAEL KOLOKOTRONES

A Dissertation Submitted to the Faculty of

The Harvard School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Science

in the Departments of Epidemiology and Biostatistics

Harvard University

Boston, Massachusetts.

May, 2020

Dissertation Advisor: Dr. James Robins Thomas Michael Kolokotrones

Topics in Causal Inference

Abstract

Methods for drawing causal inferences from observational data play a central role in fields such as the social sciences and epidemiology, in which performing experiments may be difficult or impossible for technical, practical, or ethical reasons. Though regression methods remain the most popular, many other techniques are also widely used. We focus on two of them, matching and instrumental variables.

The matching literature is fiercely divided over the optimal method for matching, with the majority of investigators advocating for either direct covariate or propensity score matching. We compare the performance of these two techniques in estimating the average effect of treatment on the treated using a variety of metrics including bias, variance, and model dependence, a measure of a bias corrected matching estimator's sensitivity to the regression model used. We find that neither method dominates the other and that which one is preferred will depend on the distribution of the covariates, the structure of the true and regression models, and the numbers of treated and untreated subjects, as well upon how many covariates are actually matched.

We also explore the use of instrumental variables when the exclusion restriction is violated, focusing particularly on the use of Egger Regression to analyze Mendelian Randomization studies. Though this estimator is widely used, it has not been rigorously analyzed. We do so here, giving conditions under which it is consistent and providing its limit under other circumstances. We also show that, when only finitely many instruments are used, it is biased, but asymptotically normal, and compute its properties in the setting in which all quantities except the causal effect are known, which provides a bound on the rate of convergence of the standard estimator, in which these quantities are estimated by linear regression.

Contents

List of Figures

List of Tables

This document contains no tables.

Chapter 1

Introduction

Since time immemorial, humans have drawn causal inferences from observational data. Though such inferences have led to dramatic advances in our understanding of the human and natural worlds, and are responsible for much of our technological and scientific progress, they are fraught with peril due to phenomena such as confounding and selection bias. Although the average person is aware that "correlation does not imply causation," they rarely fully understand why, since even many experts do not fully appreciate the threats to validity present in a particular problem. While philosophers have long been fascinated by the nature of causation, and under what assumptions it can be inferred, it was only in the twentieth century that we developed formal statistical frameworks for it, chief of which is Neyman's Potential Outcomes framework [4], which forms the foundation of work by Donald Rubin [7] and James Robins [6] (separately) that led, along with the substantial progress on graphical models by Judea Pearl [5], to the current state of the field. Though there are many approaches to controlling for confounding and drawing valid causal inferences from observational data, many of which were developed far before our modern understanding of the problem, we will focus on two, more specialized, approaches, which are widely used in the social sciences and epidemiology: matching and instrumental variables.

The first two chapters of the following work focus on matching, which attempts to control for potential confounding by matching each "treated" or "exposed" subject to a corresponding "untreated" or "unexposed" control who is as much like that subject as possible, so that the two differ meaningfully only in their treatment or exposure status. Under such circumstances, the difference in outcomes between the two will be due only to the effect of treatment or exposure and, possibly, idiosyncratic factors that are not related to whether or not they were treated or exposed. In particular, if the matching is sufficiently good (oftentimes meaning perfect) then the matched analysis will control for confounding, the potential that treatment/exposure is related to other characteristics of the individual and these characteristics affect the outcome (in the extreme case, it is only these associated characteristics, not treatment/exposure itself, that affect the outcome), since all relevant characteristics will be sufficiently similar in the treated/exposed subject and his

matched control that they will have the same effect on the outcome in both and any differences will be solely due to the effects of treatment/exposure.

However, there are many ways in which to match subjects, and, when the matching process is imperfect, as it almost always is, the choice of matching technique will affect who is matched with whom, and, thus, potentially, the estimated causal effect. Two of the most popular matching methods are matching directly on some collection of covariates, typically those that are felt to be potential confounders for the causal effect of interest, in order to minimize some distance metric that is related to the difference in covariate values (or some transformation thereof), and matching on the propensity score, the probability that an individual will be treated/exposed given his covariates, so that matched pairs differ in their propensity score as little as possible. Both approaches have strong proponents and which technique is superior is vigorously debated in the literature.

In order to address the question of whether one matching method dominates the other, we explore how the bias and variance of matching estimators, using either direct or propensity score matching, depend on the number of treated/exposed subjects and controls and the number and distribution of matched covariates. In addition to the simple matching estimators described above, we also evaluate methods that combine matching with regression in order to account for the fact that, when matching is not perfect, the covariate values of a treated/exposed subject and his matched control may differ, sometimes substantially.

Such bias-corrected matching estimators, which combine matching and regression can vary significantly in their estimated causal effects depending on the choice of regression model. Thus, choices by the analyst can affect the answer, which weakens one of the primary arguments for using matching over other methods: that matching is "less parametric" than regression, and, thus, less dependent on choices made by the analyst. To explore this question, we examine how much the estimated causal effect can change based on the choice of regression model. In particular, for a collection of regression models, we evaluate the empirical variance, over models, of the estimated causal effect, which we refer to as the Model Dependence, following King and Nielsen [3]. As with the bias and variance for a single matching estimator, we examine how this differs between direct and propensity score matching and how this difference varies with the number of treated/exposed subjects and controls and the number and distribution of matched covariates.

Having explored the properties of matching estimators in the prior two chapters, for the final

main chapter, we turn our attention to another means of estimating causal effects in the presence of potential confounding: Instrumental Variable (IV) methods. An Instrumental Variable is a quantity that is associated with the potential risk factor, but not with the outcome, except through the risk factor's effect on the outcome. If one has a true IV, then this allows one to estimate the causal effect of the potential risk factor on the outcome using a variety of standard methods. In epidemiology, a popular choice for instruments has been genetic markers, which is referred to as Mendelian Randomization (MR), based on the idea that, for many risk factors, genes that affect the level/state of the risk factor are known. However, many genes actually have multiple effects and, thus, may affect the outcome, either directly or indirectly, via pathways other than their effect on the potential risk factor, so-called pleiotropy. In response to this issue, alternative IV estimators, that can tolerate pleiotropy have been developed. Perhaps the most popular is Egger Regression, which was originally developed for evaluating meta-analyses, but was repurposed to allow the use of Mendelian Randomization, even in the presence of pleiotropy, under particular assumptions about the association between the selected genetic markers and the potential risk factor and their association with the outcome (the strength of pleiotropy), the so-called InSIDE assumption [1, 2].

However, despite this estimator being widely used, it has never been formally characterized or even been rigorously shown to be consistent (the arguments for its use are based on informally taking sequential limits as first the sample size then the number of instruments goes to infinity). We begin the process of formally characterizing its asymptotic properties in terms of the strength of the instruments (the association of the markers with the potential risk factor), the extent of pleiotropy (the association of the markers with the outcome), and violations of the InSIDE assumption.

Bibliography

[1] Jack Bowden, George Davey Smith, and Stephen Burgess. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression". *International Journal of Epidemiology* 44.2 (2015), pp. 512–525.

[2] Matthias Egger, George Davey Smith, and Christoph Minder. "Bias in meta-analysis detected by a simple, graphical test". *BMJ* 315 (1997), pp. 629–34.

[3] Gary King and Richard Nielsen. "Why Propensity Scores Should Not Be Used for Matching". *Political Analysis* 27.4 (2019).

[4] Jerzy Neyman. "Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes". *Master's Thesis* (1923).

[5] Judea Pearl. "Reverend Bayes on inference engines: A distributed hierarchical approach". *Proceedings, AAAI-82* (1982).

[6] James Robins. "A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect". *Mathematical Modelling* 7 (1986), pp. 1393–1512.

[7] Donald Rubin. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". *J. Educ. Psychol.* 66.5 (1974), pp. 688–701.

Chapter 2

Bias and Variance of Matching Estimators

Tom Kolokotrones, Max Goplerud, Richard Nielsen, Gary King, and James Robins

2.1 Introduction

In a widely-cited 2019 paper, boldly entitled "Why Propensity Scores Should not be Used for Matching," King and Nielsen argue that "propensity score matching... often accomplishes the opposite of its intended goal—thus increasing imbalance, inefficiency, model dependence, and bias," even when the propensity score is known as in a (possibly stratified) randomized clinical trial, and, thus, as the title states, should not be used for matching [5]. Given the strength of this assertion, we ask whether the superiority of direct over propensity score matching is indeed universal, or whether there may be some cases, possibly unusual, in which matching on the propensity score does, in fact, result in better performance. In order to evaluate this, we consider the bias and variance of matching estimators for the Average Effect of Treatment on the treated (ATT) when directly matching on the covariates (or some transformation of them, as in Mahlanobis matching) vs. when matching on the propensity score [6, 7]. The ATT is commonly estimated using both matching methods, and, indeed, such estimators are the focus of King and Nielsen's paper [5].

In order to sharpen our analysis, we consider an extreme case which provides an apparently severe disadvantage to the propensity score: the case in which the propensity score is a known constant. This implies that the distribution of covariates is identical among both treated and untreated subjects and, thus, treatment is not (unconditionally) confounded by either measured or unmeasured covariates. Thus, in our setting, propensity score matching cannot use the covariates at all, and, instead, matches at random; while direct matching can potentially utilize all of the information available in the observed covariates. On its face, this is very strong limitation for propensity score matching, but also highlights a major complaint King and Nielsen have about matching on the propensity score: that propensity score matching is similar to random matching in that it often matches points that are far apart in covariate space, even when much closer matches are available [5]. Therefore, using a constant propensity score only exacerbates what King and Nielsen

consider to be one of the greatest problems with the method and, thus, if this extreme version of propensity score matching is able to outperform direct covariate matching by any criterion, this will provide strong evidence that there are indeed settings in which propensity scores should be used for matching, or, at least, favored over direct covariate matching.

Further, in practice, the propensity score is typically not known and must be estimated. However, when the model for the propensity score is correctly specified, matching on the estimated propensity score obtained by fitting a properly specified parametric model is known to improve performance over using the true score, even when it is known [3]. Therefore using the true propensity score further disadvantages this version of propensity score matching vs. what is used in practice, which should further bias our results in favor of direct covariate matching.

The structure of the remainder of the paper is as follows. Section 2 introduces the notation. Section 3 introduces three types of estimators for the Average Effect of Treatment on the Treated, as described by Abadie and Imbens, and explores their bias and robustness properties [1, 2]. Section 4 further discusses the bias of matching estimators under direct vs. propensity score matching and explores how this is related to properties of the distribution of the covariates. Section 5 derives expressions for the variance of matching estimators, particularly their dependence on the matching discrepancy. Section 6 explores how the variance of matching estimators is affected by the distribution of the covariates as well as compares the relative variance of matching estimators using direct vs. propensity score matching. Section 7 concludes.

We find that there are settings in which estimators of the ATT that use propensity score matching outperform those that use direct covariate matching. The reason for this phenomenon is that, rather surprisingly, matching directly on the covariates can introduce bias, even in the total absence of confounding, as is the case under a constant propensity score.

2.2 Basics

Let Y be a continuous outcome, A a binary treatment (where $A = 1$ corresponds to treatment and $A = 0$ to no treatment), X a vector of continuous covariates, and $Y(a)$ the counterfactual outcome under treatment a. We use $[n]$ to denote the set $\{1, \ldots, n\}$, as is typical in Computer Science. Consider a collection of N_1 treated and N_0 untreated subjects. Our goal is to estimate the

expected effect of treatment, A, on the outcome, Y, among those subjects that were treated, the so-called Average Effect of Treatment on the Treated (ATT), which is given by $\tau = \mathrm{E}[Y(1) - Y(0)|A = 1]$.

Consider a matching criterion M, which is a function $M : [N_1] \to [N_0]^m$ that associates with each of the N_1 treated subjects m of the N_0 untreated controls. In what follows, we will restrict attention to the case of 1-1 matching with replacement, so that $m = 1$ and each treated subject is matched to exactly one untreated control (although multiple treated subjects may be matched to the same untreated control). In the typical case, a match is defined as the untreated control $j = M(i) \in [N_0]$ that minimizes some distance $d(X_i, X_j)$, where d is some metric, such as the Euclidean or Mahlanobis distance or the absolute difference in propensity score. Let $\tau_i = Y_i(1) - Y_i(0)$ be the individual treatment effect for subject i. Since exactly one of $Y_i(0)$ and $Y_i(1)$ will actually be observed, τ_i must be estimated. For any random variable, Z, we use Z to represent the values of the treated subjects and Z_M for the values of the matched controls and define $\Delta Z_i = Z_{M(i)} - Z_i$ and $\overline{Z} = N_1^{-1} \sum_{i=1}^{N_1} Z_i$. Following Abadie and Imbens, we refer to ΔZ as the matching discrepancy of Z [1, 2].

In simple matching, $Y_i(0)$ for treated subjects is estimated by $\widehat{Y}_i(0) = Y_{M(i)}$. We can then estimate the effect of treatment on treated individual i by, $\hat{\tau}_i = Y_i - Y_{M(i)}$ and estimate the ATT as $\hat{\tau} = N_1^{-1} \sum_{i=1}^{N_1} \left(Y_i - Y_{M(i)}\right)$. Unfortunately, as we will see below, this estimator will, in general, be biased, if the matching function is allowed to depend on the covariates, even when the propensity score is constant and there is no confounding for treatment. However, if $Y_i(0)$ satisfies $Y_i(0) = \mu_0(X_i) + \epsilon_i$, $\mathrm{E}[\epsilon_i|X_i] = 0$, then several other estimators become available [1, 2]. The first is the regression imputation estimator, which estimates the effect of treatment on treated individual i by $\hat{\tau}_i = Y_i - \hat{\mu}_0(X_i)$, and the ATT by $\hat{\tau} = N_1^{-1} \sum_{i=1}^{N_1} \left(Y_i - \hat{\mu}_0(X_i)\right)$, where $\hat{\mu}_0$ is estimated using the data from some subset of the untreated controls. The second is referred to as the bias-corrected matching estimator and combines the simple matching and regression estimators in order to account for the fact that, in general, $X_i \neq X_{M(i)}$ [8]. It does so by using a fitted regression function to correct for the matching discrepancy and so estimates $Y_i(0)$ by $\widehat{Y}_i(0) = Y_{M(i)} + \hat{\mu}_0(X_i) - \hat{\mu}_0\left(X_{M(i)}\right)$, instead of simply using $Y_{M(i)}$. One can then estimate the effect of treatment on treated individual i by $\hat{\tau}_i = Y_i - \hat{Y}_i(0) = Y_i - Y_{M(i)} + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i)$ and estimate the ATT by $\hat{\tau} = N_1^{-1} \sum_{i=1}^{N_1} \left[Y_i - Y_{M(i)} + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i)\right]$. Our primary focus

throughout the paper will be the bias-corrected estimator for the ATT, but, we will explore each of the three estimators.

Abadie and Imbens carefully analyzed the asymptotic bias, variance, and normality of matching estimators for nearest neighbor matching [1, 2]. However, our focus will be somewhat different. Abadie and Imbens specifically focus on the asymptotic performance of matching estimators after bias correction. In their first paper on the subject, they subtract the exact bias from the matching estimator, while in the second they estimate this bias using a nonparametric estimator. However, this procedure relies on the matching discrepancy shrinking sufficiently rapidly, which will not always hold in what follows. Additionally, Abadie and Imbens focus on nearest neighbor matching, while we are interested in examining the performance of several different forms of matching and so will not restrict ourselves to the nearest neighbor setting.

2.3 Bias of ATT Estimators

In this section, we will examine the bias of the three estimators of the ATT discussed above: simple matching, regression imputation, and bias-corrected matching. Our results will be stated in terms of the matching discrepancy and will not be specific to any particular matching method.

2.3.1 Simple Matching

We begin with simple matching. Let the true model be $Y_i(a) = \mu_a(X_i) + \epsilon_i$, where $\mathrm{E}[\epsilon_i|X_i] = 0$, so that the true treatment effect for subject i is $\tau_i = \mu_1(X_i) - \mu_0(X_i) = \tau(X_i)$. Then, the estimated effect of treatment on subject i is:

$$\begin{aligned}\hat{\tau}_i &= Y_i - Y_{M(i)} = \mu_1(X_i) - \mu_0\left(X_{M(i)}\right) + \epsilon_i - \epsilon_{M(i)} \\ &= \left(\mu_1(X_i) - \mu_0(X_i)\right) - \left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right) - \Delta\epsilon_i \\ &= \tau_i - \left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right) - \Delta\epsilon_i\end{aligned}$$

and the conditional bias will be,

$$\mathrm{E}\left[\hat{\tau}_i|X_i\right] - \tau_i = \mathrm{E}\left[\mu_0(X_i) - \mu_0\left(X_{M(i)}\right)\middle|X_i\right] = \mu_0(X_i) - \mathrm{E}\left[\mu_0\left(X_{M(i)}\right)\middle|X_i\right]$$

so the estimated individual treatment effect will be (conditionally) unbiased if and only if $\mathrm{E}\left[\mu_0\left(X_{M(i)}\right)\middle|X_i\right] = \mu_0(X_i)$, or, put differently, if and only if the (conditional) expected matching discrepancy of $\mu(X)$ for subject i is zero.

The results for the estimator of the ATT, are similar.

$$\hat{\tau} = N_1^{-1}\sum_{i=1}^{N_1}\hat{\tau}_i = N_1^{-1}\sum_{i=1}^{N_1}\tau_i - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right) + N_1^{-1}\sum_{i=1}^{N_1}\Delta\epsilon_i$$

$$\mathrm{E}\left[\hat{\tau}|X\right] - \bar{\tau} = N_1^{-1}\sum_{i=1}^{N_1}\mathrm{E}\left[\mu_0(X_i) - \mu_0\left(X_{M(i)}\right)\middle|X_i\right] = N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0(X_i) - \mathrm{E}\left[\mu_0\left(X_{M(i)}\right)\middle|X_i\right]\right)$$

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = N_1^{-1}\sum_{i=1}^{N_1}\mathrm{E}\left[\mu_0(X_i) - \mu_0\left(X_{M(i)}\right)\right] = \mathrm{E}\left[\mu_0(X_i) - \mu_0\left(X_{M(i)}\right)\right]$$

This leads to an interesting, and somewhat surprising, conclusion: if $\mathrm{E}\left[\mu_0\left(X\right)\right]$ is the same in the treated subjects and their matched controls, then $\hat{\tau}$ will be unbiased (either conditionally or unconditionally, respectively), regardless of any other properties of the matching scheme.

In the case in which μ_0 is an affine function of x, so $\mu_0(x) = \beta_0 + \beta^t x$,

$$\mathrm{E}\left[\hat{\tau}|X\right] - \bar{\tau} = N_1^{-1}\beta^t\sum_{i=1}^{N_1}\left(X_i - \mathrm{E}\left[X_{M(i)}\middle|X_i\right]\right) = -N_1^{-1}\beta^t\sum_{i=1}^{N_1}\mathrm{E}\left[\Delta X_i|X_i\right]$$

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = -\beta^t\mathrm{E}\left[\Delta X_i\right]$$

so the estimated ATT will be conditionally unbiased if the conditional expected matching discrepancy is zero and unconditionally unbiased if the expected matching discrepancy is zero. Note that this is equivalent to expected value of X being the same in the treated subjects and their matched controls (i.e. $\mathrm{E}\left[X_i\right] = \mathrm{E}\left[X_{M(i)}\right]$). This means that, when the true model is affine, matching does not have to be perfect, it only has to equalize the first moment between the treated and untreated subjects in order to give an unbiased estimate of the ATT (either unconditionally or conditional

on X, the values of the observed covariates among the treated subjects). This provides theoretical support for the common practice of measuring imbalance by comparing the means of covariates among the treated and untreated (before or after matching) and considering matching to have been successful when the means in both groups are similar. Indeed, when the true model is affine, this is all that is needed in order to ensure that the estimated ATT is unbiased.

2.3.2 Regression Imputation

If we have an estimator of μ_0, $\hat{\mu}_0$, the regression imputation estimator of the ATT, as defined by Abadie and Imbens [1, 2], gives:

$$\begin{aligned}\hat{\tau}_i &= Y_i - \hat{\mu}_0(X_i) = \mu_1(X_i) + \epsilon_i - \hat{\mu}_0(X_i) = (\mu_1(X_i) - \mu_0(X_i)) - (\hat{\mu}_0(X_i) - \mu_0(X_i)) + \epsilon_i \\ &= \tau_i - (\hat{\mu}_0(X_i) - \mu_0(X_i)) + \epsilon_i\end{aligned}$$

Then,

$$\mathrm{E}[\hat{\tau}_i|X_i] - \tau_i = \mathrm{E}[\mu_0(X_i) - \hat{\mu}_0(X_i)|X_i] = \mu_0(X_i) - \mathrm{E}[\hat{\mu}_0|X_i](X_i)$$

The equivalent computations for the ATT are:

$$\hat{\tau} = N_1^{-1}\sum_{i=1}^{N_1}\hat{\tau}_i = N_1^{-1}\sum_{i=1}^{N_1}\tau_i - N_1^{-1}\sum_{i=1}^{N_1}(\hat{\mu}_0(X_i) - \mu_0(X_i)) + N_1^{-1}\sum_{i=1}^{N_1}\epsilon_i$$

$$\mathrm{E}[\hat{\tau}|X] - \bar{\tau} = N_1^{-1}\sum_{i=1}^{N_1}(\mu_0(X_i) - \mathrm{E}[\hat{\mu}_0|X](X_i))$$

$$\mathrm{E}[\hat{\tau}] - \tau = \mathrm{E}[\mu_0(X_i) - \mathrm{E}[\hat{\mu}_0|X](X_i)]$$

Thus, the regression imputation estimator of the ATT will be unbiased if the regression function is conditionally unbiased. Let μ_0^* be the (stochastic) limit of $\hat{\mu}$, then we can expand the bias as

follows:

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = \mathrm{E}\left[\mu_0\left(X_i\right) - \mu_0^*\left(X_i\right)\right] + \mathrm{E}\left[\mu_0^*\left(X_i\right) - \mathrm{E}\left[\hat{\mu}_0 | X\right]\left(X_i\right)\right]$$

Thus, the bias of the regression imputation estimator arises from two sources, the asymptotic bias of the regression function and its additional finite sample bias. Abadie and Imbens assume that the estimator $\hat{\mu}_0$ is fit using only untreated controls, and, thus, will be independent of X [2]. However, many analysts would combine such a regression estimator with matching and would first match the untreated subjects to the treated ones before running the regression only on matched controls, which may mean that $\hat{\mu}_0 \not\perp\!\!\!\perp X$ if $\hat{\mu}_0$ is biased. In order to include this case, we will not assume that $\hat{\mu}_0 \perp\!\!\!\perp X$ until later.

If both μ and $\hat{\mu}$ are affine functions of x, so that $\mu(x) = \beta^t x$, $\hat{\mu}(x) = \hat{\beta}^t x$, where we set the zeroth entry of each x equal to one in order to simplify notation, as is standard, then

$$\mathrm{E}\left[\hat{\tau}|X\right] - \bar{\tau} = N_1^{-1} \sum_{i=1}^{N_1} \left(\beta - \mathrm{E}\left[\hat{\beta}\middle|X\right]\right)^t X_i = \left(\beta - \mathrm{E}\left[\hat{\beta}\middle|X\right]\right)^t \cdot N_1^{-1} \sum_{i=1}^{N_1} X_i$$

$$\begin{aligned}\mathrm{E}\left[\hat{\tau}\right] - \tau &= \mathrm{E}\left[\left(\beta - \mathrm{E}\left[\hat{\beta}\middle|X\right]\right)^t X_i\right] = \mathrm{E}\left[\left(\beta - \beta^*\right)^t X_i\right] + \mathrm{E}\left[\left(\beta^* - \mathrm{E}\left[\hat{\beta}\middle|X\right]\right)^t X_i\right] \\ &= \left(\beta - \beta^*\right)^t \mathrm{E}\left[X_i\right] - \mathrm{E}\left[\mathrm{E}\left[\hat{\beta} - \beta^*\middle|X\right]^t X_i\right] = \left(\beta - \beta^*\right)^t \mathrm{E}\left[X_i\right] + o(1)\end{aligned}$$

under mild regularity conditions such as that $\hat{\beta} \xrightarrow{\mathcal{L}_1} \beta^*$ and X is bounded or $\hat{\beta} \xrightarrow{\mathcal{L}_2} \beta^*$ and X has finite second moment. Therefore, the asymptotic bias of the regression imputation estimator of the ATT is the product of the asymptotic bias of $\hat{\beta}$ and the expected value of X_i and so, will be asymptotically unbiased if the regression estimator is.

When using the regression imputation estimator, linearity does not provide any clear advantage for obtaining an unbiased estimate of the ATT vs. the general case; the estimated ATT will be unbiased if the regression function is conditionally unbiased. In the affine case, this will occur when the regression function used to estimate $\hat{\beta}_0$ is correctly specified. However, this is what one would naively expect and it does not provide any surprising additional robustness, unlike what we saw with the simple matching estimator.

2.3.3 Bias Corrected Matching

Finally, in the case of the bias-corrected matching estimator,

$$\begin{aligned}\hat{\tau}_i &= Y_i - Y_{M(i)} + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i) \\ &= (\mu_1(X_i) + \epsilon_i) - \left(\mu_0\left(X_{M(i)}\right) + \epsilon_{M(i)}\right) + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i) \\ &= (\mu_1(X_i) - \mu_0(X_i)) + \left(\hat{\mu}_0\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - (\hat{\mu}_0(X_i) - \mu_0(X_i)) - \Delta\epsilon_i \\ &= \tau_i + \left(\hat{\mu}_0\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - (\hat{\mu}_0(X_i) - \mu_0(X_i)) - \Delta\epsilon_i\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[\hat{\tau}_i|X_i\right] - \tau_i &= \mathrm{E}\left[\mathrm{E}\left[\left(\hat{\mu}_0\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - (\hat{\mu}_0(X_i) - \mu_0(X_i))\middle|X_M, X\right]\middle|X_i\right] \\ &= \mathrm{E}\left[\left(\mathrm{E}\left[\hat{\mu}_0|X, X_M\right]\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - \left(\mathrm{E}\left[\hat{\mu}_0|X, X_M\right](X_i) - \mu_0(X_i)\right)\middle|X_i\right]\end{aligned}$$

The equivalent calculations for the ATT are:

$$\begin{aligned}\hat{\tau} = N_1^{-1}\sum_{i=1}^{N_1}\hat{\tau}_i = \bar{\tau} + N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\mu}_0\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\mu}_0(X_i) - \mu_0(X_i)\right) \\ - N_1^{-1}\sum_{i=1}^{N_1}\Delta\epsilon_i\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left[\hat{\tau}|X\right] - \bar{\tau} = N_1^{-1}\sum_{i=1}^{N_1}\mathrm{E}\left[\mathrm{E}\left[\hat{\mu}_0|X, X_M\right]\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\middle|X\right] \\ - N_1^{-1}\sum_{i=1}^{N_1}\left(\mathrm{E}\left[\hat{\mu}_0|X\right](X_i) - \mu_0(X_i)\right)\end{aligned}$$

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = \mathrm{E}\left[\mathrm{E}\left[\hat{\mu}_0|X, X_M\right]\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right] - \mathrm{E}\left[\mathrm{E}\left[\hat{\mu}_0|X\right](X_i) - \mu_0(X_i)\right]$$

So the bias-corrected estimator of the ATT will be unbiased if the estimated regression function is unbiased conditional on X, X_M or if the matching is perfect so that the first and second terms are identical. However, when μ and $\hat{\mu}$ are affine functions of x so $\mu(x) = \beta_0 + \beta^t x$, $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}^t x$

the situation simplifies. Then,

$$\mathrm{E}\left[\hat{\tau}|X\right] - \bar{\tau} = \mathrm{E}\left[\left(\mathrm{E}\left[\hat{\beta}\middle|X, X_M\right] - \beta\right)^t \overline{\Delta X}\middle| X\right]$$

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = \mathrm{E}\left[\left(\mathrm{E}\left[\hat{\beta}\middle|X, X_M\right] - \beta\right)^t \Delta X_i\right]$$

As before, if β^* is the (stochastic) limit of $\hat{\beta}$, then we can reexpress the bias as,

$$\mathrm{E}\left[\hat{\tau}\right] - \tau = (\beta^* - \beta)^t \,\mathrm{E}\left[\Delta X_i\right] + \mathrm{E}\left[\left(\mathrm{E}\left[\hat{\beta}\middle|X, X_M\right] - \beta^*\right)^t \Delta X_i\right] = (\beta^* - \beta)^t \,\mathrm{E}\left[\Delta X_i\right] + o(1)$$

under mild regularity conditions such as ΔX being bounded and $\hat{\beta} \xrightarrow{\mathcal{L}_1} \beta^*$ or ΔX having finite second moment and $\hat{\beta} \xrightarrow{\mathcal{L}_2} \beta^*$. Thus, the bias-corrected matching estimator of the ATT will be unbiased if $\hat{\beta}$ is conditionally unbiased (meaning that it is correctly specified) and asymptotically unbiased if either the expected matching discrepancy is zero or the regression function is correctly specified. Thus, the bias-corrected matching estimator combines the robustness properties of both simple matching and regression imputation.

Therefore, when both μ_0 and $\hat{\mu}_0$ are affine functions of x, the bias-corrected matching estimator exhibits a strong form of double robustness: it will be asymptotically unbiased if either the expected matching discrepancy is zero or the regression function is correctly specified. The first criterion is exactly what we saw in the case of simple matching. However it is still, in some sense, rather surprising, since it tells us that the quality of individual matches doesn't matter, just that the expected matching discrepancy is zero, so that the errors cancel each other out as the sample becomes arbitrarily large.

2.4 Bias, Matching Method, and Covariate Distribution

We now examine the practical implications of these results for both direct and propensity score matching for a variety of commonly encountered covariate distributions, focusing on how the choice of distribution affects the bias of matching estimators of the ATT.

2.4.1 Direct Matching

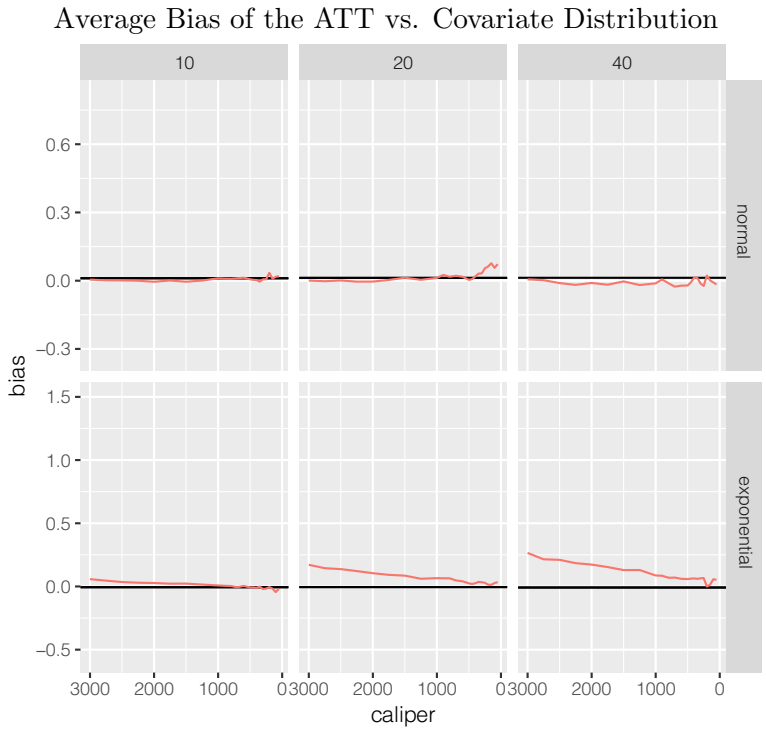
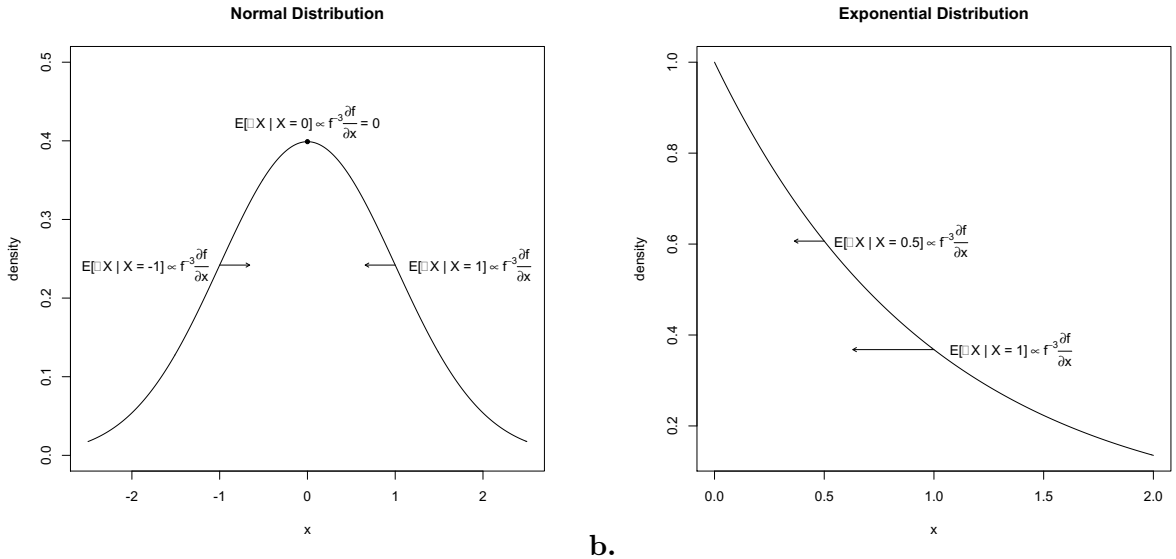


Figure 2.1: Average bias of the bias-corrected matching estimator of the ATT for Normally and Exponentially Distributed covariates. The numbers at the top of the figure give the number of covariates being matched, while caliper indicates the number of matched pairs retained after discarding the worst matches.

Figure 2.1 shows that, as we discussed above, when the matching discrepancy is small, the bias-corrected matching estimator of the ATT will be nearly unbiased, even if the regression function is misspecified. The figure shows that the empirical bias of the estimated ATT, under nearest-neighbor matching, is smaller when the covariates are normally distributed (Figure 2.1, top row) than when they are exponentially distributed (Figure 2.1, bottom row). This is because normally distributed covariates have zero expected matching discrepancy, while exponentially distributed covariates have a negative expected matching discrepancy.

Figure 2.2 illustrates the origin of the difference in matching discrepancies. When covariates are symmetrically distributed, as they are under the Normal Distribution, the conditional expected matching discrepancy at opposing points will be equal and opposite so that the unconditional expected matching discrepancy will be zero because the contributions of opposing points exactly cancel, as seen in Figure 2.2a. For the normal distribution, the matched point is more likely to lie



a. **b.**

Figure 2.2: Expected conditional matching discrepancies at several points of the **a.** Normal and **b.** Exponential Distributions. Note that, for the Normal Distribution, opposing points will have equal and opposite expected matching discrepancies, while for the Exponential Distribution, all points will have negative expected matching discrepancies.

closer to the origin than the original point, although this may not be the case for other symmetric distributions. When covariates are drawn from a skewed distribution, such as the Exponential Distribution, the matching discrepancies will no longer cancel. Indeed, for the Exponential Distribution, the expected conditional matching discrepancy will always be negative and will increase in magnitude with increasing x, meaning that the matched point is, again, more likely to be closer to the origin than the original point, as seen in Figure 2.2b.

This behavior can be seen in following analytic result originally due to Abadie and Imbens (see the appendix for an expanded version, which includes higher order terms than the original result) [2].

$$\mathrm{E}[\Delta X|X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}f(X)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x}(X)\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \tag{2.1}$$

$$\mathrm{E}[\Delta X\Delta X^t|X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}f(X)^{-\frac{2}{d}}\cdot\mathcal{I}_d\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \tag{2.2}$$

where d is the number of covariates on which matching is performed, $\mathcal{I}_d$ is the d dimensional identity matrix, and, here, X denotes the covariate values of the individual being matched.

We note that the direction of the conditional expected matching discrepancy is given by the

gradient of the density at the point X. For the Uniform Distribution, the expected matching discrepancy is always 0 at any point (ignoring edge effects). For the Normal Distribution, the matching discrepancy is always towards the origin and is symmetric, with opposite points having equal and opposite matching discrepancies, so that the unconditional matching discrepancy is zero. For the Exponential Distribution, the discrepancy again always points towards the origin, but, since exponential random variables are nonnegative, this means that the conditional matching discrepancy will always be negative, and, thus, the unconditional expected matching discrepancy will be as well.

From this, it appears that, when both the true model and the regression model are affine, symmetric covariate distributions will result in improved matching estimates of the ATT, since they naturally result in matching discrepancies with zero mean, and, based on our above results, this guarantees that the simple matching estimator will be unbiased and the bias-corrected estimator will be asymptotically unbiased, regardless of whether or not the bias correction function is correctly specified. However, while this is relatively straightforward in one dimension, it becomes less obvious how this intuition should extend to higher dimensions with correlated covariates. In particular, if our collection of covariates includes some variable $X^{(1)}$ as well as its square, $X^{(1)2}$, even if $X^{(1)}$ is symmetric or, better yet, uniform, $X^{(1)2}$ and, thus, the joint distribution $(X^{(1)}, X^{(1)2})$ will certainly not be. As Figure 2.3 shows, in general, the distribution of $X^{(1)2}$ will be skewed towards the origin, at least near 0. Thus, nearest neighbor matching on $X^{(1)2}$ will fail to produce an expected matching discrepancy of zero.

In fact, when a model contains both simple covariates, like $X^{(1)}, X^{(2)}, \ldots$ and derived covariates which are functions of the simple covariates $\mathcal{X}^{(i)} = f_i(X)$, such as interaction terms, like $X^{(1)}X^{(2)}$, and quadratic terms, like $X^{(1)2}$ (or even more general functions such as $\log X^{(1)}$), it is typical to match only on the simple covariates and ignore the presence of derived covariates in the matching process. This also typically leads to the derived covariates having nonzero expected matching discrepancies even if the simple covariates actually do have zero expected matching discrepancies, although the explanation for why this is is slightly different. The reason is that the nonlinear transformation $x \mapsto x^2$ distorts distances in a nonlinear fashion so that distances near 0 are shrunk while distances near 1 (and -1) are stretched. Thus, the conditional matching discrepancy will be altered. In particular, if it is zero before the transformation, it will typically be nonzero afterwards.

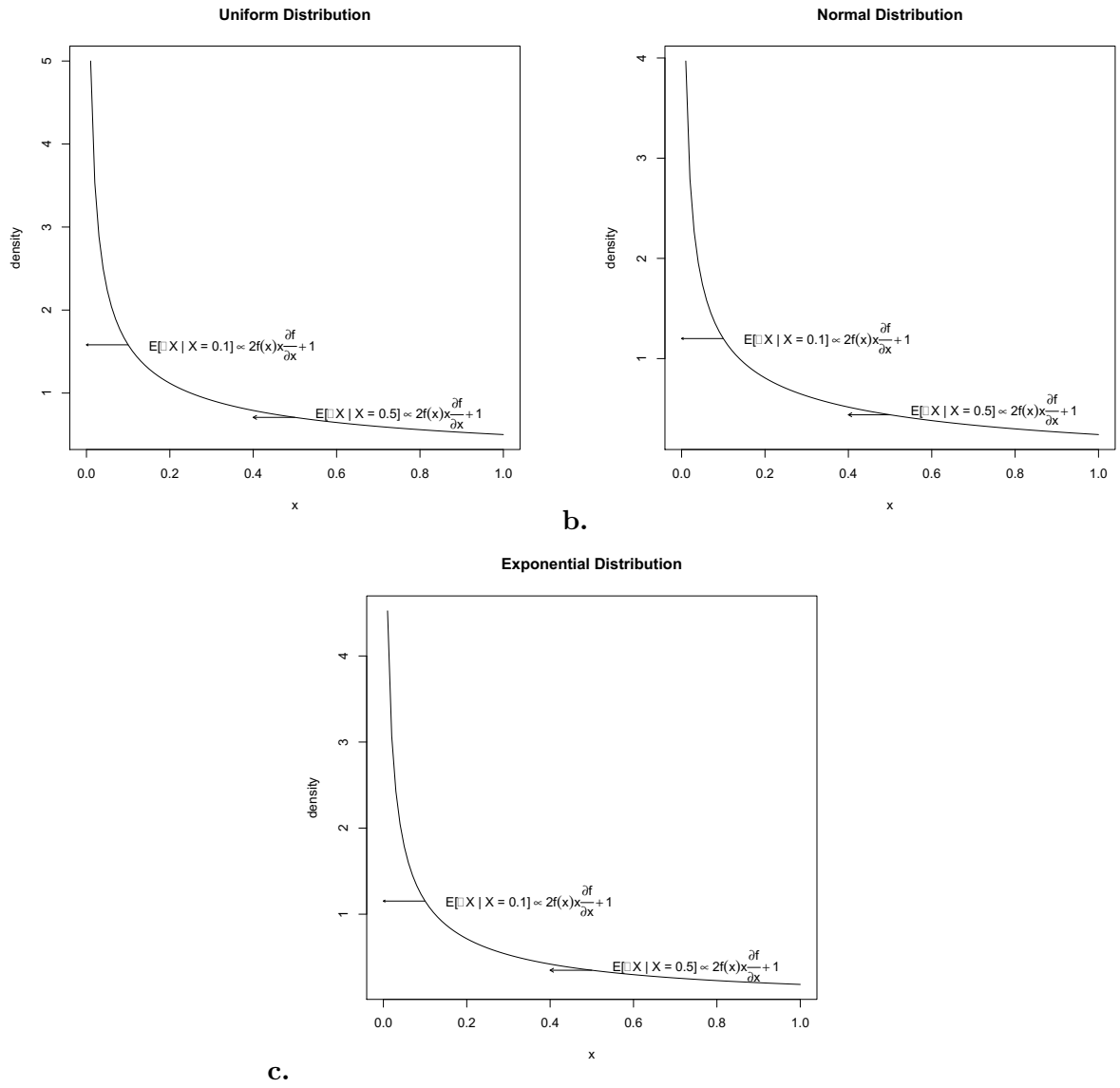
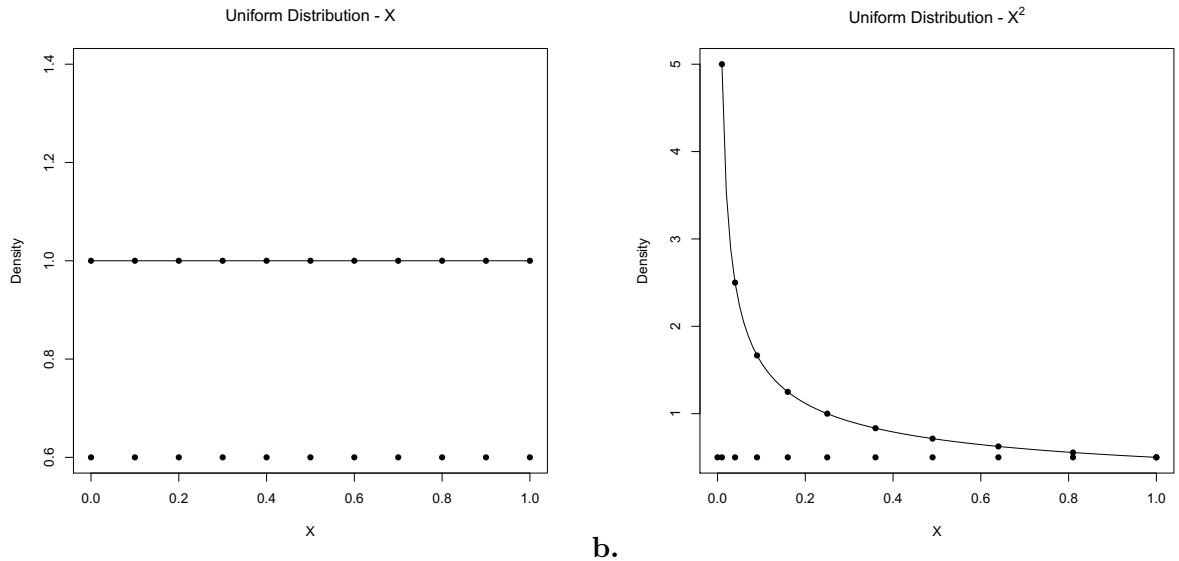


Figure 2.3: Distribution of X^2 for various distributions. **a.** Even though the base covariate is distributed Uniformly, X^2 is not, with the density concentrating around 0. This results in the conditional expected matching discrepancy being negative at every point. This effect is also seen with the **b.** Normal and **c.** Exponential Distributions.

This is most evident in Figures 2.4a and b, which show that, if X is uniformly distributed, the deciles are evenly spaced for X, but not for X^2. In particular, the deciles of X^2 are very closely spaced near 0 and spread out as X increases, and so are widely spaced near 1. Let $q(i)$ be the x value of the i^{th} decile. Since X is uniformly distributed, if $x = q(i)$, it has an equal probability of being matched to $q(i+1)$ or $q(i-1)$. Thus, the fact that $q(i+1) - q(i) > q(i) - q(i-1)$ implies that the matching discrepancy is positive. The situation is more complex for other distributions since it is no



a. **b.**

Figure 2.4: Distribution of X and X^2 when X is Uniformly Distributed. **a.** Distribution of X with points indicating the deciles. **b.** Distribution of X^2 with points indicating the deciles. Note that, while the deciles are evenly spaced for X, this is clearly not the case for X^2, with the lowest deciles tightly grouped near zero and the upper deciles widely spread out.

longer the case that matching to $q(i+1)$ and $q(i-1)$ is equally likely. In order to understand their behavior, it is more enlightening to examine an analytic expression for the conditional matching discrepancy akin to Equation 2.1. We can extend that equation to the case in which we have derived covariates $(\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \dots)$, but only match on the simple covariates $(X^{(1)}, X^{(2)}, \dots)$, using a power series expansion of the defining functions of the derived covariates, f_k. A full derivation is presented in the appendix. In this setting,

$$\begin{aligned} &\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\middle|X_i\right] \\ &= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \\ &\quad\times\sum_u\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\frac{\partial f_k}{\partial x_u}(X_i)+f(X_i)^{-\frac{2}{d}}\frac{1}{2}\frac{\partial^2 f_k}{\partial x_u^2}(X_i)\right]+o\left(N_0^{-\frac{2}{d}}\right) \end{aligned} \tag{2.3}$$

$$\begin{aligned} &\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\Delta\mathcal{X}_i^{(l)}\middle|X_i\right] \\ &= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_u f(X_i)^{-\frac{2}{d}}\frac{\partial f_k}{\partial x_u}(X_i)\frac{\partial f_l}{\partial x_u}(X_i)+o\left(N_0^{-\frac{2}{d}}\right) \end{aligned} \tag{2.4}$$

From Equation 2.3, we see that if $\mathcal{X} = X^{(1)2}$ and the distribution of $X^{(1)}$ is uniform, the first term in brackets disappears and the second term is equal to 1, so the conditional expected matching discrepancy (and, thus, also the unconditional matching discrepancy) will be positive, even though the conditional expected matching discrepancy of $X^{(1)}$ is zero; this is just what we saw in Figure 2.4. Indeed, if the distribution of $X^{(1)}$ is symmetric, then its derivative will be antisymmetric, so, since $\frac{\partial f_k}{\partial x}(x) = 2x$ is also antisymmetric, the first term will be symmetric and the conditional expected matching discrepancy will generally be nonzero (although, since the first term in brackets my be negative, as is the case when $X^{(1)}$ is normally distributed, the conditional expected matching discrepancy may be zero at particular points). Thus, even if the true model and regression model are both affine, if the regression model in the bias-corrected matching estimator is incorrectly specified, then the estimated ATT will, in general, be biased, since the covariates no longer have zero expected matching discrepancy. Therefore, when using nearest-neighbor matching on the simple covariates alone, in the presence of quadratic, and other higher order terms, one will, in general, need to properly specify the regression model in order for the estimated ATT to be asymptotically unbiased.

However, it is not necessarily the case that all derived covariates will have nonzero expected matching discrepancy. Looking at the expression for the conditional expected matching discrepancy, we can see that, if the distributions of the simple covariates are symmetric and they are mutually independent, then, if the derived covariates are antisymmetric in each of the simple covariates, the expected matching discrepancy will be zero. In particular, simple two-way interaction terms will have zero expected matching discrepancy, as will odd powers of a single covariate. However, if the simple covariates are correlated, this will no longer be the case and even simple two-way interactions may have nonzero expected matching discrepancies so that the bias-corrected matching estimator will, in general, be biased, unless the regression function is correctly specified.

2.4.2 Propensity Score Matching

As we have seen above, for many distributions, direct matching on the covariates will not result in zero expected matching discrepancy, either because the covariates being matched upon have skewed distributions, or because the model includes derived covariates. Thus, even when both the true and regression models are affine, the bias-corrected matching estimator of the ATT will, in

general, not be even asymptotically unbiased, unless the regression model is properly specified. However, we can also consider other matching methods besides direct nearest-neighbor matching.

One of the most popular alternatives to direct matching is matching on the propensity score, the probability that a subject will be treated given his covariate values. In our extreme setting, in which the propensity score is a constant (which is equal to the fraction of treated subjects among all subjects treated or untreated), this means any treated subject is equally likely to be matched to any untreated subject. Also note that, since the propensity score is constant across all subjects, the distribution of covariates among treated and untreated subjects must be the same. Thus, $\mathrm{E}\left[\Delta X_i\right] = \mathrm{E}\left[X_{M(i)} - X_i\right] = \mathrm{E}\left[X_{M(i)}\right] - \mathrm{E}\left[X_i\right] = 0$, since X is distributed identically among treated and untreated subjects and the matched subject is chosen completely at random, independent of the value of X_i.

Therefore, as long as both the true and regression models are affine, when using propensity score matching, the bias-corrected estimator of the ATT will always be asymptotically unbiased (and the simple matching estimator will be unbiased). This is in contrast to what occurs when using direct matching, since, in that case, for the expected matching discrepancy to be zero, the matched covariates must be drawn from distributions with special structure (as is the case if they are drawn from symmetric distributions) and any derived covariates must not have quadratic or higher order terms (or even interaction terms, if the simple covariates are correlated). When using the propensity score for matching, the expected matching discrepancy will always be zero, no matter what covariates are used and how they are distributed, so the bias-corrected matching estimate of the ATT will always be asymptotically unbiased.

2.4.3 Coarsened Exact Matching

A less popular alternative to direct covariate matching and propensity score matching is Coarsened Exact Matching (CEM) [4]. When using CEM, the analyst first partitions each matched covariate into bins (the so called coarsening) and then matches treated subjects to untreated controls which perfectly match their coarsened covariate values (e.g. if we match on X_1, which we coarsen into two categories $X_1 < 0$ and $X_1 \geq 0$, a treated subject with $x_1 = 1$ will be matched at random with a control that has $x_1 \geq 0$; if no such untreated control exists, that treated subject is discarded, as only perfect matches are allowed). The major benefit of CEM is that it achieves

(approximate) perfect matches between subjects, but it requires very large numbers of untreated subjects in order to ensure that all possible configurations of covariates have enough untreated representatives. In particular, if we match on d covariates, which we coarsen into just two categories each, we will have 2^d possible configurations, so it potentially requires very large numbers of untreated subjects in order to match on more than a handful of covariates, with the worst case being when the covariates are completely independent. However, if the covariates are very highly correlated, so that only a small fraction of the possible configurations are exist, it may be possible to match on many more than $O\left(\log_2 N_0\right)$ covariates.

However, if one has sufficient untreated controls, in the case of a constant propensity score, CEM shares a remarkable feature with propensity score matching: matches of subjects within any bin will have zero expected matching discrepancy, and, thus, CEM will result in zero expected matching discrepancy. The reason for this is similar to the argument for propensity score matching. Matching of two subjects within a bin is completely at random, conditional only on the fact that their covariate values fall within a certain range. Since both treated and untreated subjects have identical covariate distributions, they will also have identical covariate distributions conditional on falling within a particular bin, B. Thus, $\mathrm{E}\left[\Delta X_i \middle| X_i, X_{M(i)} \in B\right] = \mathrm{E}\left[X_{M(i)} - X_i \middle| X_i, X_{M(i)} \in B\right] = \mathrm{E}\left[X_{M(i)} \middle| X_i, X_{M(i)} \in B\right] - \mathrm{E}\left[X_i \middle| X_i, X_{M(i)} \in B\right] = 0$.

Therefore, regardless of the distribution of the covariates, when the propensity score is constant, CEM will have zero expected matching discrepancy, and, thus, if both the true and regression models are affine, the bias-corrected matching estimator of the ATT will be asymptotically unbiased regardless of whether or not the regression function is properly specified. Thus, CEM possesses the most favorable property of propensity score matching in this setting, although it may be more difficult to use in practice, when the number of covariates is large, because it requires large numbers of untreated subjects to populate all the possible configurations. However, when the propensity score is not constant, the distribution of covariates may differ between the treated and untreated within each bin (which is referred to as intrastratum confounding), so the expected matching discrepancy within each bin is no longer guaranteed to be zero and, thus, the overall expected matching discrepancy of CEM may also be nonzero and the estimated ATT may be asymptotically biased, if the regression model is not properly specified.

2.5 Variance

Having examined the effect of different matching methods on the bias, we now expand our analysis to the variance. The fact that propensity score matching guarantees that the bias-corrected matching estimator is asymptotically unbiased whenever both the true and regression models are affine appears to be an important advantage, but, since the propensity score may select matches that are far apart in covariate space, this could potentially lead to the bias-corrected matching estimator of the ATT having a much higher variance when propensity score matching is used instead of direct matching. We will explore this below after taking some time to characterize the variance of the simple matching, regression imputation, and bias-corrected matching estimators.

2.5.1 Simple Matching

We begin with the simple matching estimator. Recalling that,

$$\hat{\tau}_i = Y_i - Y_{M(i)} = \tau_i - \left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right) - \Delta\epsilon_i$$

the variance of the the estimated individual treatment effect is,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\tau}_i | X_i\right] &= \mathrm{Var}\left[\mathrm{E}\left[\hat{\tau}_i \middle| X_i, X_{M(i)}\right] \middle| X_i\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\tau}_i \middle| X_i, X_{M(i)}\right] \middle| X_i\right] \\
&= \mathrm{Var}\left[\tau_i - \left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right) \middle| X_i\right] + \mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right) + \sigma_\epsilon^2\left(X_{M(i)}\right) \middle| X_i\right] \\
&= \mathrm{Var}\left[\mu_0\left(X_{M(i)}\right) \middle| X_i\right] + \mathrm{E}\left[\sigma_\epsilon^2\left(X_{M(i)}\right) \middle| X_i\right] + \sigma_\epsilon^2\left(X_i\right)
\end{aligned}$$

The conditional variance of the individual treatment effect decomposes nicely into the sum of the conditional variances of $\mu_0\left(X_{M(i)}\right)$ and the error terms. However, the situation for the ATT is more complicated. The variance of the ATT conditional on the covariates in both the treated

subjects and their matched controls is

$$\begin{aligned}
\text{Var}\left[\hat{\tau}|X, X_M\right] &= \text{Var}\left[N_1^{-1}\sum_{i=1}^{N_1}\Delta\epsilon_i\middle|X, X_M\right] = N_1^{-2}\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)} - \sum_{i=1}^{N_1}\epsilon_i\middle|X, X_M\right] \\
&= N_1^{-2}\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right] + N_1^{-2}\sum_{i=1}^{N_1}\text{Var}\left[\epsilon_i|X_i\right] \\
&= N_1^{-2}\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right] + N_1^{-2}\sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right)
\end{aligned}$$

Note that, while $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for $i \neq j$ and $\epsilon_i \perp\!\!\!\perp \epsilon_{M(j)}$ for any i, j, the $\epsilon_{M(i)}$ may not be mutually independent given X_M, since more than one treated subject may match to the same control. Thus, the first term is not necessarily the variance of the average of iid variables, and, thus, will not simplify further without additional assumptions. The variance conditional only on the covariates of the treated subjects is

$$\begin{aligned}
\text{Var}\left[\hat{\tau}|X\right] &= \text{Var}\left[\text{E}\left[\hat{\tau}|X, X_M\right]|X\right] + \text{E}\left[\text{Var}\left[\hat{\tau}|X, X_M\right]|X\right] \\
&= \text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right)\middle|X\right] \\
&\quad + N_1^{-2}\text{E}\left[\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right] + \sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right)\middle|X\right] \\
&= N_1^{-2}\text{Var}\left[\sum_{i=1}^{N_1}\mu_0\left(X_{M(i)}\right)\middle|X\right] + N_1^{-2}\text{E}\left[\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right]\middle|X\right] + N_1^{-2}\sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right)
\end{aligned}$$

while the unconditional variance is,

$$\begin{aligned}
\text{Var}\left[\hat{\tau}\right] &= \text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0\left(X_i\right)\right)\right] + N_1^{-2}\text{E}\left[\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right]\right] \\
&\quad + N_1^{-2}\sum_{i=1}^{N_1}\text{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] \\
&= \text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_1\left(X_i\right) - \mu_0\left(X_{M(i)}\right)\right)\right] + N_1^{-2}\text{E}\left[\text{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right]\right] \\
&\quad + N_1^{-1}\text{E}\left[\sigma_\epsilon^2\left(X_i\right)\right]
\end{aligned}$$

The second term in these expressions will simplify if the $\epsilon_{M(i)}$s are mutually independent, so that

$N_1^{-1}\mathrm{E}\left[\mathrm{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle|X_M\right]\right] = \mathrm{E}\left[\sigma_\epsilon^2\left(X_{M(i)}\right)\right]$. This will occur asymptotically if the distribution of X in the treated subjects, $\mu_{X,1}$, is absolutely continuous with respect to the distribution of X in the untreated controls, $\mu_{X,0}$ ($\mu_{X,1} << \mu_{X,0}$), $\frac{\partial\mu_{X,1}}{\partial\mu_{X,0}}$ is bounded almost surely, and $N_0^{-1}N_1^{3+\epsilon} \to 0$, where $\epsilon > 0$, so that each treated subject will asymptotically be matched to a different control.

In order to simplify the first term, we first recall that $\tau_i = \mu_1(X_i) - \mu_0(X_i) = \tau(X_i)$, so τ_i depends only on X. We can also power expand $\mu_0\left(X_{M(i)}\right)$ around X_i giving $\mu_0\left(X_{M(i)}\right) - \mu_0(X_i) = \sum_{\alpha>0}\frac{1}{\alpha!}D^\alpha\mu_0(X_i)\,\Delta X^\alpha$, where α is a multiindex. Then, the first term becomes,

$$\begin{aligned}
&\mathrm{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right)\right] \\
&= \mathrm{Var}\left[\bar{\tau}(X)\right] - 2\mathrm{Cov}\left[\bar{\tau}(X), N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right)\right] \\
&\quad + N_1^{-2}\mathrm{Var}\left[\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right)\right] \\
&= N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] - 2\mathrm{Cov}\left[\bar{\tau}(X), N_1^{-1}\sum_{i=1}^{N_1}\sum_{\alpha>0}\frac{1}{\alpha!}D^\alpha\mu_0(X_i)\,\Delta X^\alpha\right] \\
&\quad + N_1^{-2}\mathrm{Var}\left[\sum_{i=1}^{N_1}\sum_{\alpha>0}\frac{1}{\alpha!}D^\alpha\mu_0(X_i)\,\Delta X^\alpha\right]
\end{aligned}$$

The variance decomposes into the sum of three pieces. The first is the finite sample variance of $\bar{\tau}$, the conditional ATT, about its expectation $\tau = \mathrm{E}\left[\tau(X)\right]$. The second is the covariance of $\bar{\tau}$ and the average matching discrepancy of μ_0. The last is the variance of the average matching discrepancy of μ_0. If we further assume that μ_0 is an affine function of X, so $\mu_0(X) = \beta_0 + \beta^t X$. Then $\mu_0\left(X_{M(i)}\right) - \mu_0(X) = \beta^t\Delta X$, so

$$\begin{aligned}
&\mathrm{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0(X_i)\right)\right] \\
&= N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] - 2\mathrm{Cov}\left[\bar{\tau}(X), \beta^t\overline{\Delta X}\right] + \mathrm{Var}\left[\beta^t\overline{\Delta X}\right] \\
&= N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] - 2\mathrm{Cov}\left[\bar{\tau}(X), \overline{\Delta X}^t\right]\beta + \beta^t\mathrm{Var}\left[\overline{\Delta X}\right]\beta
\end{aligned}$$

Under the conditions above: $\mu_{X,1} << \mu_{X,0}$, $\frac{\partial\mu_{X,1}}{\partial\mu_{X,0}}$ is bounded almost surely, and $N_0^{-1}N_1^{3+\epsilon} \to 0$,

$\epsilon > 0$, since the X_is are IID, asymptotically, this becomes,

$$\begin{aligned}
&\text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\mu_0\left(X_{M(i)}\right) - \mu_0\left(X_i\right)\right)\right] \\
&\quad \approx N_1^{-1}\left[\text{Var}\left[\tau\left(X_i\right)\right] - 2\text{Cov}\left[\tau\left(X_i\right), \Delta X_i^t\right]\beta + \beta^t \text{Var}\left[\Delta X_i\right]\beta\right]
\end{aligned}$$

This expression is again the sum of three pieces: the variance of the treatment effect across the population, the covariance of the individual level treatment effect with the matching discrepancy times the coefficient vector β, and the quadratic form of β with respect to the variance of the matching discrepancy. In the case in which $\epsilon \perp\!\!\!\perp X$, then, asymptotically, we have,

$$\text{Var}\left[\hat{\tau}\right] = N_1^{-1}\left[\text{Var}\left[\tau\left(X_i\right)\right] - 2\text{Cov}\left[\tau\left(X_i\right), \Delta X_i^t\right]\beta + \beta^t \text{Var}\left[\overline{\Delta X}\right]\beta + 2\sigma_\epsilon^2\right]$$

2.5.2 Regression Imputation

Recalling that, for regression imputation,

$$\hat{\tau}_i = Y_i - \hat{\mu}_0\left(X_i\right) = \tau_i - \left(\hat{\mu}_0\left(X_i\right) - \mu_0\left(X_i\right)\right) + \epsilon_i$$

the conditional variance of the individual treatment effect is,

$$\begin{aligned}
\text{Var}\left[\hat{\tau}_i | X_i\right] &= \text{Var}\left[\text{E}\left[\hat{\tau}_i | X_i, \hat{\mu}\right] | X_i\right] + \text{E}\left[\text{Var}\left[\hat{\tau}_i | X_i, \hat{\mu}\right] | X_i\right] \\
&= \text{Var}\left[\tau_i - \left(\hat{\mu}_0\left(X_i\right) - \mu_0\left(X_i\right)\right) | X_i\right] + \text{E}\left[\sigma_\epsilon^2\left(X_i\right) \middle| X_i\right] \\
&= \text{Var}\left[\hat{\mu}_0 | X_i\right]\left(X_i\right) + \sigma_\epsilon^2\left(X_i\right)
\end{aligned}$$

The variance of the ATT conditional on both the covariates among the treated subjects and the estimated regression function is

$$\text{Var}\left[\hat{\tau} | X, \hat{\mu}_0\right] = \text{Var}\left[N_1^{-1}\sum_{i=1}^{N_1}\epsilon_i\right] = N_1^{-2}\sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right)$$

Then,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\tau}|X\right] &= \mathrm{Var}\left[\mathrm{E}\left[\hat{\tau}|X,\hat{\mu}_0\right]|X\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\tau}|X,\hat{\mu}_0\right]|X\right] \\
&= \mathrm{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\mu}_0\left(X_i\right) - \mu_0\left(X_i\right)\right)\middle| X\right] + N_1^{-2}\sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right) \\
&= N_1^{-2}\mathrm{Var}\left[\sum_{i=1}^{N_1}\hat{\mu}_0\left(X_i\right)\middle| X\right] + N_1^{-2}\sum_{i=1}^{N_1}\sigma_\epsilon^2\left(X_i\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left[\hat{\tau}\right] &= \mathrm{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\mu}_0\left(X_i\right) - \mu_0\left(X_i\right)\right)\right] + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] \\
&= \mathrm{Var}\left[\bar{\tau}(X)\right] - 2\mathrm{Cov}\left[\bar{\tau}(X), N_1^{-1}\sum_{i=1}^{N_1}\hat{\mu}_0\left(X_i\right)\right] + \mathrm{Var}\left[N_1^{-1}\sum_{i=1}^{N_1}\hat{\mu}_0\left(X_i\right)\right] \\
&\quad + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right]
\end{aligned}$$

The terms involving $\hat{\mu}_0$ in these expressions can have a very complicated structure depending on the nature of $\hat{\mu}_0$ and how it is estimated. However, in the case in which $\hat{\mu}_0$ is an affine function of x, it simplifies dramatically. For notational convenience we let the zeroth entry of x be 1 (as is standard) so that $\hat{\mu}_0(x) = \hat{\beta}^t x$.

$$\mathrm{Var}\left[\hat{\tau}|X\right] = \mathrm{Var}\left[N_1^{-1}\sum_{i=1}^{N_1}\hat{\beta}^t X_i\right] + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] = \bar{X}^t\mathrm{Var}\left[\hat{\beta}\middle| X\right]\bar{X} + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right]$$

$$\begin{aligned}
\text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\mu}_0\left(X_i\right) - \mu_0\left(X_i\right)\right)\right] &= \text{Var}\left[\bar{\tau} - N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\beta}-\beta\right)^t X_i\right] \\
&= N_1^{-1}\text{Var}\left[\tau(X_i)\right] - 2\text{Cov}\left[\bar{\tau}(X), N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\beta}-\beta\right)^t X_i\right] + \text{Var}\left[N_1^{-1}\sum_{i=1}^{N_1}\left(\hat{\beta}-\beta\right)^t X_i\right] \\
&= N_1^{-1}\text{Var}\left[\tau(X_i)\right] - 2\text{Cov}\left[\bar{\tau}(X), \text{E}\left[\left(\hat{\beta}-\beta\right)^t \bar{X}\middle| X\right]\right] - 2\text{E}\left[\text{Cov}\left[\bar{\tau}(X), \left(\hat{\beta}-\beta\right)^t \bar{X}\middle| X\right]\right] \\
&\quad + \text{Var}\left[\text{E}\left[\left(\hat{\beta}-\beta\right)^t \bar{X}\middle| X\right]\right] + \text{E}\left[\text{Var}\left[\left(\hat{\beta}-\beta\right)^t \bar{X}\middle| X\right]\bar{X}\right] \\
&= N_1^{-1}\text{Var}\left[\tau(X_i)\right] - 2\text{Cov}\left[\bar{\tau}(X), \left(\text{E}\left[\hat{\beta}\middle| X\right]-\beta\right)^t \bar{X}\right] + \text{Var}\left[\left(\text{E}\left[\hat{\beta}\middle| X\right]-\beta\right)^t \bar{X}\right] \\
&\quad + \text{E}\left[\bar{X}^t\text{Var}\left[\hat{\beta}\middle| X\right]\bar{X}\right]
\end{aligned}$$

$$\begin{aligned}
\text{Var}\left[\hat{\tau}\right] = N_1^{-1}\text{Var}\left[\tau(X_i)\right] &- 2\text{Cov}\left[\bar{\tau}(X), \left(\text{E}\left[\hat{\beta}\middle| X\right]-\beta\right)^t \bar{X}\right] + \text{Var}\left[\left(\text{E}\left[\hat{\beta}\middle| X\right]-\beta\right)^t \bar{X}\right] \\
&+ \text{E}\left[\bar{X}^t\text{Var}\left[\hat{\beta}\middle| X\right]\bar{X}\right] + N_1^{-1}\text{E}\left[\sigma_\epsilon^2\left(X_i\right)\right]
\end{aligned}$$

If $\hat{\beta}$ is conditionally unbiased, as is the case if the model is correctly specified, then the second and third terms drop out. While if $\hat{\beta} \perp\!\!\!\perp X$, as is the case if the untreated controls used to estimate β are not matched to the treated subjects, as is the case in Abadie and Imbens, the expression simplifies to

$$\begin{aligned}
\text{Var}\left[\hat{\tau}\right] &= N_1^{-1}\text{Var}\left[\tau(X_i)\right] - 2\left(\text{E}\left[\hat{\beta}\right]-\beta\right)^t \text{Cov}\left[\bar{X}, \bar{\tau}(X)\right] + \left(\text{E}\left[\hat{\beta}\right]-\beta\right)^t \text{Var}\left[\bar{X}\right]\left(\text{E}\left[\hat{\beta}\right]-\beta\right) \\
&\quad + \text{tr}\left[\text{E}\left[\text{Var}\left[\hat{\beta}\right]\text{E}\left[\bar{X}\bar{X}^t\right]\right]\right] + N_1^{-1}\text{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] \\
&= N_1^{-1}\left[\text{Var}\left[\tau(X_i)\right] - 2\left(\text{E}\left[\hat{\beta}\right]-\beta\right)^t \text{Cov}\left[X_i, \tau(X_i)\right] + \left(\text{E}\left[\hat{\beta}\right]-\beta\right)^t \text{Var}\left[X_i\right]\left(\text{E}\left[\hat{\beta}\right]-\beta\right)\right. \\
&\quad \left. + \text{tr}\left[\text{E}\left[\text{Var}\left[\hat{\beta}\right]\text{E}\left[X_i X_i^t\right]\right]\right] + \text{E}\left[\sigma_\epsilon^2\left(X_i\right)\right]\right]
\end{aligned}$$

2.5.3 Bias-Corrected Matching

For the bias-corrected matching estimator,

$$\begin{aligned}
\hat{\tau}_i &= Y_i - Y_{M(i)} + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i) \\
&= \tau_i + \left(\hat{\mu}_0\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - \left(\hat{\mu}_0(X_i) - \mu_0(X_i)\right) - \Delta\epsilon_i
\end{aligned}$$

The conditional variance of the individual treatment effect is,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\tau}_i|X_i\right] &= \mathrm{Var}\left[\mathrm{E}\left[\hat{\tau}|X, X_M\right]|X_i\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\tau}|X, X_M\right]|X_i\right] \\
&= \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\mu}_0|X, X_M\right]\left(X_{M(i)}\right) - \mu_0\left(X_{M(i)}\right)\right) - \left(\mathrm{E}\left[\hat{\mu}_0|X, X_M\right](X_i) - \mu_0(X_i)\right)\middle|X_i\right] \\
&\quad + \mathrm{E}\left[\mathrm{Var}\left[\hat{\mu}_0|X, X_M\right](X_i)|X_i\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\mu}_0|X, X_M\right]\left(X_{M(i)}\right)\middle|X_i\right] \\
&\quad - 2\mathrm{E}\left[\mathrm{Cov}\left[\hat{\mu}_0\left(X_i\right), \hat{\mu}_0\left(X_{M(i)}\right)\middle|X, X_M\right]\middle|X_i\right] \\
&\quad\quad + \mathrm{E}\left[\mathrm{Var}\left[\epsilon_i|X_i\right]|X_i\right] + \mathrm{E}\left[\mathrm{Var}\left[\epsilon_{M(i)}\middle|X_{M(i)}\right]\middle|X_i\right] \\
&\quad + 2\mathrm{E}\left[\mathrm{Cov}\left[\hat{\mu}_0, \Delta\epsilon_i|X, X_M\right](X_i)|X_i\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\hat{\mu}_0, \Delta\epsilon_i|X, X_M\right]\left(X_{M(i)}\right)\middle|X_i\right]
\end{aligned}$$

The first term comes from the conditional bias of $\hat{\mu}$, the next three terms arise from the fact that $\hat{\mu}$ must be estimated from the data, the next two terms arise from each subject's idiosyncratic variation (ϵ_i and $\epsilon_{M(i)}$), and the last two terms come from the fact that $\hat{\mu}$ is potentially estimated using ϵ_i and $\epsilon_{M(i)}$. This expression offers little further insight into the conditional variance of the estimated individual treatment effect, so we make several simplifying assumptions. First, we assume that both μ_0 and $\hat{\mu}_0$ are affine functions of x, so that $\mu_0(x) = \beta^t x, \hat{\mu}_0(x) = \hat{\beta}^t x$, where we take the zeroth component of x to be 1 in order to simplify notation. We then have,

$$\hat{\tau}_i = \tau_i + \left(\hat{\beta} - \beta\right)^t X_{M(i)} - \left(\hat{\beta} - \beta\right)^t X_i - \Delta\epsilon_i = \tau_i + \left(\hat{\beta} - \beta\right)^t \Delta X_i - \Delta\epsilon_i$$

$$\begin{aligned}
\mathrm{Var}\left[\hat{\tau}_i|X_i\right] &= \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\beta}\middle|X, X_M\right] - \beta\right)^t \Delta X_i\middle|X_i\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\beta}^t\Delta X_i\middle|X, X_M\right]\middle|X_i\right] \\
&\quad + \sigma_\epsilon^2(X_i) + \mathrm{E}\left[\sigma_\epsilon^2\left(X_{M(i)}\right)\middle|X_i\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\hat{\beta}^t\Delta X_i, \Delta\epsilon_i\middle|X, X_M\right]\middle|X_i\right] \\
&= \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\beta}\middle|X, X_M\right] - \beta\right)^t \Delta X_i\middle|X_i\right] + \mathrm{E}\left[\Delta X_i^t \mathrm{Var}\left[\hat{\beta}\middle|X, X_M\right]\Delta X_i\middle|X_i\right] \\
&\quad + \sigma_\epsilon^2(X_i) + \mathrm{E}\left[\sigma_\epsilon^2\left(X_{M(i)}\right)\middle|X_i\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\Delta\epsilon_i, \hat{\beta}\middle|X, X_M\right]^t \Delta X_i\middle|X_i\right]
\end{aligned}$$

The first term arises from the bias of the estimated regression function, the second term comes from the fact that the regression function is estimated rather than known, the third and fourth terms are derived from idiosyncratic noise (the error terms), and the final term comes from the fact that the regression function is estimated from the same data that is used in the matching estimator.

In order to understand the magnitude of the final term, we need to more carefully specify the form of the estimator of $\hat{\beta}$. In particular, we will allow it to use a collection of both treated and untreated subjects, which we denote by $\mathcal{I}_c$ (which will technically be a sequence since a control may be included multiple times at separate indices) and may use a set of covariates that differs from those in the true model; in particular, we allow it to also include treatment as a covariate (which will be necessary if treated individuals are used in the fitting process). Because the covariates used in the regression may not be the same as those in the true model (most importantly, if it is misspecified, the regression model will omit covariates that enter nontrivially in the true model), we will use the general form $Y_i = \mu_a(X_i) + \epsilon_i$ (the reason for this will become clear momentarily). Thus, $\hat{\beta}$ may correspond to only a subset of the coefficients estimated in the linear regression (in particular, it will omit the coefficient of the treatment variable, if treatment is included in the model). Then, $\hat{\beta} = \Pi \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i Y_i = \Pi \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i \mu_0(X_i) + \Pi \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i \epsilon_i$, where Π is the matrix that projects the estimated covariates onto the subset corresponding to β (in fact, we can let Π be an arbitrary matrix).

The first term is a function of the Xs alone. Thus, if i corresponds to a treated subject, $\text{Cov}\left[\epsilon_i, \hat{\beta}\middle| X, X_M\right] = \Pi \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} X_i \sigma_\epsilon^2 (X_i) = N_1^{-1} \Pi \text{E}\left[X_i X_i^t\right]^{-1} X_i \sigma_\epsilon^2 (X_i) + o_p\left(N_1^{-1}\right)$. If $M(i)$ corresponds to a matched untreated control,
$\text{Cov}\left[\epsilon_{M(i)}, \hat{\beta}\middle| X, X_M\right] = K_{M(i)} \Pi \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} X_{M(i)} \sigma_\epsilon^2 (X_i)$, where, borrowing notation from Abadie and Imbens, $K_{M(i)}$ is the number of times that untreated control $M(i)$ is used as a match [1, 2]. Thus, in order for the last term to be $O\left(N_1^{-1}\right)$, $\text{E}\left[K_{M(i)}\middle| X_i\right] \in O(1)$.

The equivalent calculations for the ATT are,

$$\begin{aligned}
\text{Var}\left[\hat{\tau}|X\right] = \text{Var}&\left[\left(\text{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\middle| X\right] + \text{E}\left[\overline{\Delta X}^t \text{Var}\left[\hat{\beta}\middle| X, X_M\right] \overline{\Delta X}\middle| X\right] \\
&+ N_1^{-2} \sum_{i=1}^{N_1} \sigma_\epsilon^2 (X_i) + N_1^{-2} \text{E}\left[\text{Var}\left[\sum_{i=1}^{N_1} \epsilon_{M(i)}\middle| X_M\right]\middle| X\right] \\
&- 2\text{E}\left[\text{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]^t \overline{\Delta X}\middle| X\right]
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}\left[\hat{\tau}\right] &= \mathrm{Var}\left[\bar{\tau} + \left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] + \mathrm{E}\left[\overline{\Delta X}^t \mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right] \overline{\Delta X}\right] \\
&\quad + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] + N_1^{-2}\mathrm{E}\left[\mathrm{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle| X_M\right]\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]^t \overline{\Delta X}\right] \\
&= N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] + 2\mathrm{Cov}\left[\bar{\tau}(X), \left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] \\
&\quad + \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]^t \overline{\Delta X}\right] \\
&\quad + \mathrm{E}\left[\overline{\Delta X}^t \mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right] \overline{\Delta X}\right] + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] + N_1^{-2}\mathrm{E}\left[\mathrm{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle| X_M\right]\right]
\end{aligned}
$$

In order to make sense of this expression, we first compute $\mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right]$ and $\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]$. Using the notation from earlier,

$$
\begin{aligned}
\mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right] =& \mathrm{Var}\left[\Pi\left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i\left(\mu_a(X_i) + \epsilon_i\right)\middle| X, X_M\right] \\
&= \Pi\left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i^t K_i \sigma_\epsilon^2(X_i) X_i \left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \Pi^t
\end{aligned}
$$

where K_i is the number of times a particular subject occurs in X_c. For a treated subject, $K_i = 1$, while for a matched untreated control, K_i may be greater than one, since a single untreated control may be matched to multiple treated subjects.

$$
\begin{aligned}
\mathrm{Cov}\left[\epsilon_i, \hat{\beta}\middle| X, X_M\right] &= \mathrm{Cov}\left[\epsilon_i, \Pi\left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i\left(\mu_a(X_i) + \epsilon_i\right)\middle| X, X_M\right] \\
&= \Pi\left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} K_i X_i \sigma_i^2(X_i)
\end{aligned}
$$

If β is estimated using only the untreated controls, then,

$$
\begin{aligned}
\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right] &= \mathrm{Cov}\left[N_1^{-1}\sum_{i=1}^{N_1}\left(\epsilon_{M(i)} - \epsilon_i\right), \Pi\left(\sum_{i\in\mathcal{I}_c} X_i X_i^t\right)^{-1} \sum_{i\in\mathcal{I}_c} X_i\left(\mu_a(X_i) + \epsilon_i\right)\middle| X, X_M\right] \\
&= N_1^{-1}\Pi\left(\sum_{i=1}^{N_1} X_{M(i)} X_{M(i)}^t\right)^{-1} \sum_{i=1}^{N_1} K_{M(i)} X_i \sigma^2(X_{M(i)})
\end{aligned}
$$

If β is estimated using both treated and untreated subjects then,

$$\begin{aligned}\mathrm{Cov}\left[\overline{\Delta\epsilon},\hat{\beta}\middle|X,X_M\right] &= \mathrm{Cov}\left[N_1^{-1}\sum_{i=1}^{N_1}\left(\epsilon_{M(i)}-\epsilon_i\right),\Pi\left(\sum_{i\in\mathcal{I}_c}X_iX_i^t\right)^{-1}\sum_{i\in\mathcal{I}_c}X_i\left(\mu_a(X_i)+\epsilon_i\right)\middle|X,X_M\right]\\ &= N_1^{-1}\Pi\left(\sum_{i=1}^{N_1}\left(X_iX_i^t+X_{M(i)}X_{M(i)}^t\right)\right)^{-1}\\ &\quad\times\sum_{i=1}^{N_1}\left(X_{M(i)}K_{M(i)}\sigma^2(X_{M(i)})-X_i\sigma^2(X_i)\right)\end{aligned}$$

If $\mu_{X,1} << \mu_{X,0}$, $\frac{\partial\mu_{X,1}}{\partial\mu_{X,0}}$ is bounded almost surely, and $N_0^{-1}N_1^{3+\epsilon}\to 0$, $\epsilon>0$, asymptotically the untreated controls to which the treated subjects are matched will become independent, so that each treated/untreated matched pair will be independent of every other one so that $K_i\to 1$. Then, if only the matched controls are used in the regression model

$$N_1\mathrm{Var}\left[\hat{\beta}\middle|X,X_M\right]\xrightarrow{p}\Pi\mathrm{E}\left[X_{M(i)}X_{M(i)}^t\right]^{-1}\mathrm{E}\left[X_{M(i)}^t\sigma_\epsilon^2\left(X_{M(i)}\right)X_{M(i)}\right]\mathrm{E}\left[X_{M(i)}X_{M(i)}^t\right]^{-1}\Pi^t$$

$$N_1\mathrm{Cov}\left[\overline{\Delta\epsilon},\hat{\beta}\middle|X,X_M\right]\xrightarrow{p}\Pi\mathrm{E}\left[\sum_{i=1}^{N_1}X_{M(i)}X_{M(i)}^t\right]^{-1}\mathrm{E}\left[\sigma^2(X_{M(i)})\right]$$

If both the treated subjects and their matched controls are used then,

$$\begin{aligned}N_1\mathrm{Var}\left[\hat{\beta}\middle|X,X_M\right]\xrightarrow{p}\Pi\mathrm{E}&\left[X_iX_i^t+X_{M(i)}X_{M(i)}^t\right]^{-1}\mathrm{E}\left[X_i^t\sigma_\epsilon^2\left(X_{M(i)}\right)X_i+X_{M(i)}^t\sigma_\epsilon^2\left(X_{M(i)}\right)X_{M(i)}\right]\\ &\times\mathrm{E}\left[X_iX_i^t+X_{M(i)}X_{M(i)}^t\right]^{-1}\Pi^t\end{aligned}$$

$$N_1\mathrm{Cov}\left[\overline{\Delta\epsilon},\hat{\beta}\middle|X,X_M\right]\xrightarrow{p}\Pi\mathrm{E}\left[X_iX_i^t+X_{M(i)}X_{M(i)}^t\right]^{-1}\mathrm{E}\left[X_{M(i)}\sigma^2(X_{M(i)})-X_i\sigma^2(X_i)\right]$$

Let $V_\beta=\lim_{N_1\to\infty}N_1\mathrm{Var}\left[\hat{\beta}\middle|X,X_M\right]$ and $C_\beta=\lim_{N_1\to\infty}N_1\mathrm{Cov}\left[\overline{\Delta\epsilon},\hat{\beta}\middle|X,X_M\right]$. Then, defining $V'_\beta=N_1\mathrm{Var}\left[\hat{\beta}\middle|X,X_M\right]-V_\beta$, $C'_\beta=N_1\mathrm{Cov}\left[\overline{\Delta\epsilon},\hat{\beta}\middle|X,X_M\right]-C_\beta$, $V'_\beta,C'_\beta\xrightarrow{p}0$. If we further assume that $\mathrm{E}\left[\overline{\Delta X}^tV'_\beta\overline{\Delta X}\right]\in o\left(\mathrm{E}\left[\overline{\Delta X}^t\overline{\Delta X}\right]\right)$ and $\mathrm{E}\left|\overline{\Delta X}^tC'_\beta\right|\in o\left(\mathrm{E}\left\|\overline{\Delta X}\right\|_1\right)\subseteq o\left(\mathrm{E}\left\|\overline{\Delta X}\right\|_2\right)$,

then we can write,

$$\begin{aligned}\mathrm{Var}\left[\hat{\tau}\right] = N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] &+ 2\mathrm{Cov}\left[\bar{\tau}(X), \left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] \\ &+ \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] - 2N_1^{-1}C_\beta^t \mathrm{E}\left[\overline{\Delta X}\right] \\ &+ N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_{M(i)}\right)\right] \\ &+ o\left(N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t \overline{\Delta X}\right]\right) + o\left(N_1^{-1}\mathrm{E}\left\|\overline{\Delta X}\right\|_1\right)\end{aligned}$$

If $X \perp\!\!\!\perp \epsilon$, we get one final simplification so,

$$\begin{aligned}\mathrm{Var}\left[\hat{\tau}\right] = N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] &+ 2\mathrm{Cov}\left[\bar{\tau}(X), \left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] \\ &+ \mathrm{Var}\left[\left(\mathrm{E}\left[\hat{\beta}\middle| X, X_M\right] - \beta\right)^t \overline{\Delta X}\right] - 2N_1^{-1}C_\beta^t \mathrm{E}\left[\overline{\Delta X}\right] \\ &+ N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] + 2\sigma_\epsilon^2 \\ &+ o\left(N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t \overline{\Delta X}\right]\right) + o\left(N_1^{-1}\mathrm{E}\left\|\overline{\Delta X}\right\|_1\right)\end{aligned}$$

so, if $\hat{\beta}$ is unbiased, the variance of the bias corrected matching estimator will only be affected by the matching process used through the moments of the matching discrepancy, which provides us with a straightforward way to compare the effects of different matching methods.

2.6 Variance, Matching Method, and Covariate Distribution

We can now evaluate how the variances of matching estimators depend on the distribution of the covariates and the matching method used. For simplicity, we will focus on bias-corrected matching estimators in which the true and regression models are affine and correctly specified, meaning that the model specified by the regression function contains the true model as a submodel (so it can contain additional covariates not included in the true model, but must include every covariate on which the true model nontrivially depends). Additionally, we will restrict our focus to the case in which all covariates are independent and identically distributed and are matched on. The case in which derived covariates appear in the regression function will be explored in future work. If the regression function is properly specified, then the expression for the variance of the bias-corrected

matching estimator simplifies to

$$\begin{aligned}\mathrm{Var}\left[\hat{\tau}\right] = N_1^{-1}\mathrm{Var}\left[\tau(X_i)\right] + \mathrm{E}\left[\overline{\Delta X}^t \mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right]\overline{\Delta X}\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]^t \overline{\Delta X}\right] \\ + N_1^{-1}\mathrm{E}\left[\sigma_\epsilon^2\left(X_i\right)\right] + N_1^{-2}\mathrm{E}\left[\mathrm{Var}\left[\sum_{i=1}^{N_1}\epsilon_{M(i)}\middle| X_M\right]\right]\end{aligned}$$

If we further assume that $X \perp\!\!\!\perp \epsilon$ and $\mu_{X,1} << \mu_{X,0}$, $\frac{\partial\mu_{X,1}}{\partial\mu_{X,0}}$ is bounded, and $N_0^{-1}N_1^{3+\epsilon} \to 0$, $\epsilon > 0$, then, asymptotically, the matched subjects will be independent and

$$\mathrm{Var}\left[\hat{\tau}\right] \approx N_1^{-1}\left(\mathrm{Var}\left[\tau(X_i)\right] + 2\sigma_\epsilon^2\right) + \mathrm{E}\left[\overline{\Delta X}^t \mathrm{Var}\left[\hat{\beta}\middle| X, X_M\right]\overline{\Delta X}\right] - 2\mathrm{E}\left[\mathrm{Cov}\left[\overline{\Delta\epsilon}, \hat{\beta}\middle| X, X_M\right]^t \overline{\Delta X}\right]$$

so that the variance separates into two parts, the first term, which is independent of the matching scheme used, and a matching dependent part that depends explicitly on the matching discrepancy.

While the second term is manifestly positive, the third term need not be. In order to address this possibility, a simple, but less common solution would be to fit the regression function using data that is not otherwise used in the matching procedure. In that case, the final term vanishes, and every term contributes positively to the variance. In a more realistic scenario, $\hat{\beta}$ is fit using some collection of the same untreated controls to which treated subjects may be potentially matched. If the matched controls are the only ones used to fit the regression, then $\mathrm{Cov}\left[\Delta\epsilon, \hat{\beta}\right]$ will be $O\left(N_1^{-1}\right)$, which is the same order as the remaining terms. However, if additional controls are used, either the entire population of untreated subjects, N_0, or simply additional matches that are not used to estimate the ATT (perhaps r additional matches are drawn for each treated subject, after the controls that are used to compute the ATT are removed from the pool), then the covariance term will instead be $O\left(N_0^{-1}\right)$ or $O\left(\left(rN_1\right)^{-1}\right)$, respectively. So, if $N_1N_0^{-1} \to 0$, as we have already assumed, or $r \to \infty$, then the covariance term will go to zero much faster than the other terms which are of order $O\left(N_1^{-1}\right)$, and its sign will not ultimately matter. If we additionally decompose the variance, as we did in the previous section, we obtain,

$$\mathrm{Var}\left[\hat{\tau}\right] \approx N_1^{-1}\left(\mathrm{Var}\left[\tau(X_i)\right] + 2\sigma_\epsilon^2\right) + N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] + o\left(N_1^{-1}\mathrm{E}\left[\overline{\Delta X}^t\overline{\Delta X}\right]\right)$$

Thus, in order to understand the behavior of the variance under different matching methods,

we need to characterize $\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right]$. Expanding the quadratic form (and assuming that the covariates are IID) yields

$$\begin{aligned}
\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] &= \sum_{u,v} V_{\beta,uv} \mathrm{E}\left[\overline{\Delta X}^{(u)} \overline{\Delta X}^{(v)}\right] = \sum_{u,v} V_{\beta u,v} N_1^{-2} \sum_{i,j} \mathrm{E}\left[\Delta X_i^{(u)} \Delta X_j^{(v)}\right] \\
&= N_1^{-2} \sum_{u,v} V_{\beta,uv} \left[\sum_{i=j} \mathrm{E}\left[\Delta X_i^{(u)} \Delta X_j^{(v)}\right] + \sum_{i \neq j} \mathrm{E}\left[\Delta X_i^{(u)} \Delta X_j^{(v)}\right]\right] \\
&= N_1^{-2} \sum_{u,v} V_{\beta,uv} \left[\sum_{i} \mathrm{E}\left[\Delta X_i^{(u)} \Delta X_i^{(v)}\right] + \sum_{i \neq j} \mathrm{E}\left[\Delta X_i^{(u)}\right] \mathrm{E}\left[\Delta X_j^{(v)}\right]\right] \\
&= N_1^{-2} \sum_{u,v} V_{\beta,uv} \left[N_1 \mathrm{E}\left[\Delta X_i^{(u)} \Delta X_i^{(v)}\right] + N_1 \cdot (N_1 - 1) \mathrm{E}\left[\Delta X_i^{(u)}\right] \mathrm{E}\left[\Delta X_j^{(v)}\right]\right] \\
&= \sum_{u,v} V_{\beta,uv} \left[N_1^{-1} \left(\mathrm{E}\left[\Delta X_i^{(u)} \Delta X_i^{(v)}\right] - \mathrm{E}\left[\Delta X_i^{(u)}\right] \mathrm{E}\left[\Delta X_i^{(v)}\right]\right)\right. \\
&\qquad \left. + \mathrm{E}\left[\Delta X_i^{(u)}\right] \mathrm{E}\left[\Delta X_i^{(v)}\right]\right] \\
&= \sum_{u,v} V_{\beta,uv} \left[N_1^{-1} \mathrm{Var}\left[\Delta X_i\right]_{uv} + \mathrm{E}\left[\Delta X_i^{(u)}\right] \mathrm{E}\left[\Delta X_i^{(v)}\right]\right]
\end{aligned}$$

Thus, the expectation has two pieces: a portion that arises from the variance of the matching discrepancy and is $O\left(N_1^{-1}\right)$ and a part that comes from the squared bias and is $O(1)$. To complete the computation, we need to replace moments of the matching discrepancy with their actual values. For direct covariate matching, we can use the expressions,

$$\begin{aligned}
\mathrm{E}[\Delta X | X] &= d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \pi^{-1} f(X)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x}(X) \cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \\
\mathrm{E}[\Delta X \Delta X^t | X] &= d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \pi^{-1} f(X)^{-\frac{2}{d}} \cdot \mathcal{I}_d \cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

from previously, while for propensity score matching we can compute the moments of the matching discrepancy directly.

As we showed above, using a fixed propensity score is equivalent to random matching so that X and its match are independent and, thus, $\mathrm{E}\left[\Delta X_i\right] = 0$ and $\mathrm{Var}\left[\Delta X_i\right] = \mathrm{Var}\left[X_i\right] + \mathrm{Var}\left[X_{M(i)}\right] = 2\mathrm{Var}\left[X_i\right]$. We will use these results to consider how the choice of matching method and the distribution of covariates affects the variance of the bias-corrected matching estimate of the ATT.

We are ultimately interested in the unconditional expectations of these terms. Under our

assumptions that the covariates are independent and identically distributed, the expressions for the unconditional expectations can be simplified somewhat, using

$$\mathrm{E}\left[f\left(X_i\right)^{-\frac{2}{d}}\right] = \prod_u \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right] = \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^d$$

$$\begin{aligned}
\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right] &= \mathrm{E}\left[f\left(X_i^{(u)}\right)\cdot\frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\cdot\prod_{v\neq u} f\left(X_i^{(v)}\right)^{-\frac{d+2}{d}}\right] \\
&= \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\right]\cdot\prod_{v\neq u}\mathrm{E}\left[f\left(X_i^{(v)}\right)^{-\frac{2}{d}}\right] \\
&= \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\right]\cdot\mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^{d-1}
\end{aligned}$$

so, we have,

$$\begin{aligned}
\mathrm{E}[\Delta X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\right] \\
\times\mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^{d-1}\cdot N_0^{-\frac{2}{d}} \\
+ o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

$$\mathrm{E}[\Delta X\Delta X^t|X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^d\mathcal{I}_d\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)$$

Due to the absence of off diagonal terms, to highest order, in the expression for the second moment of the matching discrepancy, and the fact that the covariates are IID, the expression for $\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right]$ simplifies to

$$\begin{aligned}
\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] &= \sum_{u,v} V_{\beta,uv}\left[N_1^{-1}\mathrm{Var}\left[\Delta X_i\right]_{uv} + \mathrm{E}\left[\Delta X_i^{(u)}\right]\mathrm{E}\left[\Delta X_i^{(v)}\right]\right] \\
&= N_1^{-1}\mathrm{Var}\left[\Delta X_i^{(u)}\right]\sum_u V_{\beta,uu} + \mathrm{E}\left[\Delta X_i^{(u)}\right]^2\sum_{u,v} V_{\beta,uv}
\end{aligned}$$

Since propensity score matching has zero matching discrepancy the expression for propensity

score matching is even simpler,

$$\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] = N_1^{-1}\mathrm{Var}\left[\Delta X_i^{(u)}\right]\sum_u V_{\beta,uu}$$

It will also be useful to note that, using Stirling's approximation,

$$d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1} \sim (\pi d)^{d^{-1}}(2\pi e)^{-1}$$

2.6.1 Uniform Distribution

We will first consider the case in which the covariates are distributed uniformly ($f(x) = \mathrm{I}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$). As we discussed previously, symmetric distributions have zero expected matching discrepancy under direct nearest-neighbor matching, so the variance (or equivalently, the second moment) of the matching discrepancy will determine the effect of the matching discrepancy on the variance of the estimator of the ATT. Then for direct matching,

$$\begin{aligned}\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] &= N_1^{-1}\cdot d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}N_1^{-1}\right)\\ &\sim (\pi d)^{d^{-1}}(2\pi e)^{-1}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}}N_1^{-1} + o\left(N_0^{-\frac{2}{d}}N_1^{-1}\right)\end{aligned}$$

while for the propensity score,

$$\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] = \frac{1}{6}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_1^{-1} + o\left(N_1^{-1}\right)$$

since $\mathrm{Var}\left[\Delta X\right] = \frac{1}{6}$.

Thus, direct matching will result in a smaller value of $\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right]$ than propensity score matching if the ratio of their leading terms $3(\pi d)^{d^{-1}}(2\pi e)^{-1}N_0^{-\frac{2}{d}} < 1$, which will always occur if $d \geq 2$ or $N_0 \geq 2$. Thus, direct matching will always be preferred to propensity score matching for estimating the bias-corrected matching ATT, but this advantage will be less prominent for smaller values of N_0 and will lessen as d, since this decreases the effect of N_0.

2.6.2 Normal Distribution

Next, we consider the case in which the covariates are normally distributed ($f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2\sigma^2}}$). Since the normal distribution is also symmetric, the value of $\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right]$ will again be determined solely by the variance of matching discrepancy.

For direct matching,

$$\begin{aligned}\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] &= 2\sigma^2 d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\left(\frac{d}{d-2}\right)^{\frac{d}{2}}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}} N_1^{-1} + o\left(N_0^{-\frac{2}{d}} N_1^{-1}\right) \\ &\sim \sigma^2 e^{-1} d^{-1}(\pi d)^{d^{-1}}\left(\frac{d}{d-2}\right)^{\frac{d}{2}}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}} N_1^{-1} + o\left(N_0^{-\frac{2}{d}} N_1^{-1}\right)\end{aligned}$$

For propensity score matching,

$$\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] = 2\sigma^2 \cdot \mathrm{tr}\left[V_\beta\right]\cdot N_1^{-1} + o\left(N_1^{-1}\right)$$

So, direct matching will be preferred if the ratio of the leading terms $2^{-1}e^{-1}d^{-1}(\pi d)^{d^{-1}}\left(\frac{d}{d-2}\right)^{\frac{d}{2}} N_0^{-\frac{2}{d}} < 1$, which will always be the case for $d > 2$. Thus, as in the uniform case, direct matching will be preferred to propensity score matching, because direct matching is also unbiased for symmetric distributions, but its matching discrepancy has a lower variance and this advantage will grow with the number of available controls, N_0. However, since the ratio scales like $N_0^{-\frac{2}{d}}$, this advantage will rapidly become smaller as d increases, since the value of an additional control in decreasing the variance of the matching discrepancy rapidly diminishes with d.

2.6.3 Exponential Distribution

Finally, we consider the case in which the covariates are exponentially distributed ($f(x) = \lambda^{-1}e^{-\lambda x}$). Notably, unlike the uniform and normal distributions, the exponential distribution is not symmetric, and, thus, direct matching will not be unbiased. Therefore, we can no longer simply compare the ratio of the variances in order to determine which matching technique performs better.

Using direct matching,

$$\begin{aligned}
\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] &= \lambda^{-2}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\left(\frac{d}{d-2}\right)^{d}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}}N_1^{-1} \\
&\quad + \lambda^{-2}\pi^{-2}d^{-2}\Gamma\left(\frac{d+2}{d}\right)^{2}\Gamma\left(\frac{d+2}{2}\right)^{\frac{4}{d}}\left(\frac{d}{d-2}\right)^{d}\cdot \sum_{u,v} V_{\beta,uv}\cdot N_0^{-\frac{4}{d}} \\
&\quad + o\left(N_0^{-\frac{4}{d}}\right) + o\left(N_0^{-\frac{2}{d}}N_1^{-1}\right) \\
&\sim \lambda^{-2}(\pi d)^{d^{-1}}(2\pi e)^{-1}\left(\frac{d}{d-2}\right)^{d}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}}N_1^{-1} \\
&\quad + \lambda^{-2}(\pi d)^{2d^{-1}}(2\pi e)^{-2}\left(\frac{d}{d-2}\right)^{d}\cdot \sum_{u,v} V_{\beta,uv}\cdot N_0^{-\frac{4}{d}} \\
&\quad + o\left(N_0^{-\frac{4}{d}}\right) + o\left(N_0^{-\frac{2}{d}}N_1^{-1}\right) \\
&\sim \lambda^{-2}(2\pi)^{-1}e\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_0^{-\frac{2}{d}}N_1^{-1} + \lambda^{-2}(2\pi)^{-2}e^2\cdot \sum_{u,v} V_{\beta,uv}\cdot N_0^{-\frac{4}{d}} \\
&\quad + o\left(N_0^{-\frac{4}{d}}\right) + o\left(N_0^{-\frac{2}{d}}N_1^{-1}\right)
\end{aligned}$$

Using propensity score matching gives,

$$\mathrm{E}\left[\overline{\Delta X}^t V_\beta \overline{\Delta X}\right] = 2\lambda^{-2}\cdot \mathrm{tr}\left[V_\beta\right]\cdot N_1^{-1} + o\left(N_1^{-1}\right)$$

since $\mathrm{Var}\left[\Delta X\right] = 2\lambda^{-2}$.

If direct matching were unbiased, it would again have the advantage. The ratio of the leading terms of the variance of the matching discrepancy is $(4\pi)^{-1}eN_0^{-\frac{2}{d}}$, so, under direct matching, the matching discrepancy has a lower variance. However, direct matching also results in biased matching and so contributes an additional $O\left(N_0^{-\frac{4}{d}}\right)$ term, which, notably, does not decline as N_1 grows. Thus, as N_1 increases, propensity score matching will outperform direct covariate matching. This will happen most rapidly when d is larger, since the effect of using more controls will be greatly diminished in high dimensions since $N_0^{\frac{4}{d}}$ will grow very slowly.

2.7 Conclusion

In this work, we examined the effect of the choice of matching method on the performance of matching based estimators of the Average Effect of Treatment. In particular, we found that, when both the true outcome model and the regression model are affine, the ATT is unbiased under simple matching and asymptotically unbiased under bias-corrected matching, if either the regression model is correctly specified or the expected matching discrepancy is zero. The latter criterion is equivalent to the means of the treated subjects and matched untreated controls being equal, after matching. This is a strong form of double robustness, which is much more flexible than the weaker result, seen in the general case, that the ATT will be unbiased if either the regression model is properly specified or the matching is perfect, particularly since the second condition can never hold when the covariates are continuous. Further, it provides theoretical justification for the standard practice of measuring imbalance between two groups by comparing the differences in means.

We further showed that, under the above linearity assumption, when the propensity score is constant, propensity score matching is always unbiased and, therefore, results in asymptotically unbiased estimators of the ATT, even when the regression model is misspecified. This provides a basis for preferring propensity score matching over direct covariate matching for estimating the ATT using matching estimators, although this ignores the difference in the variance between the two methods. An interesting compromise is Coarsened Exact Matching, which will also result in zero expected matching discrepancy when the propensity score is constant, but will better control the variance of the matching discrepancy, since matches can occur only within the same coarsened level of each covariate. However, since CEM is a form of perfect matching on the coarsened covariates, it will require a large number of controls in order to avoid throwing away many cases that cannot be exactly matched.

We also showed that, when the covariates are symmetrically distributed, direct matching is also unbiased, which would lead to it being preferred over propensity score matching because of its lower variance. However, it is often the case that the true or regression models may contain derived covariates which are functions of the covariates that are actually matched upon. In general, these derived covariates will not be uniform or symmetric even if the original covariates from which they were derived were. For example, if $X^{(1)}$ is symmetric, or even uniform, $X^{(1)2}$ will, in general,

not be. Thus, when using direct covariate matching, the matching discrepancy for these derived covariates will be nonzero and the estimated ATT will be biased. However, under propensity score matching, the matching discrepancy of all covariates, even derived ones, will be zero, and so the ATT will be asymptotically unbiased.

Finally we explored how, under some mild regularity conditions on the data used to fit the regression function, the variance of the ATT behaves under direct covariate matching vs. propensity score matching when the regression model is correctly specified. When the distribution of the covariates is symmetric, direct matching is preferred since it is unbiased and its matching discrepancy has a lower variance. However, when the distribution of the covariates is not symmetric, when the number of treated subjects, N_1, is sufficiently large, propensity score matching will result in a lower variance of the estimated ATT, because the deleterious effects of the bias of direct covariate matching will overwhelm the advantage of its matching discrepancy having lower variance. The sample size at which this occurs will depend on the number of available controls (a larger number of which favors direct matching by reducing its variance) and the number of covariates being matched, which significantly attenuates the benefit of additional control subjects. Thus, when the underlying distribution of the covariates is not symmetric, we encounter a form of bias-variance tradeoff in which direct matching may perform better with smaller numbers of subjects, but is eventually overwhelmed by its bias and, so, is outperformed by propensity score matching, as the number of treated subjects grows, an effect that only becomes more pronounced as the number of covariates increases.

Taken together, these results show that neither direct covariate matching, nor propensity score matching uniformly dominates the other. The optimal method for any given setting will depend intimately on the distribution of the covariates being matched and, thus, the structure of the problem. Selecting the best technique remains, as it always has, a reflection of the analyst's understanding of the nature of the problem and the available data.

Bibliography

[1] Alberto Abadie and Guido Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects". *Journal of the American Statistical Association* 29.1 (2011), pp. 1–11.

[2] Alberto Abadie and Guido Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica* 74.1 (2006), pp. 237–267.

[3] Keisuke Hirano, Guido W. Imbens, and Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score". *Econometrica* 71.4 (2003), pp. 1161–1189.

[4] Stefano M. Iacus, Gary King, and Giuseppe Porro. "Causal Inference Without Balance Checking: Coarsened Exact Matching". *Political Analysis* 20.1 (2012), pp. 1–24.

[5] Gary King and Richard Nielsen. "Why Propensity Scores Should Not Be Used for Matching". *Political Analysis* 27.4 (2019).

[6] Paul Rosenbaum and Donald Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika* 70.1 (1983), pp. 41–55.

[7] Donald Rubin. "Matching to Remove Bias in Observational Studies". *Biometrics* 29 (1973), pp. 159–183.

[8] Donald Rubin. "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies". *Biometrics* 29 (1973), pp. 185–203.

Chapter 3

Model Dependence of Matching Estimators

Tom Kolokotrones, Max Goplerud, Richard Nielsen, Gary King, and James Robins

3.1 Introduction

Matching is widely used across a variety of fields, particularly in the health and social sciences and, in order to estimate causal effects, such as the Average Effect of Treatment on the treated, as well as to reduce the variance of estimates. However, the optimal form of matching is fiercely debated, with the two most popular candidates being matching directly on the covariates (or some simple transformation thereof, such as Mahalanobis matching, which first standardizes and decorrelates the covariates before matching) or matching on the propensity score, the probability that a subject will be treated conditional on his covariates [5–7].

In a provocatively titled recent article, "Why Propensity Scores Should not be Used for Matching," King and Nielsen compare the use of several different forms of matching to estimate causal treatment effects from observational or randomized data and conclude that matching on the propensity score is inferior to other forms of matching [3]. In order to reach this conclusion, they compare the performance of propensity score matching against other forms of matching, primarily direct covariate matching, using a combination of simulated and real-world data.

In the central analysis of the paper, they focus on how the choice of matching method, specifically direct covariate (Mahalanobis) matching vs. propensity score matching, affects the sensitivity of estimates of the Average Effect of Treatment on the Treated (ATT) to the investigator's choice of regression model. To do so, they simulate 100 data sets and then, for each data set, proceed as follows.

They first select a distance measure (e.g. Mahalanobis distance or the absolute difference in propensity score) and then assign to each treated subject an untreated control, which minimizes this distance. Next, they prune the M worst matches, where M is allowed to vary from 0 to the number of treated subjects and estimate the ATT by using the remaining pairs to fit a collection of K linear regressions, where the regression functions are affine functions of treatment, several

covariates, and higher order functions of the covariates including quadratic terms and interactions between the covariates, but not with the treatment itself.

For each M, the authors then calculate two quantities, the empirical variance and the maximum estimate of the ATT over the K models. They then plot the average of each these measures across simulations vs. M, finding that matching on Mahalanobis distance strictly dominates propensity score matching for all values of M, with respect to the empirical variance, and that it is generally superior with respect to minimizing the bias of the maximum estimated ATT, becoming strictly superior past a certain value of M. They conclude from this, and the consideration of several forms of covariate imbalance, that Mahalanobis matching is to be strictly preferred over propensity score matching.

The empirical variance and maximum estimated ATT over a class of models can be thought of as measures of how sensitive the estimated ATT is to the choices made by the analyst [3]. If the variance is high, or the maximum estimated ATT differs significantly from the truth, then the reported ATT will be heavily dependent on the the model the analyst selects, which enables him, consciously, or unconsciously, to bias the presented results towards his desired conclusion [3]. A high variance across models also means that investigators with identical data, but different preferences, may interpret the results in vastly different ways by choosing models that favor their positions. Thus, it is reasonable to consider such quantities to be measures of Model Dependence and to favor matching methods that minimize this dependence. In this work, we will focus on the empirical variance of the ATT over models, which we will call Model Dependence, since both the empirical variance of the ATT across models and maximum observed ATT measure similar things, but the empirical variance is much more analytically tractable since the maximum is typically not a smooth function of the inputs.

In what follows we will compare direct matching on the covariates to an extreme version of propensity score matching in which the propensity score is constant across individuals. This puts the propensity score at an apparently extreme disadvantage since it cannot use any information about the covariates, while direct matching has complete access to the covariate values, and, therefore, propensity score matching in this setting is equivalent to matching at random. This is particularly interesting since King and Nielsen note that propensity score matching behaves similarly to random matching, meaning that, even when the propensity score is not fixed, matching on the propensity

score results in matches that are often far apart using conventional measures of distance, such as Euclidean or Mahalanobis distance, even when much closer matches are available. They believe that this phenomenon is, at least partially, responsible for the poor performance of propensity score matching in their analysis. Therefore, particularly since using a constant propensity score emphasizes one of King and Nielsen's most significant problems with the propensity score, finding that the propensity score is able to outperform direct matching in this setting would then provide strong evidence that propensity score matching is not uniformly dominated by direct matching.

The remainder of the paper is is structured as follows. Section 2 introduces the notation and defines Model Dependence. Section 3 derives analytic expressions for the Model Dependence and specializes them to both direct and propensity score matching, using both simple and derived covariates in the regression functions. Section 4 compares the Model Dependence of direct and propensity score matching for estimating the ATT and characterizes situations in which one may be preferred to the other. Section 5 examines when pruning "bad" matches decreases model dependence for direct matching and when it is harmful. Section 6 shows that performing regression after matching can be viewed as a type of bias-corrected estimator, thereby extending our results to this widely used approach. Section 7 concludes.

3.2 Basics

Let Y be a continuous outcome, A a binary treatment (where $A = 1$ corresponds to treatment and $A = 0$ to no treatment), X a vector of continuous covariates, and $Y(a)$ the counterfactual outcome under treatment a. We use $[n]$ to denote the set $\{1, \ldots, n\}$, as is typical in Computer Science. Consider a collection of N_1 treated and N_0 untreated subjects. Our goal is to estimate the expected effect of treatment, A, on the outcome, Y, among those subjects that were treated, the so-called Average Effect of Treatment on the Treated (ATT), which is given by $\tau = \mathrm{E}[Y(1) - Y(0)|A = 1]$.

Consider a matching criterion M, which is a function $M : [N_1] \to [N_0]^m$ that associates with each of the N_1 treated subjects m of the N_0 untreated controls. In what follows, we will restrict attention to the case of 1-1 matching with replacement, so that $m = 1$ and each treated subject is matched to exactly one untreated control (although multiple treated subjects may be matched to

the same untreated control). In the typical case, a match is defined as the untreated control $j = M(i) \in [N_0]$ that minimizes some distance $d(X_i, X_j)$, where d is some metric, such as the Euclidean or Mahlanobis distance or the absolute difference in propensity score. Let $\tau_i = Y_i(1) - Y_i(0)$ be the individual treatment effect for subject i. Since exactly one of $Y_i(0)$ and $Y_i(1)$ will actually be observed, τ_i must be estimated. For any random variable, Z, we use Z to represent the values of the treated subjects and Z_M for the values of the matched controls and define $\Delta Z_i = Z_{M(i)} - Z_i$ and $\overline{Z} = N_1^{-1} \sum_{i=1}^{N_1} Z_i$. Following Abadie and Imbens, we refer to ΔZ as the matching discrepancy of Z [1, 2].

In what follows, we will focus on so-called bias-corrected matching estimators as defined by Abadie and Imbens [1, 2] and originally described by Rubin [8]. Simple matching estimators estimate the individual effect of treatment on the treated by $\hat{\tau}_i = Y_i - Y_{M(i)}$. However, since, untreated subject $M(i)$, the control matched to treated subject i, may not have the exact same covariate values as subject i, this estimator may be biased. If $Y_i(0) = \mu_0(X_i) + \epsilon_i$, where $\mathrm{E}[\epsilon_i | X_i] = 0$, then we could use μ_0 to correct for the fact that $X_i \neq X_{M(i)}$. However, μ_0 is typically unknown so we must use a substitute, $\hat{\mu}_0$, that is estimated using some subset of the population of untreated subjects. The bias-corrected matching estimator is then $\hat{\tau}_i = Y_i - Y_{M(i)} + \mu_0\left(X_{M(i)}\right) - \mu_0(X_i)$ and the corresponding estimate of the ATT is $\hat{\tau} = N^{-1} \sum_{i=1}^{N_1} \left[Y_i - Y_{M(i)} + \hat{\mu}_0\left(X_{M(i)}\right) - \hat{\mu}_0(X_i)\right]$.

Since the form of μ_0 is typically unknown, and must be postulated by the analyst, we would like to know how sensitive such bias-corrected matching estimators are to the choice of regression model that is used. Consider a family of K regression functions $\{\mu_0^k\}_{k=1}^K$, such that, for each k, $\mu_0^k(x)$ gives a predicted outcome, $\hat{y}$ for an untreated subject with covariates x, (which may not correspond to the true relationship). Let $\mathcal{I}_0 \subseteq [N_0]$ and denote the least squares estimate of μ_0^k, using the data from the untreated subjects in $\mathcal{I}_0$ by $\hat{\mu}_0^k$. Then, the k^{th} estimator of the individual treatment effect for subject i is

$$\hat{\tau}_i^k = Y_i - Y_{M(i)} + \hat{\mu}_0^k\left(X_{M(i)}\right) - \hat{\mu}_0^k(X_i)$$

so the k^{th} estimate of the ATT is

$$\hat{\tau}^k = N_1^{-1} \sum_{i=1}^{N_1} \left[Y_i - Y_{M(i)} + \hat{\mu}_0^k \left(X_{M(i)} \right) - \hat{\mu}_0^k (X_i) \right]$$

and the average ATT across all regression functions is given by

$$\begin{aligned} \overline{\tau} &= (N_1 K)^{-1} \sum_{i,k=1}^{N_1,K} \left[Y_i - Y_{M(i)} + \hat{\mu}_0^k \left(X_{M(i)} \right) - \hat{\mu}_0^k (X_i) \right] \\ &= N_1^{-1} \sum_{i=1}^{N_1} (Y_i - Y_{M(i)}) + (N_1 K)^{-1} \sum_{i,k=1}^{N_1,K} \left[\hat{\mu}_0^k \left(X_{M(i)} \right) - \hat{\mu}_0^k (X_i) \right] \end{aligned}$$

Henceforth, we shall focus on linear regression functions of the form $\mu_0^k(x) = \alpha + x^t \beta^k$ with corresponding estimators $\hat{\mu}_0^k(x) = \hat{\alpha}^k + x^t \hat{\beta}^k$. To simplify notation we define $\Delta Z_i = Z_{M(i)} - Z_i$, $\overline{Z} = N_1^{-1} \sum_{i=1}^{N_1} Z_i$, and $\bar{\beta} = K^{-1} \sum_{k=1}^{K} \hat{\beta}^k$. Following Abadie and Imbens, we refer to ΔZ as the matching discrepancy of Z [1, 2]. Then, the above simplifies to

$$\hat{\tau}^k = -\overline{\Delta Y} + \overline{\Delta X}^t \hat{\beta}^k$$

$$\overline{\tau} = -\overline{\Delta Y} + \overline{\Delta X}^t \bar{\beta}$$

$$\hat{\tau}^k - \overline{\tau} = \overline{\Delta X}^t \left(\hat{\beta}^k - \bar{\beta} \right)$$

In what follows, we will be particularly interested in the quantity,

$$K^{-1} \sum_{i=1}^{K} \left(\hat{\tau}^k - \overline{\tau} \right)^2 = \overline{\Delta X}^t \left(\hat{\beta}^k - \bar{\beta} \right)^{\otimes 2} \overline{\Delta X}$$

the empirical variance of the estimated ATT over the family of selected regression functions, which we call Model Dependence. It provides a measure of the sensitivity of the estimated ATT to the regression function, μ_0^k, used to perform the bias correction.

Since different regression functions may have different covariates, the βs in the expression above represent a universal β that has entries for each possible covariate used by any of the regression

functions. If the regression function μ_0^k is a function of a covariate, say X_i, the corresponding entry in β^k is simply the value of the coefficient in μ_0^k; the remaining entries will be zero. Likewise, all matrices will be padded by zeros, as necessary, to ensure that they conform to β.

3.3 Model Dependence

In previous work, we considered the bias and variance of bias-corrected matching estimators for the Average effect of Treatment on the Treated (ATT), noting that, at least in the case of linear models, this bias is controlled by the expected matching discrepancy [4]. When using nearest-neighbor matching, the expected matching discrepancy will naturally be 0 for symmetric distributions in one dimension and will tend to be nonzero for skewed distributions. However, the situation becomes complicated in higher dimensions since correlations between covariates may destroy symmetry, and, even if a covariate has a symmetric, or uniform, distribution, functions of it, such as its square, will, in general, not be symmetric or uniform. Since the matching discrepancy appears prominently in the definition of Model Dependence, it has the potential to be affected by all of these phenomena.

In order to explore how the structure of the covariates affects Model Dependence, we begin by expanding its definition.

$$\begin{aligned} K^{-1}\sum_k \left(\hat{\tau}^k - \overline{\tau}\right)^2 &= K^{-1}\sum_k \left[\overline{\Delta X}^t\left(\hat{\beta}^k - \bar{\beta}\right)\right]^2 = \overline{\Delta X}^t K^{-1}\sum_k \left(\hat{\beta}^k - \bar{\beta}\right)^{\otimes 2}\overline{\Delta X} \\ &= \overline{\Delta X}^t\left[K^{-1}\sum_k \left(\hat{\beta}^k - \beta\right)^{\otimes 2} - \left(\bar{\beta} - \beta\right)^{\otimes 2}\right]\overline{\Delta X} \end{aligned}$$

We now consider the expectation conditional on X, the collection of covariates for both the treated (X^{tr}) and untreated (X^{un}), which gives,

$$\mathrm{E}\left[K^{-1}\sum_k \left(\hat{\tau}^k - \overline{\tau}\right)^2 \middle| X\right] = \overline{\Delta X}^t\left\{K^{-1}\sum_k \mathrm{E}\left[\left(\hat{\beta}^k - \beta\right)^{\otimes 2}\middle| X\right] - \mathrm{E}\left[\left(\bar{\beta} - \beta\right)^{\otimes 2}\middle| X\right]\right\}\overline{\Delta X}$$

The first term is relatively straightforward to compute, but, since $\bar{\beta} = \sum_k \hat{\beta}^k$, the second term will contain cross terms of the form $(\hat{\beta}^k - \beta)(\hat{\beta}^l - \beta)^t$ when expanded. If the $\hat{\beta}^k$s were independent, then these cross terms would vanish in expectation and the sum would simplify. However, since the $\hat{\beta}^k$s are estimated using the same data, the coefficients will be correlated and, therefore, the cross

terms will not vanish, in general.

Let m be the number of points used to fit the regression models, which will typically be N_1, the number of treated subjects, since the regression for bias correction is usually fit using only the matched controls, of which there is one for each treated subject (where an untreated subject may be used multiple times if it is matched to more than one treated subject). Then we have,

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\middle|X\right] = K^{-1}\sum_k \Delta X^t \mathrm{E}\left[\left(\hat{\beta}^k-\bar{\beta}\right)^{\otimes 2}\middle|X\right]\Delta X \\
&= m^{-1}\overline{\Delta X}^t A\left(\{\mu_0^k\}_{k=1}^K, \sigma^2_{Y|X}, F_X\right)\overline{\Delta X} + m^{-1}\overline{\Delta X}^t B\left(X, \{\mu_0^k\}_{k=1}^K, \sigma^2_{Y|X}, F_X\right)\overline{\Delta X}
\end{aligned}$$

where A is a matrix valued function of the regression functions used, the conditional variance of Y among the untreated, and the distribution of the covariates (but not their actual values) and $B(X, \sigma^2_{Y|X}) \in o_p(1)$ is a matrix valued function of the realized values of the treated $\{X_i^{tr}\}_{i=1}^{N_1}$, untreated $\{X_i^{un}\}_{i=1}^{N_0}$, their distribution, F_X, $\sigma^2_{Y|X}$, and the regression functions $\{\mu_0^k\}_{k=1}^K$. If the models are correctly specified, then it is straightforward to write down explicit expressions for A and B, but, under mild regularity conditions, they will exist, in general.

Expanding this expression, and denoting the u^{th} covariate vector by $X^{(u)}$, gives,

$$\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\middle|X\right] = m^{-1}\sum_{u,v} A_{uv}\overline{\Delta X}^{(u)}\overline{\Delta X}^{(v)} + m^{-1}\sum_{u,v} B_{uv}(X)\overline{\Delta X}^{(u)}\overline{\Delta X}^{(v)}$$

Taking the expectation, in order to remove the conditioning, gives:

$$\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right] = m^{-1}\sum_{u,v} A_{uv}\mathrm{E}\left[\overline{\Delta X}^{(u)}\overline{\Delta X}^{(v)}\right] + m^{-1}\sum_{u,v}\mathrm{E}\left[B_{uv}(X)\overline{\Delta X}^{(u)}\overline{\Delta X}^{(v)}\right]$$

Ideally, ΔX_i and ΔX_j would be independent for $i \neq j$, but this is, unfortunately, not the case, although it will be true in the correct limit and will hold approximately more generally. Clearly, if two treated subjects have similar covariate values, they are more likely to be matched with the same control, so the ΔX_is will not be independent given X^{tr}. However, if the covariates are continuous, so that each subject will have different covariate values with probability one, and if the number of untreated subjects is large compared to the number of treated subjects, so that no two are matched to the same control, the ΔX_is will be approximately conditionally independent, and, thus,

unconditionally independent since the X_i^{tr}s are IID. Therefore, in the following, we will assume that the ΔX_is are independent We can achieve this asymptotically if $\mu_{X,1}$, the distribution of X among the treated subjects is absolutely continuous with respect to $\mu_{X,0}$, the distribution of X among the untreated controls ($\mu_{X,1} << \mu_{X,0}$), $\frac{\partial \mu_{X,1}}{\partial \mu_{X,0}}$ is bounded almost surely, and $\lim_{N_1 \to \infty} N_0^{-1} N_1^{3+\epsilon} = 0$, where $\epsilon > 0$. We will also make the assumption that all the covariates used in the regression were directly matched on (for example, if the regression included both $X^{(1)}$ and $X^{(1)2}$, then we would match on both the original and derived covariate, which is not typically done). We will relax this assumption to allow the covariates to be (sufficiently smooth) functions of the matching parameters later.

Then,

$$\begin{aligned}
\mathrm{E}\left[\overline{\Delta X}^{(u)}\overline{\Delta X}^{(v)}\right] &= N_1^{-2}\sum_{i,j}\mathrm{E}\left[\Delta X_i^{(u)}\Delta X_j^{(v)}\right] \\
&= N_1^{-2}\sum_{i,j=i}\mathrm{E}\left[\Delta X_i^{(u)}\Delta X_j^{(v)}\right] + N_1^{-2}\sum_{i,j\neq i}\mathrm{E}\left[\Delta X_i^{(u)}\Delta X_j^{(v)}\right] \\
&= N_1^{-2}\sum_{i}\mathrm{E}\left[\Delta X_i^{(u)}\Delta X_i^{(v)}\right] + N_1^{-2}\sum_{i,j\neq i}\mathrm{E}\left[\Delta X_i^{(u)}\right]\mathrm{E}\left[\Delta X_j^{(v)}\right] \\
&= N_1^{-1}\mathrm{E}\left[\Delta X_i^{(u)}\Delta X_i^{(v)}\right] + \left(1 - N_1^{-1}\right)\mathrm{E}\left[\Delta X_i^{(u)}\right]\mathrm{E}\left[\Delta X_i^{(v)}\right] \\
&= N_1^{-1}\mathrm{Var}\left[\Delta X_i\right]_{uv} + \mathrm{E}\left[\Delta X_i^{(u)}\right]\mathrm{E}\left[\Delta X_i^{(v)}\right]
\end{aligned}$$

Note that, this expression consists of a variance term, which has a leading constant N_1^{-1} which goes to zero as N_1 goes to infinity and a squared bias term, which has a leading constant of 1. Thus, if the expected matching discrepancy is nonzero, increasing the number of treated subjects will reduce the sample variance of the ATT estimates only up to a point, which will be determined by the magnitude of the expected matching discrepancy.

In order to complete the computation we must select a matching function. We will begin with direct nearest neighbor matching on the covariates.

3.3.1 Direct Matching

Nearest neighbor matching, either directly on the covariates, or some suitable transformation of them, such as Mahalanobis matching, which standardizes, normalizes, and decorrelates the

covariates before matching, remains the most popular method of matching in many fields. We will focus on nearest-neighbor Euclidean matching, since most other direct matching methods can be interpreted as some transformation of the covariates followed by nearest neighbor matching using Euclidean distance.

We draw upon the following result due to Abadie and Imbens (see the appendix of [4] for an expanded version, which includes higher order terms than the original result) [1, 2].

$$\mathrm{E}[\Delta X|X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}f(X)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x}(X)\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)$$
$$\mathrm{E}[\Delta X\Delta X^t|X] = d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}f(X)^{-\frac{2}{d}}\cdot I\cdot N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)$$

where d is the number of factors on which the matching is performed and here X denotes the covariate values of the individual being matched.

Taking the expectation of the above expressions and substituting them into our previous result yields,

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right]\\
&\quad= m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\right]\sum_u A_{uu}\\
&\qquad+ m^{-1}N_0^{-\frac{4}{d}}d^{-2}\Gamma\left(\frac{d+2}{d}\right)^2\Gamma\left(\frac{d+2}{2}\right)^{\frac{4}{d}}\pi^{-2}\\
&\qquad\quad\sum_{u,v}A_{uv}\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right]\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_v}(X_i)\right]\\
&\qquad+ o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right)
\end{aligned}$$

under sufficient regularity conditions that $\mathrm{E}\left[\overline{\Delta X}^t B(X)\overline{\Delta X}\right]\in o\left(\mathrm{E}\left[\overline{\Delta X}^t\overline{\Delta X}\right]\right)$
$= o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right)$.

We summarize this as the following:

Theorem 3.1. *Given collections of treated* $\{X_i^{tr}\}_{i=1}^{N_1}$ *and untreated subjects* $\{X_i^{un}\}_{i=1}^{N_0}$ *and regression functions* $\{\mu_0^k\}_{k=1}^K$*, which are affine functions of the covariates, if each covariate is matched*

upon, the sample variance, with respect to k, of the collection of the corresponding bias-corrected matching estimators $\{\hat{\tau}^k\}_{k=1}^K$, $\mathrm{Var}_k(\hat{\tau}^k)$, *is given by:*

$$
\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_{k=1}^{K}\left(\hat{\tau}^k-\overline{\tau}\right)^2\right] \\
&= m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\right]\sum_u A_{uu} \\
&\quad + m^{-1}N_0^{-\frac{4}{d}}d^{-2}\Gamma\left(\frac{d+2}{d}\right)^2\Gamma\left(\frac{d+2}{2}\right)^{\frac{4}{d}}\pi^{-2} \\
&\quad\quad \sum_{u,v}A_{uv}\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right]\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_v}(X_i)\right] \\
&\quad + o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right) \\
&\sim m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}(2\pi e)^{-1}(\pi d)^{d^{-1}}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\right]\sum_u A_{uu} \\
&\quad + m^{-1}N_0^{-\frac{4}{d}}(2\pi e)^{-2}(\pi d)^{2d^{-1}}\sum_{u,v}A_{uv}\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right]\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_v}(X_i)\right] \\
&\quad + o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right)
\end{aligned}
$$

where the approximation follows by applying Stirling's Theorem and noting that $\Gamma\left(\frac{d+2}{d}\right)\to 1$ *as* $d\to\infty$.

Proof. Sketch as above. See appendix for details. □

Although these expressions appear unwieldy, we will see shortly that, when specialized to particular distributions, they will allow us to compare the model sensitivity of direct matching to propensity score matching, in a straightforward manner.

3.3.2 General Covariates

In the above, we explicitly assumed that all possible covariates were used in the matching (including higher order terms if they were included). However, it is often the case that only the first order terms are used to match, but that higher order, or otherwise derived, terms are used in the regression function. This complicates things somewhat.

For monomial terms, we can use the following expansion,

$$\begin{aligned}\Delta\left(\prod_l X_i^{(k_l)\alpha_l}\right) &= \prod_l X_{M(i)}^{(k_l)\alpha_l} - \prod_l X_i^{(k_l)\alpha_l} \\ &= \prod_l \left(\Delta X_i^{(k_l)} + X_i^{(k_l)}\right)^{\alpha_l} - \prod_l X_i^{(k_l)\alpha_l} \\ &= \prod_l \sum_{n_l=0}^{\alpha_l} \binom{\alpha_l}{n_l} X_i^{(k_l)(\alpha_l - n_l)} \Delta X_i^{(k_l) n_l} - \prod_l X_i^{(k_l)\alpha_l}\end{aligned}$$

However, it is also possible to derive a result for any function that admits a power series expansion of a sufficient order (at least third order for the following result).

Let N_0 and N_1 be the number of untreated and treated subjects, respectively. Let $\left\{X^{(i)}\right\}_{i=1}^d$ be a collection of covariates that will be used for matching and define a collection of derived covariates $\left\{\mathcal{X}^{(i)}\right\}_{i=1}^p$ such that $\mathcal{X}^{(i)} = f_i(X)$. We will refer to the original collection of matched covariates, as simple covariates. Then, expanding f as a power series, we get:

$$\begin{aligned}\Delta\mathcal{X}_i^{(k)} = \Delta f_k(X_i) &= f_k\left(X_{M(i)}\right) - f_k\left(X_i\right) = f_k\left(X_i + \Delta X_i\right) - f_k\left(X_i\right) \\ &= \sum_{|\alpha|\geq 0} \frac{1}{\alpha!}\partial^\alpha f(X_i)\Delta X_i^\alpha - f_k(X_i) = \sum_{|\alpha|>0} \frac{1}{\alpha!}\partial^\alpha f(X_i)\Delta X_i^\alpha\end{aligned}$$

Note that, the most common case, derived covariates that are (possibly multivariate) monomial functions of the matched covariates, is easily treated within this general framework.

Then, proceeding via a similar (but more intricate) analysis to the one above, we find that,

Corollary 3.2. *Given collections of treated* $\{X_i^{tr}\}_{i=1}^{N_1}$ *and untreated subjects* $\{X_i^{un}\}_{i=1}^{N_0}$ *and regression functions* $\{\mu_0^k\}_{k=1}^K$*, which are affine functions of possibly derived covariates* $\{\mathcal{X}^{(i)}\}_{i=1}^p$ *such that* $\mathcal{X}^{(i)} = f_i(X)$*, where* X *is the vector of simple covariates* $\{X^{(i)}\}_{i=1}^d$*, upon which matching was performed, the sample variance of the collection of the corresponding bias-corrected matching*

estimators $\{\hat{\tau}^k\}_{k=1}^K$, $\mathrm{Var}_k(\hat{\tau}^k)$, *is given by:*

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right] \\
&\quad= m^{-1}\sum_{u,v}\left(A_{uv}+o(1)\right)\left(N_1^{-1}\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\Delta\mathcal{X}_i^{(v)}\right]+\left(1-N_1^{-1}\right)\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_i^{(v)}\right]\right) \\
&\quad= m^{-1}\sum_{u,v}\left(A_{uv}+o(1)\right)\left(N_1^{-1}\mathrm{Var}\left[\Delta\mathcal{X}_i\right]_{uv}+\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_i^{(v)}\right]\right)
\end{aligned}$$

where

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\middle|X_i\right] \\
&\quad= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_u\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\frac{\partial f_k}{\partial x_u}(X_i)+f(X_i)^{-\frac{2}{d}}\frac{1}{2}\frac{\partial^2 f_k}{\partial x_u^2}(X_i)\right] \\
&\qquad+o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\Delta\mathcal{X}_i^{(l)}\middle|X_i\right] \\
&\quad= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_u f(X_i)^{-\frac{2}{d}}\frac{\partial f_k}{\partial x_u}(X_i)\frac{\partial f_l}{\partial x_u}(X_i)+o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

Proof. See appendix. □

Comparing this to the above, we see that the results have the same form as in the case of simple covariates, with the same functional dependence on m, N_1, N_0 and d and similar asymptotic behavior, although the coefficients of each term have changed.

From the above, expressions, we see that up to $o\left(m^{-1}N_0^{-\frac{2}{d}}\left(N_1^{-1}+N_0^{-\frac{2}{d}}\right)\right)$, the Model Dependence is controlled by two terms, the squared bias and the variance of the matching discrepancy, which scale like $O\left(m^{-1}N_0^{-\frac{4}{d}}\right)$ and $O\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)$, respectively. So, as the number of dimensions or treated subjects grows large, the squared bias term will dominate. In particular, as the number of treated subjects goes to infinity, the contribution of the variance term will become negligible compared to the contribution from the squared bias.

3.3.3 Propensity Score Matching

We now consider the case of propensity score matching using a known constant propensity score. Then, if $X \sim F$, for both treated and untreated subjects, $\mathcal{X}$, will also have the same distribution among the treated and the untreated, so the density of $\Delta\mathcal{X}$ is given by

$$f_{\Delta\mathcal{X}}(x) = \int f_{\mathcal{X}}(x+y) f_{\mathcal{X}}(y) dy$$

Recall that the Model Dependence is given by

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_{k}\left(\hat{\tau}^k - \overline{\tau}\right)^2\right] \\
&\quad = m^{-1}\sum_{u,v}\left(A_{uv} + o(1)\right)\left(N_1^{-1}\mathrm{Var}\left[\Delta\mathcal{X}_i\right]_{uv} + \mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_i^{(v)}\right]\right)
\end{aligned}$$

Since all matches are equally good given a fixed propensity score, it is easy to see that $\mathrm{E}[\Delta\mathcal{X}_i] = \mathrm{E}[\mathcal{X}_i] - \mathrm{E}[\mathcal{X}_{M(i)}] = 0$, so propensity score matching is unbiased and the second term vanishes so

$$\begin{aligned}
\mathrm{E}\left[K^{-1}\sum_{k}\left(\hat{\tau}^k - \overline{\tau}\right)^2\right] &= m^{-1}N_1^{-1}\sum_{u,v}\left(A_{uv} + o(1)\right)\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\Delta\mathcal{X}_i^{(v)}\right] \\
&= m^{-1}N_1^{-1}\sum_{u,v}\left(A_{uv} + o(1)\right)\mathrm{Var}\left[\Delta\mathcal{X}_i\right]_{uv} \\
&= m^{-1}N_1^{-1}\mathrm{tr}\left[\left(A + o(1)\right)\mathrm{Var}\left[\Delta\mathcal{X}_i\right]\right]
\end{aligned}$$

In contrast to the situation for direct matching, this expression always goes to zero as the number of treated subjects goes to infinity, since propensity score matching is unbiased. However, propensity score matching will also result in a larger squared matching discrepancy than matching directly on the covariates, so that, for smaller numbers of treated subjects, when the bias term is small relative to the variance term, direct matching may perform better. This can be viewed as a form of bias-variance tradeoff, in which, as usual, variance plays a larger role at small sample sizes, while bias dominates for larger ones (since typically variance declines with increasing sample size while bias remains fixed).

3.4 Comparing Direct and Propensity Score Matching

3.4.1 Simple Covariates

We now examine the relative performance of direct and propensity score matching for a number of commonly encountered distributions. To simplify the analysis, we will begin with the case in which all the covariates are independent, equally distributed, and are used in the matching process. Before we begin, we collect the distributions and variances of ΔX for propensity score matching for some important distributions, which will arise in our comparisons.

Uniform: $f(x) = \mathrm{I}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$, $f_{\Delta X}(x) = (1-|x|)\mathrm{I}_{[-1,1]}(x)$, $\mathrm{Var}\left[\Delta X\right] = \frac{1}{6}$.
Normal: $f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2\sigma^2}}$, $f_{\Delta X}(x) = (4\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{x^2}{4\sigma^2}}$, $\mathrm{Var}[\Delta X] = 2\sigma^2$
Exponential: $f(x) = \lambda e^{-\lambda x}$, $f_{\Delta X}(x) = \frac{\lambda}{2} e^{-\lambda|x|}$, $\mathrm{Var}\left[\Delta X\right] = 2\lambda^{-2}$.

To make the comparisons easier to interpret, we first simplify the expression for the Model Dependence for direct matching.

$$\mathrm{E}\left[f\left(X_i\right)^{-\frac{2}{d}}\right] = \prod_u \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right] = \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^d$$

$$\begin{aligned}
\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x_u}(X_i)\right] &= \mathrm{E}\left[f\left(X_i^{(u)}\right) \cdot \frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right) \cdot \prod_{v \neq u} f\left(X_i^{(v)}\right)^{-\frac{d+2}{d}}\right] \\
&= \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\right] \cdot \prod_{v \neq u} \mathrm{E}\left[f\left(X_i^{(v)}\right)^{-\frac{2}{d}}\right] \\
&= \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x_u}\left(X_i^{(u)}\right)\right] \cdot \mathrm{E}\left[f\left(X_i^{(u)}\right)^{-\frac{2}{d}}\right]^{d-1}
\end{aligned}$$

The assumptions also yield an additional simplification in the propensity score case: $\mathrm{tr}\left[A\mathrm{Var}\left[\Delta\mathcal{X}_i\right]\right] = \mathrm{tr}\left[A\right]\mathrm{Var}\left[\Delta\mathcal{X}_i^{(u)}\right]$, which makes comparison of direct and propensity score matching straightforward.

Recall that, in our discussion of direct matching, we obtained the following expression for the

Model Dependence:

$$
\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right]\\
&\quad=m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\pi^{-1}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\right]\sum_u A_{uu}\\
&\qquad+m^{-1}N_0^{-\frac{4}{d}}d^{-2}\Gamma\left(\frac{d+2}{d}\right)^2\Gamma\left(\frac{d+2}{2}\right)^{\frac{4}{d}}\pi^{-2}\\
&\qquad\quad\times\sum_{u,v}A_{uv}\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right]\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_v}(X_i)\right]\\
&\qquad+o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right)\\
&\quad\sim m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}(2\pi e)^{-1}(\pi d)^{d^{-1}}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\right]\sum_u A_{uu}\\
&\qquad+m^{-1}N_0^{-\frac{4}{d}}(2\pi e)^{-2}(\pi d)^{2d^{-1}}\sum_{u,v}A_{uv}\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\right]\mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_v}(X_i)\right]\\
&\qquad+o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)+o\left(m^{-1}N_0^{-\frac{4}{d}}\right)
\end{aligned}
$$

3.4.1.1 Uniform Distribution

For the Uniform Distribution, direct matching gives a Model Dependence of

$$\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right]\sim m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}(\pi d)^{d^{-1}}(2\pi e)^{-1}\Gamma\left(\frac{d+2}{d}\right)\mathrm{tr}(A)+o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right)$$

Since $\mathrm{Var}\left[\Delta X\right]=\frac{1}{6}$, using propensity score matching, the Model Dependence is

$$\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k-\overline{\tau}\right)^2\right]=\frac{1}{6}m^{-1}N_1^{-1}\mathrm{tr}(A)+o(m^{-1}N_1^{-1})$$

As we noted previously, matching directly on a uniformly distributed covariate leads to unbiased matches and, therefore, the bias term disappears for direct matching as well. The ratio of the leading terms for direct and propensity score matching is $3(\pi e)^{-1}(\pi d)^{d^{-1}}\Gamma\left(\frac{d+2}{d}\right)N_0^{-\frac{2}{d}}\sim 3(\pi e)^{-1}N_0^{-\frac{2}{d}}$, so, direct matching will always result in lower Model Dependence, but this advantage will decrease as d increases. (If we instead use the full expression instead of its asymptotic approximation, we find that direct matching is preferred if either $N_0\geq 2$ or $d>1$). This is clearly seen in the top row of

Figure 3.1.

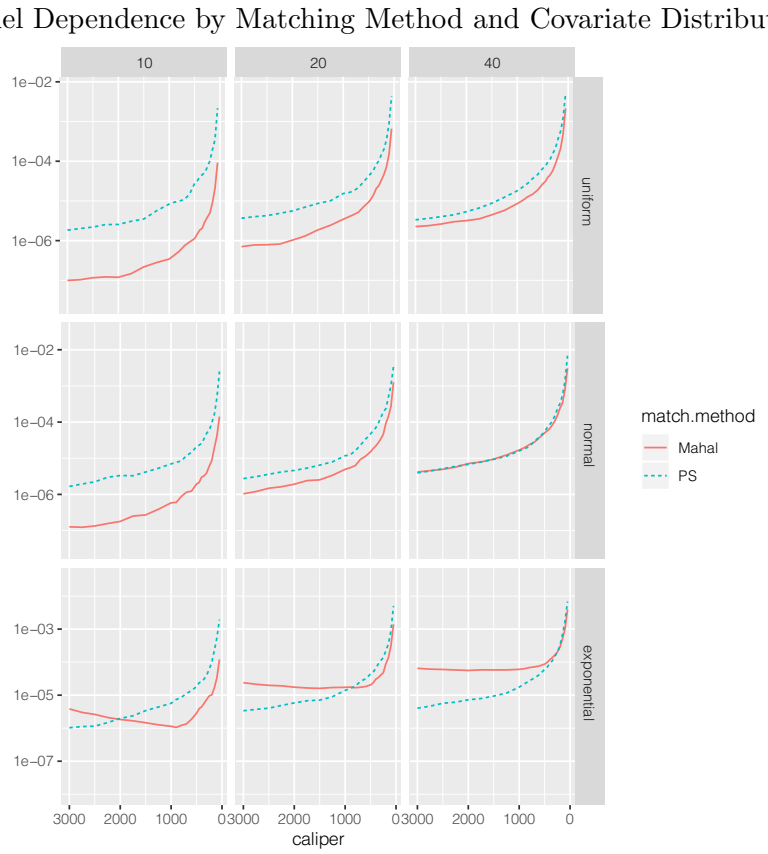


Figure 3.1: Model Dependence for direct vs. propensity score matching under the Uniform, Normal, and Exponential distributions. The numbers at the top of the figure give the number of covariates being matched, while caliper indicates the number of matched pairs retained after discarding the worst matches.

3.4.1.2 Normal Distribution

For the normal distribution, the Model Dependence using direct matching is

$$\mathrm{E}\left[K^{-1}\sum_{k}\left(\hat{\tau}^{k}-\overline{\tau}\right)^{2}\right]$$
$$\sim \sigma^{2}e^{-1}m^{-1}N_{1}^{-1}N_{0}^{-\frac{2}{d}}(\pi d)^{d^{-1}}\Gamma\left(\frac{d+2}{d}\right)\left(\frac{d}{d-2}\right)^{\frac{d}{2}}\mathrm{tr}(A)+o\left(m^{-1}N_{1}^{-1}N_{0}^{-\frac{2}{d}}\right)$$

Since $\text{Var}[\Delta X] = 2\sigma^2$, using propensity score matching results in a Model Dependence of

$$\text{E}\left[K^{-1}\sum_k \left(\hat{\tau}^k - \overline{\tau}\right)^2\right] = 2\sigma^2 m^{-1}N_1^{-1}\text{tr}(A) + o(m^{-1}N_1^{-1})$$

We again note that matching on the covariates is unbiased because the normal distribution is symmetric. The ratio of the leading terms of the Model Dependence for direct vs. propensity score matching is $2^{-1}e^{-1}(\pi d)^{d^{-1}}\Gamma\left(\frac{d+2}{d}\right)\left(\frac{d}{d-2}\right)^{\frac{d}{2}} N_0^{-\frac{2}{d}} \sim 2^{-1}N_0^{-\frac{2}{d}}$, so, direct matching will always be preferred (indeed, this is true even if we use the full expression rather than its asymptotic approximation). We can see this clearly in the middle row of Figure 3.1.

3.4.1.3 Exponential Distribution

For the exponential distribution, direct matching yields a Model Dependence of

$$\begin{aligned}
\text{E}\left[K^{-1}\sum_k \left(\hat{\tau}^k - \overline{\tau}\right)^2\right] &\sim m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}(\pi d)^{d^{-1}}(2\pi e)^{-1}\Gamma\left(\frac{d+2}{d}\right)\left(\frac{d}{d-2}\right)^{d}\lambda^{-2}\text{tr}(A) \\
&\quad + m^{-1}N_0^{-\frac{4}{d}}(\pi d)^{2d^{-1}}(2\pi e)^{-2}\Gamma\left(\frac{d+2}{d}\right)^2\left(\frac{d}{d-2}\right)^{2d}\lambda^{-2}\sum_{u,v}A_{uv} \\
&\quad + o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right) + o\left(m^{-1}N_0^{-\frac{4}{d}}\right) \\
&\sim m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}(2\pi)^{-1}e\lambda^{-1}\text{tr}(A) + m^{-1}N_0^{-\frac{4}{d}}(2\pi)^{-2}e^2\lambda^{-2}\sum_{u,v}A_{uv} \\
&\quad + o\left(m^{-1}N_1^{-1}N_0^{-\frac{2}{d}}\right) + o\left(m^{-1}N_0^{-\frac{4}{d}}\right)
\end{aligned}$$

Since $\text{Var}\left[\Delta X\right] = 2\lambda^{-2}$, using propensity score matching gives a Model Dependence of

$$\text{E}\left[K^{-1}\sum_k \left(\hat{\tau}^k - \overline{\tau}\right)^2\right] = 2\lambda^{-2}m^{-1}N_1^{-1}\text{tr}(A) + o(m^{-1}N_1^{-1})$$

Comparing the variance terms, we see that the term is smaller in the case of direct matching, as we would expect, by a factor of $(4\pi)^{-1}eN_0^{-\frac{2}{d}} \approx .22N_0^{-\frac{2}{d}}$. However, while direct matching has a nonzero squared bias term, which does not scale with N_1, propensity score matching is unbiased and so lacks such a term. Thus, for large N_1, propensity score matching will outperform direct matching, particularly in high dimensions, since, as d grows large, the effect of N_0 becomes negligible. Thus,

for large N_1 and d, matching on the propensity score will perform significantly better because of the substantial bias of direct matching. This is apparent in the bottom row of Figure 3.1.

3.4.2 Derived Covariates

We now consider the case in which derived covariates may be included in the regression functions. Since derived covariates may be functions of multiple simple covariates, as is the case for a two-way interaction term like $X^{(u)}X^{(v)}$, they may not be independent. However, we will continue to assume that the covariates that are actually matched upon are independent. As before, let $\left\{X^{(i)}\right\}_{i=1}^{d}$ be a collection of covariates, which are mutually independent, and on which matching is performed, and let $\left\{\mathcal{X}^{(i)}\right\}_{i=1}^{p}$ be a collection of derived covariates such that $\mathcal{X}^{(i)} = f_i(X)$ where $\mathcal{X}^{(i)} \neq \mathcal{X}^{(j)}$ for $i \neq j$.

In what follows, we will restrict our analysis to derived covariates which are monomials of order at most two in the matched covariates, which includes constants, simple covariates like $X^{(u)}$, two-way interaction terms of the form $X^{(u)}X^{(v)}$, and quadratic terms such as $X^{(u)2}$.

Under our model of propensity score matching, since treated and untreated subjects are matched at random due to the propensity score being identical for everyone, we will still have $\mathrm{E}\left[\Delta\mathcal{X}\right] = 0$ and $\mathrm{Var}\left[\Delta\mathcal{X}\right] = 2\mathrm{Var}\left[\mathcal{X}\right]$. Thus, $\mathrm{Var}\left[\Delta\mathcal{X}\right]_{uv} = 2\mathrm{Cov}\left[\mathcal{X}^{(u)}, \mathcal{X}^{(v)}\right] = 2\mathrm{E}\left[\mathcal{X}^{(u)}\mathcal{X}^{(v)}\right] - 2\mathrm{E}\left[\mathcal{X}^{(u)}\right]\mathrm{E}\left[\mathcal{X}^{(v)}\right]$. If both $\mathcal{X}^{(u)}$ and $\mathcal{X}^{(v)}$ are quadratic, then, since $\mathcal{X}^{(u)} \neq \mathcal{X}^{(v)}$ for $u \neq v$, they must be functions of two different matched covariates, which are, by assumption, independent, so $\mathcal{X}^{(u)}$ and $\mathcal{X}^{(v)}$, must be independent as well and $\mathrm{Cov}\left[\mathcal{X}^{(u)}, \mathcal{X}^{(v)}\right] = 0$. If $\mathcal{X}^{(u)} = X^{(s)2}$ and $\mathcal{X}^{(v)} = X^{(s)}X^{(t)}$ then $\mathrm{Cov}\left[\mathcal{X}^{(u)}, \mathcal{X}^{(v)}\right] = \mathrm{E}\left[X^{(t)}\right]\left(\mathrm{E}\left[X^{(s)3}\right] - \mathrm{E}\left[X^{(s)2}\right]\mathrm{E}\left[X^{(s)}\right]\right)$ and if $\mathcal{X}^{(u)} = X^{(s)2}$ and $\mathcal{X}^{(v)} = X^{(s)}$ then $\mathrm{Cov}\left[\mathcal{X}^{(u)}, \mathcal{X}^{(v)}\right] = \mathrm{E}\left[X^{(s)3}\right] - \mathrm{E}\left[X^{(s)2}\right]\mathrm{E}\left[X^{(s)}\right]$, both of which will be zero if $X^{(s)}$ is symmetric around zero. Clearly, if $\mathcal{X}^{(u)}$ and $\mathcal{X}^{(v)}$ have no terms in common, then they will also have zero covariance. Thus, if the matched covariates are symmetric about zero, then, up to second order, the off diagonal terms of $\mathrm{Var}\left[\Delta\mathcal{X}\right]$ vanish.

For direct matching we have,

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\Delta\mathcal{X}_i^{(l)}\middle|X_i\right] \\
&\quad = N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_u f(X_i)^{-\frac{2}{d}}\frac{\partial f_k}{\partial x_u}(X_i)\frac{\partial f_l}{\partial x_u}(X_i) + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

so, if the matched covariates are symmetrically distributed, then the off diagonal terms of $\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\Delta\mathcal{X}_i^{(l)}\right]$ will be $o\left(N_0^{-\frac{2}{d}}\right)$, similar to the case of propensity score matching. However, since

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\middle|X_i\right] \\
&\quad = N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_u\left[f(X_i)^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_u}(X_i)\frac{\partial f_k}{\partial x_u}(X_i) + f(X_i)^{-\frac{2}{d}}\frac{1}{2}\frac{\partial^2 f_k}{\partial x_u^2}(X_i)\right] \\
&\qquad + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

even if the matched covariates are symmetrically distributed, the expected matching discrepancy, $\mathrm{E}\left[\Delta\mathcal{X}_i^{(k)}\right]$, may be $\Omega\left(N_0^{-\frac{2}{d}}\right)$ since, for quadratic terms (although not for two-way interactions or lower order terms), the final term in brackets will be 1, and, thus, the bias will be nonzero, unless this is exactly cancelled by the first term in the summand.

3.4.2.1 Uniform Distribution

Since the uniform distribution on $\left[-\frac{1}{2},\frac{1}{2}\right]$ is symmetric around zero, from the discussion above, we expect that if we include only two-way interactions and lower-order terms, the behavior of direct versus propensity score matching will be similar to the case of simple covariates. This is clearly seen in the top two rows of Figure 3.2, in which direct matching always outperforms propensity score matching. However, as soon as we include quadratic terms, the behavior changes, and, as N_1 and d become large, propensity score matching performs better, similarly to the exponential case for simple covariates (Figure 3.2, bottom row).

For direct matching we have,

$$\begin{aligned}
\mathrm{E}\left[\Delta X_i^{(k)2}\right] &= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim (2\pi e)^{-1}N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \approx \frac{1}{17}N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

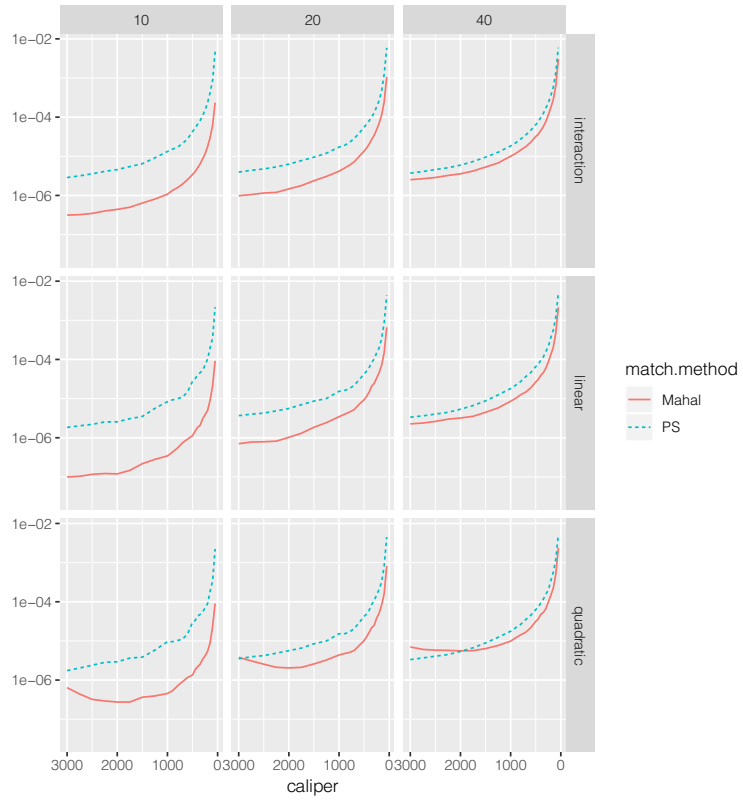


Figure 3.2: Model Dependence for the Uniform Distribution under direct vs. propensity score matching with only simple covariates, two-way interactions, and quadratic terms. The numbers at the top of the figure give the number of covariates being matched, while caliper indicates the number of matched pairs retained after discarding the worst matches.

$$
\begin{aligned}
&\mathrm{E}\left[\left(\Delta\left(X_i^{(k)}X_i^{(l)}\right)\right)^2\right] \\
&= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\left(X_i^{(k)2}+X_i^{(l)2}\right)\right]+o\left(N_0^{-\frac{2}{d}}\right) \\
&= N_0^{-\frac{2}{d}}\cdot\frac{1}{6}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}+o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim (12\pi e)^{-1}N_0^{-\frac{2}{d}}+o\left(N_0^{-\frac{2}{d}}\right)\approx\frac{1}{103}N_0^{-\frac{2}{d}}+o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}
$$

$$\begin{aligned}
\mathrm{E}\left[\left(\Delta\left(X_i^{(k)2}\right)\right)^2\right] &= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\cdot 4\ X_i^{(k)2}\right] + o\left(N_0^{-\frac{2}{d}}\right) \\
&= N_0^{-\frac{2}{d}}\cdot\frac{1}{3}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim (6\pi e)^{-1}N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right) \approx \frac{1}{51}N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

For propensity score matching, we have,

$$\mathrm{E}\left[\left(\Delta\left(X_i^{(k)}X_i^{(l)}\right)\right)^2\right] = 2\mathrm{E}\left[X_i^{(k)2}X_i^{(l)2}\right] = \frac{1}{72}$$

$$\begin{aligned}
\mathrm{E}\left[\left(\Delta\left(X_i^{(k)2}\right)\right)^2\right] &= \mathrm{Var}\left[\left(\Delta\left(X_i^{(k)2}\right)\right)\right] = 2\left(\mathrm{E}\left[\left(X_i^{(k)4}\right)\right] - \mathrm{E}\left[\left(X_i^{(k)2}\right)\right]^2\right) \\
&= 2\left(\frac{1}{80} - \frac{1}{144}\right) = \frac{1}{90}
\end{aligned}$$

From the above, and the expression for Model Dependence,

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_k\left(\hat{\tau}^k - \overline{\tau}\right)^2\right] \\
&= m^{-1}\sum_{u,v}\left(A_{uv} + o(1)\right)\left(N_1^{-1}\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\Delta\mathcal{X}_i^{(v)}\right] + \left(1 - N_1^{-1}\right)\mathrm{E}\left[\Delta\mathcal{X}_i^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_i^{(v)}\right]\right)
\end{aligned}$$

we see that, while the contributions to the first term will be similar for both direct and propensity score matching, except for the additional factor of $N_0^{-\frac{2}{d}}$, the second, squared bias term will be zero for propensity score matching, while it is $\Omega\left(N_0^{-\frac{4}{d}}\right)$ for direct matching. Therefore, as N_1 grows, and the first term becomes small, Model Dependence for direct matching will be dominated by the second term, particularly in higher dimensions, so that propensity score matching will perform better for larger N_1, particularly when d is large.

3.4.2.2 Normal Distribution

As with the uniform distribution, the normal distribution is symmetric around 0 so that the relative performance of direct and propensity score matching will be similar when including either only the matched covariates or with the addition of two-way interactions (Figure 3.3, top two rows).

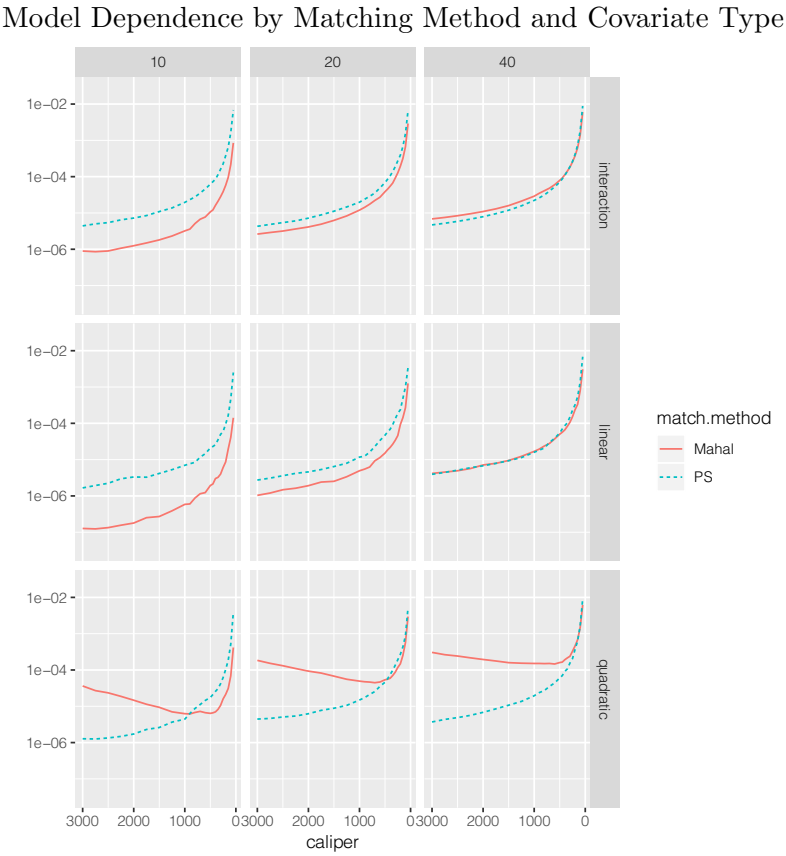


Figure 3.3: Model Dependence for the Normal Distribution under direct vs. propensity score matching with only simple covariates, two-way interactions, and quadratic terms. The numbers at the top of the figure give the number of covariates being matched, while caliper indicates the number of matched pairs retained after discarding the worst matches.

For direct matching we have,

$$
\begin{aligned}
&\mathrm{E}\left[\Delta X_i^{(k)2}\right] \\
&= N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \mathrm{E}\left[2X_i^{(k)} f(X_i)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x_k}(X_i) + f(X_i)^{-\frac{2}{d}}\right] + o\left(N_0^{-\frac{2}{d}}\right) \\
&= -2\sigma^2 N_0^{-\frac{2}{d}} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \cdot \frac{d+2}{d-2} \left(\frac{d}{d-2}\right)^{\frac{d}{2}} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim -\sigma^2 N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\left[\left(\Delta\left(X_i^{(k)}X_i^{(l)}\right)\right)^2\right] \\
&= N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\mathrm{E}\left[f(X_i)^{-\frac{2}{d}}\left(X_i^{(k)2}+X_i^{(l)2}\right)\right]+o\left(N_0^{-\frac{2}{d}}\right) \\
&= 4\sigma^4 N_0^{-\frac{2}{d}}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\frac{d}{d-2}\left(\frac{d}{d-2}\right)^{\frac{d}{2}}+o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim 2\sigma^4 N_0^{-\frac{2}{d}}+o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\left[\left(\Delta X_i^{(k)2}\right)^2\right] \\
&= 4N_0^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\mathrm{E}\left[X_i^{(k)2}f(X_i)^{-\frac{2}{d}}\right]+o\left(N_0^{-\frac{2}{d}}\right) \\
&= 8\sigma^4 N_0^{-\frac{2}{d}}d^{-1}\Gamma\left(\frac{d+2}{d}\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\frac{d}{d-2}\left(\frac{d}{d-2}\right)^{\frac{d}{2}}+o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim 4\sigma^4 N_0^{-\frac{2}{d}}+o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}
$$

For propensity score matching, we obtain,

$$
\mathrm{E}\left[\left(\Delta\left(X_i^{(k)}X_i^{(l)}\right)\right)^2\right] = 2\mathrm{E}\left[X_i^{(k)2}X_i^{(l)2}\right] = 2\sigma^4
$$

$$
\begin{aligned}
&\mathrm{E}\left[\left(\Delta\left(X_i^{(k)2}\right)\right)^2\right] = \mathrm{Var}\left[\left(\Delta\left(X_i^{(k)2}\right)\right)\right] = 2\left(\mathrm{E}\left[\left(X_i^{(k)4}\right)\right]-\mathrm{E}\left[\left(X_i^{(k)2}\right)\right]^2\right) \\
&= 2\left(3\sigma^4-\sigma^4\right) = 4\sigma^4
\end{aligned}
$$

It is interesting to note that, the variance terms for direct and propensity score matching differ only by a factor of $N_0^{-\frac{2}{d}}$. Additionally, as was the case with the uniform distribution, under propensity score matching, the bias term is zero, while this is not the case for direct matching with quadratic terms. Thus, as N_1 increases, propensity score matching has a lower Model Dependence, particularly when d is large, as seen in the bottom row of Figure 3.3.

3.4.2.3 Exponential Distribution

The exponential distribution differs from the uniform and normal distributions in several major ways. First, it does not have mean zero, so the variance term in the expression for the Model Dependence will have off diagonal terms. Second, since it is not symmetric about the origin, the expected matching discrepancy will have nonzero mean, even when using simple covariates. Due to the complex form of the variance of the matching discrepancy, which arises because the off diagonal terms are nonzero due to the asymmetry of the exponential distribution, we will focus our attention on the expectation of the matching discrepancy and its effects on the Model Dependence.

For direct matching,

$$\begin{aligned}
&\mathrm{E}\left[\Delta X_i^{(k)} X_i^{(l)}\right] \\
&= N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \mathrm{E}\left[-\lambda f(X_i)^{-\frac{2}{d}} \left(X_i^{(k)} + X_i^{(l)}\right)\right] + o\left(N_0^{-\frac{2}{d}}\right) \\
&= -2\lambda^{-2} N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \left(\frac{d}{d-2}\right)^{d+1} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim -\pi^{-1} e \lambda^{-2} N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta X_i^{(k)2}\right] \\
&= N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \mathrm{E}\left[-2\lambda X^{(k)} f(X_i)^{-\frac{2}{d}} + f(X_i)^{-\frac{2}{d}}\right] + o\left(N_0^{-\frac{2}{d}}\right) \\
&= -\lambda^{-2} N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \cdot \frac{d+2}{d-2} \cdot \left(\frac{d}{d-2}\right)^{d} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim -\frac{1}{2} \pi^{-1} e \lambda^{-2} N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

For comparison,

$$\begin{aligned}
\mathrm{E}\left[\Delta X_i^{(k)}\right] &= N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \mathrm{E}\left[f(X_i)^{-\frac{d+2}{d}} \frac{\partial f}{\partial x_k}(X_i)\right] + o\left(N_0^{-\frac{2}{d}}\right) \\
&= -\lambda^{-1} N_0^{-\frac{2}{d}} \pi^{-1} d^{-1} \Gamma\left(\frac{d+2}{d}\right) \Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \left(\frac{d}{d-2}\right)^{d} + o\left(N_0^{-\frac{2}{d}}\right) \\
&\sim -\frac{1}{2} \pi^{-1} e \lambda^{-1} N_0^{-\frac{2}{d}} + o\left(N_0^{-\frac{2}{d}}\right)
\end{aligned}$$

Thus, for the exponential distribution, all covariates of order at most two have nonzero expected matching discrepancies, all of which are of similar magnitude and identical order, differing by at most a factor of λ. Therefore, we expect the relative performance of direct and propensity score matching to be similar in all cases, so that, for large N_1, propensity score matching will outperform direct matching, particularly for larger values of d. This is exactly what we see in Figure 3.4.

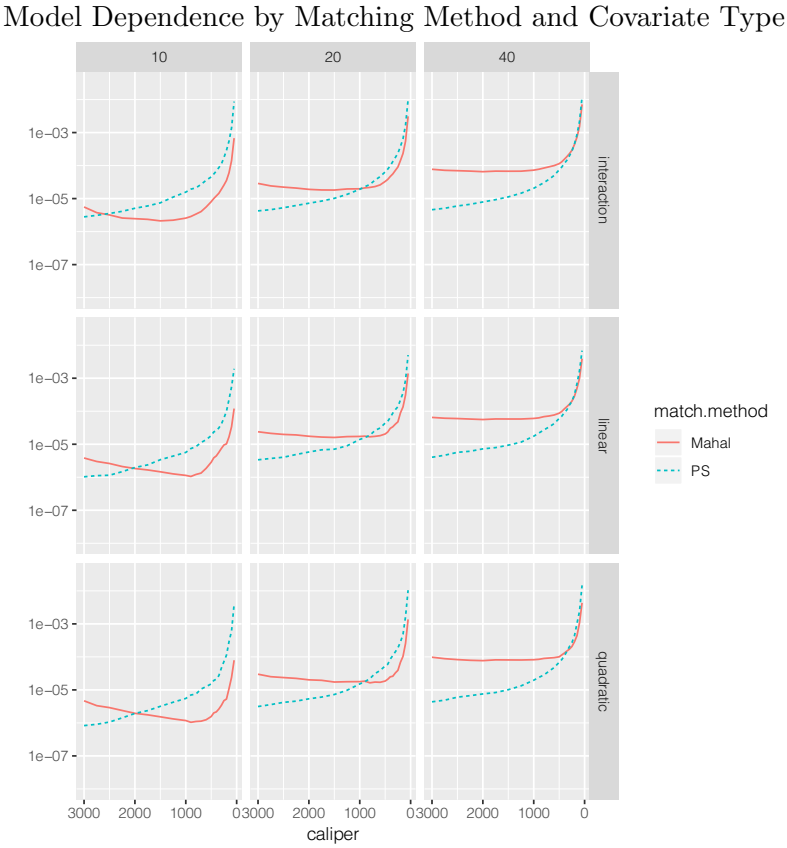


Figure 3.4: Model Dependence for the Exponential Distribution under direct vs. propensity score matching with only simple covariates, two-way interactions, and quadratic terms. The numbers at the top of the figure give the number of covariates being matched, while caliper indicates the number of matched pairs retained after discarding the worst matches.

3.5 Pruning Matches

Thus far, after matching, we have used all of the treated subjects that were successfully matched in the analysis. However, typical practice is to only retain those matched pairs that are of a specified level of quality. One common approach is to rank the matches by the magnitude of their matching

discrepancies and then keep only the K best. Another commonly used criterion is to only retain matched pairs with matching discrepancies that are less than some fixed standard, which is typically referred to as caliper matching. The idea behind throwing away badly matched pairs is that poor matches may lead to bias. However, this also reduces the number of pairs used in estimation, which will increase the variance of the estimated treatment effect. Thus, the process of pruning can be seen as trying to optimize a bias-variance tradeoff. However, such a tradeoff may not exist in all cases and, under some circumstances, all pruning is deleterious.

In some cases, removing matched pairs only increases the Model Dependence, while in other scenarios, pruning poor matches initially decreases and then increases Model Dependence as more and more matched pairs are removed. This phenomenon is closely related to whether direct or propensity score matching performs better than propensity score matching for a given distribution/regression model. Recall that, if the expected matching discrepancy was zero for all covariates, then direct matching was always preferred over propensity score matching because direct matching resulted in the matching discrepancy having lower variance. However, in cases in which the expected bias of the matching discrepancy is nonzero, as N_1 increases, propensity score matching performs better because the variance term in the Model Dependence falls off inversely with N_1, while the bias term does not. Exactly the same phenomenon arises here, although from a slightly different source. Here, pruning bad matches reduces the magnitude of the bias term, while also increasing the size of the variance term by reducing the effective N_1. In the appendix, we present a detailed derivation of this tradeoff using a nonstandard version of the Laplace expansion. If we let $r_{\max}$ be the maximum allowed matching discrepancy, then this alters the expression for the Model Dependence by replacing the standard gamma functions with generalized incomplete gamma functions that depend on $\rho(r)$, an increasing function of r whose form is computed in the appendix. Additionally, this makes N_1 a function of $r_{\max}$, since not all treated subjects may have

good matches, so that some will not be included. The modified expression is,

$$\begin{aligned}
&\mathrm{E}\left[K^{-1}\sum_{k}\left(\hat{\tau}^{k}-\overline{\tau}\right)^{2}\right] \\
&\quad = m^{-1}\sum_{u,v}\left(A_{uv}+o(1)\right)\left(N_{1}^{-1}\left(r_{\max}\right)\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(u)}\Delta\mathcal{X}_{i}^{(v)}\right]+\left(1-N_{1}^{-1}\left(r_{\max}\right)\right)\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(v)}\right]\right) \\
&\quad = m^{-1}\sum_{u,v}\left(A_{uv}+o(1)\right)\left(N_{1}^{-1}\left(r_{\max}\right)\mathrm{Var}\left[\Delta\mathcal{X}_{i}\right]_{uv}+\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(u)}\right]\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(v)}\right]\right)
\end{aligned}$$

where

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(k)}\middle|X_{i}\right] \\
&\quad = N_{0}^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d},0,\rho\left(r_{\max}\right)\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}} \\
&\qquad \times\sum_{u}\left[f(X_{i})^{-\frac{d+2}{d}}\frac{\partial f}{\partial x_{u}}(X_{i})\frac{\partial f_{k}}{\partial x_{u}}(X_{i})+f(X_{i})^{-\frac{2}{d}}\frac{1}{2}\frac{\partial^{2}f_{k}}{\partial x_{u}^{2}}(X_{i})\right] \\
&\qquad + o\left(N_{0}^{-\frac{2}{d}}\right)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta\mathcal{X}_{i}^{(k)}\Delta\mathcal{X}_{i}^{(l)}\middle|X_{i}\right] \\
&\quad = N_{0}^{-\frac{2}{d}}\pi^{-1}d^{-1}\Gamma\left(\frac{d+2}{d},0,\rho\left(r_{\max}\right)\right)\Gamma\left(\frac{d+2}{2}\right)^{\frac{2}{d}}\sum_{u}f(X_{i})^{-\frac{2}{d}}\frac{\partial f_{k}}{\partial x_{u}}(X_{i})\frac{\partial f_{l}}{\partial x_{u}}(X_{i})+o\left(N_{0}^{-\frac{2}{d}}\right)
\end{aligned}$$

Since the generalized incomplete gamma function is increasing in its final argument, as $r_{\max}$ increases, the bias term will increase in magnitude. The effect on the variance term depends on the number of treated subjects available, as well as the current value of $r_{\max}$. $N_{1}\left(r_{\max}\right)$ is increasing in r, so whether the magnitude of the variance term increases or decreases depends on the relative rates of growth of $N_{1}^{-1}\left(r_{\max}\right)$ and $\Gamma\left(\frac{d+2}{d},0,\rho\left(r_{\max}\right)\right)$. Often, the combined effects of these changes will be for the Model Dependence to initially decrease as matched pairs are dropped until it reaches an optimal value of $r_{\max}$, at which point the Model Dependence will increase ever after. However, in order for this tradeoff to occur, the bias term must be nonzero.

3.6 Regression After Matching and Bias-Corrected Matching Estimators

Above, we explored how bias-corrected matching estimators of the ATT may depend on the choice of bias-correction function and how the degree of such sensitivity, the Model Dependence, depends on the distribution of the covariates. However, King and Nielsen used a different technique for estimating the ATT [3]. They first matched each treated subject to an untreated subject and then fit a regression model to the collection of treated subjects and matched controls. As we will show below, this is, in fact, a form of bias-corrected matching estimator, so the above discussion applies equally well to this procedure as it does to explicitly constructed bias-corrected matching estimators.

To be explicit, the procedure works as follows. First, match each treated subject to a single untreated control. Next, fit a regression of the form $Y_i = \alpha + \tau A + X^t\beta + \epsilon_i$ to the collection of cases and matched controls. Finally, use $\hat{\tau}$ as an estimate of the ATT. Somewhat surprisingly, this estimator can also be written as a bias-corrected matching estimator, by manipulating the normal equations of the regression. If we match first and then fit the regression model $Y = \alpha+\tau A+X^t\beta+\epsilon$, using both the N_1 treated subjects and their matched controls, the normal equations give,

$$
\begin{aligned}
0 &= \sum_{i=1}^{N_1} \left(Y_i - \alpha - \tau A_i - X_i^t\beta\right) + \sum_{i=1}^{N_1} \left(Y_{M(i)} - \alpha - \tau A_{M(i)} - X_{M(i)}^t\beta\right) \\
0 &= \sum_{i=1}^{N_1} A_i \left(Y_i - \alpha - \tau A_i - X_i^t\beta\right) + \sum_{i=1}^{N_1} A_{M(i)} \left(Y_{M(i)} - \alpha - \tau A_{M(i)} - X_{M(i)}^t\beta\right) \\
&= \sum_{i=1}^{N_1} \left(Y_i - \alpha - \tau A_i - X_i^t\beta\right) \\
0 &= \sum_{i=1}^{N_1} X_i \left(Y_i - \alpha - \tau A_i - X_i^t\beta\right) + \sum_{i=1}^{N_1} X_{M(i)} \left(Y_{M(i)} - \alpha - \tau A_{M(i)} - X_{M(i)}^t\beta\right)
\end{aligned}
$$

where the third equality is due to the fact that controls have $A = 0$.

After some manipulation (detailed in the appendix), the normal equations yield,

$$\begin{aligned}
\hat{\alpha} &= \overline{Y}_M - \overline{X}_M^t \hat{\beta} \\
\hat{\tau} &= \overline{Y}_1 - \hat{\alpha} - \overline{X}_1^t \hat{\beta} = \left(\overline{Y}_1 - \overline{Y}_M\right) + \left(\overline{X}_M - \overline{X}_1\right)^t \hat{\beta} = -\overline{\Delta Y} + \overline{\Delta X}^t \hat{\beta} \\
\hat{\beta} &= \left(\sum_{i=1}^{N_1} \left[\left(X_i - \overline{X}_1\right)^{\otimes 2} + \left(X_{M(i)} - \overline{X}_M\right)^{\otimes 2} \right] \right)^{-1} \\
&\quad \times \sum_{i=1}^{N_1} \left[\left(X_i - \overline{X}_1\right) \left(Y_i - \overline{Y}_1\right) + \left(X_{M(i)} - \overline{X}_M\right) \left(Y_{M(i)} - \overline{Y}_M\right) \right]
\end{aligned}$$

$\hat{\tau}$ once again takes the form $\hat{\tau} = -\overline{\Delta Y} + \overline{\Delta X}^t \hat{\beta}$ of a bias-corrected matching estimator. If we interpret τ as the average treatment effect, the result for $\hat{\tau}$ holds even if the effect of treatment is allowed to vary based on the covariates, X, or the identity of the individual (see the appendix for a detailed derivation). Thus, the only difference between this approach and the standard bias corrected estimator is the value of β, and the data which is used to estimate it.

3.7 Conclusion

In this work we compared the performance of matching directly on covariates vs. on the propensity score, using the Model Dependence of the bias-corrected matching estimator of the ATT. Which method is preferred depends on the number and distribution of the matched covariates, as well as the form of any derived covariates used in the regression function, with direct matching proving superior for symmetric covariates, which result in zero expected matching discrepancy, and propensity score matching preferred when matching on covariates with asymmetric distributions. Which matching method performs better is intimately connected to a form of bias-variance tradeoff: while direct matching leads to lower variances for the matching discrepancy than propensity score matching, the fact that propensity score matching is always unbiased means that Model Dependence may be lower when using propensity score matching. Since the variance contribution to Model Dependence falls off with the inverse number of treated subjects, while the bias term does not, when many treated subjects are available, propensity score matching will generally be preferred, especially when matching on a large number of covariates, in which case, reducing the bias of direct matching is very difficult, since it scales like $N_0^{-\frac{2}{d}}$, and, so, will tend to be far from zero.

Further, we showed that a similar bias-variance tradeoff guides the question of whether poor matches should be pruned in order to reduce Model Dependence. Here, if the bias of matching is zero, then pruning is never beneficial, but, if the expected matching discrepancy is nonzero, then pruning the worst matches, in order to reduce the bias, is initially beneficial, and will decrease Model Dependence. However, this is eventually outweighed by the increase in variance, as the number of matched pairs used in estimation decreases, leading to Model Dependence actually increasing when the pruning becomes excessive.

We also showed that, regression after matching leads to a form of bias-corrected matching estimator. This allows us to connect our results on bias-corrected matching estimators, which are not commonly used in the applied literature, to regression after matching, which is the dominant methodology in many fields in the social sciences. Thus, our conclusions about how Model Dependence is affected by the distribution of the covariates carry over to this important setting, as well.

Overall, we have demonstrated that the optimal approach to matching will depend on the nature and number of the covariates on which matching occurs, with direct matching being preferred for symmetrically distributed covariates, or when the number of covariates or treated subjects is small. However, as the number of treated subjects increases, in cases in which direct matching gives a nonzero expected matching discrepancy, propensity score matching will be preferred, especially when the number of matched covariates is large. Thus, it is particularly important for the analyst to think carefully about the structure of the problem before selecting an approach to matching, since neither direct nor propensity score matching dominates the other.

Bibliography

[1] Alberto Abadie and Guido Imbens. "Bias-Corrected Matching Estimators for Average Treatment Effects". *Journal of the American Statistical Association* 29.1 (2011), pp. 1–11.

[2] Alberto Abadie and Guido Imbens. "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica* 74.1 (2006), pp. 237–267.

[3] Gary King and Richard Nielsen. "Why Propensity Scores Should Not Be Used for Matching". *Political Analysis* 27.4 (2019).

[4] Thomas Kolokotrones et al. "Bias and Variance of Matching Estimators" (2020).

[5] Prasanta Chandra Mahalanobis. "On the generalised distance in statistics". *Proceedings of the National Institute of Sciences of India* 2.1 (1936), pp. 49–55.

[6] Paul Rosenbaum and Donald Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika* 70.1 (1983), pp. 41–55.

[7] Donald Rubin. "Matching to Remove Bias in Observational Studies". *Biometrics* 29 (1973), pp. 159–183.

[8] Donald Rubin. "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies". *Biometrics* 29 (1973), pp. 185–203.

Chapter 4

Mendelian Randomization and Egger Regression

Tom Kolokotrones, Rajarshi Mukherjee, Qingyuan Zhao, and James Robins

4.1 Introduction

Mendelian Randomization, the practice of using genetic markers as instrumental variables to elucidate the causal effect of putative risk factors on health outcomes, has become a popular tool in modern epidemiology [5]. The method was first proposed by Katan in 1986, who wished to determine whether the association between low serum cholesterol and cancer risk was causal, in particular, whether low cholesterol led to an elevated risk [4]. Since some ApoE alleles are associated with elevated serum cholesterol, and which alleles a subject possesses are determined at conception, Katan proposed that by comparing the cancer risk in bearers of high vs. low cholesterol alleles, one could determine whether or not low cholesterol causally increases cancer risk. However, he never actually performed the analysis.

The first use of the term "Mendelian Randomization" was due to Gray and Wheatley in 1991 who used the term to refer to a somewhat different type of instrumental variable [3]. The two wished to compare the efficacy of allogenic Bone Marrow Transplant (BMT) vs. chemotherapy, in the treatment of leukemia, but, because they were unable to randomize treatment since allogenic BMT, was, by that time, the standard of care, if available, they used the presence of an HLA (Human Leukocyte Antigen) matched sibling (who could provide bone marrow) as their instrument. Since that time, a number of other treatment studies have used the same instrument. The use of Mendelian Randomization in modern genetic epidemiology is now widespread as a means for estimating the causal effect of a given risk factor on an outcome of interest [5].

However, despite the popularity of Mendelian Randomization, the assumptions that are required for its valid application are stringent and often violated. As is the case for other instruments, the locus of interest, Z, must be associated with the outcome, Y, only through the putative risk factor, X, the so-called exclusion restriction in the terminology of Instrumental Variables (IV). Unfortunately, this assumption is frequently violated, since genes often have pleiotropic effects, so

the chosen locus may also affect Y via other pathways, in which case the causal effect of X on Y is not identifiable without further assumptions.

In 2015, Bowden, Davey-Smith, and Burgess introduced the use of Egger Regression, a technique previously developed by Egger, Davey-Smith, and Minder in 1997 for meta-analysis, which allows the use of Mendelian Randomization, even in the presence of pleiotropy under an additional assumption, the so-called Instrument Strength Independent of Direct Effect, or InSIDE, condition [1, 2]. However, while they sketched informal arguments justifying the use of Egger Regression, the estimator has never been formally analyzed or even shown to be consistent. In the following work, we provide an analysis of its use under a variety of IV assumptions, ranging from few strong to many weak instruments.

4.2 Basics

Let Y be a continuous outcome, X be a continuous predictor, U be a collection of unmeasured confounders, Z be a collection of p (possibly broken) instruments (which satisfy $Z \perp\!\!\!\perp U$, but not necessarily $Z \perp\!\!\!\perp Y|U, X$), ϵ_x and ϵ_y be independent error terms ($\epsilon \perp\!\!\!\perp Z, U$, $\epsilon_x \perp\!\!\!\perp \epsilon_y$), and n be the number of subjects. Let $\mathrm{E}[\epsilon_x] = \mathrm{E}[\epsilon_y] = 0$, $\mathrm{Var}[\epsilon_x] = \sigma_x^2$, $\mathrm{Var}[\epsilon_y] = \sigma_y^2$, $\mathrm{E}[U] = 0$, $\mathrm{Var}[U] = \Sigma_u$. (In the Mendelian Randomization setting, the Z are trinary and represent the number of minor alleles present at the selected locus.) We wish to estimate the causal effect of X on Y. We observe n IID copies of (Y, X, Z), which satisfy the following structural equation model,

$$\begin{aligned} Y &= \beta X + \gamma_y^t Z + \delta_y^t U + \epsilon_y \\ X &= \gamma_x^t Z + \delta_x^t U + \epsilon_x \end{aligned}$$

The corresponding reduced form expression for Y is,

$$\begin{aligned} Y &= \beta X + \gamma_y^t Z + \delta_y^t U + \epsilon_y \\ &= \beta\left(\gamma_x^t Z + \delta_x^t U + \epsilon_x\right) + \gamma_y^t Z + \delta_y^t U + \epsilon_y \\ &= \left(\beta\gamma_x^t + \gamma_y^t\right) Z + \left(\beta\delta_x^t + \delta_y^t\right) U + (\beta\epsilon_x + \epsilon_y) \\ &= \Gamma^t Z + \Delta^t U + \epsilon_r \end{aligned}$$

where $\Gamma = \beta\gamma_x + \gamma_y$, $\Delta = \beta\delta_x + \delta_y$, and $\epsilon_r = \beta\epsilon_x + \epsilon_y$. In order to simplify some expressions, we also define $\epsilon'_x = \delta_x^t U + \epsilon_x$, $\epsilon'_y = \delta_y^t U + \epsilon_y$ and $\epsilon'_r = \beta\epsilon'_x + \epsilon'_y = \Delta^t U + \epsilon_r$.

If $\gamma_y = 0$, so that $Z \perp\!\!\!\perp Y|U, X$, then these would be the standard structural equations for an instrumental variables model, and the causal effect could be estimated by a variety of methods including Two-Stage Least Squares (TSLS). However, if $\gamma_y \neq 0$, then $Z \not\perp\!\!\!\perp Y|U, X$ and the exclusion restriction does not hold. It is well known that, without additional assumptions, this model is not identified. Before we explore solutions to this problem, it is useful to examine the bias of the standard TSLS estimator in the presence of pleiotropy.

4.3 Bias of Two Stage Least Squares with Pleiotropy

Two Stage Least Squares proceeds in two steps. First, one uses the first stage equation $X = \gamma_x^t Z + \left(\delta_x^t U + \epsilon_x\right)$ to estimate $\hat{\gamma}_x$, using least squares, and then uses that estimate to compute the predicted value of X given Z, $\hat{X}(Z) = \hat{\gamma}_x^t Z = \mathrm{P}_Z X$, the projection of X onto the subspace spanned by Z. Next, one substitutes $\hat{X}$ into the second stage equation to give $Y = \beta\hat{X}(Z) + \gamma_y^t Z + \left(\delta_y^t U + \epsilon_y\right)$, and then estimates β using least squares so that $\hat{\beta} = \left(\hat{\mathbb{X}}^t\hat{\mathbb{X}}\right)^{-1}\hat{\mathbb{X}}^t\mathbb{Y}$, where $\mathbb{X}, \mathbb{Y}$ are the design matrix and vector of outcomes, respectively, and $\mathbb{Z}$ is the corresponding matrix of instruments. With some manipulation (see appendix) this yields:

Lemma 4.1. *If γ_x, γ_y are fixed, length p, vectors, then the Two Stage Least Squares estimator of β, $\hat{\beta} \xrightarrow{P} \beta + \left(\gamma_x^t \mathrm{E}\left[ZZ^t\right]\gamma_x\right)^{-1}\gamma_x^t \mathrm{E}\left[ZZ^t\right]\gamma_y$.*

Proof.

$$
\begin{aligned}
\hat{\beta} &= \left(\hat{\mathbb{X}}^t\hat{\mathbb{X}}\right)^{-1}\hat{\mathbb{X}}^t\mathbb{Y} \\
&= \beta + \left(n^{-1}\mathbb{X}^t\mathbb{Z}\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}n^{-1}\mathbb{Z}^t\mathbb{X}\right)^{-1} \\
&\qquad \times \left(n^{-1}\left(\mathbb{Z}\gamma_x + \mathbb{U}\delta_x + \epsilon_x\right)^t\mathbb{Z}\right)\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\left(n^{-1}\mathbb{Z}^t\left(\mathbb{Z}\gamma_y + \mathbb{U}\delta_y + \epsilon_y\right)\right) \\
&\xrightarrow{P} \beta + \left(\mathrm{E}\left[XZ^t\right]\mathrm{E}\left[ZZ^t\right]^{-1}\mathrm{E}\left[ZX^t\right]\right)^{-1} \\
&\qquad \times \mathrm{E}\left[\left(\gamma_x^t Z + \delta_x^t U + \epsilon_x\right)Z^t\right]\mathrm{E}\left[ZZ^t\right]^{-1}\mathrm{E}\left[Z\left(\gamma_y^t Z + \delta_y^t U + \epsilon_y\right)^t\right] \\
&= \beta + \left(\mathrm{E}\left[XZ^t\right]\mathrm{E}\left[ZZ^t\right]^{-1}\cdot\mathrm{E}\left[ZZ^t\right]\cdot\mathrm{E}\left[ZZ^t\right]^{-1}\mathrm{E}\left[ZX^t\right]\right)^{-1}\gamma_x^t\mathrm{E}\left[ZZ^t\right]\mathrm{E}\left[ZZ^t\right]^{-1}\mathrm{E}\left[ZZ^t\right]\gamma_y \\
&= \beta + \left(\gamma_x^t\mathrm{E}\left[ZZ^t\right]\gamma_x\right)^{-1}\gamma_x^t\mathrm{E}\left[ZZ^t\right]\gamma_y
\end{aligned}
$$

□

Thus, if γ is fixed and there is no pleiotropy, so that the exclusion restriction holds and $\gamma_y = 0$, then TSLS consistently estimates β. In fact, as long as $\gamma_x \neq 0$ and $\gamma_x^t \mathrm{E}\left[Z^t Z\right] \gamma_y = 0$, then TSLS is still consistent.

4.4 Egger Regression

Although TSLS will not consistently estimate β, in general, in the presence of pleiotropy, there exist other potential estimators. One of these is Egger Regression, a two stage estimation procedure that was originally developed for meta-analysis, which Bowden and colleagues used to estimate β, even when $\gamma_y \neq 0$ [1, 2]. Egger Regression relies on an additional assumption, which they call the Instrument Strength Independent of Direct Effect, or InSIDE, condition which assumes that γ_x and γ_y are independently distributed random variables.

In Egger Regression, the random variables used are not X, Y, and Z, but rather γ_x and $\Gamma = \beta\gamma_x + \gamma_y$, where γ_x and Γ are the coefficients that appear in the first stage and reduced form equations, respectively,

$$X = \gamma_x^t Z + \delta_x^t U + \epsilon_x$$
$$Y = \Gamma^t Z + \Delta^t U + \epsilon_r$$

However, since γ_x and Γ are unknown, we must estimate them from the above equations, and, so, we will actually use $\hat{\gamma}_x$ and $\hat{\Gamma}$, which are estimated with error. Some refinements of Egger Regression, such as one due to Zhao and colleagues take this measurement error into account [6, 7], but the original version, which is the most widely used, does not [1]. Egger Regression then regresses $\hat{\Gamma}$ on $\hat{\gamma}_x$, entry by entry using the regression model $\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_{x,j} + \epsilon_j$. This gives, $\hat{\beta}_{0E} = \overline{\hat{\Gamma}} - \hat{\beta}_E \overline{\hat{\gamma}}_x$ and $\hat{\beta}_E = \left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\Gamma} - \overline{\hat{\Gamma}}\right)$. These are the standard expressions for the coefficients in simple linear regression; a derivation using the normal equations is presented in the appendix.

4.4.1 Consistency

Despite wide use, Egger regression has not been formally shown to consistently estimate the causal effect and its use is typically justified using informal sequential asymptotic arguments (first letting n then p go to infinity) [1]. We present a formal proof of the consistency of Egger Regression here.

Theorem 4.2. *Let $\hat{\beta}_{E,p}$ be the Egger Regression estimate of β, the causal effect of X on Y, using p (possibly broken) instruments (Z). Let the elements of $\gamma_{x,p}, \gamma_{y,p}$ be independently and identically distributed like $f_x(p)\gamma_{x,0}, f_y(p)\gamma_{y,0}$, respectively, where $\gamma_0 \sim F_0$. If $p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(n^{-1}Z^tZ\right)^{-1}\right]\right] \in O(1)$ (which is implied by $\mathrm{E}\left[\lambda_{\min}\left(n^{-1}Z^tZ\right)^{-1}\right] \in O(1)$) and $f_x(p) \in \omega\left(n^{-\frac{1}{2}}\right), f_y(p) \in O\left(f_x(p)\right)$, then, $\beta_E \xrightarrow{P} \beta + \lim_{p\to\infty}\frac{f_y(p)}{f_x(p)} \cdot \mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$, so that Egger Regression consistently estimates β, if either $\mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right] = 0$ or $\lim_{p\to\infty}\frac{f_y(p)}{f_x(p)} = 0$.*

Proof. We let the entries of γ_x and γ_y vary with p, which we will denote by $\gamma_{x,p}$ and $\gamma_{y,p}$. Specifically, let the elements of $\gamma_{x,p}$ and $\gamma_{y,p}$ be independently and identically distributed like $f_x(p)\gamma_{x,0}$ and $f_y(p)\gamma_{y,0}$, respectively, where $\gamma_0 \sim F_0$, $\mathrm{Var}\left[\gamma_{x,0}\right] = \sigma^2_{x,0}, \mathrm{Var}\left[\gamma_{y,0}\right] = \sigma^2_{y,0}$. Also, let $\Gamma_p = \beta\gamma_{x,p} + \gamma_{y,p}$, $\gamma'_{x,p} = f_x(p)^{-1}\gamma_{x,p}, \gamma'_{y,p} = f_y(p)^{-1}\gamma_{y,p}$, and $\hat{\gamma}_{y,p} = \hat{\Gamma}_p - \beta\hat{\gamma}_{x,p}$. Then,

$$\begin{aligned}\hat{\beta}_{E,p} &= \left(\left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)^t\left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)\right)^{-1}\left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)^t\left(\hat{\Gamma}_p - \overline{\hat{\Gamma}}_p\right)\\ &= \beta + \left(\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\hat{\gamma}_{x,p}\right)^{-1}\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}\end{aligned}$$

We now expand $\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}$. Recall that $\hat{\gamma}_{x,p} = \left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\mathbb{X}$, $\hat{\Gamma}_p = \left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\mathbb{Y}$. Then,

$$\hat{\gamma}_{y,p} = \hat{\Gamma}_p - \beta\hat{\gamma}_{x,p} = \left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\left(\mathbb{Y} - \beta\mathbb{X}\right)$$

$$\hat{\gamma}_{x,p} - \gamma_{x,p} = \left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\left(\mathbb{U}\delta_x + \epsilon_x\right)$$

$$\hat{\Gamma}_p - \Gamma_p = \left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\left(\mathbb{U}\delta_y + \epsilon_y\right)$$

Note that this means that $\hat{\gamma}_{x,p} - \gamma_{x,p} \perp\!\!\!\perp (\gamma_{x,p}, \gamma_{y,p})$, $\hat{\gamma}_{y,p} - \gamma_{y,p} \perp\!\!\!\perp (\gamma_{x,p}, \gamma_{y,p})$. Then,

$$\begin{aligned}
p^{-1} f_x(p)^{-1} f_y(p)^{-1} \hat{\gamma}_{x,p}^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}_{y,p} \\
&= p^{-1} \hat{\gamma}_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}'_{y,p} \\
&= p^{-1} \left(\left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right) + \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \\
&\quad \times \left(\left(\hat{\gamma}'_{y,p} - \gamma'_{y,p}\right) + \gamma'_{y,p}\right) \\
&= p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{y,p} - \gamma'_{y,p}\right) \\
&\quad + p^{-1} \gamma_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{y,p} - \gamma'_{y,p}\right) \\
&\quad + p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p} \\
&\quad + p^{-1} \gamma_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p}
\end{aligned}$$

The last term goes to $\mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$. Additionally, if the middle terms go to zero, the first term will as well. This requires some condition so that the norm of the error terms will be effectively controlled. It is sufficient that $p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(n^{-1}Z^tZ\right)^{-1}\right]\right] \in O(1)$.

In the first part of the proof, we will assume that $f_y(p) \in \omega\left(n^{-\frac{1}{2}}\right)$, as well as $f_x(p)$. This will make γ_x and γ_y interchangeable in what follows, which will reduce the number of quantities that we need to calculate. Using Holder's Inequality,

$$\begin{aligned}
&\mathrm{E}\left|p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p}\right| \\
&\quad \leq \mathrm{E}\left[p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)\right]^{\frac{1}{2}} \\
&\qquad \times \mathrm{E}\left[p^{-1} \gamma_{y,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p}\right]^{\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&\mathrm{E}\left[p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)\right]\\
&\quad=p^{-1}f_x(p)^{-2}\mathrm{E}\left[\mathrm{tr}\left[\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\mathbb{Z}^t\left(\delta_x^t\Sigma_u\delta_x+\sigma_x^2\right)\mathcal{I}_n\mathbb{Z}\left(\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]\\
&\quad=n^{-1}f_x(p)^{-2}\left(\delta_x^t\Sigma_u\delta_x+\sigma_x^2\right)p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]\\
&\quad\leq n^{-1}f_x(p)^{-2}\left(\delta_x^t\Sigma_u\delta_x+\sigma_x^2\right)p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]\\
&\quad\leq n^{-1}f_x(p)^{-2}\left(\delta_x^t\Sigma_u\delta_x+\sigma_x^2\right)\cdot O(1)\\
&\quad\in o(1)
\end{aligned}
$$

since $f_x(p)\in\omega\left(n^{-\frac{1}{2}}\right)$ by assumption.

Then, $p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)\xrightarrow{\mathcal{L}^1}0$ Since this also holds replacing x by y, we have,

$$
\begin{aligned}
p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)\xrightarrow{\mathcal{L}^1}0\\
p^{-1}\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)\xrightarrow{\mathcal{L}^1}0
\end{aligned}
$$

Since,

$$
\begin{aligned}
&p^{-1}\mathrm{E}\left[\gamma'^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{x,p}\right]=p^{-1}\mathrm{E}\left[\left(\gamma'^t_{x,p}-\bar{\gamma}'^t_{x,p}\right)\left(\gamma'_{x,p}-\bar{\gamma}'_{x,p}\right)\right]\\
&\quad=p^{-1}\mathrm{E}\left[\gamma'^t_{x,p}\gamma'_{x,p}-\bar{\gamma}'^t_{x,p}\bar{\gamma}'_{x,p}\right]=p^{-1}\sum_{i=1}^{p}\mathrm{E}\left[\gamma'^2_{x,p,i}-\bar{\gamma}'^2_{x,p}\right]=p^{-1}(p-1)\sigma^2_{x,0}\\
&\quad\to\sigma^2_{x,0}
\end{aligned}
$$

and a similar expression holds with y replacing x, we have

$$
\begin{aligned}
\gamma'^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{x,p}\xrightarrow{\mathcal{L}^1}\sigma^2_{x,0}\\
\gamma'^t_{y,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{y,p}\xrightarrow{\mathcal{L}^1}\sigma^2_{y,0}
\end{aligned}
$$

Finally, invoking Holder's Inequality and that $\mathcal{L}^2$ convergence implies convergence in probability,

we have,

$$
\begin{aligned}
p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right) &\xrightarrow{P} 0\\
p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right) &\xrightarrow{P} 0\\
p^{-1}\gamma'^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right) &\xrightarrow{P} 0\\
p^{-1}\gamma'^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right) &\xrightarrow{P} 0\\
p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{y,p} &\xrightarrow{P} 0
\end{aligned}
$$

Then, since,

$$
p^{-1}\gamma'^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{y,p}\xrightarrow{P}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]
$$

by the Continuous Mapping Theorem,

$$
\begin{aligned}
p^{-1}f_x(p)^{-1}f_y(p)^{-1}\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}&\\
\xrightarrow{P}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]&
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
p^{-1}f_x(p)^{-2}\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\hat{\gamma}_{x,p}&\\
\xrightarrow{P}\mathrm{Var}\left[\gamma_{x,0}\right]&
\end{aligned}
$$

Thus, by a final application of the Continuous Mapping Theorem,

$$
\hat{\beta}_{E,p}\xrightarrow{P}\beta+\lim_{p\to\infty}\frac{f_y(p)}{f_x(p)}\cdot\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]
$$

Finally, we treat the more general case in which $f_x(p)\in\omega\left(n^{-\frac{1}{2}}\right), f_y(p)\in O\left(n^{-\frac{1}{2}}\right)$, but not necessarily $f_y(p)\in\omega\left(n^{-\frac{1}{2}}\right)$. Instead of analyzing
$p^{-1}f_x(p)^{-1}f_y(p)^{-1}\hat{\gamma}^t_{x,p}\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}$, we instead analyze

$f_x(p)^{-1}f_y(p)\cdot p^{-1}f_x(p)^{-1}f_y(p)^{-1}\hat{\gamma}_{x,p}^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}$. Then,

$$\begin{aligned}
&f_x(p)^{-1}f_y(p)\cdot p^{-1}f_x(p)^{-1}f_y(p)^{-1}\hat{\gamma}_{x,p}^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\hat{\gamma}_{y,p}\\
&\quad=p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\cdot f_x(p)^{-1}f_y(p)\cdot\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)\\
&\qquad+p^{-1}\gamma_{x,p}'^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\cdot f_x(p)^{-1}f_y(p)\cdot\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)\\
&\qquad+p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\cdot f_x(p)^{-1}f_y(p)\cdot\gamma'_{y,p}\\
&\qquad+f_x(p)^{-1}f_y(p)\cdot p^{-1}\gamma_{x,p}'^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{y,p}
\end{aligned}$$

When we apply Holder's inequality, we will always include the $f_x(p)^{-1}f_y(p)$ with the term containing γ_y. Thus, we need to compute two new quantities,

$$\begin{aligned}
&\mathrm{E}\left[f_x(p)^{-2}f_y(p)^2\cdot p^{-1}\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)\right]\\
&\quad=f_x(p)^{-2}f_y(p)^2\cdot n^{-1}f_y(p)^{-2}\left(\delta_y^t\Sigma_u\delta_y+\sigma_y^2\right)p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]\\
&\quad\leq n^{-1}f_x(p)^{-2}\left(\delta_y^t\Sigma_u\delta_y+\sigma_y^2\right)p^{-1}\mathrm{E}\left[\mathrm{tr}\left[\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]\\
&\quad=n^{-1}f_x(p)^{-2}\left(\delta_y^t\Sigma_u\delta_y+\sigma_y^2\right)\cdot O(1)\\
&\quad\in o(1)
\end{aligned}$$

since, $f_x(p)\in\omega\left(n^{-\frac{1}{2}}\right)$.

$$\begin{aligned}
&p^{-1}\mathrm{E}\left[f_x(p)^{-2}f_y(p)^2\cdot\gamma_{x,p}'^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{x,p}\right]\\
&\quad=f_x(p)^{-2}f_y(p)^2\cdot p^{-1}\mathrm{E}\left[\left(\gamma_{x,p}'^t-\bar{\gamma}_{x,p}'^t\right)\left(\gamma'_{x,p}-\bar{\gamma}'_{x,p}\right)\right]\\
&\quad\to\sigma_{x,0}^2\cdot\lim_{p\to\infty}f_x(p)^{-2}f_y(p)^2\in O(1)
\end{aligned}$$

Then, combined with the quantities we have already calculated, Holder's Inequality gives,

$$\begin{aligned}
f_x(p)^{-1}f_y(p)\cdot p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)&\xrightarrow{P}0\\
f_x(p)^{-1}f_y(p)\cdot p^{-1}\gamma_{x,p}'^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\hat{\gamma}'_{y,p}-\gamma'_{y,p}\right)&\xrightarrow{P}0\\
f_x(p)^{-1}f_y(p)\cdot p^{-1}\left(\hat{\gamma}'_{x,p}-\gamma'_{x,p}\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma'_{y,p}&\xrightarrow{P}0
\end{aligned}$$

Thus, the Continuous Mapping Theorem yields,

$$\begin{aligned}
&f_x(p)^{-1} f_y(p) \cdot p^{-1} f_x(p)^{-1} f_y(p)^{-1} \hat{\gamma}_{x,p}^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}_{y,p} \\
&\quad = f_x(p)^{-1} f_y(p) \cdot p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{y,p} - \gamma'_{y,p}\right) \\
&\qquad + f_x(p)^{-1} f_y(p) \cdot p^{-1} \gamma_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{y,p} - \gamma'_{y,p}\right) \\
&\qquad + f_x(p)^{-1} f_y(p) \cdot p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p} \\
&\qquad + f_x(p)^{-1} f_y(p) \cdot p^{-1} \gamma_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \gamma'_{y,p} \\
&\quad \xrightarrow{P} \lim_{p\to\infty} \frac{f_y(p)}{f_x(p)} \mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]
\end{aligned}$$

Then, since $p^{-1} f_x(p)^{-2} \hat{\gamma}_{x,p}^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}_{x,p} \xrightarrow{P} \mathrm{Var}\left[\gamma_{x,0}\right]$, as before, a final application of the Continuous Mapping Theorem gives,

$$\begin{aligned}
\hat{\beta}_{E,p} &= \left(\left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)^t \left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)\right)^{-1} \left(\hat{\gamma}_{x,p} - \overline{\hat{\gamma}}_{x,p}\right)^t \left(\hat{\Gamma}_p - \overline{\hat{\Gamma}}_p\right) \\
&= \beta + \frac{f_y(p)}{f_x(p)} \left(\hat{\gamma}_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}'_{x,p}\right)^{-1} \\
&\qquad \times \hat{\gamma}_{x,p}'^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \hat{\gamma}'_{y,p} \\
&\xrightarrow{P} \beta + \lim_{p\to\infty} \frac{f_y(p)}{f_x(p)} \cdot \mathrm{Var}\left[\gamma_{x,0}\right]^{-1} \mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]
\end{aligned}$$

This completes the proof.

□

Corollary 4.3. *If the entries of Z are IID with zero mean and finite fourth moment, $p, n \to \infty$, $n^{-1} p(n) \to \lambda \in [0, 1)$, and the assumptions of Theorem 4.2 hold (except for the trace condition), then $\beta_E \xrightarrow{P} \beta$, so that Egger Regression consistently estimates β.*

Proof. Let σ_z^2 be the variance of an entry of Z. Then, we can apply the Marchenko-Pasteur Theorem in order to write

$$\mathrm{tr}\left[\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right] = p \int s^{-1} dF_{n^{-1}Z^tZ}(s) \to p \int s^{-1} dF_{\gamma,\sigma_z^{-2}} = p\sigma_z^{-2} (1-\lambda)^{-1}$$

Then,

$$n^{-1} f_x(p)^{-2} \left(\delta_x^t \Sigma_u \delta_x + \sigma_x^2\right) p^{-1} \mathrm{E}\left[\mathrm{tr}\left[\left(n^{-1}\mathbb{Z}^t\mathbb{Z}\right)^{-1}\right]\right]$$
$$\to n^{-1} f_x(p)^{-2} (1-\lambda)^{-1} \left(\delta_x^t \Sigma_u \delta_x + \sigma_x^2\right) \sigma_z^{-2} \in o(1)$$

so

$$\mathrm{E}\left| p^{-1} \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \left(\hat{\gamma}'_{x,p} - \gamma'_{x,p}\right)\right| \in o(1)$$

and Theorem 4.2 holds.

□

4.4.2 Asymptotic Expansion

Although we have shown that the Egger Regression estimator is consistent for $\beta + \lim_{p\to\infty} \frac{f_y(p)}{f_x(p)} \cdot \mathrm{Var}\left[\gamma_{x,0}\right]^{-1} \mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$, we would like to characterize it more fully. In particular, we would like to know its rate of convergence and (scaled) asymptotic distribution. In order to do so, we begin by computing an asymptotic expansion for the case when p is fixed and γ is deterministic, but unknown.

Theorem 4.4. *Let p be fixed and γ be a deterministic sequence. Then, if $\hat{\beta}_{E,p} = \left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\Gamma} - \overline{\hat{\Gamma}}\right)$ is the causal effect estimated using Egger regression,*

$$\sqrt{n}\left(\hat{\beta}_{E,p} - \beta_{E,p}\right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_{\beta_{E,p}} + o_P(1)$$

with $\mathrm{E}\left[\psi_{\beta_E,p}\right] = 0$, *where*

$$\begin{aligned}\psi_{\beta_{E,p}} = &\left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \Big[\Big(\gamma_y^t - 2\left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_y - \overline{\gamma}_y\right)\right) \gamma_x^t\Big) \\ &\times \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(X_i - \gamma_x^t Z_i\right) \\ &+\gamma_x^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(Y_i - \beta X_i - \gamma_y^t Z_i\right)\Big]\end{aligned}$$

$$\beta_{E,p} = \beta + \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_y - \overline{\gamma}_y\right)$$

Proof. Let $\hat{\gamma}_y = \hat{\Gamma} - \beta\hat{\gamma}_x$. Since γ_x and Γ are estimated by linear regression, their influence functions are,

$$\begin{aligned}\psi_{\gamma_x}(X_i, Z_i) &= \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(X_i - \gamma_x^t Z_i\right) \\ \psi_{\Gamma}(Y_i, Z_i) &= \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(Y_i - \Gamma^t Z_i\right)\end{aligned}$$

Then, the influence function for γ_y will be,

$$\psi_{\gamma_y}(X_i, Y_i, Z_i) = \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(Y_i - \beta X_i - \gamma_y^t Z_i\right)$$

Using these expressions, we can now proceed. After some manipulation we get,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_{E,p} - \beta_{E,p}\right) &= \sqrt{n}\left(\left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\Gamma} - \overline{\hat{\Gamma}}\right) - \beta_p\right) \\ &= \left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \\ &\quad \Big[\sqrt{n}\left(\hat{\gamma}_x - \gamma_x\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\hat{\gamma}_y - \gamma_y\right) \\ &\quad\quad + \sqrt{n}\left(\hat{\gamma}_x - \gamma_x\right)^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\gamma_y \\ &\quad\quad + \gamma_x^t \left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1} 1_{p\times p}\right)\sqrt{n}\left(\hat{\gamma}_y - \gamma_y\right)\Big] \\ &\quad - \left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \\ &\quad\quad \times \sqrt{n}\left[\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right) - \left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right] \\ &\quad\quad \times \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_y - \overline{\gamma}_y\right)\right)\end{aligned}$$

$$\sqrt{n}\left[\left(\hat{\gamma}_x-\overline{\hat{\gamma}}_x\right)^t\left(\hat{\gamma}_x-\overline{\hat{\gamma}}_x\right)-\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right]$$
$$=2n^{-\frac{1}{2}}\sum_{i=1}^{n}\gamma_x^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\mathrm{E}\left[ZZ^t\right]^{-1}Z_i\left(X_i-\gamma_x^tZ_i\right)+o_P(1)$$

Expanding this term not only helps to simplify the numerator, but also tells us that the first part of the expression, the sample variance of $\hat{\gamma}_x$ converges to the sample variance of γ_x, in probability, so we can also simplify the denominator.

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_{E,p}-\beta_{E,p}\right)&=\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right)^{-1}\\&\quad\left[o_P(1)+\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\psi_{\gamma_x}(X_i,Z_i)+o_P(1)\right)^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\gamma_y\right.\\&\quad\left.+\gamma_x^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\psi_{\gamma_y}(X_i,Y_i,Z_i)+o_P(1)\right)\right]\\&\quad-\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right)^{-1}\\&\quad\times2\gamma_x^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\psi_{\gamma_x}(X_i,Z_i)+o_P(1)\right)\\&\quad\times\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right)^{-1}\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_y-\overline{\gamma}_y\right)\right)+o_P(1)\\&=n^{-\frac{1}{2}}\sum_{i=1}^{n}\left[\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right)^{-1}\right.\\&\quad\left(\gamma_y^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\mathrm{E}\left[ZZ^t\right]^{-1}Z_i\left(X_i-\gamma_x^tZ_i\right)\right.\\&\quad+\gamma_x^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\mathrm{E}\left[ZZ^t\right]^{-1}Z_i\left(Y_i-\beta X_i-\gamma_y^tZ_i\right)\\&\quad-2\gamma_x^t\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p-p^{-1}1_{p\times p}\right)\mathrm{E}\left[ZZ^t\right]^{-1}Z_i\left(X_i-\gamma_x^tZ_i\right)\\&\quad\left.\left.\times\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_x-\overline{\gamma}_x\right)\right)^{-1}\left(\left(\gamma_x-\overline{\gamma}_x\right)^t\left(\gamma_y-\overline{\gamma}_y\right)\right)\right)\right]+o_P(1)\end{aligned}$$

From the above expansion, we can directly read out the influence function,

$$\begin{aligned}
\psi_{\beta_{E,p}} = &\left((\gamma_x - \overline{\gamma}_x)^t (\gamma_x - \overline{\gamma}_x)\right)^{-1} \Big[\gamma_y^t \left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(X_i - \gamma_x^t Z_i\right) \\
&+ \gamma_x^t \left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(Y_i - \beta X_i - \gamma_y^t Z_i\right) \\
&- 2\gamma_x^t \left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(X_i - \gamma_x^t Z_i\right) \\
&\times \left((\gamma_x - \overline{\gamma}_x)^t (\gamma_x - \overline{\gamma}_x)\right)^{-1} \left((\gamma_x - \overline{\gamma}_x)^t (\gamma_y - \overline{\gamma}_y)\right)\Big] \\
= &\left((\gamma_x - \overline{\gamma}_x)^t (\gamma_x - \overline{\gamma}_x)\right)^{-1} \Big[\Big(\gamma_y^t - 2\left((\gamma_x - \overline{\gamma}_x)^t (\gamma_x - \overline{\gamma}_x)\right)^{-1} \left((\gamma_x - \overline{\gamma}_x)^t (\gamma_y - \overline{\gamma}_y)\right)\gamma_x^t\Big) \\
&\times \left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(X_i - \gamma_x^t Z_i\right) \\
&+\gamma_x^t \left(\mathcal{I}_p - p^{-1}1_{p\times p}\right)\left(\mathcal{I}_p - p^{-1}1_{p\times p}\right) \mathrm{E}\left[ZZ^t\right]^{-1} Z_i \left(Y_i - \beta X_i - \gamma_y^t Z_i\right)\Big]
\end{aligned}$$

$$\mathrm{E}\left[Z_i\left(X_i - \gamma_x^t Z_i\right)\right] = \mathrm{E}\left[Z_i \mathrm{E}\left[\left(X_i - \gamma_x^t Z_i\right)\right]\middle| Z_i\right] = \mathrm{E}\left[Z_i \mathrm{E}\left[\delta_x^t U + \epsilon_x\right]\middle| Z_i\right] = 0$$

$$\mathrm{E}\left[Z_i\left(Y_i - \beta X - \gamma_y^t Z_i\right)\right] = \mathrm{E}\left[Z_i \mathrm{E}\left[\left(Y_i - \beta X - \gamma_y^t Z_i\right)\right]\middle| Z_i\right] = \mathrm{E}\left[Z_i \mathrm{E}\left[\delta_y^t U + \epsilon_y\right]\middle| Z_i\right] = 0$$

so $\mathrm{E}\left[\psi_{\beta_{E,p}}\right] = 0$.

This completes the proof.

□

Corollary 4.5. *For fixed p and deterministic γ, the Egger Regression estimator of the causal effect, $\hat{\beta}_{E,p}$, is asymptotically biased, with bias $\beta_{E,p} - \beta_E$, is $n^{\frac{1}{2}}$ consistent for $\beta_{E,p}$, and is asymptotically normal with limiting variance $\mathrm{E}\left[\psi_{\beta_{E,p}} \psi_{\beta_{E,p}}^t\right]$.*

Proof. This follows directly from the expansion in Theorem 4.4.

□

We would like to extend this result to the setting in which we allow p to grow as a function of n. If we let $\beta_E = \beta + \mathrm{Var}\left[\gamma_{x,0}\right]^{-1} \mathrm{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$, which we previously showed to be the limit in probability of $\hat{\beta}_E$. Then, the identity

$$\hat{\beta}_{E,p} - \beta_E = \left(\beta_{E,p} - \beta_E\right) + \left(\hat{\beta}_E - \beta_{E,p}\right)$$

suggests that a starting point would be to develop an expansion of $\sqrt{p}\left(\beta_p - \beta_E\right)$. We present such an expansion below, but note that there are several problems with such an approach. The first is that, in Theorem 4.4, we assumed that γ was a fixed sequence, whereas, in such an expansion, it must be random. The second is that the $o_P(1)$ term in Theorem 4.4 may actually be dependent on p. When p is fixed, simply applying the union bound shows that this term will remain $o_P(1)$ no matter how large p is, as long as it does not scale with n. However, when p scales with n, this argument no longer holds.

However, it is still worthwhile to perform the expansion,

Theorem 4.6. *Let* $\sqrt{p}\left(1 - f_x(p)^{-1} f_y(p)\right) \xrightarrow{P} 1$ *and* $f_x(p), f_y(p) \in \omega\left(n^{-\frac{1}{2}}\right)$ *so that* $\beta_{E,p} = \beta + \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_y - \overline{\gamma}_y\right)$ *and* $\beta_E = \beta + \operatorname{Var}\left[\gamma_{x,0}\right]^{-1} \operatorname{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$ *are the fixed* p *and asymptotic limits of the Egger Regression estimator,* $\hat{\beta}_{E,p} = \left(\left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)\right)^{-1} \left(\hat{\gamma}_x - \overline{\hat{\gamma}}_x\right)^t \left(\hat{\Gamma} - \overline{\hat{\Gamma}}\right)$*, respectively, where* $\hat{\gamma}$ *and* $\hat{\Gamma}$ *are estimated by linear regression. Then,*

$$\sqrt{p}\left(\beta_{E,p} - \beta_E\right) = p^{-\frac{1}{2}} \sum_{i=1}^{p} \psi_{\beta_E} + o_P(1)$$

where

$$\begin{aligned} \psi_{\beta_E} = \operatorname{Var}\left[\gamma_{x,0}\right]^{-1} \Big[& \left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)\left(\gamma'_{y,i} - \mu_{\gamma_{y,0}}\right) - \operatorname{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]\right) \\ & - \left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)^2 - \operatorname{Var}\left[\gamma_{x,0}\right]\right) \cdot \operatorname{Var}\left[\gamma_{x,0}\right]^{-2} \operatorname{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right] \Big] \end{aligned}$$

Note that $\mathrm{E}\left[\psi_{\beta_E}\right] = 0$.

Proof.

$$
\begin{aligned}
&\sqrt{p}\left(\beta_p - \beta_E\right) \\
&\quad= \sqrt{p}\left(\left(\left(\gamma_x - \overline{\gamma}_x\right)^t\left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1}\left(\gamma_x - \overline{\gamma}_x\right)^t\left(\gamma_y - \overline{\gamma}_y\right) - \mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) \\
&\quad= \sqrt{p}\cdot f_x(p)^{-1}f_y(p)\left(\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right)\right)^{-1}\left(\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_y - \overline{\gamma}'_y\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) \\
&\qquad+ \sqrt{p}\left(f_x(p)^{-1}f_y(p)\left(\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right)\right)^{-1} - \mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\right)\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right] \\
&\quad= \left(\left(\gamma_x - \overline{\gamma}_x\right)^t\left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \\
&\qquad\times\left(f_x(p)^{-1}f_y(p)\sqrt{p}\left(\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_y - \overline{\gamma}'_y\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right)\right. \\
&\qquad\left.-\sqrt{p}\left(\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right) - f_x(p)^{-1}f_y(p)\mathrm{Var}\left[\gamma_{x,0}\right]\right)\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) \\
&\quad= \left(p^{-1}\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right)\right)^{-1} \\
&\qquad\times\left(f_x(p)^{-1}f_y(p)\cdot p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \overline{\gamma}'_x\right)\left(\gamma'_{y,i} - \overline{\gamma}'_y\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right)\right. \\
&\qquad\left.- p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \overline{\gamma}'_x\right)^2 - f_x(p)^{-1}f_y(p)\mathrm{Var}\left[\gamma_{x,0}\right]\right)\cdot\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) \\
&\quad= \left(p^{-1}\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right)\right)^{-1} \\
&\qquad\times\left(p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)\left(\gamma'_{y,i} - \mu_{\gamma_{y,0}}\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) - p^{\frac{1}{2}}\left(\overline{\gamma}'_x - \mu_{\gamma_{x,0}}\right)\left(\overline{\gamma}'_y - \mu_{\gamma_{y,0}}\right)\right. \\
&\qquad\left.- \left(p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)^2 - \mathrm{Var}\left[\gamma_{x,0}\right]\right) - p^{\frac{1}{2}}\left(\overline{\gamma}'_x - \mu_{\gamma_{x,0}}\right)^2\right)\cdot\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) \\
&\qquad+ \left(p^{-1}\left(\gamma'_x - \overline{\gamma}'_x\right)^t\left(\gamma'_x - \overline{\gamma}'_x\right)\right)^{-1}\sqrt{p}\left(f_x(p)^{-1}f_y(p) - 1\right)\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right] \\
&\quad= \mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\left(p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)\left(\gamma'_{y,i} - \mu_{\gamma_{y,0}}\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right)\right. \\
&\qquad\left.- \left(p^{-\frac{1}{2}}\sum_{i=1}^{p}\left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)^2 - \mathrm{Var}\left[\gamma_{x,0}\right]\right)\right)\cdot\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right) + o_P(1) \\
&\quad= p^{-\frac{1}{2}}\sum_{i=1}^{p}\mathrm{Var}\left[\gamma_{x,0}\right]^{-1}\left[\left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)\left(\gamma'_{y,i} - \mu_{\gamma_{y,0}}\right) - \mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right)\right. \\
&\qquad\left.- \left(\left(\gamma'_{x,i} - \mu_{\gamma_{x,0}}\right)^2 - \mathrm{Var}\left[\gamma_{x,0}\right]\right)\cdot\mathrm{Var}\left[\gamma_{x,0}\right]^{-2}\mathrm{Cov}\left[\gamma_{x,0},\gamma_{y,0}\right]\right] + o_P(1)
\end{aligned}
$$

□

Corollary 4.7. *If γ is observed, $\beta_{E,p}$ is $p^{\frac{1}{2}}$ consistent for β_E, and is asymptotically normal with limiting variance* $\mathrm{E}\left[\psi_{\beta_E}\psi_{\beta_E}^t\right]$.

Proof. This follows directly from the expansion in Theorem 4.6.

□

For the aforementioned reasons, even combined with Theorem 4.4, this is not enough to show that the Egger regression estimator has a valid asymptotic expansion in the case in which p grows with n and γ is random. However, it does suggest that the variance estimated for any fixed sequence of γs, will incorrectly estimate the true variance of the estimator. In particular, since the actual γ associated with any genetic locus is fixed, variance estimates will not properly account for the fact that γ varies across loci.

This also suggests that the meaningful parameter for the asymptotics is really p, not n. No matter how large p is, so long as it is fixed, the Egger regression estimator will be biased for the true causal effect, even if n goes to infinity. Further, since $\beta_{E,p}$ is only $p^{\frac{1}{2}}$ consistent for β_E, Egger regression cannot hope to be more than $p^{\frac{1}{2}}$ consistent. Therefore, unless $n^{-1}p \to \lambda \in (0,1)$, Egger Regression cannot be $n^{\frac{1}{2}}$ consistent.

4.5 Simulations

In order to further explore the behavior of Egger Regression, we turn to simulation. In our simulations, we use the following structural equations,

$$Y = \beta X + \gamma_y^t Z + \epsilon_y'$$
$$X = \gamma_x^t Z + \epsilon_x'$$

$Z_{ij} \sim \mathcal{N}(0,1)$, $\epsilon_i' \sim \mathcal{N}(0, \Sigma_\epsilon)$, $\gamma_{x,j} \sim \mathcal{N}\left(0, f_x(p)^2\right), \gamma_{y,j} \sim \mathcal{N}\left(0, f_y(p)^2\right)$, where $\Sigma_\epsilon = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $\rho = \frac{1}{4}$, and all random variables are IID. We let $f_x(p) = p^{-\theta_x}, f_y(p) = p^{-\theta_y}$. Each figure represents 1000 replications. We use a $p : n$ ratio of $1 : 10$ for $p = 3, 10, 31, 100, 316 = \left\lfloor 10^{\frac{i}{2}} \right\rfloor, i = 1, 2, 3, 4, 5$. For each combination of θ values, the densities for each p are estimated using kernel smoothing and are plotted together in a single figure.

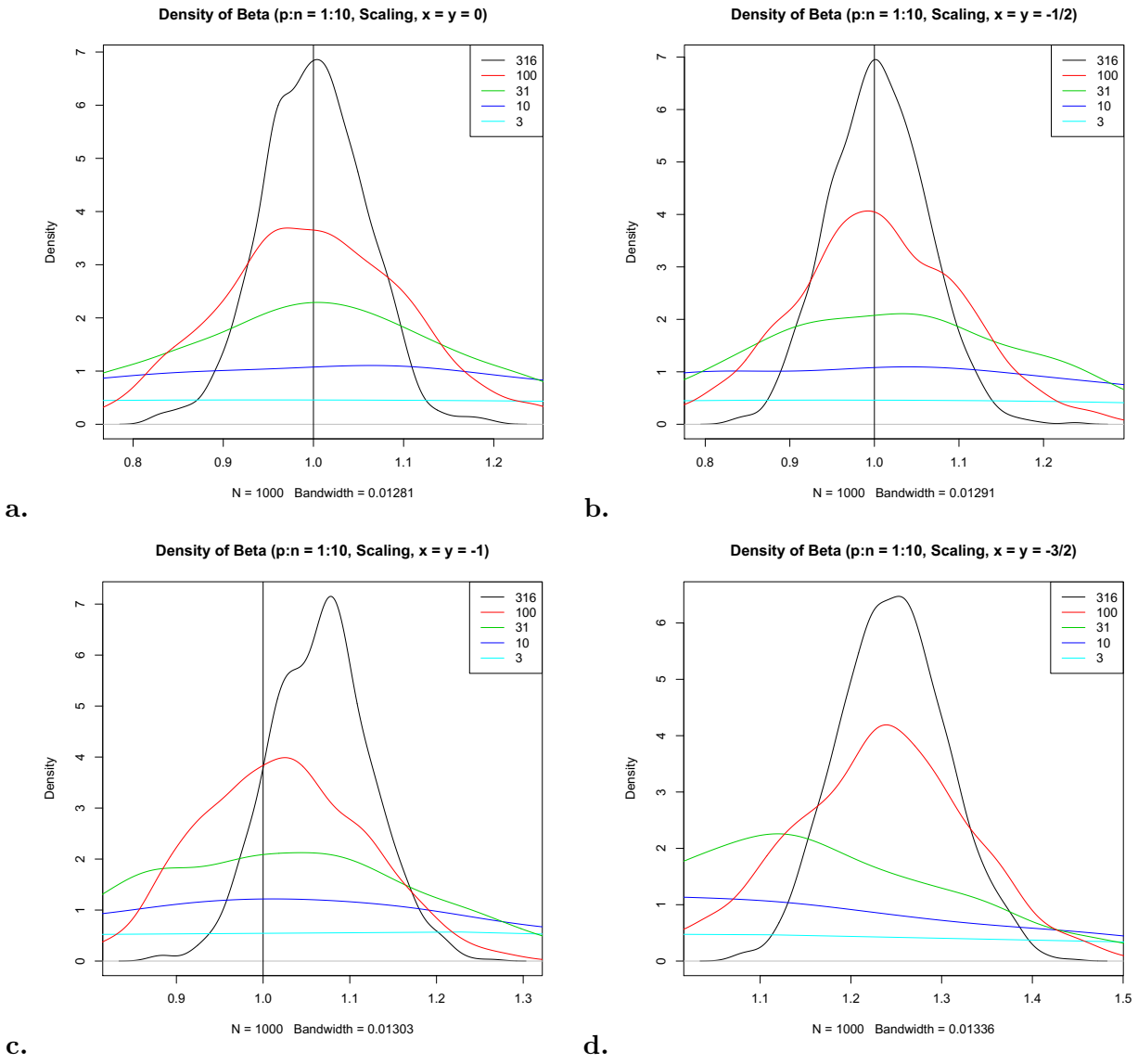
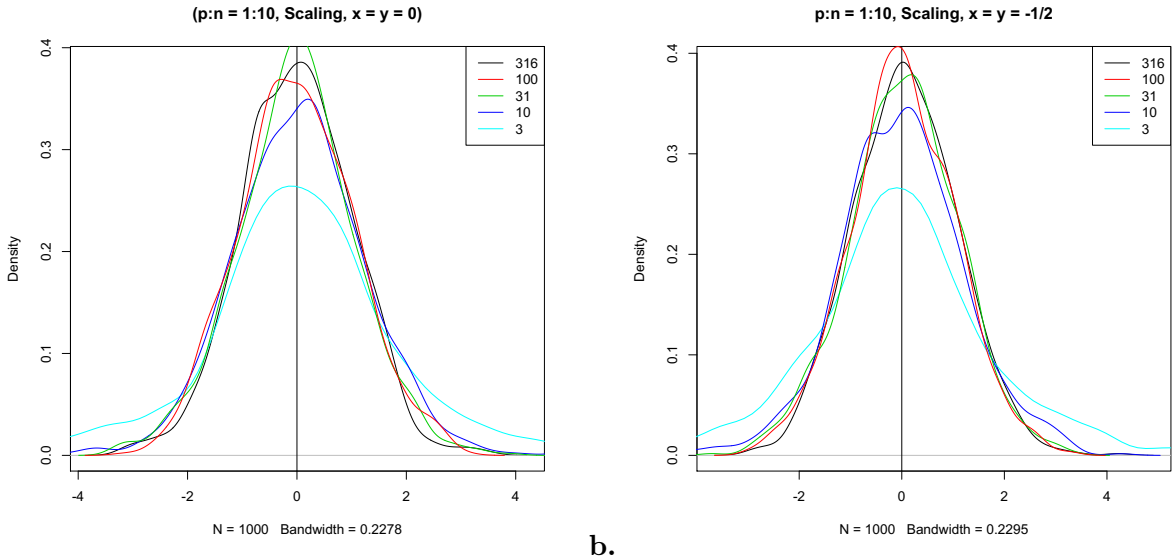


Figure 4.1: Density of $\hat{\beta}$ $(\theta_x = \theta_y)$. Simulated results from Egger Regression with $p : n = 1 : 10$ for **a.** $\theta_x = \theta_y = 0$, **b.** $\theta_x = \theta_y = \frac{1}{2}$, **c.** $\theta_x = \theta_y = 1$, **d.** $\theta_x = \theta_y = \frac{3}{2}$. Lines correspond to $p = 3, 10, 31, 100, 316$.

From Figure 4.1, we see that for $\theta_x = \theta_y \leq \frac{1}{2}$, Egger regression appears to be consistent and unbiased. While for $\theta_x = \theta_y > \frac{1}{2}$, it is biased. This is consistent with our analytic results, which required that $f_x(p) \in \omega\left(n^{-\frac{1}{2}}\right)$ in order to be able to prove convergence, and suggests that, at least in this respect, they cannot be improved.

In Figure 4.2 we also see that $p^{\frac{1}{2}}\left(\hat{\beta} - \beta\right)$ appears to converge to a fixed normal distribution. In combination with the above, this suggests that, when $\lim n^{-1}p \to \lambda \in (0, 1)$, $\hat{\beta}$ is $p^{\frac{1}{2}}$ consistent



a. **b.**

Figure 4.2: Density of $p^{\frac{1}{2}}\left(\hat{\beta}-\beta\right)$. Simulated results from Egger Regression with $p:n=1:10$ for **a.** $\theta_x=\theta_y=0$, **b.** $\theta_x=\theta_y=\frac{1}{2}$. Lines correspond to $p=3,10,31,100,316$.

and asymptotically normal.

If $\theta_x \neq \theta_y$, then, if $\theta_x \leq \frac{1}{2}$ and $\theta_y > \theta_x$, $\hat{\beta} \to \beta$, as seen in Figures 4.3a, b, and c. However, if $\theta_x > \frac{1}{2}$, $\hat{\beta}$ will be biased as seen in Figure 4.3d. This is again consistent with the fact that we required $f_x(p) \in \omega\left(n^{-\frac{1}{2}}\right)$, in order to be able to prove convergence and suggests that this condition is required for Egger Regression to be able to consistently estimate the causal effect.

If $\theta_x \leq \frac{1}{2}$ and $\theta_y = \theta_x - \frac{1}{2}$, then $\hat{\beta}$ is not consistent, but appears to converge to a fixed distribution and is unbiased, as seen in Figure 4.4a. This agrees with the fact that Figure 4.2 suggests that the Egger Regression estimator is $p^{\frac{1}{2}}$ consistent. Interestingly, this convergence to a nondegenerate normal distribution also appears to occur when $\theta_x = 1, \theta_y = \frac{1}{2}$, as seen in Figure 4.4c. When $\theta_x > 1$ and $\theta_y = \theta_x - \frac{1}{2}$, $\hat{\beta}$ is biased, as seen in Figure 4.4d, which is what we would expect. Finally, when $\theta_y < \theta_x - \frac{1}{2}$, the distribution of $\hat{\beta}$ diverges as p increases, as seen in Figure 4.4b. This is also not surprising, since the instrument is extremely weak relative to the pleiotropy and so is overwhelmed.

4.6 Conclusion

In this work, we have shown that if $n^{-1}p \to \lambda \in [0,1)$ and the strength of the instruments, $f_x(p)$, declines strictly slower than $n^{-\frac{1}{2}}$ and no slower than the strength of the pleiotropic effects, $f_y(p)$, the Egger regression estimator $\hat{\beta}_{E,p} \xrightarrow{P} \beta_E = \beta + \lim_{p\to\infty} \frac{f_y(p)}{f_x(p)} \cdot \text{Var}\left[\gamma_{x,0}\right]^{-1} \text{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right]$.

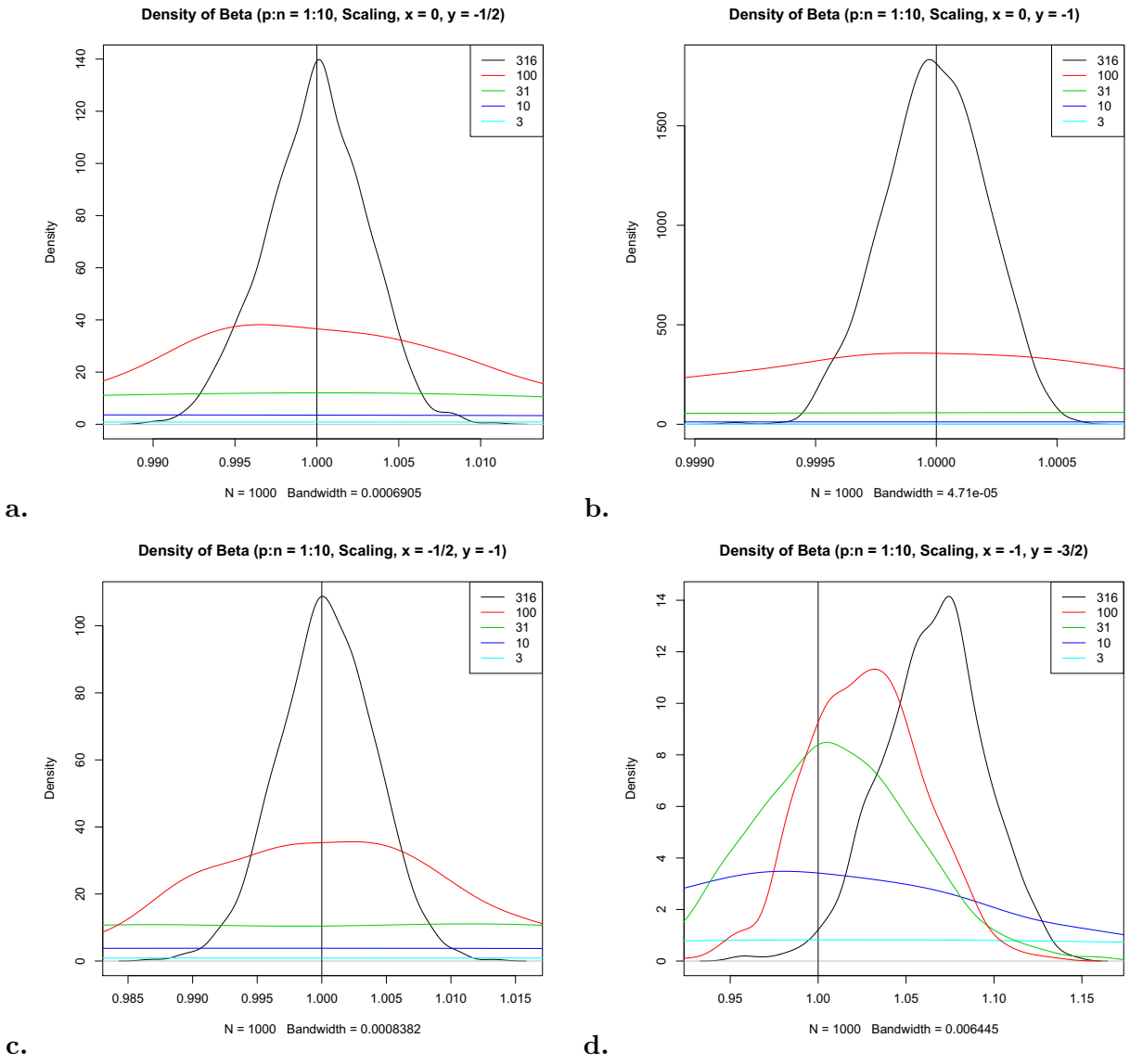


Figure 4.3: Density of $\hat{\beta}$ ($\theta_x < \theta_y$). Simulated results from Egger Regression with $p : n = 1 : 10$ for **a.** $\theta_x = 0, \theta_y = \frac{1}{2}$, **b.** $\theta_x = 0, \theta_y = 1$, **c.** $\theta_x = \frac{1}{2}, \theta_y = 1$, **d.** $\theta_x = 1, \theta_y = \frac{3}{2}$. Lines correspond to $p = 3, 10, 31, 100, 316$.

Therefore, it consistently estimates the true causal effect of X on Y, β, if either $\text{Cov}\left[\gamma_{x,0}, \gamma_{y,0}\right] = 0$ or the strength of pleiotropy declines faster than the strength of the instruments ($f_y(p) \in o\left(f_x(p)\right)$).

We have also shown that, for fixed p (and fixed γ), the Egger Regression estimator is biased and converges in probability to $\beta_{E,p} = \beta + \left(\left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_x - \overline{\gamma}_x\right)\right)^{-1} \left(\gamma_x - \overline{\gamma}_x\right)^t \left(\gamma_y - \overline{\gamma}_y\right)$ and is $n^{\frac{1}{2}}$ consistent for $\beta_{E,p}$ and asymptotically normal. Further, if γ is directly observed, and does not need to be estimated, then we demonstrated that $\beta_{E,p} \xrightarrow{P} \beta_E$ and, further, it is $p^{\frac{1}{2}}$ consistent for β_E and

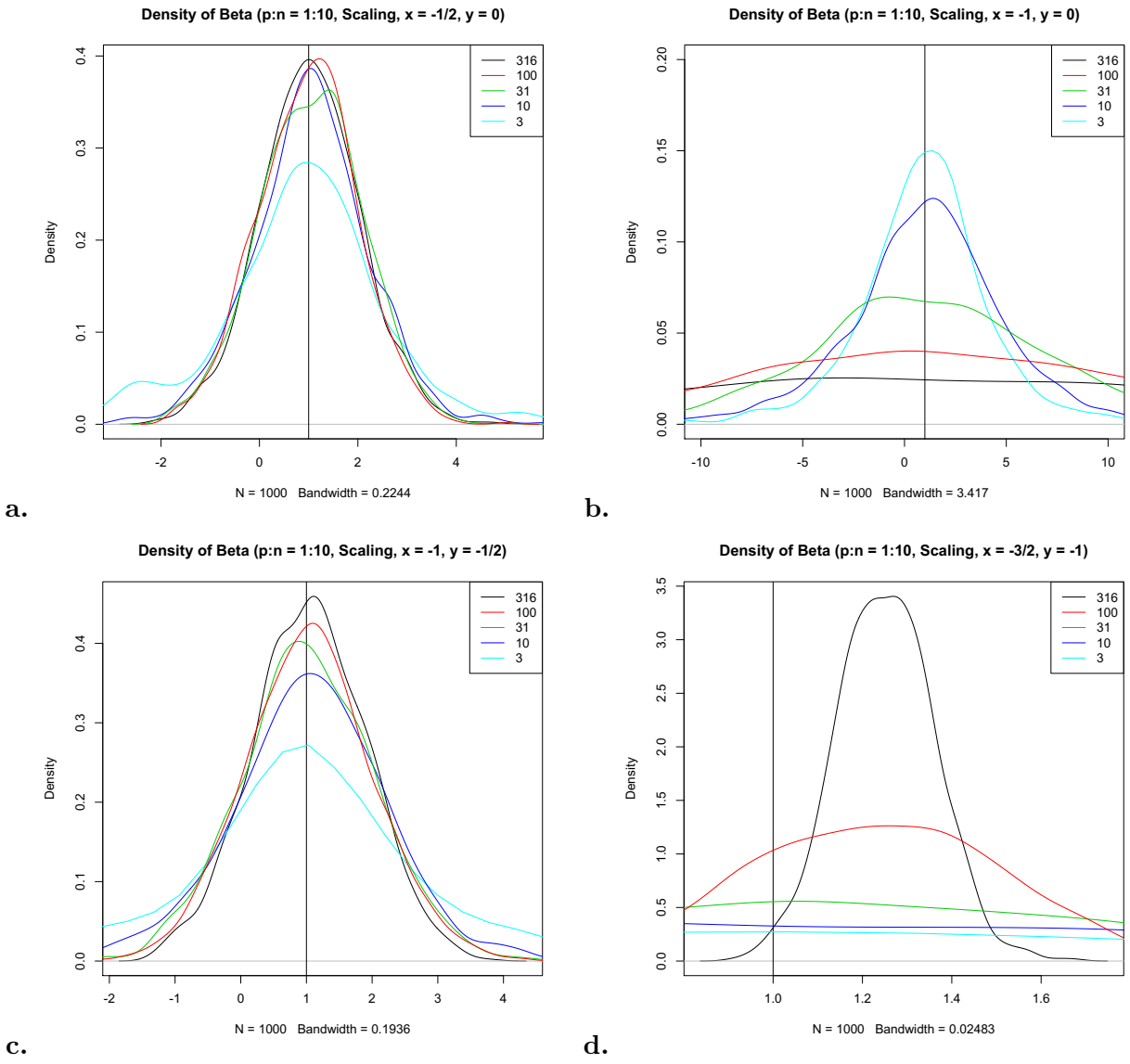


Figure 4.4: Density of $\hat{\beta}$ ($\theta_x > \theta_y$). Simulated results from Egger Regression with $p : n = 1 : 10$ for **a.** $\theta_x = \frac{1}{2}, \theta_y = 0$, **b.** $\theta_x = 1, \theta_y = 0$, **c.** $\theta_x = 1, \theta_y = \frac{1}{2}$, **d.** $\theta_x = \frac{3}{2}, \theta_y = 1$. Lines correspond to $p = 3, 10, 31, 100, 316$.

asymptotically normal. This suggests that, $\hat{\beta}_{E,p}$, the standard Egger Regression estimator, which must estimate $\hat{\gamma}$, since γ is unknown, cannot converge faster to β_E than $p^{\frac{1}{2}}$ and, thus, will not be $n^{\frac{1}{2}}$ consistent unless $n^{-1}p \to \lambda \in (0, 1)$.

However, several important results remain. First, it is important to show that the Egger Regression estimator $\hat{\beta}_{E,p}$ is $p^{\frac{1}{2}}$ consistent, meaning that for $n^{-1}p \to \lambda \in (0, 1)$ it is $n^{\frac{1}{2}}$ consistent. Second, it would be ideal to show that $\hat{\beta}_{E,p}$ is asymptotically normal, although, as discussed above,

there are several structural challenges involved in doing so. Both of these properties are implied by our simulation results, and so we expect to be able to prove that the Egger Regression estimator is $p^{\frac{1}{2}}$ consistent and asymptotically normal in future work.

Bibliography

[1] Jack Bowden, George Davey Smith, and Stephen Burgess. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression". *International Journal of Epidemiology* 44.2 (2015), pp. 512–525.

[2] Matthias Egger, George Davey Smith, and Christoph Minder. "Bias in meta-analysis detected by a simple, graphical test". *BMJ* 315 (1997), pp. 629–34.

[3] Richard Gray and Keith Wheatley. "How to avoid bias when comparing bone marrow transplantation with chemotherapy". *Bone Marrow Transplant* 7.Suppl. 3 (1991), pp. 9–12.

[4] Martjin B Katan. "Apolipoprotein E isoforms, serum cholesterol, and cancer". *Lancet* 327.8479 (1986), pp. 507–8.

[5] George Davey Smith and Shah Ebrahim. "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?" *International Journal of Epidemiology* 32.1 (2003), pp. 1–22.

[6] Qingyuan Zhao et al. "Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score". *Annals of Statistics* (2020).

[7] Qingyuan Zhao et al. "Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples". *Statistical Science* 34.2 (2019), pp. 317–333.

Chapter 5

Conclusion

In this work, we explored various aspects of matching and Instrumental Variable estimators. In particular, we showed that neither direct covariate matching nor propensity score matching strictly dominates the other and that, under a variety of metrics, specifically the mean and variance of the matching estimator itself, and the Model Dependence of families of matching estimators, each tends to perform better in different regimes. In particular, when the propensity score is known and fixed, and the true and regression models are linear, when the covariate distribution is symmetric, and higher order terms do not appear in either the true or regression models, then direct matching results in unbiased matches and is universally preferred, although its advantage declines as more covariates are matched upon. However, when the covariate distribution is not symmetric, or when higher order terms appear in either the true or regression model, then direct matching results in biased matches, and, thus, while it may be advantageous in smaller samples, particularly when the number of matched covariates is small, as the numbers of treated subjects and matched covariates becomes large, propensity score matching performs better because it results in unbiased matches.

We also demonstrated several asymptotic properties of Egger Regression. First, we showed that, as long as the instruments are not too weak, either in an absolute sense or relative to the strength of pleiotropy, if the number of instruments and the sample size go to infinity, and their ratio converges to a quantity less than 1, Egger Regression consistently estimates the true causal effect when either the InSIDE assumption holds or when the level of pleiotropy is strictly weaker than the strength of the instruments. When InSIDE fails to hold and the instruments are not too weak, then the estimator still converges in probability, but the estimate is biased. We also showed that, for a fixed, finite number of instruments, the Egger Regression estimator is asymptotically normal, but its limit is biased. Additionally, in the case in which the instrument strength and pleiotropy are known and do not need to be estimated, the corresponding estimator is asymptotically normal and is $p^{\frac{1}{2}}$-consistent for its limiting value, where p is the number of instruments. Therefore, the standard Egger Estimator cannot converge faster and, so, can only be $n^{\frac{1}{2}}$ consistent if, $n^{-1}p$ goes to a constant strictly between 0 and 1.

These findings affirm the idea that the matching method should be selected based on the structure of the data, and that direct and propensity score matching each have settings in which they outperform the alternative. Likewise, these initial asymptotic results validate the use of Egger Regression in the presence of pleiotropy, as well as demonstrating its limits. Hopefully, these results will provide assistance on deciding when to use direct vs. propensity score matching, as well as when Egger Regression is an appropriate technique for estimating causal effects.