Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects

Akanksha Negi, Jeffrey M. Wooldridge[†]

December 22, 2019

Regression adjustment with covariates in experiments is intended to improve precision over a simple difference in means between the treated and control outcomes. The efficiency argument in favor of regression adjustment has come under criticism lately, where papers like Freedman (2008a,b) find no systematic gain in asymptotic efficiency of the covariate adjusted estimator. In this paper, we verify that when treatment effects are heterogeneous, additively controlling for covariates in a regression is not guaranteed to bring additional efficiency gain. We then show that, like in Lin (2013), estimating separate regressions for the control and treated groups is guaranteed to do no worse than both the simple difference-in-means estimator and just including the covariates in additive fashion. Usually, the estimator that includes a full set of interactions strictly improves asymptotic efficiency. Unlike Imbens and Rubin (2015), who assume full knowledge of the population means of the covariates in a random sampling context, we show that the fully interacted estimator improves asymptotic efficiency even when one accounts for the sampling variation in the sample means of the covariates. This result appears to be new, and simulations show that the efficiency gains can be substantial. We also show that in some important cases – applicable to binary, fractional, count, and other nonnegative responses – nonlinear regression adjustment is consistent without any restrictions on the conditional mean functions.

JEL Classification Codes: C21, C25

Keywords: Experiment, Regression Adjustment, Heterogeneous Treatment Effects

1 Introduction

The role of covariates in randomized experiments has been studied since the early 1930s [Fisher (1935)][1]. When compared with the simple difference-in-means (SDM) estimator, the main benefit of adjusting for covariates is that the precision of the estimated average treatment effect (ATE) can be improved if the covariates are sufficiently predictive of the outcome [Cochran (1957); Lin (2013)]. Nevertheless, regression adjustment is not uniformly accepted as being preferred over the SDM estimator. For example, Freedman (2008a,b) argues against using regression adjustment because it is not guaranteed to be unbiased unless one makes the strong assumption that the conditional expectation functions are correctly specified and linear in parameters.

It is important to understand that there are two different, potentially valid criticisms of regression adjustment (RA). The first is the issue of bias just mentioned: unlike the SDM estimator, which is unbiased conditional on having some units in both the control and treatment groups, RA estimators are only guaranteed to be consistent, not unbiased. Therefore, in experiments with small sample sizes, one might be willing to forego potential efficiency gains in order to ensure an unbiased estimator of the average treatment effect. As Bruhn and McKenzie (2009) points out, samples of 100 to 500 individuals or 20 to 100 schools or health clinics is fairly common in experiments conducted in development economics. In situations where unbiased estimation is the highest priority, the current paper has little to add, other than to provide simulation evidence that using our preferred RA procedure often results in small bias. Henceforth, we are not interested in small-sample problems as a criticism of RA. More and more economic experiments, especially those conducted online, include enough units to make consistency and asymptotic efficiency a relevant criteria for evaluating estimators of ATEs. In cases where effect sizes are small (but important in the aggregate), improving precision can be important even when sample sizes seem fairly large.

A second criticism of RA methods, and the one most relevant for this paper, is that RA methods may not improve over the SDM estimator even if we focus on asymptotic efficiency. Freedman (2008a,b) and Lin (2013) level this criticism of RA when the covariates are simply added as controls along with a treatment indicator in a linear regression analysis. Freedman (2008a), for example, finds no systematic efficiency gain from using covariates. Lin (2013) provides an in-depth discussion about how simply adding covariates will not necessarily produce efficiency gains when treatment effects are heterogeneous. Both Freedman and Lin operate under a finite population paradigm where all population units are observed in the sample. Therefore, uncertainty in the estimators is due to the assignment into treatment and control, and not due to sampling from a population [Abadie et al. (2017b), Abadie et al. (2017a) and Rosenbaum (2002) discuss similar settings].

[1] In the statistics and biomedical literature, such variables are also known as prognostic factors or concomitant variables. In the econometrics of program evaluation literature they are known as pre-treatment covariates. As the name suggests, they are ideally measured before the treatment is administered.

Random sampling from a population is still an important setting for empirical work, and the findings in Freedman (2008a) and Lin (2013) do not extend to the random sampling scenario. One would require a framework that explicitly accounts for both kinds of uncertainties: sampling-based and design-based. Our paper will not have more to say about the differences between these two types of randomness. For a deeper discussion of sampling-based and design-based uncertainty, see Abadie et al. (2017b). Our primary motivation for studying this problem comes from Imbens and Rubin (2015), who study linear regression adjustment in the same sampling setting that we use here: independent and identically distributed (i.i.d.) draws from a population. However, Imbens and Rubin state efficiency results only in the case that the population means of the covariates are known, even though only a random sample is available. In addition, Imbens and Rubin only consider *linear* regression adjustment, and, in fact, warn against using nonlinear regression adjustment.

After writing the initial draft of this paper, we discovered that regression adjustment in experiments with random sampling has been studied in the statistics literature. These include the papers by Yang and Tsiatis (2001), Tsiatis et al. (2008), Ansel et al. (2018), and Pitkin et al. (2017) for the case of linear adjustment and by Rosenblum and Van Der Laan (2010) and Bartlett (2018) for the case of nonlinear regression adjustment. While several of the results derived in these papers overlap with what we show, our approach is more transparent, and the representations of the various estimators make the source of efficiency gains easy to understand when using separate versus pooled methods, and whether or not one knows that population means of the covariates. Our arguments in the linear case are based on population linear projections. Moreover, in the nonlinear case, we appeal to results on doubly robust estimation by Wooldridge (2007), relying on population properties of pseudo-true parameters. At the same time, we fill a gap in the literature by studying pooled nonlinear methods. Namely, we show that, for settings where separate nonlinear regression adjustment is consistent, the pooled version is also consistent. Such a result is practically important in cases where a researcher may lack sufficient degrees of freedom to feel comfortable with estimating separate models for the control and treatment groups. If one uses a pooled nonlinear method and finds an improvement in precision, the researcher may be satisfied with that.

For emphasis, in the case of linear regression adjustment, we study four estimators: the SDM estimator, the pooled regression adjusted (PRA) estimator, the full regression adjusted (FRA) estimator – which uses separate regressions for the control and treatment groups – and what we call the infeasible regression adjusted (IRA) estimator, which is like the FRA estimator but assumes the population means of the covariates are known. We include IRA for completeness, as it is studied in Imbens and Rubin (2015), and doing so allows us to characterize the lost efficiency due to having to estimate the population means.

Our most important results in the linear regression case can be easily summarized. First, even when accounting for the sampling error of the covariates in estimating the population means, using separate linear regressions for the control and treatment groups leads to an ATE estimator that is never less precise (asymptotically) than the SDM estimator and the PRA estimator. Unless small sample bias is a concern, there is no reason not to use full regression adjustment. Further, there are two interesting cases when there will be no precision gain when using full RA compared with pooled RA. The first is when there is no heterogeneity in the slopes of the linear projections of the potential outcomes (although there could be in the unobservables). In this case, it is not surprising that using pooled RA is sufficient to capture the efficiency gains of using covariates. The second important case where there is no additional gain from FRA is when the design is balanced: the probability of being in the treatment group is equal to 0.5. Therefore, if one has imposed a balanced design and is considering only linear regression adjustment, the pooled method is probably preferred (due to conserving degrees of freedom). A final result, which is pretty obvious, is that there is no efficiency gain when the covariates are not predictive of the potential outcomes; then, SDM is asymptotically efficient. We want to emphasize that there is no (asymptotic) cost in doing the regression adjustment, whether PRA or FRA: each estimator has the same asymptotic variance. In situations where one has good predictors of the outcome, regression adjustment can be attractive. Our simulation study illustrates the special cases derived from our theortical results.

We also simply characterize situations where nonlinear regression adjustment preserves consistency of average treatment effect estimators without imposing additional assumptions. In particular, we show that when the response is binary, fractional, count, or some other non-negative outcome, certain kinds of nonlinear regression adjustment consistently estimates the average treatment effect. Our simulations for the case of binary and non-negative response suggest that nonlinear RA, especially the full version, can produce sampling variances that are substantially smaller than SDM and also linear regression adjustment. Plus, in terms of bias the nonlinear FRA (NFRA) methods are comparable to the SDM estimator, which we know is unbiased.

The rest of the paper is organized as follows. Section 2 briefly introduces the potential outcomes setting and defines the population average treatment effect – the parameter of interest in this paper. Section 3 discusses the random assignment mechanism and the random sampling assumption. Section 4 is important and describes linear regression adjustment in the population in terms of linear projections, which are consistently estimated by ordinary least squares (OLS) given a random sample. Importantly, we need not impose any assumptions on the conditional mean functions of the potential outcomes. Section 5 presents the asymptotic variances of the four linear estimators and ranks them on the basis of asymptotic efficiency. We also characterize the cases under which estimating two separate regressions does not improve efficiency over SDM or PRA (or both). Section 6 presents Monte Carlo simulations that compare the bias and root mean squared error (RMSE) of the estimators for eight different data generating processes. In section 7 we draw on results from the doubly robust estimation literature and characterize the nonlinear RA estimators – both pooled and full – that produce consistent estimators of the ATE. Our simulations in this section show that nonlinear methods have modest bias while considerably improving efficiency compared with both SDM and linear RA methods. Section 8 concludes the paper with a discussion of some future research topics. All proofs along with figures and tables are included in the appendix.

2 Potential Outcomes and Parameter of Interest

Our framework is the standard Neyman-Rubin causal model, involving potential (or counterfactual) outcomes. Let $\{Y(0), Y(1)\}$ be the two potential outcomes corresponding to the control and treatment states, respectively, where $\{Y(0), Y(1)\}$ has a joint distribution in the population. The setup is nonparametric in that we make no assumptions about the distribution of $\{Y(0), Y(1)\}$ other than finite moment conditions needed to apply standard asymptotic theory. In particular, $\{Y(0), Y(1)\}$ may be discrete, continuous, or mixed random variables. For example, $Y(0)$ and $Y(1)$ can be binary employment indicators for nonparticipation and participation in a job training program. Or, they could be the fraction of assets held in the stock market, or counts of the number of hospital visits taken by a patient.

Define the means of the potential outcomes as

$$\begin{aligned}\mu_0 &= \mathbb{E}\left[Y(0)\right]\\ \mu_1 &= \mathbb{E}\left[Y(1)\right]\end{aligned}$$

The parameter of interest is the population average treatment effect (PATE),

$$\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = \mu_1 - \mu_0$$

As has been often noted in the literature, the problem of causal inference is essentially a missing data problem. We only observe one of the the outcomes, $Y(0)$ or $Y(1)$, once the treatment, represented by the Bernoulli random variable W, is determined. Specifically, the observed Y is defined by

$$Y = \begin{cases} Y(0), & \text{if } W = 0 \\ Y(1), & \text{if } W = 1 \end{cases} \tag{1}$$

It is also useful to write Y as

$$Y = (1 - W) \cdot Y(0) + W \cdot Y(1). \tag{2}$$

3 Random Assignment and Random Sampling

In determining an appropriate method to estimate τ, we need to know how the treatment, W, is assigned. In this paper, we assume that W is independent of the potential outcomes as well as observed covariates, which we write as $\mathbf{X} = (X_1, X_2, \ldots, X_K)$. Formally, the random assignment assumption is as follows.

Assumption 3.1. *The binary assignment indicator, W, is a Bernoulli random variable and* *is independent of* $\{Y(0), Y(1), \mathbf{X}\}$*, where* $\mathbf{X} = (X_1, X_2, \ldots, X_K)$*. Mathematically,*

$$W \perp \{Y(0), Y(1), \mathbf{X}\}.$$

Letting $\rho = \mathbb{P}\,(W = 1)$ *be the probability of being assigned into treatment, assume that* $0 < \rho < 1$*.*

The assumption of random assignment implies that there are many consistent estimators of τ. The goal in this paper is to rank, as much as possible, commonly used estimators of τ in terms of asymptotic efficiency.

As mentioned in the introduction, both early [Neyman (1923) and Fisher (1935)] and recent [Freedman (2008a) Freedman (2008b) and Lin (2013)] approaches to estimating ATEs assume that the entire population is the sample. Therefore, the only stochastic element of the setup is the assignment of the treatment, which is randomized. Such a perspective rules out any uncertainty stemming from unobservability of the entire population (also termed sampling-based uncertainty) and only allows uncertainty that arises due to the experimental design (also known as design-based uncertainty). Here, we adopt the assumption commonly used in studying various estimators in statistics and econometrics.

Assumption 3.2. *For a nonrandom integer* N*,* $\{(Y_i(0), Y_i(1), W_i, \mathbf{X}_i) : i = 1, 2, \ldots, N\}$ *are independent and identically distributed draws from the population.*

Given the random sampling assumption, standard asymptotic theory for i.i.d. sequences of random vectors can be applied, where N tends to infinity. We assume in what follows that at least second moments of the potential outcomes and covariates are finite so that, when we use regression methods, we can apply the law of large numbers and central limit theorem. We do not state these moment assumptions explicitly.

For each unit i drawn from the population, the treatment effect is

$$Y_i(1) - Y_i(0),$$

which we can write as

$$Y_i(1) - Y_i(0) = \tau + [V_i(1) - V_i(0)]$$

where $V_i\,(w) = Y_i(w) - \mu_w$ for $w \in \{0, 1\}$. The treatment effects are homogeneous when the unit-specific components, $V_i(1) - V_i(0)$, are identically zero for any random draw i.

4 Estimators

We now carefully describe the estimators that we use in the linear regression context.

4.1 Simple Difference in Means (SDM)

Random assignment provides the luxury of using an estimator available from basic statistics. This estimator dates back to Neyman (1923) in the context of causal inference using potential outcomes. Let W_i be the treatment indicator for unit i. Then $N_0 = \sum_{i=1}^{N}(1 - W_i)$ and $N_1 = \sum_{i=1}^{N} W_i$ are the number of control and treated units in the sample, respectively. These are random variables. When N_0, $N_1 > 0$ we can define the sample averages for the control and treated units:

$$\bar{Y}_0 = N_0^{-1} \sum_{i=1}^{N} (1 - W_i) Y_i \tag{3}$$

$$\bar{Y}_1 = N_1^{-1} \sum_{i=1}^{N} W_i Y_i, \tag{4}$$

where Y_i is the observed outcome for unit i. The simple difference-in-means estimator is

$$\hat{\tau}_{SDM} = \bar{Y}_1 - \bar{Y}_0. \tag{5}$$

Under random assignment and conditional on N_0, $N_1 > 0$, $\hat{\tau}_{SDM}$ is unbiased for τ – see, for example, Imbens and Rubin (2015). Further, $\hat{\tau}_{SDM}$ is consistent as $N \to \infty$ for τ when $0 < \rho < 1$, as we assume. As is well know, $\hat{\tau}_{SDM}$ can be obtained as the coefficient on W_i in the simple regression

$$Y_i \text{ on } 1, W_i,\ i = 1, \ldots, N. \tag{6}$$

See, for example, Imbens and Rubin (2015).

4.2 Pooled Regression Adjustment (PRA)

When we have covariates that (hopefully) predict the outcome Y, the simplest way to use those covariates is to add them to the simple regression in (6). As documented in Słoczyński (2018), adding covariates along with a binary treatment indicator is still very common in estimating treatment effects, whether one has randomized assignment or assumes unconfoundedness conditional on the covariates. Specifically, the regression is

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i,\ i = 1, 2, \ldots, N.$$

The coefficient on W_i is the estimator of τ, and we called this estimator "pooled regression adjustment" (PRA) and denote it $\hat{\tau}_{PRA}$. The name "pooled" emphasizes that we are pooling across the control and treatment groups in imposing common coefficients on the vector of covariates, $\mathbf{X}_i$. In other words, the slopes are the same for $W = 0$ and $W = 1$. It is important to understand that we are making no assumption about whether the coefficients in an underlying linear model in the population are the same, or even whether there is an underlying linear model representing a conditional expectation. This will become clear in the next subsection when we formally describe linear projections.

As is well known, adding the variables $\mathbf{X}_i$ to the simple regression does not change the probability limit provided W_i and $\mathbf{X}_i$ are uncorrelated, which follows under random assignment. However, it is not always the case that adding $\mathbf{X}_i$ improves asymptotic efficiency: it can lead to a treatment effect estimator that actually has a larger asymptotic variance than the SDM estimator.

4.3 Linear Projections and Infeasible Regression Adjustment (IRA)

The SDM estimator is an example of an estimator that can be written as

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0,$$

where, in the case of SDM, $\hat{\mu}_0$ and $\hat{\mu}_1$ are the sample averages of the control and treated groups, respectively. But there are other ways to consistently estimate μ_0 and μ_1 when we have covariates $\mathbf{X}$, represented as a $1 \times K$ vector. In particular, define the linear projections of the potential outcomes on the vector of covariates as

$$\mathbb{L}\left[Y(0)|1, \mathbf{X}\right] = \alpha_0 + \mathbf{X}\boldsymbol{\beta}_0 \tag{7}$$

$$\mathbb{L}\left[Y(1)|1, \mathbf{X}\right] = \alpha_1 + \mathbf{X}\boldsymbol{\beta}_1, \tag{8}$$

where the expressions for α_0, α_1, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$ can be found in Wooldridge (2010) Section 2.3. As discussed in Wooldridge, the linear projections always exist and are well defined provided $Y(w)$ and the elements of $\mathbf{X}$ have finite second moments. Any of the random variables can be discrete, continuous, or mixed. The elements of $\mathbf{X}$ can include the usual functional forms, such as logarithms, squares, and interactions. The requirement for the coefficients in the linear projections (LPs) to be unique is simply that the variance-covariance matrix of $\mathbf{X}$, $\boldsymbol{\Omega}_{\mathbf{X}}$, is nonsingular, an assumption that rules out perfect collinearity in the population.

It is often helpful to slightly rewrite the LPs. Define the $1 \times K$ vector of population means of $\mathbf{X}$ as $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}\left(\mathbf{X}\right)$, and let $\dot{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$ be the deviations from the population mean. Then we can write the linear projection in terms of the means μ_0 and μ_1 as

$$\mathbb{L}\left[Y(0)|1, \mathbf{X}\right] = \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 \tag{9}$$

$$\mathbb{L}\left[Y(1)|1, \mathbf{X}\right] = \mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 \tag{10}$$

The two representations make it clear that the PATE, τ, can be expressed as

$$\tau = \mu_1 - \mu_0 = (\alpha_1 - \alpha_0) + \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

Therefore, if we have consistent estimators of α_0, α_1, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\mu}_{\mathbf{X}}$, then we can consistently estimate τ. Importantly, as discussed in Wooldridge (2010) Chapter 4, ordinary least squares estimation using a random sample always consistently estimates the parameters in a population linear projection (subject to the mild finite second moment assumptions and the non-singularity of $\boldsymbol{\Omega}_{\mathbf{X}}$). This is true regardless of the nature of $Y(w)$ or $\mathbf{X}$. This insight is critical for understanding why regression adjustment produces consistent estimators of τ, and for the asymptotic efficiency arguments later on. Unlike in Imbens and Wooldridge (2009), we do not assume that the linear projection is the same as the conditional mean. We are silent on the conditional mean functions $\mathbb{E}\left[Y(0)|\mathbf{X}\right]$ and $\mathbb{E}\left[Y(1)|\mathbf{X}\right]$.

Given random assignment, consistent estimators of the LP coefficients are obtained from the separate regressions

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 0 \tag{11}$$

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 1 \tag{12}$$

Wooldridge (2010) Chapter 19 formally shows that the linear projections are consistently estimated under the assumption that selection – in this case, W_i – is independent of $[\mathbf{X}_i, Y_i(w)]$. This is sometimes called the "missing completely at random" (MCAR) assumption in the missing data literature [for example, Little and Rubin (2002)].

If we assume that the vector of population means $\boldsymbol{\mu}_{\mathbf{X}}$ is known, a consistent estimator of τ is

$$\hat{\tau}_{IRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \boldsymbol{\mu}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0),$$

where $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\boldsymbol{\beta}}_1$ are the OLS estimators from the separate regressions. We call this the "infeasible regression adjustment" (IRA) estimator because it depends on $\boldsymbol{\mu}_{\mathbf{X}}$, which is likely to be unknown in our context with random sampling.

From the algebra of ordinary least squares (OLS) it is easily shown that $\hat{\tau}_{IRA}$ can be obtained as the coefficient on W_i in the regression that includes a full set of interactions between W_i and $\dot{\mathbf{X}}_i$, namely,

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i \cdot \dot{\mathbf{X}}_i,\ i = 1, \ldots, N.$$

The demeaning of the covariates ensures that the coefficient on W_i is $\hat{\tau}_{IRA}$. This regression is also convenient for obtaining a valid standard error for $\hat{\tau}_{IRA}$, as the usual Eicker-Huber-White heteroskedasticity-robust standard error is asymptotically valid.

In the case where the linear projections are also the conditional expectations – that is, $\mathbb{E}\left[Y(w)|\mathbf{X}\right] = \mathbb{L}\left[Y(w)|1, \mathbf{X}\right]$, $w \in \{0, 1\}$ – $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\boldsymbol{\beta}}_1$ are unbiased conditional on $\{\mathbf{X}_i : i = 1, 2, \ldots, N\}$, provided we rule out perfect collinearity in the control and treated subsamples. Then, $\hat{\tau}_{IRA}$ would also be unbiased conditional on $\{\mathbf{X}_i : i = 1, 2, \ldots, N\}$, and unbiased unconditionally if its expectation exists. But linearity of the conditional expectations is much too strong an assumption, and it is clearly not needed for unbiasedness or consistency of the

SDM estimator. Therefore, in what follows, we make no assumptions about $\mathbb{E}\left[Y(w)|\mathbf{X}\right]$. We simply assume enough moments are finite and rule out perfect collinearity in $\mathbf{X}$ in order for the linear projections to exist.

4.4 Full Regression Adjustment (FRA)

We can easily make the IRA estimator feasible by replacing $\boldsymbol{\mu}_{\mathbf{X}}$ with the sample average, $\bar{\mathbf{X}} = N^{-1}\sum_{i=1}^{N}\mathbf{X}_i$. This leads to what we will call the "full regression adjustment" (FRA) estimator:

$$\hat{\tau}_{FRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{X}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0).$$

This estimator can also be obtained as the OLS coefficient on W_i but from the regression

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i\cdot\ddot{\mathbf{X}}_i,\ i = 1,2,\ldots,N,$$

where

$$\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}},\ i = 1,2,\ldots,N$$

are the demeaned covariates using the sample average. This estimator is always available given a sample $\{(Y_i, W_i, \mathbf{X}_i) : i = 1,2,\ldots,N\}$. Generally, $\hat{\tau}_{FRA} \neq \hat{\tau}_{IRA}$. Like $\hat{\tau}_{IRA}$, we can only conclude that $\hat{\tau}_{FRA}$ is consistent, although it will be unbiased under essentially the same assumptions discussed for $\hat{\tau}_{IRA}$. In the next section, we will rank the four estimators, to the extent possible, in terms of asymptotic efficiency.

5 Asymptotic Variances and Efficiency Comparisons

We first derive the asymptotic variances of the SDM, PRA, IRA and FRA estimators in the general case of heterogeneous treatment effects. Naturally, the formulas include homogeneous treatment effects as a special case. We then compare the asymptotic variances in general and in special cases.

In order to obtain the asymptotic variances, we need to study the linear projections of the potential outcomes on the covariates more closely. Recall that we can write the potential outcomes as

$$Y(0) = \mu_0 + V(0) \tag{13}$$

$$Y(1) = \mu_1 + V(1), \tag{14}$$

where $V(0)$ and $V(1)$ have zero means, by construction. Following the discussion in Section 4, we linearly project each of $V(0)$ and $V(1)$ onto the population demeaned covariates, $\dot{\mathbf{X}}$:

$$V(0) = \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) \tag{15}$$

$$V(1) = \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1) \tag{16}$$

where the intercepts are necessarily zero. Then

$$Y(0) = \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) \tag{17}$$

$$Y(1) = \mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1) \tag{18}$$

By definition of the linear projection,

$$\mathbb{E}\left[U(0)\right] = \mathbb{E}\left[U(1)\right] = 0$$
$$\mathbb{E}\left[\dot{\mathbf{X}}'U(0)\right] = \mathbb{E}\left[\dot{\mathbf{X}}'U(1)\right] = \mathbf{0}$$

It follows that

$$\mathbb{V}\left[Y(0)\right] = \boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0 + \sigma_0^2$$
$$\mathbb{V}\left[Y(1)\right] = \boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1 + \sigma_1^2$$

where $\sigma_0^2 = \mathbb{V}\left[U(0)\right]$ and $\sigma_1^2 = \mathbb{V}\left[U(1)\right]$.

We can write the observed outcome, Y, as

$$Y = (1-W)\left[\mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0)\right] + W\left[\mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1)\right] \tag{19}$$

$$= \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) + \tau W + \left(W \cdot \dot{\mathbf{X}}\right)\boldsymbol{\delta} + W \cdot \left[U(1) - U(0)\right] \tag{20}$$

where $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$. The following lemma is a precursor to the efficiency comparisons.

Lemma 5.1. *Under the assumptions of random assignment given in 3.1, random sampling 3.2, and finite moment assumptions, the following asymptotic distributions hold:*

$$\sqrt{N}\left(\hat{\tau}_{SDM} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega_{SDM}^2\right) \tag{21}$$

$$\omega_{SDM}^2 = \frac{\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1}{\rho} + \frac{\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0}{(1-\rho)} + \frac{\sigma_1^2}{\rho} + \frac{\sigma_0^2}{(1-\rho)} \tag{22}$$

$$\sqrt{N}\left(\hat{\tau}_{PRA} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega_{PRA}^2\right) \tag{23}$$

$$\omega_{PRA}^2 = \left(\frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)}\right)\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + \frac{\sigma_1^2}{\rho} + \frac{\sigma_0^2}{(1-\rho)} \tag{24}$$

$$\sqrt{N}\left(\hat{\tau}_{FRA}-\tau\right)\xrightarrow{d}\mathcal{N}\left(0,\omega^2_{FRA}\right) \tag{25}$$

$$\omega^2_{FRA}=\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\frac{\sigma_1^2}{\rho}+\frac{\sigma_0^2}{(1-\rho)} \tag{26}$$

$$\sqrt{N}\left(\hat{\tau}_{IRA}-\tau\right)\xrightarrow{d}\mathcal{N}\left(0,\omega^2_{IRA}\right) \tag{27}$$

$$\omega^2_{IRA}=\frac{\sigma_1^2}{\rho}+\frac{\sigma_0^2}{(1-\rho)}.\ \square \tag{28}$$

The asymptotic variance expressions allow us to determine asymptotic efficiency under various scenarios. Not surprisingly, all four asymptotic variances depend on the error variances, σ_0^2 and σ_1^2. Generally, the asymptotic variances of the three feasible estimators can depend on $\boldsymbol{\Omega}_{\mathbf{X}}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$.

By comparing the formulas in Lemma 5.1 we have the following result, which ranks the asymptotic variances of the four different estimators in the general case of heterogeneous treatments and $\rho\in(0,1)$.

Theorem 5.2. *Under the assumptions of Lemma 5.1,*

(i)

$$\omega^2_{FRA}\le\omega^2_{SDM} \tag{29}$$

$$\omega^2_{FRA}\le\omega^2_{PRA} \tag{30}$$

$$\omega^2_{IRA}\le\omega^2_{FRA} \tag{31}$$

(ii) If $\boldsymbol{\beta}_0=\boldsymbol{\beta}_1=\mathbf{0}$ then all asymptotic variances are the same.

(iii) If $\boldsymbol{\beta}_0=\boldsymbol{\beta}_1=\boldsymbol{\beta}$ then $\omega^2_{PRA}=\omega^2_{FRA}=\omega^2_{IRA}$ and if $\boldsymbol{\beta}\neq\mathbf{0}$ then ω^2_{SDM} is strictly larger.

(iv) If $\rho=1/2$ then $\omega^2_{PRA}=\omega^2_{FRA}\le\omega^2_{SDM}$, with strict inequality in the latter case unless $\boldsymbol{\beta}_1=-\boldsymbol{\beta}_0$.

Many of the results in Theorem 5.2 follow from inspection of the asymptotic variance formulas, although some are more subtle. For example, (31) is immediate because the first term in (26) is nonnegative. Part (iii) is also immediate because all asymptotic variances equal $\sigma_1^2/\rho+\sigma_0^2/(1-\rho)$ when $\boldsymbol{\beta}_0=\boldsymbol{\beta}_1$. For part (iv), the function

$$g(\rho)\equiv\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)},\rho\in(0,1)$$

can be shown to have a minimum value of unity, uniquely achieved when $\rho = 1/2$.

The most difficult inequality to establish, and the one that is most important, is (29). Straightforward matrix multiplication shows that

$$\omega_{SDM}^2 - \omega_{FRA}^2 = \frac{\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1}{\rho} + \frac{\boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0}{(1-\rho)} - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \boldsymbol{\Omega}_{\mathbf{X}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) = \boldsymbol{\lambda}' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\lambda},$$

where

$$\boldsymbol{\lambda} = \sqrt{\left(\frac{1-\rho}{\rho}\right)} \boldsymbol{\beta}_1 + \sqrt{\left(\frac{\rho}{1-\rho}\right)} \boldsymbol{\beta}_0.$$

Because $\boldsymbol{\Omega}_{\mathbf{X}}$ is assumed positive definite, $\omega_{SDM}^2 = \omega_{FRA}^2$ if and only if $\boldsymbol{\lambda} = \mathbf{0}$. One case where $\boldsymbol{\lambda} = \mathbf{0}$ is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \mathbf{0}$, in which case the covariates do not predict the potential outcomes. It can happen in other cases but all of the slope coefficients would have to have opposite signs in the linear projections of the two potential outcomes. For example, if $\rho = 1/2$, we would need $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$, which means the slopes in the linear projection of $Y(1)$ on $1, \mathbf{X}$ would be the opposite signs of the slope coefficients in the linear projection of $Y(0)$ on $1, \mathbf{X}$. This seems highly unlikely. For example, we would expect pre-training education to have a positive effect on earnings whether or not someone participates in a job training program. In the homogenous case $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 \neq \mathbf{0}$, $\omega_{SDM}^2 > \omega_{FRA}^2 = \omega_{PRA}^2$.

We can never know for sure whether $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$, but we should know whether $\rho = 1/2$ based on the design of the experiment. If $\rho = 1/2$ then 5.2 suggests that the pooled estimator is probably preferred: it is as asymptotically efficient as the full RA estimator and conserves on degrees of freedom, which may be important if N is not large and the potential K (number of covariates) is somewhat large. For $\rho \neq 1/2$, Theorem 5.2 shows that the full RA estimator is attractive provided small-sample issues are not important. In particular, $\hat{\tau}_{FRA}$ is always more asymptotically efficient that both $\hat{\tau}_{SDM}$ and $\hat{\tau}_{PRA}$ in the presence of heterogenous slopes, and there is no (asymptotic) price to pay if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$ or even if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \mathbf{0}$. Estimating the $2K$ parameters is, asymptotically, harmless when it comes to the precision in estimating τ.

It may be helpful to provide intuition as to why $\hat{\tau}_{FRA}$ is more efficient than $\hat{\tau}_{SDM}$. Consider estimating the mean of the potential outcome in the treated state, μ_1. The FRA estimator is

$$\hat{\mu}_{1,FRA} = \hat{\alpha}_1 + \bar{\mathbf{X}} \hat{\boldsymbol{\beta}}_1,$$

where $\bar{\mathbf{X}}$ is the sample average across the entire sample. By the simple mechanics of OLS,

$$\bar{Y}_1 = \hat{\alpha}_1 + \bar{\mathbf{X}}_1 \hat{\boldsymbol{\beta}}_1$$

where

$$\bar{\mathbf{X}}_1 = N_1^{-1} \sum_{i=1}^{N} W_i \mathbf{X}_i$$

is the sample average of the $\mathbf{X}_i$ over the treated units. Because of random assignment, $\bar{\mathbf{X}}_1$ is also a $\sqrt{N}$-consistent, asymptotically normal estimator of $\boldsymbol{\mu}_{\mathbf{X}}$. But it is inefficient compared with $\bar{\mathbf{X}}$ because the latter uses the entire sample. The same is true of $\hat{\mu}_{0,FRA}$ and $\bar{Y}_0$. This does not quite prove that $\hat{\tau}_{FRA}$ is asymptotically more efficient because $\hat{\mu}_{1,FRA}$ and $\hat{\mu}_{0,FRA}$ are not (asymptotically) uncorrelated, but it does provide some intuition. The same sort of intuition indicates why $\hat{\tau}_{IRA}$ is asymptotically more efficient than $\hat{\tau}_{FRA}$: $\hat{\tau}_{IRA}$ is not subject to the sampling error in estimating $\boldsymbol{\mu}_{\mathbf{X}}$.

6 Simulations

In this section we study the finite sample properties of the four estimators just discussed. We evaluate the estimators primarily in terms of root mean squared error (RMSE), since this accounts for bias as well as sampling variance. Since bias has been cited as a concern with covariate adjustment estimators, especially in small-scale experiments, looking at the trade offs between bias and efficiency through RMSE is key to studying the finite sample performance of these estimators. In order to compute the RMSE, we generate a population of 10,000 observations corresponding to each data generating process, outlined below, to approximate an "infinite" population setting. We then repeatedly draw random samples of sizes 100, 500 and 1,000 from the population a thousand times. For a comprehensive assessment, we report the RMSE across these different sample sizes and treatment probability combinations where the treatment probabilities range from 0.1 to 0.9. To keep the tables simple, we report results only for the odd treatment probabilities even though the graphs are plotted for all values. The reported simulation results are for the case of heterogeneous treatment effects in the population, both in terms of the slopes on the linear projections and in the distribution of the projection errors, $U(0)$ and $U(1)$.

6.1 Design Details

The treatment, W, is a binary variable, and so it has a Bernoulli distribution with

$$\mathbb{P}\left(W = 1\right) = \rho,$$

and we vary the value of ρ. For the potential outcomes, we consider continuous and discrete responses. In the first, the potential outcomes are conditionally normally distributed, with means linear in a quadratic in two covariates, X_1 and X_2. Specifically,

$$\begin{aligned}
Y(0) &= \gamma_{00} + \gamma_{01}X_1 + \gamma_{02}X_2 + \gamma_{03}X_1^2 + \gamma_{04}X_2^2 + \gamma_{05}X_1X_2 + R(0) \equiv \mathbf{Z}\boldsymbol{\gamma}_0 + R(0) \\
Y(1) &= \gamma_{10} + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_1^2 + \gamma_{14}X_2^2 + \gamma_{15}X_1X_2 + R(1) \equiv \mathbf{Z}\boldsymbol{\gamma}_1 + R(1),
\end{aligned}$$

where

$$\mathbf{Z} = \begin{pmatrix} 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1X_2 \end{pmatrix}$$

and

$$\begin{aligned} R(0)|\,(X_1, X_2) &\sim \mathcal{N}(0, \sigma_0^2) \\ R(1)|\,(X_1, X_2) &\sim \mathcal{N}(0, \sigma_1^2) \end{aligned}$$

We allow the γ_{wj} to differ across $w \in \{0, 1\}$, and so there is heterogeneity in the treatment effects in terms of the observables, $\mathbf{X}$, and the unobservables, $R(0)$ and $R(1)$ which are allowed to be correlated.[2] We also allow σ_0^2 and σ_1^2 to differ.

It is important to understand that, in order to be realistic, we do not assume that the quadratic conditional mean function is known. Instead, the researcher uses only linear regression on a constant, X_1, and X_2. In a traditional view of econometrics, these regressions would be "misspecified." Of course, to ensure we have the best mean squared error predictors of $Y(0)$ and $Y(1)$, we would use the correct specifications of $\mathbb{E}\left[Y(0)|\mathbf{X}\right]$ and $\mathbb{E}\left[Y(1)|\mathbf{X}\right]$. But it would be unusual for us to know the exact specification of the conditional mean functions. One can argue that most empirical researchers would include simple functions, such as squares as interactions. But then the true mean function could depend on higher order polynomials, or other more exotic functions. In fact, the mean might not even be linear in parameters. We take our setup as reflecting the realistic case that the researcher uses a linear regression that does not correspond to the correct conditional mean.

Our second design generates the potential outcomes as binary variables. Remember, when W is randomly assigned, we can use any kind of linear regression adjustment to improve asymptotic efficiency, regardless of the nature of $Y(0)$, $Y(1)$. Specifically, for $\mathbf{Z}$ defined above,

$$\begin{aligned} Y(0) &= 1[\mathbf{Z}\boldsymbol{\gamma}_0 + R(0) > 0] \\ Y(1) &= 1[\mathbf{Z}\boldsymbol{\gamma}_1 + R(1) > 0], \end{aligned}$$

where $R(0)$ and $R(1)$ are again independent of (X_1, X_2) and normally distributed, this time each with unit variances. As before, $R(0)$ and $R(1)$ are allowed to be correlated. In the binary response case, one might traditionally think of two forms of "misspecification" in using linear regression adjustment on $(1, X_1, X_2)$. First, we are using what is traditionally called a "linear probability model" rather than the correct probit model. Second, we are omitting the terms X_1^2, X_2^2, and X_1X_2. Thus, there are two kinds of functional form "misspecification." Our view is that, to make a case for linear regression adjustment, it should produce notable efficiency gains even when the potential outcomes are discrete (although we return to this issue in the next section).

We consider two different designs for generating the covariates. Both are based on an

[2] $R(0)$ and $R(1)$ are generated to be affine transformations of the same standard normal variable.

underlying bivariate normal distribution:

$$\mathbf{X}^{*\prime} = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}\right].$$

In the first design, $\mathbf{X} = \mathbf{X}^*$. In the second design, $X_1 = X_1^*$ and

$$X_2 = 1[X_2^* > 0],$$

so that X_2 is binary (in which case X_2^2 is redundant in the mechanism generating the potential outcomes).

With the linear and probit data we consider two different levels of heterogeneity across the coefficients $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$, which we label "mild" and "strong." For the mild heterogeneity with continuous covariates

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 2 & 2 & -2 & -0.05 & -0.02 & 0.3 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & 1 & -1 & -0.05 & -0.03 & 0.6 \end{pmatrix}$$

and for the strong heterogeneity

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -1 & -0.05 & 0.02 & 0.6 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 1 & -1 & 1.5 & 0.03 & -0.02 & -0.6 \end{pmatrix}$$

With the binary regressor, for mild heterogeneity we have

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -1 & 0.05 & 0.2 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & 3 & 1 & 0.05 & 0.9 \end{pmatrix}$$

and for strong heterogeneity we have

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -2 & -0.05 & 0.2 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & -1 & 1 & 0.05 & -0.9 \end{pmatrix}$$

Combining the linear and probit designs for the potential outcomes with two different levels of heterogeneity and two different covariate compositions leads to a total of eight scenarios. We allow the treatment probability, ρ, to range between 0.1 and 0.9 in increments of 0.1 . We consider three sample sizes, 100, 500, and 1,000. Note that when $N = 100$ and $\rho = 0.1$, we expect only 10 treated units and 90 control units. The covariates are generated to ensure that they are predictive of the potential outcomes, with the population R-squared ranging between $(0.1, 0.6)$. The variances σ_0^2 and σ_1^2 are allowed to be different for the two potential outcomes and across four of the eight different data generation processes.

We assess the relative finite sample performance of the four estimators under each such scenario, which we term a DGP. The Table below describes each of the DGP's in detail.

Table 6.1: Description of the Data Generating Processes

DGP	Design	Covariates	Heterogeneity	R_0^2	R_1^2	σ_0^2	σ_1^2	PATE
1	Quadratic	$\mathbf{X}$	Mild	0.52	0.44	16	9	2.68
2	Quadratic	$\mathbf{X}$	Strong	0.31	0.46	16	9	0.93
3	Quadratic	$X_2 = 1[X_2^* > 0]$	Mild	0.59	0.33	1	4	7.46
4	Quadratic	$X_2 = 1[X_2^* > 0]$	Strong	0.27	0.34	9	4	2.92
5	Probit	$\mathbf{X}$	Mild	0.59	0.38	1	1	0.28
6	Probit	$\mathbf{X}$	Strong	0.51	0.45	1	1	0.09
7	Probit	$X_2 = 1[X_2^* > 0]$	Mild	0.45	0.28	1	1	0.35
8	Probit	$X_2 = 1[X_2^* > 0]$	Strong	0.38	0.40	1	1	0.43

6.2 Discussion of Simulation Findings

In the eight different DGPs, we see that FRA performs better than SDM and PRA in terms of RMSE. This behavior seems to be more pronounced at larger sample sizes as seen from the figures. Two things are worth pointing. One, the difference in IRA and FRA is less prominent for cases of mild heterogeneity. In such cases, PRA also performs comparably. This makes sense since pooling slopes in the treatment and control groups when the slopes are not very different should produces estimates that are close to the ones estimated by the separate slopes regression. Second, as was clear from Theorem 5.2, PRA and FRA have approximately the same RMSE at $\rho = 0.5$. This is not surprising to see in the graph because at larger sample sizes, biases in these estimators are negligible which means that RMSE is approximately the same as the variance.

Overall we see that the finite sample performance of FRA is superior to SDM and PRA for a variety of data generating processes (see figures 1 and 2 for quadratic design with mild and strong levels of heterogeneity, 3 and 4 for quadratic design with one binary covariate with mild and strong levels of heterogeneity, 5 and 6 for a probit design with mild and strong levels of heterogeneity and finally 7 and 8 for a probit design with one binary covariate with mild and strong levels of heterogeneity. For tables, see 1, 2 and 3).

7 Nonlinear Regression Adjustment

If the outcome Y – more precisely, the potential outcomes, $Y(0)$ and $Y(1)$ – are discrete, or have limited support, using nonlinear conditional mean functions, chosen to ensure fitted values are logically consistent with $\mathbb{E}[Y|\mathbf{X}]$, have considerable appeal. Intuitively, getting better approximations to $\mathbb{E}[Y(0)|\mathbf{X}]$ and $\mathbb{E}[Y(1)|\mathbf{X}]$ can yield estimators with smaller asymptotic variances when compared with the SDM estimator and linear regression adjustment. However, as cautioned by Imbens and Rubin (2015) page 128, one should not sacrifice consistency in order to obtain an asymptotically more efficient estimator. Imbens and Rubin (2015) leave the impression that all nonlinear models should be avoided because consistency cannot be ensured.

In this section, we use the features of the linear exponential family class of distributions, combined with particular conditional mean models, to show that if one is careful in choosing the combination of conditional mean function and quasi-log likelihood (QLL) function, one can preserve consistency. Unfortunately, we cannot formally show that using this particular set of nonlinear models is more efficient than the SDM estimator, but our simulations suggest the efficiency gains can be substantial. (And we have found no cases where it is worse to do nonlinear RA.)

In deciding on nonlinear RA methods, the key is to remember is that

$$\tau_{ate} = \mu_1 - \mu_0,$$

and so we need to consistently estimate μ_1 and μ_0 without imposing additional assumptions. Earlier we showed how linear regression adjustment does just that. And, linear RA, when done separately to estimate μ_0 and μ_1, is asymptotically more efficient than the SDM estimator. Our goal here is to summarize the nonlinear methods that produce consistent estimators of τ_{ate} without additional assumptions (except for standard regularity conditions). We start with pooled methods.

7.1 Pooled Nonlinear Regression Adjustment

In the generalized linear models (GLM) literature, it is well known that certain combinations of QLLs in the linear exponential family (LEF) and link functions lead to first order conditions where, in the sample, the residuals average to zero and are uncorrelated with every explanatory variable. To state the precise results, let $g(\cdot)$ be a strictly increasing function on $\mathbb{R}$, with range that can be a subset of $\mathbb{R}$. The inverse, $g^{-1}(\cdot)$, is known as the "link function" in the GLM literature. In the context of treatment effect estimation with mean function $g(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)$, when using the so-called canonical link function [McCullagh and Nelder (1989)] the first order conditions (FOCs) are of the form

$$\sum_{i=1}^{N}\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = 0$$
$$\sum_{i=1}^{N} W_i\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = 0$$
$$\sum_{i=1}^{N} \mathbf{X}_i'\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = \mathbf{0}$$

When $g(z) = z$, these equations produce the first order conditions for the pooled OLS estimator. The leading cases where these conditions hold for nonlinear estimation are for the Bernoulli QLL when $g(z) = \Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logistic function and for the Poisson QLL when $g(z) = \exp(z)$.

Under random sampling and weak regularity conditions, the probability limits of the estimators solve the population versions of the sample moment conditions. Let α^*, $\boldsymbol{\beta}^*$, and γ^* denote the probability limits of $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}$, respectively. Importantly, as argued in White (1982), these plims exist very generally without assuming that mean function is correctly specified – just as the parameters in the linear projection exist under very weak assumptions. The first two FOCs in the population are

$$\mathbb{E}\left[Y - g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right] = 0 \tag{32}$$

$$\mathbb{E}\left\{W\left[Y - g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right]\right\} = 0; \tag{33}$$

we will not need the last set of conditions obtained from the gradient with respect to $\boldsymbol{\beta}$. As before, we assume that $\rho = \mathbb{P}\left(W = 1\right)$ satisfies $0 < \rho < 1$.

Now, recall that $Y = (1 - W)Y(0) + WY(1)$. Then, by random assignment,

$$\mathbb{E}\left(Y\right) = \mathbb{E}\left(1 - W\right)\mathbb{E}\left[Y(0)\right] + \mathbb{E}\left(W\right)\mathbb{E}\left[Y(1)\right] = (1 - \rho)\mu_0 + \rho\mu_1.$$

Therefore, we can write (32) as

$$(1 - \rho)\mu_0 + \rho\mu_1 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right]$$

By iterated expectations,

$$\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right] = \mathbb{E}\left\{\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)|\mathbf{X}\right]\right\}$$

and, because W is independent of $\mathbf{X}$ with $\mathbb{P}\left(W = 1\right) = \rho$,

$$\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)|\mathbf{X}\right] = (1 - \rho)g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right) + \rho g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right).$$

Therefore,

$$(1 - \rho)\mu_0 + \rho\mu_1 = (1 - \rho)\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right] + \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right]. \tag{34}$$

Further, using and $WY = WY(1)$, from (33),

$$\mathbb{E}\left[WY(1)\right] = \mathbb{E}\left[Wg\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right].$$

Again using random assignment and iterated expectations,

$$\rho\mu_1 = (1 - \rho) \cdot 0 + \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right] = \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right].$$

Because $\rho > 0$, we have

$$\mu_1 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right]. \tag{35}$$

Also, because $\rho < 1$, (34) now implies

$$\mu_0 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right] \tag{36}$$

It follows that

$$\tau_{ate} = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right] - \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right]. \tag{37}$$

This equation essentially is the basis for proving that pooled regression adjustment, where we use a QLL in the linear exponential family and the conditional mean implied by the canonical link function, is consistent. Consistency follows because the estimated ATE using the pooled method is

$$\hat{\tau}_{ate,pooled} = N^{-1}\sum_{i=1}^{N}\left[g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}\right) - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]. \tag{38}$$

By Wooldridge (2010) question 12.17, this converges in probability to (37). As a practical matter, it is convenient to note that (38) is the exact quantity reported by standard software packages when one requests the average "marginal" (or "partial") effect of the binary variable W after a standard GLM estimation. Packages that have this pre-programmed also provide a valid standard error, although one must be sure to use a "robust" option during the GLM estimation so that a sandwich estimator is used for the asymptotic variance of the parameter estimators $\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}', \hat{\gamma}\right)'$.

Table 7.1 summarizes the useful combinations of the QLL and mean functions that lead to consistent estimation of τ_{ate} without additional assumptions.

Table 7.1: QLL and mean function combinations

Restrictions on Support of Response	Quasi-Log Likelihood Function	Conditional Mean Function
None	Gaussian (Normal)	Linear
$Y(w) \in [0,1]$	Bernoulli	Logistic
$Y(w) \in [0,B]$	Binomial	Logistic
$Y(w) \geq 0$	Poisson	Exponential
$Y_g(w) \geq 0$, $\sum_{g=0}^{G} Y_g(w) = 1$	Multinomial	Logistic

The Bernoulli/logistic case applies to binary or fractional outcomes, without change. When Y is fractional, it can have probability mass at zero, one, or anywhere else. See Papke and Wooldridge (1996) for further discussion. In any case, we treat the problem as quasi-MLE because we do not wish to assume either that the distribution or mean function is correct.

The Poisson/exponential combination is very useful for nonnegative outcomes without a natural upper bound, although it can be applied to any nonnegative outcome. This includes count outcomes but also continuous outcomes and outcomes with corner solutions at zero (or other focal points). In the latter case, it is important to understand that commonly used models, such as Tobit, do not provide any known robustness to misspecification of the Tobit model. By contrast, the Poisson QMLE with an exponential mean provides full robustness. Remember, we are not trying to estimate the conditional mean functions; we are trying to consistently estimate the unconditional means, μ_0 and μ_1. Other than linear regression, the Poisson QMLE with an exponential mean is the only sensible choice for nonnegative, unbounded responses.

If the outcome has a natural, known upper bound, say B_i, which may vary by unit i, the binomial QMLE can be used in conjunction with the mean function

$$m(b, \mathbf{x}, w, \boldsymbol{\theta}) = b\left[\frac{\exp(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)}{1 + \exp(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)}\right],$$

as this is known to be the mean associated with the canonical link for the binomial distribution. The data then consists of $(Y_i, B_i, \mathbf{X}_i, W_i)$. Again, it does not matter whether Y_i is an integer response or is continuous, or even has a corner at zero, B_i, or both: using the binomial QMLE with logistic mean is simply a way to possibly improve over SDM or linear RA. The estimated ATE is

$$N^{-1}\sum_{i=1}^{N} B_i\left[\frac{\exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma})}{1 + \exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma})} - \frac{\exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}})}{1 + \exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}})}\right];$$

again, this is simple the average partial effect with respect to the binary variable W.

The last entry in Table 7.1 extends the Bernoulli QLL/logistic mean and is relevant in two general situations. The first is when the support of the response is finite (and greater than two; otherwise one would use the logistic mean function with the Bernoulli QLL). For example, $Y_g(w)$ could be an ordered response, such as a measure of health on a Lichert scale, or an unordered response, such as the choice of a health plan. A second situation is when the response consists of fractions that sum to unity, such as proportions of wealth in different investment categories, in which case the model has been called “multinomial fractional logit” [Mullahy (2015)]. If there are $G+1$ possible outcomes then there are $G+1$ means each for the control and treated states. The conditional mean functions for a pooled estimation would be

$$m_g(\mathbf{x}, w, \theta) = \frac{\exp(\alpha_g + \mathbf{x}\boldsymbol{\beta}_g + \gamma_g w)}{\left[1 + \sum_{h=1}^{G} \exp(\alpha_h + \mathbf{x}\boldsymbol{\beta}_h + \gamma_h w)\right]},\ g = 0, 1, ..., G$$

with $\alpha_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$. Then the estimated means are

$$\hat{\mu}_{wg} = N^{-1} \sum_{i=1}^{N} \frac{\exp(\hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_g + \hat{\gamma}_g w)}{\left[1 + \sum_{h=1}^{G} \exp(\hat{\alpha}_h + \mathbf{X}_i\hat{\boldsymbol{\beta}}_h + \hat{\gamma}_h w)\right]},\ w \in \{0, 1\},\ g \in \{0, 1, ..., G\}$$

and the estimated average treatment effect for each g is

$$\hat{\tau}_{ate,g} = \hat{\mu}_{1g} - \hat{\mu}_{0g}.$$

Because for each w, $\sum_{g=0}^{G} \hat{\mu}_{wg} = 1$, the sum over g of the $\hat{\tau}_{ate,g}$ is necessarily zero.

7.2 Full Nonlinear Regression Adjustment

As in the linear case, consistency is preserved if we estimate two separate regression functions for the control and treatment cases. This follows from Wooldridge (2007) results on doubly robust estimators, where, in the current setting, the propensity score, $\mathbb{P}(W = 1|\mathbf{X} = \mathbf{x}) = \rho$, is not a function of $\mathbf{x}$. But a direct argument is easier to follow. For example, consider using a QLL with the canonical link function using only the treatments. The FOC for $\hat{\alpha}_1$, the intercept inside the conditional mean function, is simply

$$\sum_{i=1}^{N} W_i \left[Y_i - g\left(\hat{\alpha}_1 + \mathbf{X}_i\hat{\boldsymbol{\beta}}_1\right)\right] = 0.$$

Notice again how the treatment indicator serves to select the subsample of treated units. The population analog is

$$\mathbb{E}\left[WY(1)\right] = \mathbb{E}\left[Wg\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^*\right)\right]$$

or, because of random assignment,

$$\rho\mu_1 = \rho\mathbb{E}\left[g\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^*\right)\right].$$

It follows that

$$\mu_1 = \mathbb{E}\left[g\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^*\right)\right]$$

The same argument works for the untreated case, where W_i is replaced with $(1 - W_i)$, and $\left(\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1'\right)'$ are replaced with $\left(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0'\right)'$. The conclusion is

$$\mu_0 = \mathbb{E}\left[g\left(\alpha_0^* + \mathbf{X}\boldsymbol{\beta}_0^*\right)\right].$$

Remember, the parameters with a "*" now indicate the probability limits from the two separate estimations, rather than there being the same parameters as in the pooled estimation. It follows under general regularity conditions that a consistent and asymptotically normal estimator of τ_{ate} is

$$\hat{\tau}_{ate,full} = N^{-1} \sum_{i=1}^{N} \left[g\left(\hat{\alpha}_1 + \mathbf{X}_i\hat{\boldsymbol{\beta}}_1\right) - g\left(\hat{\alpha}_0 + \mathbf{X}_i\hat{\boldsymbol{\beta}}_0\right)\right].$$

As a practical matter, some packages, such as Stata, have built-in commands for some full RA estimators, including the Bernoulli/logistic and the Poisson/exponential combinations, and so a standard error is computed along with the estimate. Again, one must be sure to use a robust variance matrix estimator for the parameters. Alternatively, using a bootstrap routine is very efficient for these kinds of estimators.

In deciding on a procedure to use – linear versus nonlinear, pooled versus full – it is important to understand that all methods studied in this paper produce consistent estimators of τ_{ate}. In the linear case, we have the result that full RA is asymptotically efficient compared with SDM and pooled RA. As mentioned earlier, a proof that full nonlinear RA is asymptotically more efficient than the pooled version is elusive. Also, we have not proven that full nonlinear RA is always at least as asymptotically efficient as SDM. We now report representative simulations that show the nonlinear methods can improve precision substantially in some cases without introducing bias, even in pretty small sample sizes.

7.3 Simulations

For non-linear simulations we only consider continuous covariates which means that for both binary and non-negative data generating processes, $\mathbf{X} = \mathbf{X}^*$ where,

$$\mathbf{X}^{*\prime} = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}\right].$$

As with the linear simulations,

$$\mathbf{Z} = \begin{pmatrix} 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1X_2 \end{pmatrix}.$$

The tables report bias and standard deviation for sample sizes of $N = 500$ and $N = 1,000$ which are drawn repeatedly thousand times from a population of size 10,000 as in the linear regression adjustment simulations. To keep the tables simple, we only report bias and standard deviation values for treatment probabilities $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$ even though the graphs are plotted for ρ ranging between 0.1 to 0.9.

7.3.1 Binary Response

For the binary response case, the outcomes have been generated using a probit mean function as given below

$$\begin{aligned} Y(0) &= 1[\mathbf{Z}\boldsymbol{\gamma}_0 + R(0) > 0] \\ Y(1) &= 1[\mathbf{Z}\boldsymbol{\gamma}_1 + R(1) > 0], \end{aligned}$$

and

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} -2 & 1 & 2 & 0.05 & 0.02 & 0.1 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 0 & 3 & 1 & -0.05 & 0.03 & 0.9 \end{pmatrix}$$

and

$$\begin{aligned} R(0)|\,(X_1, X_2) &\sim \mathcal{N}(0,1) \\ R(1)|\,(X_1, X_2) &\sim \mathcal{N}(0,1) \end{aligned}$$

where $R(0)$ and $R(1)$ are allowed to be correlated like before. While estimating the nonlinear pooled and separate slopes estimators we use case (ii) in Table 7.1.

We find that separate slopes nonlinear estimator (NFRA) has the lowest root mean squared error compared with the linear estimators and the pooled nonlinear estimator (NPRA) for all treatment probabilities (see figure 9). The tables show that nonlinear estimators have bias that is comparable to the SDM estimator (see table 4).

7.3.2 Nonnegative Response

For the non-negative response, the outcomes have been generated using a log normal distribution as given below

$$\begin{aligned} Y(0) &= \exp\left(\frac{\mathbf{Z}\boldsymbol{\gamma_0} + R(0)}{10} + 0.3 \cdot \mathcal{N}(0,1)\right) \\ Y(1) &= \exp\left(\frac{\mathbf{Z}\boldsymbol{\gamma_1} + R(1)}{10} + 0.4 \cdot \mathcal{N}(0,1)\right), \end{aligned}$$

where

$$\gamma_0' = \begin{pmatrix} 0 & 2 & -1 & -0.05 & 0.02 & 0.6 \end{pmatrix}, \quad \gamma_1' = \begin{pmatrix} 1 & -1 & 1.5 & 0.03 & -0.02 & -0.6 \end{pmatrix}$$

and

$$R(0) | (X_1, X_2) \sim \mathcal{N}(0, 4)$$
$$R(1) | (X_1, X_2) \sim \mathcal{N}(0, 9)$$

where $R(0)$ and $R(1)$ are allowed to be correlated. While estimating the nonlinear pooled and separate slopes estimators we use case (iv) in Table 7.1.

Similar to the binary response simulations, we see that that NFRA again has the lowest root mean squared error compared with both linear and pooled nonlinear estimators across all treatment probabilities. In fact, NFRA peforms better than both SDM and FRA. The NPRA and linear PRA are very similar in terms of RMSE; see table 5 and figure 10.

8 Concluding Remarks

We have studied linear and nonlinear regression adjustment estimators of the average treatment effect in an experimental framework. For linear regression adjustment, this paper makes some key contributions to the econometrics literature on randomized experiments. First, by considering a previously ignored aspect of the separate slopes estimator, this paper is able to fill a clear gap in the literature by showing the full RA estimator is always the most efficient even when the population means of the covariates is estimated using the sample sample. Second, in obtaining our results, we rely only on linear projections, and so no extra assumptions are used in establishing asymptotic efficiency. Third, the setup allows us to determine when using full RA, or RA at all, is unecessary to achieve efficiency. Our simulation findings support the theory and show that substantial efficiency gains are possible when we have good predictors of the response. Obtaining the correct standard errors for the full RA estimator is particularly simple in commonly used software packages. For example, Stata®, with its built-in "teffects" command, provides the correct standard errors for the FRA estimator

As an interesting complement to our work, Słoczyński (2018) studies pooled versus full RA when assignment is unconfounded conditional on covariates. Assuming that the conditional means are linear in parameters, Słoczyński (2018) shows that using pooled RA when the treatment effects are heterogeneous is inconsistent for the ATE in a way that is particularly troublesome in designs that are heavily unbalanced. In particular, the pooled RA estimator consistently estimates the weighted average $(1 - \rho) \cdot \tau_{ATT} + \rho \cdot \tau_{ATU}$, where τ_{ATT} is the average treatment effect on the treated ($W = 1$) and τ_{ATU} is the ATE on the untreated ($W = 0$). The ATE can be expressed as $\tau_{ATE} = \rho \cdot \tau_{ATT} + (1 - \rho) \cdot \tau_{ATU}$, and so the PRA estimator, in the limit, gets the weights reversed. Under random assignment, there is no difference between τ_{ATE}, τ_{ATT}, and τ_{ATU}, and so consistency of PRA for τ_{ATE} is not the issue. But as we showed, the pooled RA estimator is generally inefficient when treatment effects are heterogeneous. Also, when $\rho = 1/2$, there is no inconsistency in the pooled RA estimator when unconfoundedness holds. As we have shown in this paper, in the random assignment case $\rho = 1/2$ is precisely the condition that implies no efficiency gain from full RA even when there is arbitrary heterogeneity in the treatment effects. Our findings mesh well with those of Słoczyński (2018), with the conclusion that in moderate samples, FRA should be used unless ρ is known to be close to $1/2$.

In addition to the linear estimators, we also propose nonlinear regression adjustment estimators, characterizing the combinations of quasi-log-likelihood functions and conditional means functions that ensure consistency regardless of misspecification. We believe this paper is the first to do so. We do not have theoretical results to show when the nonlinear RA methods unambiguously improve asymptotic efficiency, and this is a good topic for future research. However, our simulations suggest that nonlinear adjustment estimators can have bias comparable to that of simple difference in means (SDM) and can produce sampling variances that are considerably smaller than that of SDM and, in majority of cases, substantially smaller than linear feasible regression adjustment.

Going forward, there are a lot of natural extensions. One is to study an assignment scheme that is different from the one considered here. This paper assumes independence across treatment assignments but a more common design, known as the completely randomized experiment, induces dependence across units by fixing the number of treated units before sampling from the population. Also, because most randomized experiments in economics are plagued with issues of nonparticipation or nonrandom attrition, it is also fruitful to study regression adjustment in conjunction with an Instrumental Variables (IV). Comparing the efficiency of standard regression adjustment estimators under random assignment to estimators based on stratified assignment schemes is also a good area for future research.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. When Should You Adjust Standard Errors for Clustering? Technical report, National Bureau of Economic Research, 2017a.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. Sampling-based vs. Design-based Uncertainty in Regression Analysis. *Working Paper*, 2017b.

Ansel, J., Hong, H., and Li, J. OLS and 2SLS in Randomized and Conditionally Randomized Experiments. *Jahrbücher für Nationalökonomie und Statistik*, 238(3-4):243–293, 2018.

Bartlett, J. W. Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical statistics*, 17(5):648–666, 2018.

Bruhn, M. and McKenzie, D. In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232, 2009.

Cochran, W. G. Analysis of covariance: its nature and uses. *Biometrics*, 13(3):261–281, 1957.

Fisher, R. A. The design of experiments. 1935. *Oliver and Boyd, Edinburgh*, 1935.

Freedman, D. A. On Regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008a.

Freedman, D. A. On regression adjustments in experiments with several treatments. *The annals of applied statistics*, pages 176–196, 2008b.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Imbens, G. W. and Wooldridge, J. M. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Lin, W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

McCullagh, P. and Nelder, J. *Generalized Linear Models*. London, Chapman and Hall, 1989.

Mullahy, J. Multivariate fractional regression estimation of econometric share models. *Journal of econometric methods*, 4(1):71–100, 2015.

Neyman, J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.(Translated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

Papke, L. E. and Wooldridge, J. M. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of applied econometrics*, 11 (6):619–632, 1996.

Pitkin, E., Berk, R., Brown, L., Buja, A., George, E., Zhang, K., and Zhao, L. An Asymptotically Powerful Test for the Average Treatment Effect. *Available at "http://www-stat.wharton.upenn.edu/ lbrown/"*, 2017.

Rosenbaum, P. R. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

Rosenblum, M. and Van Der Laan, M. J. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics*, 6(1), 2010.

Słoczyński, T. A general weighted average representation of the ordinary and two-stage least squares estimands. *arXiv preprint arXiv:1810.01576*, 2018.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.

White, H. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Wooldridge, J. M. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301, 2007.

Wooldridge, J. M. *Econometric analysis of cross section and panel data.* MIT press, 2010.

Yang, L. and Tsiatis, A. A. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

9 Proofs

9.1 Proof of Lemma 5.1

Proof. Asymptotic variance of SDM
Consider the difference-in-means estimator. We can write the sample average for the treated as

$$\begin{aligned}\bar{Y}_1 &= N_1^{-1}\sum_{i=1}^{N} W_i Y_i = N_1^{-1}\sum_{i=1}^{N} W_i \left[\mu_1 + \dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right] \\ &= \mu_1 + N_1^{-1}\sum_{i=1}^{N} W_i \left[\dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right]\end{aligned}$$

Therefore,

$$\begin{aligned}\sqrt{N}\left(\bar{Y}_1 - \mu_1\right) &= (N/N_1)N^{-1/2}\sum_{i=1}^{N} W_i \left[\dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right] \\ &= (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_i \left[\dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right] + o_p(1)\end{aligned}$$

because $N_1/N \xrightarrow{p} \rho$. By the CLT,

$$\begin{aligned}N^{-1/2}\sum_{i=1}^{N} W_i \left[\dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right] &\xrightarrow{d} Normal(0, c_1^2) \\ \text{where, } c_1^2 = E\left\{W_i \left[\dot{\mathbf{X}}_i \boldsymbol{\beta}_1 + U_i(1)\right]^2\right\} \\ = \rho\left(\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1 + \sigma_1^2\right)\end{aligned}$$

where W_i independent of $(\mathbf{X}_i, U_i(1))$ is used. It follows that

$$Avar\left[\sqrt{N}\left(\bar{Y}_1 - \mu_1\right)\right] = (1/\rho)^2\rho\left(\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1 + \sigma_1^2\right) = \left(\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1 + \sigma_1^2\right)/\rho \tag{39}$$

Similarly,

$$Avar\left[\sqrt{N}\left(\bar{Y}_0 - \mu_0\right)\right] = \left(\boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0 + \sigma_0^2\right)/(1-\rho). \tag{40}$$

Combining results from eq(39) and eq(40), we have:

$$\sqrt{N}\left(\hat{\tau}_{SDM} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega_{SDM}^2\right)$$

Since the sample averages are asymptotically uncorrelated, therefore

$$\omega_{SDM}^2 = \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0/(1-\rho) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho)$$

□

Proof. Asymptotic variance of P-RA
To find the asymptotic variance of $\check{\tau}$, note that it can be obtained from

$$Y_i \text{ on } 1,\ W_i,\ \dot{\mathbf{X}}_i.$$

Note that $\dot{\mathbf{X}}_i$ is orthogonal to $(1, W_i)$ because $E(\dot{\mathbf{X}}_i) = \mathbf{0}$ and W_i is independent of $\mathbf{X}_i$. We know that

$$L(Y_i|1, W_i) = \mu_0 + \tau W_i$$

because $\tau = E(Y_i|W_i = 1) - E(Y_i|W_i = 0)$. Therefore,

$$L(Y_i|1, W_i, \dot{\mathbf{X}}_i) = \mu_0 + \tau W_i + \dot{\mathbf{X}}_i \boldsymbol{\beta}$$

By orthogonality,

$$\boldsymbol{\beta} = \left[E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\right]^{-1} E\left(\dot{\mathbf{X}}_i'Y_i\right)$$

Now

$$\begin{aligned} Y_i &= (1 - W_i)\mu_0 + (1 - W_i)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1 - W_i)U_i(0) \\ &+ W_i\mu_1 + W_i\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1) \end{aligned}$$

Therefore,

$$\begin{aligned} E\left(\dot{\mathbf{X}}_i'Y_i\right) &= E\left[(1 - W_i)\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\boldsymbol{\beta}_0\right] + E\left[W_i\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\boldsymbol{\beta}_1\right] \\ &= (1 - \rho)E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\boldsymbol{\beta}_0 + \rho E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\boldsymbol{\beta}_1 \end{aligned}$$

where we use the linear projection properties of the errors, $E\left(\dot{\mathbf{X}}_i\right) = \mathbf{0}$, and independence of W_i and $[\mathbf{X}_i, U_i(0), U_i(1)]$. Plugging in gives

$$\boldsymbol{\beta} = (1 - \rho)\boldsymbol{\beta}_0 + \rho\boldsymbol{\beta}_1$$

Now we can write the projection error as

$$\begin{aligned} U_i &= (1 - W_i)\mu_0 + (1 - W_i)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1 - W_i)U_i(0) \\ &+ W_i\mu_1 + W_i\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1) \\ &- \mu_0 - (\mu_1 - \mu_0)W_i - \dot{\mathbf{X}}_i\left[(1 - \rho)\boldsymbol{\beta}_0 + \rho\boldsymbol{\beta}_1\right] \\ &= -(W_i - \rho)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1 - W_i)U_i(0) \\ &+ (W_i - \rho)\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1). \end{aligned}$$

Because $(1, W_i)$ is orthogonal to $\dot{\mathbf{X}}_i$, it follows as in the previous section that

$$\begin{aligned} \sqrt{N}(\hat{\tau}_{PRA} - \tau) &= \left[E(\dot{W}_i^2)\right]^{-1}\left(N^{-1/2}\sum_{i=1}^{N}(W_i - \rho)U_i\right) + o_p(1) \\ &= [\rho(1 - \rho)]^{-1}\left(N^{-1/2}\sum_{i=1}^{N}(W_i - \rho)U_i\right). \end{aligned}$$

Then using asymptotic equivalence lemma and CLT, we have:

$$\sqrt{N}(\hat{\tau}_{PRA} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \omega_{PRA}^2\right)$$

where $\omega_{PRA}^2 = Var\left(\left(W_i - \rho\right)U_i\right) / \left[\rho(1 - \rho)\right]^2$.

Now we need to find the asymptotic variance of $N^{-1/2}\sum_{i=1}^{N}(W_i-\rho)U_i$. The term $(W_i-\rho)U_i$ has zero mean by the linear projection property. Further,

$$\begin{aligned}(W_i-\rho)U_i &= -(W_i-\rho)^2\dot{\mathbf{X}}_i\boldsymbol{\beta}_0+(W_i-\rho)^2\dot{\mathbf{X}}_i\boldsymbol{\beta}_1\\ &\quad +(W_i-\rho)(1-W_i)U_i(0)+(W_i-\rho)W_iU_i(1)\end{aligned}$$

The covariance between the last two terms is zero as $(1-W_i)W_i=0$. The last two terms can be written as

$$-\rho(1-W_i)U_i(0)+(1-\rho)W_iU_i(1)$$

and so

$$Var\left[-\rho(1-W_i)U_i(0)+(W_i-\rho)W_iU_i(1)\right]=\rho^2(1-\rho)\sigma_0^2+(1-\rho)^2\rho\sigma_1^2.$$

Write the first two terms as

$$(W_i-\rho)^2\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right).$$

The variance is

$$E\left[(W_i-\rho)^4\right]\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right).$$

Combining all of the terms gives

$$\begin{aligned}Avar\left[\sqrt{N}(\hat{\tau}_{PRA}-\tau)\right] &= [\rho(1-\rho)]^{-2}\left\{E\left[(W_i-\rho)^4\right]\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\rho^2(1-\rho)\sigma_0^2+(1-\rho)^2\rho\sigma_1^2\right\}\\ &= \frac{E\left[(W_i-\rho)^4\right]}{[\rho(1-\rho)]^2}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\frac{\sigma_0^2}{(1-\rho)}+\frac{\sigma_1^2}{\rho}\end{aligned}$$

Note that we can write

$$\frac{E\left[(W_i-\rho)^4\right]}{[\rho(1-\rho)]^2}=\frac{E\left[(W_i-\rho)^4\right]}{[Var(W_i)]^2}$$

and Jensen's inequality tells us this is greater than unity: take $Z_i=(W_i-\rho)^2$. We can also show

$$E\left[(W_i-\rho)^4\right]=(1-\rho)^4\rho+\rho^4(1-\rho)$$

and so the scale factor is

$$\frac{(1-\rho)^4\rho+\rho^4(1-\rho)}{[\rho(1-\rho)]^2}=\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)}.$$

Hence,

$$Avar\left[\sqrt{N}\left(\hat{\tau}_{PRA}-\tau\right)\right]=\left(\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)}\right)\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\frac{\sigma_0^2}{(1-\rho)}+\frac{\sigma_1^2}{\rho}$$

□

Proof. Asymptotic variance of F-RA
Now consider the full regression adjustment estimator. Let $\hat{\alpha}_1$ and $\hat{\boldsymbol{\beta}}_1$ be the OLS estimates from the $W_i=1$ sample:

$$Y_i \text{ on } 1,\ \mathbf{X}_i \quad W_i=1$$

and then

$$\hat{\mu}_{1,FRA}=\hat{\alpha}_1+\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1$$

where $\bar{\mathbf{X}}$ is the sample average over the entire sample. (For intuition, it is useful to note that

$\bar{Y}_1 = \hat{\alpha}_1 + \bar{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1$, and so $\hat{\mu}_1$ uses a more efficient estimator of $\boldsymbol{\mu}_{\mathbf{X}}$.) By least squares mechanics, $\hat{\mu}_1$ is the intercept in the regression

$$Y_i \text{ on } 1,\, \mathbf{X}_i - \bar{\mathbf{X}},\;\; W_i = 1.$$

Let $\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}$ and

$$\ddot{\mathbf{R}}_i = (1, \ddot{\mathbf{X}}_i).$$

Define

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_1 = \begin{pmatrix}\hat{\mu}_1\\ \hat{\boldsymbol{\beta}}_1\end{pmatrix} &= \left(\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' Y_i\right)\\ &= \left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' Y_i(1)\right).\end{aligned}$$

Now write

$$\begin{aligned}Y_i(1) &= \mu_1 + \dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1) = \mu_1 + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_1 + (\dot{\mathbf{X}}_i - \ddot{\mathbf{X}}_i)\boldsymbol{\beta}_1 + U_i(1)\\ &= \mu_1 + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_1 + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + U_i(1) = \ddot{\mathbf{R}}_i\boldsymbol{\gamma}_1 + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + U_i(1)\end{aligned}$$

Plugging in gives

$$N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' Y_i(1) = \left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)\boldsymbol{\gamma}_1 + \left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\right)(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1)$$

Now we can write

$$\hat{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1 + \left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\right)(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1)\right]$$

and so

$$\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right) = \left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\right)\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + N^{-1/2}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1)\right]$$

Next, because $\bar{\mathbf{X}} \xrightarrow{p} \boldsymbol{\mu}_{\boldsymbol{X}}$, the law of large numbers and Slutsky's Theorem imply

$$N^{-1}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i = N^{-1}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i + o_p(1)$$

where

$$\dot{\mathbf{R}}_i = (1, \dot{\mathbf{X}}_i) = (1, \mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}})$$

Further,

$$N^{-1}\sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i \xrightarrow{p} E\left(W_i \dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right) = \rho E\left(\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right).$$

Note that

$$\mathbf{A} \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right) \end{pmatrix}$$

The terms $\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1$ and $N^{-1/2}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i'U_i(1)$ are $O_p(1)$, and so

$$\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right) = (1/\rho)\mathbf{A}^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i'\right)\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + N^{-1/2}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i'U_i(1)\right] + o_p(1).$$

Consider the first element of $N^{-1}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i'$:

$$N^{-1}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i' = N^{-1}\sum_{i=1}^{N} W_i\begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i \end{pmatrix}$$

and so the first element is

$$N^{-1}\sum_{i=1}^{N} W_i = N_1/N = \hat{\rho} \xrightarrow{p} \rho.$$

Also,

$$N^{-1/2}\sum_{i=1}^{N} W_i\ddot{\mathbf{R}}_i'U_i(1) = N^{-1/2}\sum_{i=1}^{N} W_i\begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i \end{pmatrix}U_i(1)$$

and so the first element is

$$N^{-1/2}\sum_{i=1}^{N} W_iU_i(1).$$

Because of the block diagonality of $\mathbf{A}$, the first element of, $\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right)$, $\sqrt{N}\left(\hat{\mu}_1 - \mu_1\right)$ satisfies

$$\begin{aligned}\sqrt{N}\left(\hat{\mu}_{1,FRA} - \mu_1\right) &= (1/\rho)\rho\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_iU_i(1) + o_p(1) \\ &= \sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_iU_i(1) + o_p(1).\end{aligned}$$

We can also write

$$\sqrt{N}\left(\hat{\mu}_{1,FRA} - \mu_1\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_1 + W_iU_i(1)/\rho\right] + o_p(1)$$

A similar argument gives

$$\sqrt{N}\left(\hat{\mu}_{0,FRA} - \mu_0\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_0 + (1 - W_i)U_i(0)/(1 - \rho)\right] + o_p(1)$$

and so

$$\sqrt{N}\left(\hat{\tau}_{FRA} - \tau\right) = N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + W_iU_i(1)/\rho - (1 - W_i)U_i(0)/(1 - \rho)\right] + o_p(1)$$

Again, by asymptotic equivalence lemma and CLT, we have:

$$\sqrt{N}\left(\hat{\tau}_{FRA}-\tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega_{FRA}^2\right)$$

where $\omega_{FRA}^2 = Var\left(\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+W_iU_i(1)/\rho-(1-W_i)U_i(0)/(1-\rho)\right)$

Now consider the above expression inside the variance. The three terms are pairwise uncorrelated, the second and third because $W_i(1-W_i)=0$, and the first with the other two because, for example,

$$E\left[\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\dot{\mathbf{X}}_i'W_iU_i(1)\right] = E(W_i)\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'E\left[\dot{\mathbf{X}}_i'U_i(1)\right]=0$$

because $E\left[\dot{\mathbf{X}}_i'U_i(1)\right]=\mathbf{0}$ by linear projection properties. It follows that

$$\begin{aligned}Avar\left[\sqrt{N}\left(\hat{\tau}_{FRA}-\tau\right)\right] &= \left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+(1/\rho^2)E(W_i)E\left[U_i^2(1)\right]+\\ &\quad(1/(1-\rho)^2)E(1-W_i)E\left[U_i^2(0)\right]\\ \\ &= \left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\sigma_1^2/\rho+\sigma_0^2/(1-\rho).\end{aligned}$$

□

Proof. Asymptotic variance of I-RA
The derivation for τ^* follows closely that for $\hat{\tau}$, with the important difference that $\ddot{\mathbf{X}}_i=\mathbf{X}_i-\bar{\mathbf{X}}$ is replaced with $\dot{\mathbf{X}}_i=\mathbf{X}_i-\boldsymbol{\mu}_{\mathbf{X}}$. This means that the terms $\sqrt{N}(\bar{\mathbf{X}}-\boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1$ and $\sqrt{N}(\bar{\mathbf{X}}-\boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_0$ terms will not appear. Therefore,

$$Avar\left[\sqrt{N}\left(\hat{\tau}_{IRA}-\tau\right)\right]=\sigma_1^2/\rho+\sigma_0^2/(1-\rho).$$

□

9.2 Proof of Theorem 5.2

Proof. **CLAIM 1** : $\omega_{FRA}^2 \leq \omega_{SDM}^2$

For this consider consider the left hand side,

$$\begin{aligned}&Avar\left[\sqrt{N}(\hat{\tau}_{SDM}-\tau)\right]-Avar\left[\sqrt{N}(\hat{\tau}_{FRA}-\tau)\right]\\ &=\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho+\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0/(1-\rho)-\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)\end{aligned}$$

The last term in the above expression can be written as:

$$\begin{aligned}&\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho+\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0/(1-\rho)-\left[\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1+\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0-2\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1\right]\\ &=\left(\frac{1-\rho}{\rho}\right)\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1+\left(\frac{\rho}{1-\rho}\right)\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0+2\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1\\ &\equiv\boldsymbol{\delta}'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\delta}\end{aligned}$$

where

$$\boldsymbol{\delta} = \sqrt{\left(\frac{1-\rho}{\rho}\right)}\boldsymbol{\beta}_1 + \sqrt{\left(\frac{\rho}{1-\rho}\right)}\boldsymbol{\beta}_0.$$

Because $\boldsymbol{\Omega}_{\mathbf{X}}$ is positive definite, this proves the claim. One case where there is no efficiency gain is when $\rho = 1/2$ and $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$. The second condition seems unrealistic unless both vectors are zero.

CLAIM 2 : $\omega^2_{FRA} \leq \omega^2_{PRA}$

For this consider the left hand side of the expression above,

$$\begin{aligned} Avar\left[\sqrt{N}(\hat{\tau}_{PRA} - \tau)\right] - Avar\left[\sqrt{N}(\hat{\tau}_{FRA} - \tau)\right] &= \left[\frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)} - 1\right](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\,\boldsymbol{\Omega}_{\mathbf{X}}\,(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\ &\geq 0 \end{aligned}$$

CLAIM 3 : $\omega^2_{IRA} \leq \omega^2_{FRA}$

It is easy to see why this holds true since the L.H.S just equals

$$Avar\left[\sqrt{N}(\hat{\tau}_{FRA} - \tau)\right] - Avar\left[\sqrt{N}(\hat{\tau}_{IRA} - \tau)\right] = (\boldsymbol{\beta_1} - \boldsymbol{\beta_0})'\,\boldsymbol{\Omega_X}\,(\boldsymbol{\beta_1} - \boldsymbol{\beta_0})$$

Because $\boldsymbol{\Omega_X}$ is psd and the above is just a quadratic form which will be greater than or equal to zero.

Combing the results from CLAIM 1, 2 and 3 we have the result. □

10 Figures

10.1 Root mean squared error across different sample sizes

Figure 1: **Quadratic Design, Continuous covariates (Mild Heterogeneity)**

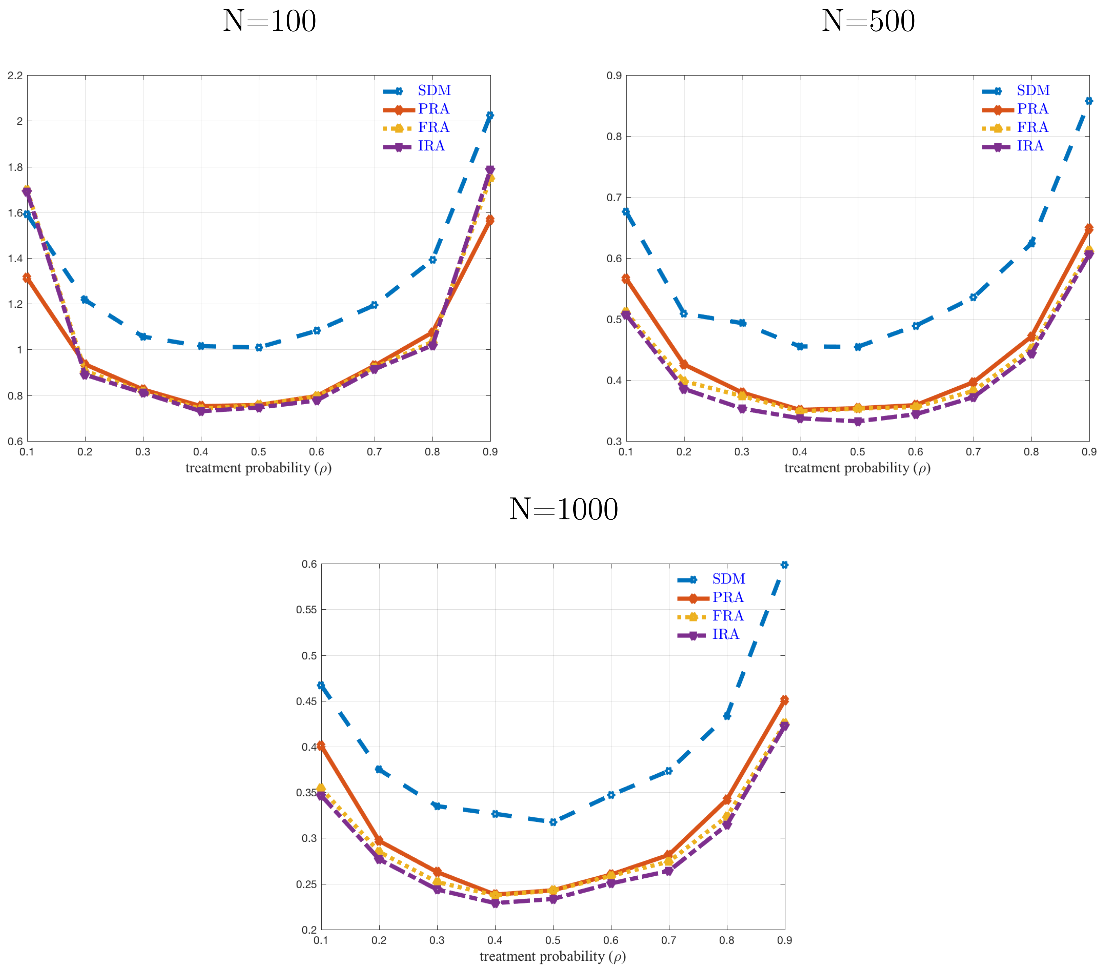
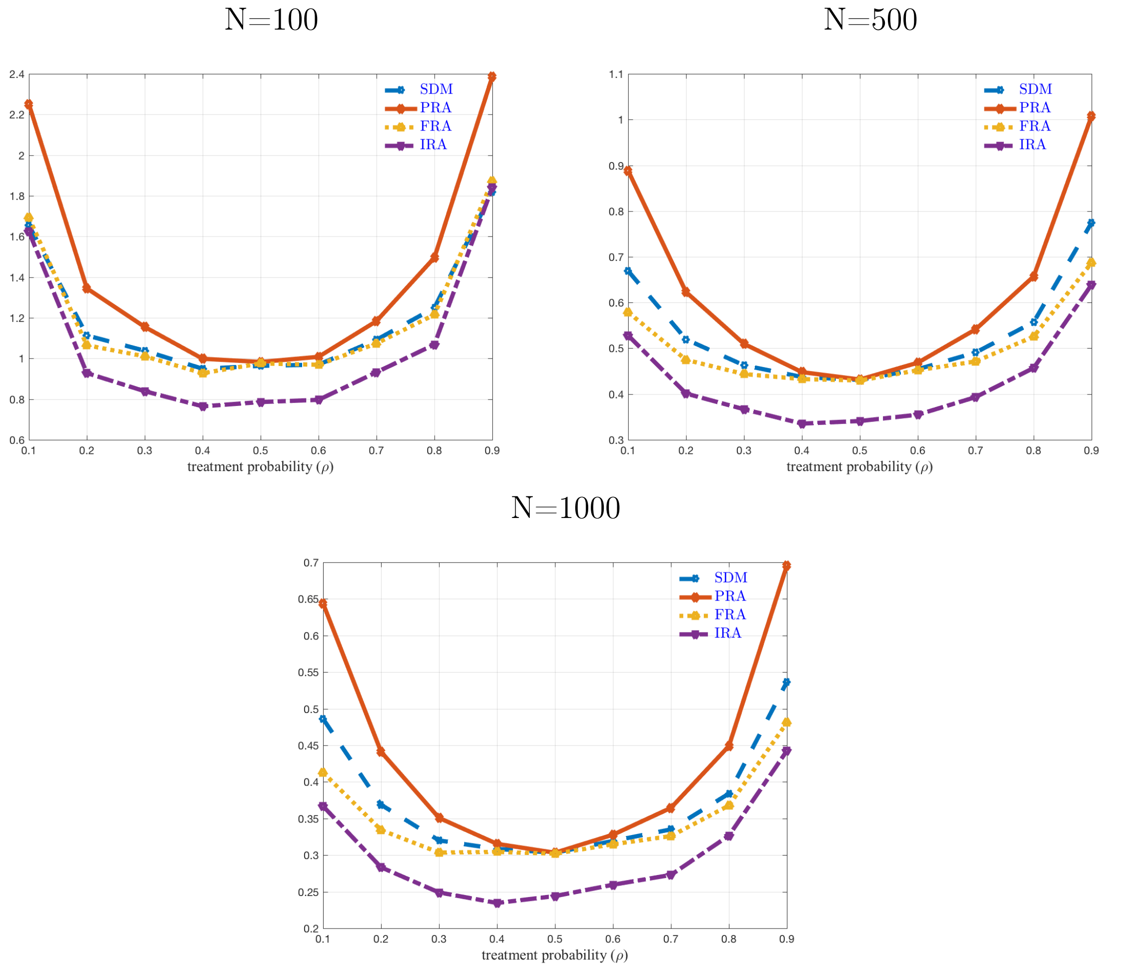


Figure 2: **Quadratic Design, continuous covariates (Strong Heterogeneity)**

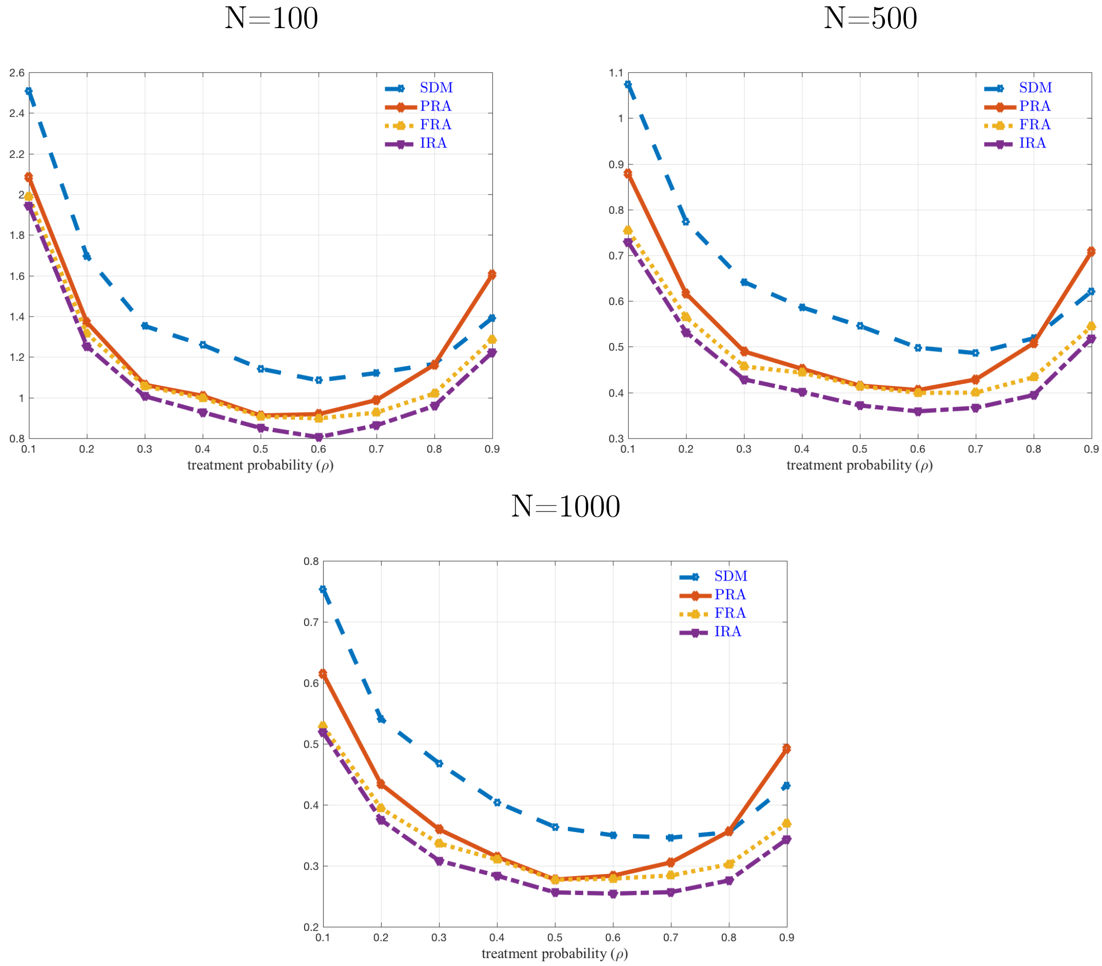


N=100

N=500

N=1000

Figure 3: **Quadratic Design, one binary covariates (Mild Heterogeneity)**

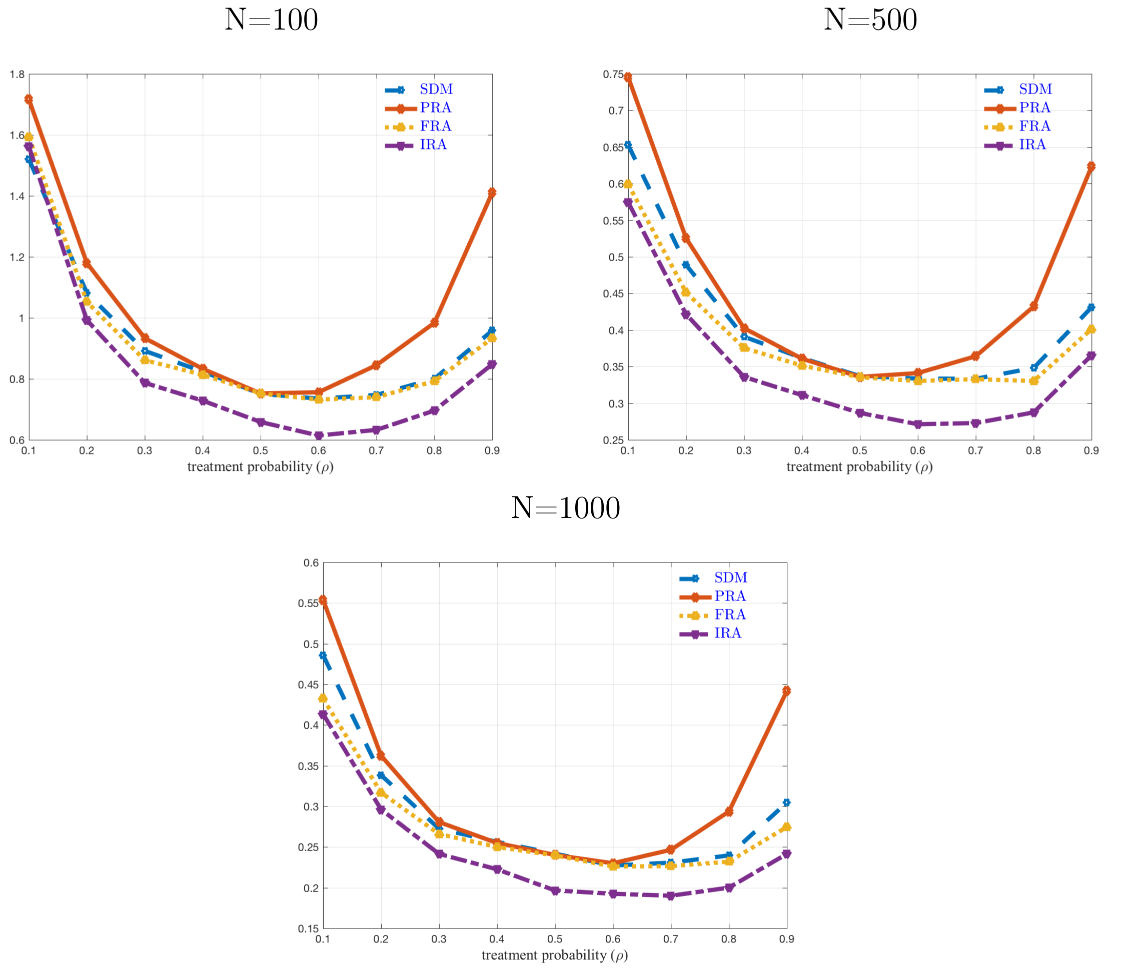


N=100

N=500

N=1000

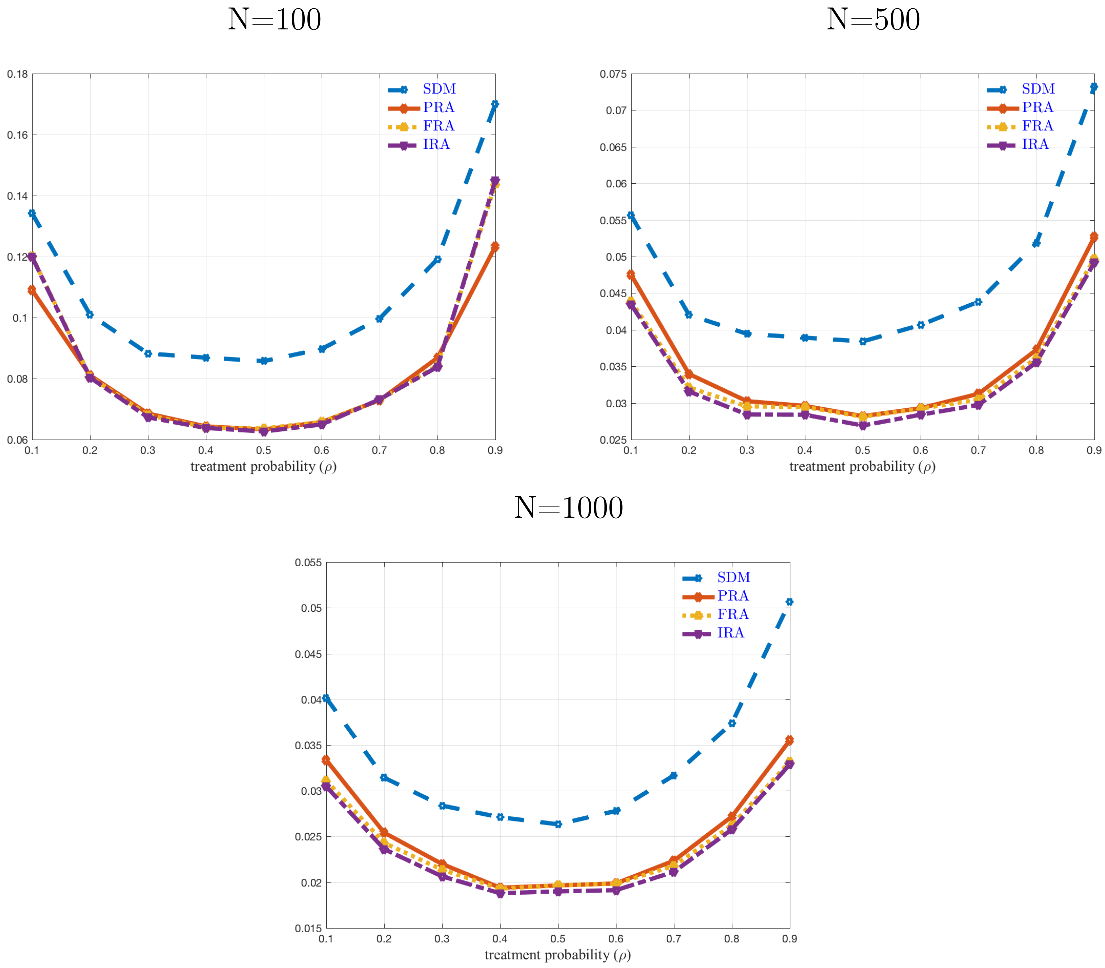
Figure 4: **Quadratic Design, one binary covariate (Strong Heterogeneity)**


N=100

N=500

N=1000

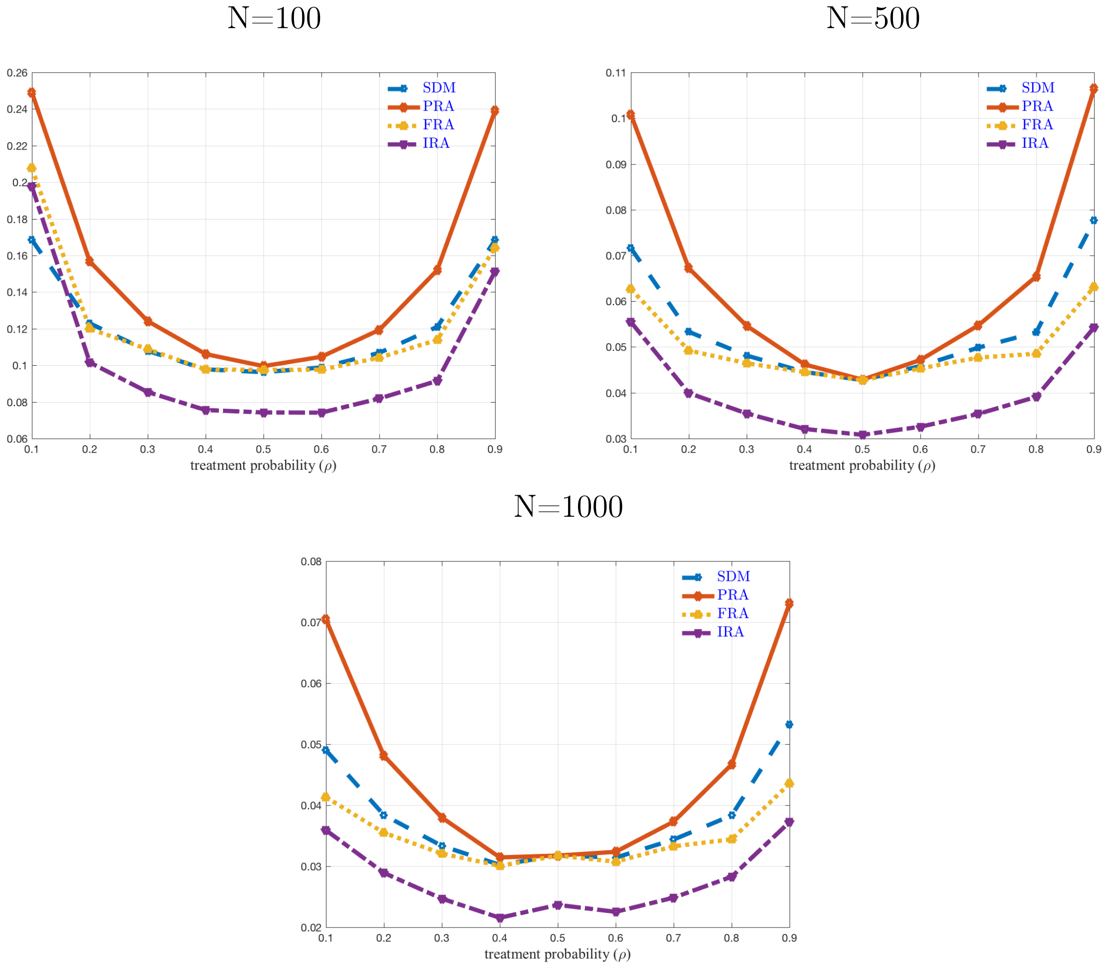
Figure 5: **Probit Design, continuous covariates (Mild Heterogeneity)**


N=100

N=500

N=1000

Figure 6: **Probit Design, continuous covariates (Strong Heterogeneity)**

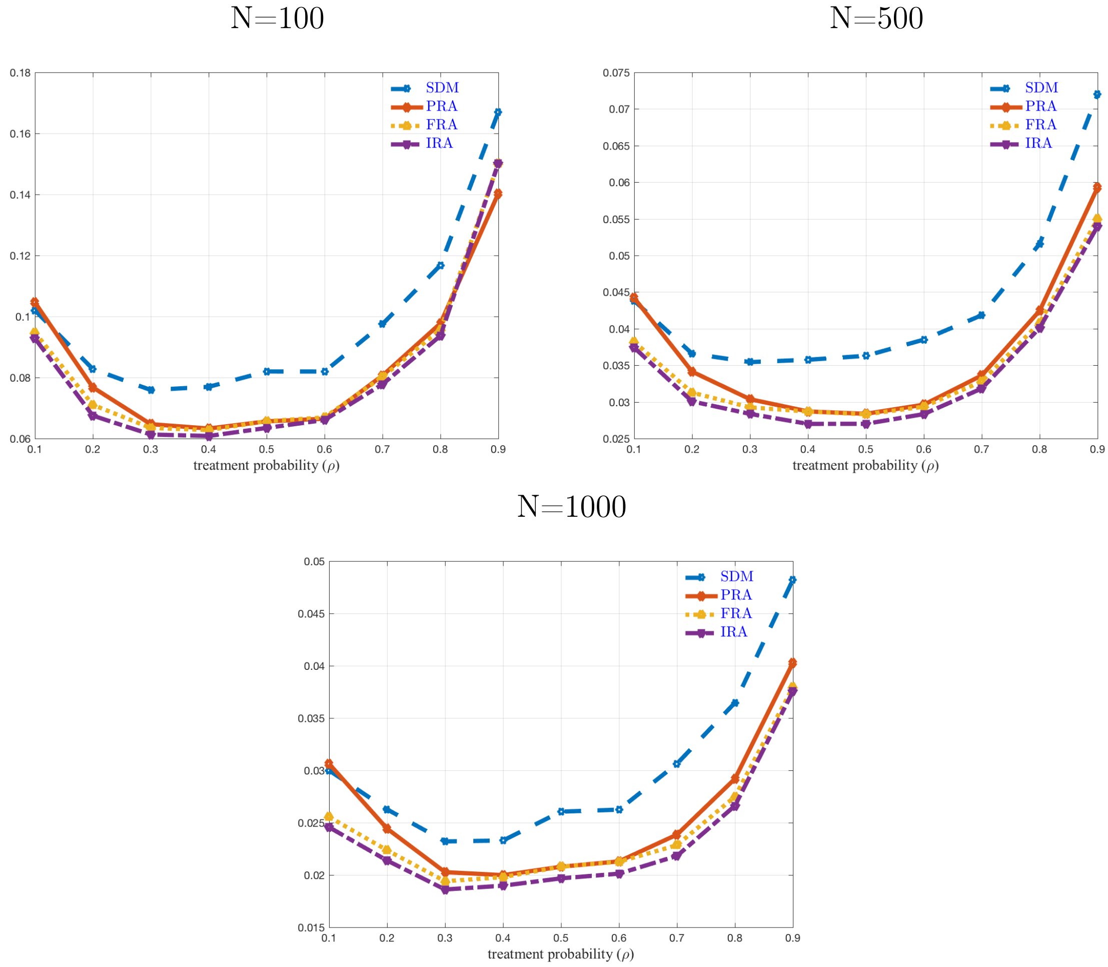


N=100

N=500

N=1000

Figure 7: **Probit Design, one binary covariate (Mild Heterogeneity)**

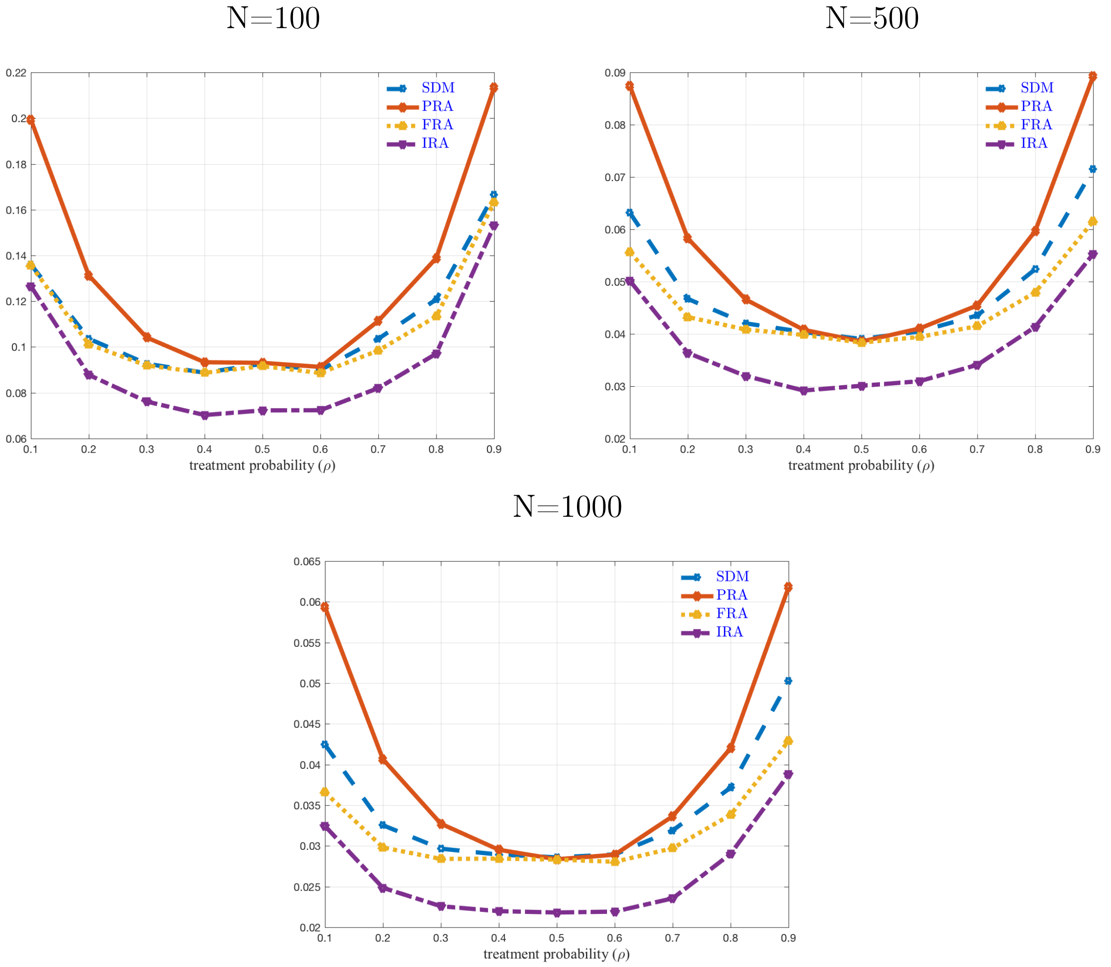


N=100

N=500

N=1000

Figure 8: **Probit Design, one binary covariate (Strong Heterogeneity)**



N=100

N=500

N=1000

Figure 9: **Binary outcome, Bernoulli QLL with Logistic mean**

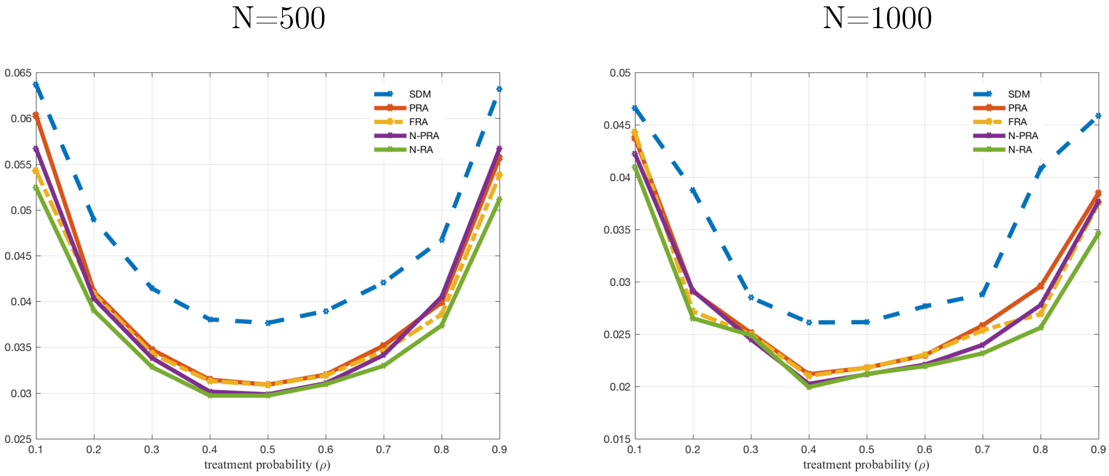
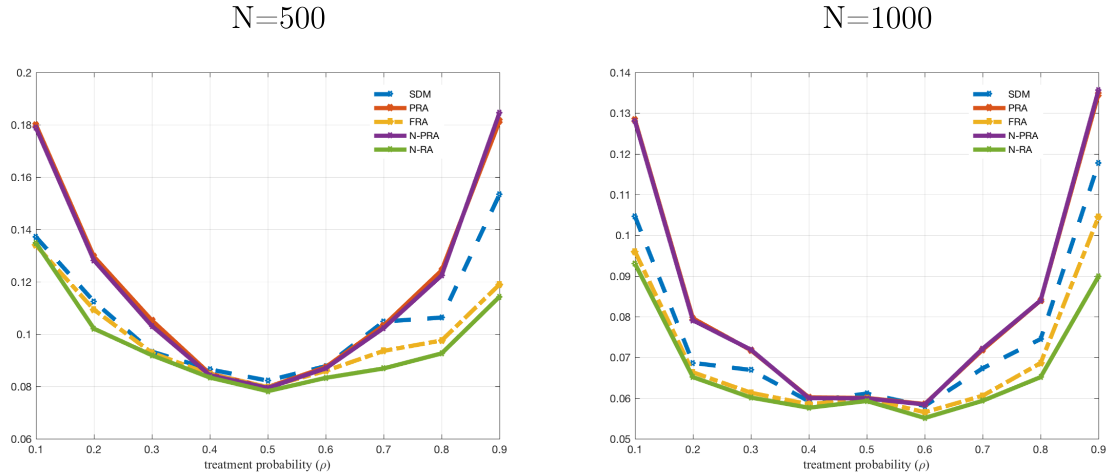


Figure 10: **Non-negative outcome, Poisson QLL with exponential mean**



11 Tables

Table 1: Bias and Standard deviation for N=100

					DGP1					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.045	1.590	0.025	1.056	0.034	1.008	-0.035	1.194	0.070	2.023
PRA	0.047	1.312	-0.022	0.825	0.039	0.756	-0.031	0.929	0.039	1.566
FRA	0.017	1.697	-0.023	0.815	0.042	0.757	-0.022	0.922	-0.021	1.750
IRA	0.004	1.690	-0.014	0.810	0.026	0.746	-0.025	0.914	-0.022	1.786
					DGP2					
SDM	0.045	1.590	0.025	1.056	0.034	1.008	-0.035	1.194	0.070	2.023
PRA	0.047	1.312	-0.022	0.825	0.039	0.756	-0.031	0.929	0.039	1.566
FRA	0.017	1.697	-0.023	0.815	0.042	0.757	-0.022	0.922	-0.021	1.750
IRA	0.004	1.690	-0.014	0.810	0.026	0.746	-0.025	0.914	-0.022	1.786
					DGP3					
SDM	-0.058	2.508	-0.085	1.350	0.041	1.141	-0.038	1.120	-0.039	1.391
PRA	-0.045	2.083	-0.069	1.061	0.054	0.910	0.038	0.987	-0.094	1.602
FRA	0.051	1.988	-0.100	1.052	0.030	0.907	0.003	0.926	0.043	1.286
IRA	0.046	1.944	-0.085	1.003	0.019	0.850	0.014	0.864	0.072	1.221
					DGP4					
SDM	0.094	1.517	-0.040	0.891	0.005	0.751	0.014	0.747	-0.031	0.958
PRA	0.013	1.716	-0.047	0.932	0.007	0.752	0.004	0.845	-0.034	1.410
FRA	0.042	1.593	-0.050	0.860	0.002	0.752	0.015	0.739	0.058	0.931
IRA	0.022	1.561	-0.072	0.783	0.003	0.658	0.003	0.632	0.019	0.848
					DGP5					
SDM	0.002	0.134	0.002	0.088	-0.002	0.086	-0.002	0.100	0.003	0.170
PRA	0.003	0.109	-0.001	0.069	0.000	0.063	0.000	0.073	0.003	0.123
FRA	0.025	0.117	0.003	0.068	0.000	0.064	0.000	0.073	-0.001	0.144
IRA	0.026	0.117	0.005	0.067	0.001	0.063	0.002	0.073	-0.001	0.145
					DGP6					
SDM	-0.002	0.169	0.000	0.108	0.000	0.096	0.001	0.107	0.003	0.168
PRA	0.000	0.249	0.001	0.124	0.003	0.099	0.004	0.119	0.007	0.239
FRA	0.028	0.206	0.009	0.108	0.004	0.097	0.004	0.104	0.004	0.164
IRA	0.030	0.195	0.013	0.084	0.008	0.074	0.008	0.081	0.004	0.151
					DGP7					
SDM	0.000	0.102	0.000	0.076	-0.003	0.082	-0.005	0.097	-0.008	0.167
PRA	-0.001	0.105	0.002	0.065	0.001	0.066	-0.001	0.081	-0.005	0.140
FRA	0.019	0.093	0.005	0.063	0.000	0.066	-0.004	0.080	-0.004	0.150
IRA	0.020	0.091	0.005	0.061	0.001	0.063	-0.002	0.078	-0.003	0.150
					DGP8					
SDM	0.004	0.136	0.006	0.092	0.004	0.093	-0.003	0.103	-0.022	0.165
PRA	-0.005	0.199	0.009	0.104	0.008	0.093	0.002	0.111	-0.015	0.213
FRA	0.020	0.134	0.011	0.091	0.007	0.091	0.002	0.098	0.010	0.163
IRA	0.022	0.125	0.011	0.075	0.010	0.072	0.006	0.082	0.013	0.152

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.

[b] Simulation across 1000 replications.

Table 2: Bias and Standard deviation for N=500

					DGP1					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.035	0.675	0.023	0.493	-0.029	0.453	-0.005	0.535	0.049	0.856
PRA	0.025	0.566	0.009	0.379	-0.019	0.353	0.007	0.396	0.029	0.648
FRA	0.023	0.511	0.008	0.374	-0.019	0.353	0.008	0.382	0.011	0.612
IRA	0.021	0.507	0.007	0.353	-0.019	0.332	0.009	0.372	0.013	0.606
					DGP2					
SDM	0.035	0.675	0.023	0.493	-0.029	0.453	-0.005	0.535	0.049	0.856
PRA	0.025	0.566	0.009	0.379	-0.019	0.353	0.007	0.396	0.029	0.648
FRA	0.023	0.511	0.008	0.374	-0.019	0.353	0.008	0.382	0.011	0.612
IRA	0.021	0.507	0.007	0.353	-0.019	0.332	0.009	0.372	0.013	0.606
					DGP3					
SDM	0.054	1.073	-0.003	0.642	-0.009	0.546	-0.013	0.486	0.001	0.621
PRA	0.031	0.878	0.002	0.490	0.003	0.415	0.010	0.428	0.025	0.707
FRA	-0.014	0.755	-0.004	0.457	-0.003	0.414	0.000	0.400	0.007	0.544
IRA	-0.011	0.729	0.003	0.429	-0.005	0.372	0.004	0.366	0.011	0.518
					DGP4					
SDM	-0.034	0.652	0.012	0.391	-0.003	0.337	0.006	0.333	0.006	0.431
PRA	-0.051	0.744	0.013	0.402	-0.004	0.336	-0.001	0.364	-0.001	0.624
FRA	-0.007	0.599	0.012	0.375	-0.004	0.335	0.007	0.333	0.001	0.401
IRA	-0.010	0.574	0.001	0.336	-0.005	0.287	0.000	0.273	0.000	0.365
					DGP5					
SDM	0.001	0.056	0.001	0.039	0.000	0.038	0.000	0.044	0.006	0.073
PRA	0.000	0.047	0.000	0.030	0.001	0.028	0.001	0.031	0.004	0.053
FRA	0.003	0.044	0.000	0.030	0.001	0.028	0.001	0.030	0.002	0.050
IRA	0.003	0.043	0.001	0.028	0.001	0.027	0.001	0.030	0.002	0.049
					DGP6					
SDM	0.000	0.072	0.001	0.048	-0.001	0.043	-0.001	0.050	0.006	0.077
PRA	0.001	0.101	0.001	0.055	-0.001	0.043	-0.001	0.055	0.008	0.106
FRA	0.006	0.062	0.003	0.046	0.000	0.043	0.001	0.048	0.005	0.063
IRA	0.006	0.055	0.004	0.035	0.000	0.031	0.001	0.035	0.004	0.054
					DGP7					
SDM	-0.001	0.044	-0.001	0.035	0.000	0.036	0.001	0.042	0.001	0.072
PRA	-0.001	0.044	0.000	0.030	0.001	0.028	0.001	0.034	0.002	0.059
FRA	0.003	0.038	0.000	0.029	0.001	0.028	0.001	0.033	0.001	0.055
IRA	0.004	0.037	0.001	0.028	0.001	0.027	0.000	0.032	0.001	0.054
					DGP8					
SDM	-0.002	0.063	-0.002	0.042	0.001	0.039	0.001	0.044	0.002	0.071
PRA	-0.002	0.087	-0.001	0.047	0.001	0.039	0.002	0.045	0.003	0.089
FRA	0.002	0.056	-0.001	0.041	0.001	0.038	0.002	0.041	0.005	0.061
IRA	0.004	0.050	0.000	0.032	0.002	0.030	0.001	0.034	0.005	0.055

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.
[b] Simulation across 1000 replications.

Table 3: Bias and Standard deviation for N=1000

					DGP1					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.016	0.467	-0.006	0.335	0.008	0.317	-0.009	0.373	0.018	0.599
PRA	-0.002	0.401	-0.009	0.263	-0.001	0.243	-0.008	0.281	0.013	0.451
FRA	-0.001	0.354	-0.010	0.252	-0.001	0.243	-0.009	0.274	0.011	0.425
IRA	0.000	0.347	-0.009	0.244	-0.001	0.233	-0.007	0.264	0.011	0.423
					DGP2					
SDM	0.016	0.467	-0.006	0.335	0.008	0.317	-0.009	0.373	0.018	0.599
PRA	-0.002	0.401	-0.009	0.263	-0.001	0.243	-0.008	0.281	0.013	0.451
FRA	-0.001	0.354	-0.010	0.252	-0.001	0.243	-0.009	0.274	0.011	0.425
IRA	0.000	0.347	-0.009	0.244	-0.001	0.233	-0.007	0.264	0.011	0.423
					DGP3					
SDM	0.019	0.753	0.001	0.468	0.010	0.363	-0.002	0.346	0.006	0.432
PRA	0.015	0.615	0.000	0.360	0.006	0.277	0.004	0.306	0.001	0.492
FRA	-0.001	0.529	-0.006	0.337	0.003	0.277	0.000	0.284	0.001	0.369
IRA	0.001	0.519	-0.004	0.308	0.003	0.256	0.002	0.257	0.000	0.344
					DGP4					
SDM	-0.007	0.486	-0.006	0.272	0.002	0.242	0.004	0.231	0.004	0.305
PRA	-0.010	0.554	-0.006	0.281	0.002	0.240	0.003	0.247	0.007	0.442
FRA	0.002	0.432	-0.004	0.266	0.001	0.240	0.003	0.226	-0.003	0.275
IRA	-0.002	0.413	-0.006	0.241	0.002	0.196	0.001	0.190	-0.004	0.241
					DGP5					
SDM	0.001	0.040	0.001	0.028	0.001	0.026	-0.001	0.032	0.001	0.051
PRA	-0.001	0.033	0.000	0.022	0.000	0.020	0.000	0.022	0.001	0.036
FRA	0.001	0.031	0.001	0.021	0.000	0.020	-0.001	0.022	0.000	0.033
IRA	0.001	0.030	0.001	0.021	0.000	0.019	0.000	0.021	0.001	0.033
					DGP6					
SDM	-0.001	0.049	0.001	0.033	0.000	0.032	-0.001	0.034	0.002	0.053
PRA	-0.002	0.070	0.001	0.038	0.001	0.032	-0.001	0.037	0.002	0.073
FRA	0.003	0.041	0.001	0.032	0.001	0.032	0.000	0.033	0.001	0.044
IRA	0.003	0.036	0.001	0.025	0.001	0.024	0.000	0.025	0.002	0.037
					DGP7					
SDM	0.002	0.030	0.001	0.023	-0.001	0.026	-0.001	0.031	-0.001	0.048
PRA	0.000	0.031	0.000	0.020	-0.001	0.021	0.000	0.024	0.000	0.040
FRA	0.003	0.025	0.001	0.019	-0.001	0.021	0.000	0.023	-0.001	0.038
IRA	0.003	0.024	0.001	0.019	-0.001	0.020	0.000	0.022	-0.001	0.038
					DGP8					
SDM	0.001	0.042	0.000	0.030	0.000	0.029	-0.001	0.032	0.001	0.050
PRA	0.000	0.059	0.000	0.033	0.000	0.028	-0.001	0.034	0.000	0.062
FRA	0.004	0.036	0.001	0.028	0.000	0.028	-0.001	0.030	0.002	0.043
IRA	0.004	0.032	0.001	0.023	0.001	0.022	0.000	0.024	0.002	0.039

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.

[b] Simulation across 1000 replications.

Table 4: Bias and Standard deviation for Binary outcome

					N=500					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.014	0.062	0.002	0.041	-0.007	0.037	0.003	0.042	0.001	0.063
PRA	0.023	0.056	0.000	0.035	-0.002	0.031	0.003	0.035	0.015	0.054
FRA	0.018	0.051	0.000	0.034	-0.002	0.031	0.002	0.035	0.009	0.053
N-PRA	0.013	0.055	0.001	0.034	-0.001	0.030	0.004	0.034	0.014	0.055
N-RA	0.006	0.052	0.001	0.033	-0.002	0.030	0.004	0.033	0.007	0.051
					N=1000					
SDM	-0.017	0.044	0.008	0.027	0.000	0.026	0.003	0.029	-0.016	0.043
PRA	-0.021	0.038	0.010	0.023	0.000	0.022	0.009	0.024	-0.006	0.038
FRA	-0.024	0.037	0.010	0.023	0.000	0.022	0.009	0.024	-0.010	0.036
N-PRA	-0.019	0.038	0.010	0.022	-0.001	0.021	0.006	0.023	-0.003	0.038
N-RA	-0.020	0.036	0.012	0.022	-0.001	0.021	0.006	0.022	-0.008	0.034

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator, N-PRA refers to pooled non-linear regression adjustment and N-RA refers to separate nonlinear regression adjustment.

[b] Simulation across 1000 replications.

[c.] True ATE is 0.037, $R_0^2 = 0.491$ and $R_1^2 = 0.457$.

Table 5: Bias and Standard deviation for Non-negative outcome

					N=500					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.010	0.137	-0.005	0.093	0.017	0.080	-0.027	0.101	-0.067	0.138
PRA	-0.003	0.180	-0.024	0.103	0.015	0.078	-0.023	0.101	-0.074	0.166
FRA	0.024	0.132	-0.006	0.093	0.015	0.078	-0.013	0.093	-0.041	0.112
N-PRA	0.000	0.179	-0.022	0.101	0.015	0.078	-0.024	0.100	-0.078	0.168
N-RA	0.027	0.132	-0.006	0.092	0.011	0.077	-0.013	0.086	-0.039	0.107
					N=1000					
SDM	-0.055	0.089	0.020	0.064	0.006	0.061	-0.014	0.066	-0.022	0.116
PRA	-0.059	0.114	0.028	0.066	0.004	0.060	-0.023	0.068	-0.023	0.133
FRA	-0.044	0.086	0.008	0.061	0.003	0.060	-0.002	0.061	-0.022	0.102
N-PRA	-0.056	0.115	0.028	0.066	0.004	0.060	-0.024	0.068	-0.025	0.133
N-RA	-0.040	0.084	0.006	0.060	0.006	0.059	-0.001	0.059	-0.013	0.089

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator, N-PRA refers to pooled non-linear regression adjustment and N-RA refers to separate nonlinear regression adjustment.

[b] Simulation across 1000 replications.

[c.] True ATE is 0.012, $R_0^2 = 0.435$ and $R_1^2 = 0.233$.