Robust and Efficient Estimation of Potential Outcome Means Under Random Assignment

Akanksha Negi, Jeffrey M. Wooldridge[†]

November 10, 2019

Abstract

We consider the problem of regression adjustment in experiments with two or more assignment levels. Building on Negi and Wooldridge (2019), which looked at the case of the average treatment effect with one control and one treatment group, we study improvements in efficiency in estimating the vector of potential outcome means using linear regression adjustment. We show that using separate regression adjustments for each assignment level is never worse, asymptotically, than using the subsample averages. We also show that separate regression adjustment generally improves over pooled regression adjustment, except in the obvious case where slope parameters in the linear projections are identical across the different assignment levels. An especially promising direction is to use certain nonlinear regression adjustment methods, which we show to be fully robust in that the conditional means can be arbitrarily misspecified and yet we still obtain consistent estimators of the potential outcome means. We apply this general potential outcomes framework to a contingent valuation study which seeks to estimate the lower bound mean willingness to pay (WTP) for an oil spill prevention program in California. We find that using separate regression adjustments for the estimating the potential outcome means leads to more precise estimates of the lower bound mean WTP estimate than the ABERS estimator which simply uses subsample averages.

JEL Classification Codes: C21, C25

Keywords: Multiple Treatments, Randomized Experiment, Regression Adjustment, Heterogeneous Treatment Effects

[†]Department of Economics, Michigan State University. Email: negiakan@msu.edu, wooldri1@msu.edu

1 Introduction

In the past several decades, the potential outcomes framework has become a staple of causal inference in statistics, econometrics, and related fields. Envisioning each unit in a population under different states of intervention or treatment allows one to define treatment or causal effects without referencing to a model. One merely needs the means of the potential outcomes, or perhaps the potential outcome (PO) means for subpopulations.

When interventions are randomized – whether the assignment is to control and treatment groups in a clinical trial (Hirano and Imbens (2001)), assignment to participate in a job training program (Calónico and Smith (2017)), receiving a school voucher when studying the effects of private schooling on educational outcomes (Angrist et al. (2006)), or contingent valuation studies, where different bid values are randomized among people (Carson et al. (2004)) – one can simply used the subsample means for each treatment level in order to obtain unbiased and consistent estimators of the PO means. In some cases, the precisions of the subsample means will be sufficient. Nevertheless, with the availability of good predictors of the outcome or response, it is appealing to think that the precision can be improved, thereby shrinking confidence intervals and making conclusions about interventions more reliable.

In this paper we build on Negi and Wooldridge (2019), who studied the problem of estimating the average treatment effect under random assignment with one control group and one treatment group. In the context of random sampling, we showed that performing separate linear regressions for the control and treatment groups in estimating the average treatment effect never does worse, asymptotically, than the simple difference in means estimator or a pooled regression adjustment estimator. In addition, we characterized the class of nonlinear regression adjustment methods that produce consistent estimators of the ATE without any additional assumptions (except regularity conditions). The simulation findings for both the linear and nonlinear cases are quite promising when covariates are available that predict the outcomes.

In the current paper, we consider any number of "treatment" levels and consider the problem of joint estimation of the vector of potential outcome means. We assume that the assignment to the

treatment is random – that is, independent of both potential outcomes and observed predictors of the POs. Importantly, other than standard regularity conditions (such as finite second moments of the covariates), we impose no additional assumptions. In other words, the full RA estimators are consistent under essentially the same assumptions as the subsample means with, generally, smaller asymptotic variance. Interestingly, even if the predictors are unhelpful, or the slopes in the linear projections are the same across all groups, no asymptotic efficiency is lost by using the most general RA method.

We also extend the nonlinear RA results in Negi and Wooldridge (2019) to the general case of G assignment levels. We show that for particular kinds of responses such as binary, fractional or nonnegative, it is possible to consistently estimate PO means using pooled and separate regression adjustment. Unlike the linear regression adjustment case, we do not have any general asymptotic results to compare full nonlinear RA with pooled nonlinear RA.

Finally, we apply the full RA estimator to data from a contingent valuation study obtained from Carson et al. (2004). This data is used to elicit a lower bound on mean willingness to pay (WTP) for a program that would prevent future oil spills along California's central coast. Our results show that the PO means for the five different bid amounts that were randomly assigned to California residents are estimated more efficiently using separate regression adjustment than just using subsample averages. This efficiency result is preserved for estimating the lower bound since it is a linear combination of PO means. Hence, using separate RA also delivers a more precise lower bound on mean WTP for the oil spill prevention program than the ABERS estimator which uses subsample averages to construct the estimate.

The rest of the paper is organized as follows: Section 2 discusses the potential outcomes framework extended to the case of G treatment levels along with a discussion of the crucial random sampling and random assignment assumptions. Section 3 derives the asymptotic variances of the different linear regression adjustment estimators, namely, subsample means, pooled regression adjustment and feasible regression adjustment. Section 4 compares the asymptotic variances of the entire vector of subsample means, pooled and feasible regression adjustment. Section 5 considers a class of nonlinear regression adjustment estimators that ensure consistency of the subsample means

without imposing additional assumptions. Section 6 discusses applications of this framework to randomized experiments, differences in differences settings and contingent valuation studies. This section also applies full regression adjustment estimator for estimating the lower bound mean WTP for the California Oil Spill study using data from Carson et al. (2004). Section 7 concludes.

2 Potential Outcomes, Random Assignment, and Random Sampling

The goal is to estimate the population means of G potential (counterfactual) outcomes, $Y(g)$, $g = 1, ..., G$. Define

$$\mu_g = \mathbb{E}\left[Y(g)\right], \, g = 1, ..., G.$$

The vector of assignment indicators is

$$\mathbf{W} = (W_1, ..., W_G),$$

where each W_g is binary and

$$W_1 + W_2 + \cdots + W_G = 1.$$

In other words, the groups are exhaustive and mutually exclusive. The setup applies to many situations, including the standard control-treatment group setup, with $G = 2$, multiple treatment levels (with $g = 1$ the control group), the basic difference-in-differences setup with $G = 4$, and in contingent valuation studies where subjects are presented with a set of G prices or bid values.

We assume that each group g has positive probability of being assigned:

$$\rho_g \equiv \mathbb{P}(W_g = 1) > 0, \, g = 1, ..., G$$

$$\rho_1 + \rho_2 + \cdots + \rho_G = 1$$

Next, let

$$\mathbf{X} = (X_1, X_2, ..., X_K)$$

be a vector of observed covariates.

Assumption 1 (**Random Assignment**). *Assignment is independent of the potential outcomes and observed covariates:*

$$\mathbf{W} \perp \left[Y(1), Y(2), ..., Y(G), \mathbf{X}\right].$$

Further, the assignment probabilities are all strictly positive. □

Assumption 1 is what puts us in the framework of experimental interventions. It would be much too strong for an observational study.

Assumption 2 (**Random Sampling**). *For a nonrandom integer N,*

$$\{[\mathbf{W}_i, Y_i(1), Y_i(2), ..., Y_i(G), \mathbf{X}_i] : i = 1, 2, ..., N\}$$

is independent and identically distributed. □

The IID assumption is not the only one we can make. For example, we could allow for a sampling-without-replacement scheme given a fixed sample size N. This would complicate the analysis because it generates slight correlation within draws. As discussed in Negi and Wooldridge (2019), Assumption 2 is traditional in studying the asymptotic properties of estimators and is realistic as an approximation. Importantly, it forces us to account for the sampling error in the sample average, $\bar{\mathbf{X}}$, as an estimator of $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}(\mathbf{X})$.

For each draw i from the population, we only observe

$$Y_i = W_{i1}Y_i(1) + W_{i2}Y_i(2) + \cdots + W_{iG}Y_i(G),$$

and so the data we have to work with is

$$\{(\mathbf{W}_i, Y_i, \mathbf{X}_i) : i = 1, 2, ..., N\}.$$

Definition of population quantities only requires us to use the random vector $(\mathbf{W}, Y, \mathbf{X})$, which represents the population.

Assumptions 1 and 2 are the only substantive restrictions used in this paper. Subsequently, we assume that linear projections exist and that the central limit theorem holds for properly standardized sample averages of IID random vectors. Therefore, we are implicitly imposing at least finite second moment assumptions on the $Y(g)$ and the X_j. We do not make this explicit in what follows.

3 Subsample Means and Linear Regression Adjustment

In this section we derive the asymptotic variances of three estimators: the subsample means, full (separate) regression adjustment, and pooled regression adjustment.

3.1 Subsample Means

The simplest estimator of μ_g is the sample average within treatment group g:

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} W_{ig} Y_i = N_g^{-1} \sum_{i=1}^{N} W_{ig} Y_i(g),$$

where

$$N_g = \sum_{i=1}^{N} W_{ig}$$

is a random variable in our setting. In expressing $\bar{Y}_g$ as a function of the $Y_i(g)$ we use $W_{ih} W_{ig} = 0$ for $h \neq g$. Under random assignment and random sampling,

$$\begin{aligned}
\mathbb{E}\left(\bar{Y}_g | W_{1g}, ..., W_{Ng}, N_g > 0\right) &= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mathbb{E}\left[Y_i(g) | W_{1g}, ..., W_{Ng}, N_g > 0\right] \\
&= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mathbb{E}\left[Y_i(g)\right] \\
&= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mu_g = \mu_g,
\end{aligned}$$

and so $\bar{Y}_g$ is unbiased conditional on observing a positive number of units in group g.

By the law of large numbers, a consistent estimator of ρ_g is

$$\hat{\rho}_g = N_g / N,$$

the sample share of units in group g. Therefore, by the law of large numbers and Slutsky's Theorem,

$$\begin{aligned}\bar{Y}_g &= \left(\frac{N}{N_g}\right) N^{-1} \sum_{i=1}^{N} W_{ig} Y_i(g) \xrightarrow{p} \rho_g^{-1} \mathbb{E}\left[W_g Y(g)\right] \\ &= \rho_g^{-1} \mathbb{E}\left(W_g\right) \mathbb{E}\left[Y(g)\right] = \mu_g,\end{aligned}$$

and so $\bar{Y}_g$ is consistent for μ_g.

By the central limit theorem, $\sqrt{N}\left(\bar{Y}_g - \mu_g\right)$ is asymptotically normal. We need an asymptotic representation of $\sqrt{N}\left(\bar{Y}_g - \mu_g\right)$ that allows us to compare its asymptotic variance with those from regression adjustment estimators. To this end, write

$$\begin{aligned}Y(g) &= \mu_g + V(g) \\ \dot{\mathbf{X}} &= \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}},\end{aligned}$$

where $\dot{\mathbf{X}}$ is $\mathbf{X}$ demeaned using the population mean, $\mu_{\mathbf{X}}$. Now project each $V(g)$ linearly onto $\dot{\mathbf{X}}$:

$$V(g) = \dot{\mathbf{X}}\boldsymbol{\beta_g} + U(g),\ g = 1, ..., G.$$

By construction, the population projection errors $U(g)$ have the properties

$$\begin{aligned}\mathbb{E}\left[U(g)\right] &= 0,\ g = 1, ..., G \\ \mathbb{E}\left[\dot{\mathbf{X}}' U(g)\right] &= \mathbf{0},\ g = 1, ..., G.\end{aligned}$$

Plugging in gives

$$Y(g) = \mu_g + \dot{\mathbf{X}}\boldsymbol{\beta_g} + U(g),\ g = 1, ..., G$$

Importantly, by random assignment, $\mathbf{W}$ is independent of $\left[U(1), ..., U(G), \dot{\mathbf{X}}\right]$. The observed outcome can be written as

$$Y = \sum_{g=1}^{G} W_g \left[\mu_g + \dot{\mathbf{X}}\boldsymbol{\beta_g} + U(g)\right].$$

Using (aa), write $\bar{Y}_g$ as

$$\begin{aligned}\bar{Y}_g &= N_g^{-1}\sum_{i=1}^{N} W_{ig}Y_i(g) = N_g^{-1}\sum_{i=1}^{N} W_{ig}\left[\mu_g + \dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] \\ &= \mu_g + N_g^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] = \mu_g + (N/N_g)N^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] \\ &= \mu_g + \left(\frac{1}{\hat{\rho}_g}\right)N^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]\end{aligned}$$

Therefore,

$$\sqrt{N}\left(\bar{Y}_g - \mu_g\right) = \hat{\rho}_g^{-1}\left\{N^{-1/2}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]\right\} \tag{1}$$

By random assignment,

$$E\left(W_{ig}\dot{\mathbf{X}}_i\right) = E(W_{ig})E\left(\dot{\mathbf{X}}_i\right) = \mathbf{0}$$

$$E\left[W_{ig}U_i(g)\right] = E\left(W_{ig}\right)E\left[U_i(g)\right] = \rho_g E\left[U_i(g)\right] = 0,$$

and so the CLT applies to the standardized average in (1). Now use $\hat{\rho}_g = \rho_g + o_p(1)$ to obtain the following first-order representation:

$$\sqrt{N}\left(\bar{Y}_g - \mu_g\right) = \rho_g^{-1}\left\{N^{-1/2}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]\right\} + o_p(1).$$

Our goal is to be able to make efficiency statements about both linear and nonlinear functions of the vector of means $\mu = (\mu_1, \mu_2, ..., \mu_G)'$, and so we stack the subsample means into the $G \times 1$ vector $\bar{\mathbf{Y}}$. For later comparison, it is helpful to remember that $\bar{\mathbf{Y}}$ is the vector of OLS coefficients in the regression

$$Y_i \text{ on } W_{i1},\ W_{i2},\ ...,\ W_{iG},\ i = 1, 2, ..., N.$$

We have proven the following result.

Theorem 3 (Asymptotic variance of Subsample means estimator of PO means)**.** *Under Assump-*

tions 1, 2, and finite second moments,

$$\begin{aligned}\sqrt{N}\left(\bar{\mathbf{Y}}-\boldsymbol{\mu}\right) &= \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N}\left[W_{i1}\dot{\mathbf{X}}_i\boldsymbol{\beta_1}/\rho_1 + W_{i1}U_i(1)/\rho_1\right] \\ N^{-1/2}\sum_{i=1}^{N}\left[W_{i2}\dot{\mathbf{X}}_i\boldsymbol{\beta_2}/\rho_2 + W_{i2}U_i(2)/\rho_2\right] \\ \vdots \\ N^{-1/2}\sum_{i=1}^{N}\left[W_{iG}\dot{\mathbf{X}}_i\boldsymbol{\beta_G}/\rho_G + W_{iG}U_i(G)/\rho_G\right] \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{L}_i+\mathbf{Q}_i\right)+o_p(1)\end{aligned}$$

where

$$\mathbf{L}_i \equiv \begin{pmatrix} W_{i1}\dot{\mathbf{X}}_i\beta_1/\rho_1 \\ W_{i2}\dot{\mathbf{X}}_i\beta_2/\rho_2 \\ \vdots \\ W_{iG}\dot{\mathbf{X}}_i\beta_G/\rho_G \end{pmatrix} \tag{2}$$

and

$$\mathbf{Q}_i \equiv \begin{pmatrix} W_{i1}U_i(1)/\rho_1 \\ W_{i2}U_i(2)/\rho_2 \\ \vdots \\ W_{iG}U_i(G)/\rho_G \end{pmatrix} \tag{3}$$

By random assignment and the linear projection property, $\mathbb{E}\left(\mathbf{L}_i\right)=\mathbb{E}\left(\mathbf{Q}_i\right)=\mathbf{0}$, and $\mathbb{E}\left(\mathbf{L}_i\mathbf{Q}_i'\right)=\mathbf{0}$. Also, because $W_{ig}W_{ih}=0$, $g\neq h$, the elements of $\mathbf{L}_i$ are pairwise uncorrelated; the same is true of the elements of $\mathbf{Q}_i$.

3.2 Full Regression Adjustment

To motivate full regression adjustment, write the linear projection for each g as

$$\begin{aligned} Y(g) &= \alpha_g + \mathbf{X}\boldsymbol{\beta_g} + U(g) \\ \mathbb{E}\left[U(g)\right] &= 0 \\ \mathbb{E}\left[\mathbf{X}'U(g)\right] &= \mathbf{0}\end{aligned}$$

It follows immediately that

$$\mu_g = \alpha_g + \boldsymbol{\mu}_{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}}.$$

Consistent estimators of α_g and $\boldsymbol{\beta}_g$ are obtained from the regression

$$Y_i \text{ on } 1,\ \mathbf{X}_i, \text{ if } W_{ig} = 1,$$

which produces intercept and slopes $\hat{\alpha}_g$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$. Letting $\breve{\mathbf{X}}_i = (1, \mathbf{X}_i)$, the probability limit of $\left(\hat{\alpha}_g, \hat{\boldsymbol{\beta}}_g'\right)'$ is

$$\begin{aligned}
\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'Y_i\right)\right] &= \rho_g^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'Y_i(g)\right)\right] \\
&= \rho_g^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\rho_g\mathbb{E}\left(\breve{\mathbf{X}}_i'Y_i(g)\right)\right] \\
&= \left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'Y_i(g)\right)\right] = \begin{pmatrix}\alpha_g \\ \boldsymbol{\beta}_{\boldsymbol{g}}\end{pmatrix}
\end{aligned}$$

where random assignment is used so that W_{ig} is independent of $[\mathbf{X}_i, Y_i(g)]$. It follows that $\left(\hat{\alpha}_g, \hat{\boldsymbol{\beta}}_{\boldsymbol{g}}'\right)'$ is consistent for $\left(\alpha_g, \boldsymbol{\beta}_{\boldsymbol{g}}'\right)$, and so a consistent estimator of μ_g is

$$\hat{\mu}_g = \hat{\alpha}_g + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}.$$

Note that this estimator, which we refer to as full (or separate) regression adjustment (FRA), is the same as an imputation procedure. Given $\hat{\alpha}_g$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$, impute a value of $Y_i(g)$ for each i in the sample, whether or not i is assigned to group g:

$$\hat{Y}_i(g) = \hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\boldsymbol{g}},\ i = 1, 2, ..., N.$$

Averaging these imputed values across all i produces $\hat{\mu}_g$. In order to derive the asymptotic variance of $\hat{\mu}_g$, it is helpful to obtain it as the intercept from the regression

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

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

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

Define

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

Now write

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

Plugging in for $Y_i(g)$ gives

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

and so

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

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

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

where

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

Further, by random assignment,

$$N^{-1}\sum_{i=1}^{N} W_{ig}\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i \xrightarrow{p} \mathbb{E}\left(W_{ig}\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right) = \rho_g \mathbb{E}\left(\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right) = \rho_g \mathbf{A},$$

where

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

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

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

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

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

and so the first element is

$$N^{-1}\sum_{i=1}^{N} W_{ig} = N_g/N = \hat{\rho}_g \xrightarrow{p} \rho_g.$$

Also,

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

and so the first element is

$$N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g).$$

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

$$\begin{aligned} \sqrt{N}\left(\hat{\mu}_g - \mu_g\right) &= (1/\rho_g)\rho_g\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + (1/\rho_g)N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g) + o_p(1) \\ &= \sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + (1/\rho_g)N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g) + o_p(1). \end{aligned}$$

We can also write

$$\sqrt{N}\left(\hat{\mu}_g - \mu_g\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_{\boldsymbol{g}} + W_{ig}U_i(g)/\rho_g\right] + o_p(1)$$

The above representation holds for all g. Then, stacking the RA estimates gives us the following theorem

Theorem 4 (Asymptotic variance of Full regression adjustment estimator of PO means)**.** *Under assumptions 1 and 2, and finite second moments,*

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) &= \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{1}} + W_{i1}U_i(1)/\rho_1\right] \\ N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{2}} + W_{i2}U_i(2)/\rho_2\right] \\ \vdots \\ N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{G}} + W_{iG}U_i(G)/\rho_G\right] \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{K}_i + \mathbf{Q}_i\right) + o_p(1)\end{aligned}$$

where

$$\mathbf{K}_i = \begin{pmatrix} \dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{1}} \\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{2}} \\ \vdots \\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{G}} \end{pmatrix} \tag{4}$$

and $\mathbf{Q}_i$ *is given in (3).*

Both $\mathbf{K}_i$ and $\mathbf{Q}_i$ have zero means, the latter by random assignment. Further, by random assignment and the linear projection property, $\mathbb{E}\left(\mathbf{K}_i\mathbf{Q}_i'\right) = \mathbf{0}$ because

$$\mathbb{E}\left[\dot{\mathbf{X}}_i'W_{ig}U_i(g)\right] = \mathbb{E}(W_{ig})\mathbb{E}\left[\dot{\mathbf{X}}_i'U_i(g)\right] = \mathbf{0}.$$

However, unlike the elements of $\mathbf{L}_i$, we must recognize that the elements of $\mathbf{K}_i$ are correlated except in the trivial case that all but one of the $\boldsymbol{\beta}_{\boldsymbol{g}}$ are zero.

3.3 Pooled Regression Adjustment

Now consider the pooled RA estimator, $\check{\mu}$, which can be obtained as the vector of coefficients on $\mathbf{W}_i = (W_{i1}, W_{i2}, ..., W_{iG})$ from the regression

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

We refer to this as a pooled method because the coefficients on $\ddot{\mathbf{X}}_i$, say, $\check{\boldsymbol{\beta}}$, are assumed to be the same for all groups. Compared with subsample means, we add the controls $\ddot{\mathbf{X}}_i$, but unlike FRA, the pooled method imposes the same coefficients across all g.

In order to find a useful first order representation of $\sqrt{N}(\check{\mu} - \mu)$, we first characterize the probability limit of $\check{\boldsymbol{\beta}}$. Under random assignment,

$$\mathbb{E}\left(\mathbf{W}'\dot{\mathbf{X}}\right) = \mathbb{E}(\mathbf{W})'\,\mathbb{E}\left(\dot{\mathbf{X}}\right) = \mathbf{0},$$

which means that the coefficients on $\mathbf{W}$ in the linear projections $\mathbb{L}(Y|\mathbf{W})$ and $\mathbb{L}\left(Y|\mathbf{W}, \dot{\mathbf{X}}\right)$ are the same and equal to μ. This essentially proves that adding the demeaned covariates still consistently estimates $\boldsymbol{\mu}$. Moreover, we can find the coefficients on $\dot{\mathbf{X}}$ in $\mathbb{L}\left(Y|\mathbf{W}, \dot{\mathbf{X}}\right)$ by finding $\mathbb{L}\left(Y|\dot{\mathbf{X}}\right)$. Let $\boldsymbol{\beta}$ be the the linear projection of Y on $\dot{\mathbf{X}}$. Then

$$\boldsymbol{\beta} = \left[\mathbb{E}\left(\dot{\mathbf{X}}'\dot{\mathbf{X}}\right)\right]^{-1}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right) = \boldsymbol{\Omega}_{\mathbf{X}}^{-1}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right)$$

Now use

$$Y = \sum_{g=1}^{G} W_g\left[\mu_g + \dot{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}} + U(g)\right]$$

so that

$$\begin{aligned}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right) &= \sum_{g=1}^{G}\left\{\mathbb{E}\left(\dot{\mathbf{X}}'W_g\mu_g\right) + \mathbb{E}\left(\dot{\mathbf{X}}'W_g\dot{\mathbf{X}}\right)\boldsymbol{\beta}_{\boldsymbol{g}} + \mathbb{E}\left[\dot{\mathbf{X}}'W_gU(g)\right]\right\} \\ &= \sum_{g=1}^{G}\left\{\mathbf{0} + \rho_g\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}} + \mathbf{0}\right\} = \boldsymbol{\Omega}_{\mathbf{X}}\left(\sum_{g=1}^{G}\rho_g\boldsymbol{\beta}_{\boldsymbol{g}}\right),\end{aligned}$$

where we again use random assignment, $\mathbb{E}\left(\dot{\mathbf{X}}\right) = \mathbf{0}$, and $\mathbb{E}\left[\dot{\mathbf{X}}' U(g)\right] = \mathbf{0}$. It follows that

$$\boldsymbol{\beta} = \boldsymbol{\Omega}_{\dot{\mathbf{X}}}^{-1} \boldsymbol{\Omega}_{\mathbf{X}} \left(\sum_{g=1}^{G} \rho_g \boldsymbol{\beta}_{\boldsymbol{g}} \right) = \left(\sum_{g=1}^{G} \rho_g \boldsymbol{\beta}_{\boldsymbol{g}} \right).$$

Therefore, the $\boldsymbol{\beta}$ in the linear projection $\mathbb{L}\left(Y|\dot{\mathbf{X}}\right)$ is simply a weighted average of the coefficients from the separate linear projections using the potential outcomes.

Now we can write

$$Y_i = \mathbf{W}_i \mu + \dot{\mathbf{X}}_i \boldsymbol{\beta} + U_i$$

where the linear projection error U_i is

$$\begin{aligned} U_i &= \sum_{g=1}^{G} W_{ig} \left[\mu_{ig} + \dot{\mathbf{X}}_i \boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g) \right] - \mathbf{W}_i \boldsymbol{\mu} - \dot{\mathbf{X}}_i \left(\sum_{g=1}^{G} \rho_g \boldsymbol{\beta}_{\boldsymbol{g}} \right) \\ &= \sum_{g=1}^{G} (W_{ig} - \rho_g) \dot{\mathbf{X}}_i \boldsymbol{\beta}_{\boldsymbol{g}} + \sum_{g=1}^{G} W_{ig} U_i(g) \end{aligned}$$

We can now obtain the asymptotic representation for $\sqrt{N}\left(\check{\mu} - \mu\right)$. Write $\boldsymbol{\theta} = (\mu', \boldsymbol{\beta}')'$, $\dot{\mathbf{R}}_i = \left(\mathbf{W}_i, \dot{\mathbf{X}}_i\right)$, $\ddot{\mathbf{R}}_i = \left(\mathbf{W}_i, \ddot{\mathbf{X}}_i\right)$, and $\check{\boldsymbol{\theta}} = \left(\check{\mu}', \check{\boldsymbol{\beta}}'\right)'$ as the OLS estimators. The asymptotic variance of $\sqrt{N}\left(\check{\mu} - \mu\right)$ is not the same as replacing $\ddot{\mathbf{X}}_i$ with $\dot{\mathbf{X}}_i$ (even though for $\check{\boldsymbol{\beta}}$ it is). Write

$$\begin{aligned} Y_i &= \mathbf{W}_i \mu + \ddot{\mathbf{X}}_i \boldsymbol{\beta} + \left(\dot{\mathbf{X}}_i - \ddot{\mathbf{X}}_i\right) \boldsymbol{\beta} + U_i = \mathbf{W}_i \mu + \ddot{\mathbf{X}}_i \boldsymbol{\beta} + \left(\bar{\mathbf{X}} - \mu_{\mathbf{X}}\right) \boldsymbol{\beta} + U_i \\ &= \ddot{\mathbf{R}}_i \boldsymbol{\theta} + \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right) \boldsymbol{\beta} + U_i. \end{aligned}$$

Now

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

and so

$$\begin{aligned}\sqrt{N}\left(\check{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) &= \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i\right)'\left[\sqrt{N}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta}+N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'U_i\right] \\ &= \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\begin{pmatrix}N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \\ \mathbf{0}\end{pmatrix}\left[\sqrt{N}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta}+N^{-1/2}\sum_{i=1}^{N}\begin{pmatrix}\mathbf{W}_i \\ \ddot{\mathbf{X}}_i\end{pmatrix}'U_i\right]\end{aligned}$$

because $N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{X}}_i'=\mathbf{0}$. Further, the terms in $[\cdot]$ are $O_p(1)$ and

$$N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i \xrightarrow{p} \begin{pmatrix}\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathbf{X}}\end{pmatrix}$$

by random assignment and $E\left(\dot{\mathbf{X}}_i\right)=\mathbf{0}$. Therefore,

$$\sqrt{N}\left(\check{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)=\begin{pmatrix}\left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathbf{X}}^{-1}\end{pmatrix}\left[\begin{pmatrix}N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \\ \mathbf{0}\end{pmatrix}\left[\sqrt{N}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta}+N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'U_i\right].$$

We can now look at $\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)$, the first G elements of $\sqrt{N}\left(\check{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$. But

$$N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \xrightarrow{p} \begin{pmatrix}\rho_1 \\ \rho_2 \\ \vdots \\ \rho_G\end{pmatrix}$$

and so

$$\sqrt{N}\left(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)=\left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1}\left[\begin{pmatrix}\rho_1 \\ \rho_2 \\ \vdots \\ \rho_G\end{pmatrix}N^{-1/2}\sum_{i=1}^{N}\dot{\mathbf{X}}_i\boldsymbol{\beta}+N^{-1/2}\sum_{i=1}^{N}\mathbf{W}_i'U_i\right]+o_p(1).$$

Note that

$$\mathbf{W}_i'\mathbf{W}_i = \begin{pmatrix} W_{i1} & 0 & \cdots & 0 \\ 0 & W_{i2} & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & W_{iG} \end{pmatrix}$$

and so

$$\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right) = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G \end{pmatrix}.$$

Therefore,

$$\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \mathbf{j}_G N^{-1/2} \sum_{i=1}^{N} \dot{\mathbf{X}}_i \boldsymbol{\beta} + \left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1} N^{-1/2} \sum_{i=1}^{N} \mathbf{W}_i' U_i + o_p(1) \qquad ((\text{k1}))$$

where $\mathbf{j}_G = (1, 1, ..., 1)'$. Now write

$$\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \mathbf{j}_G N^{-1/2} \sum_{i=1}^{N} \dot{\mathbf{X}}_i \boldsymbol{\beta} + \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G \end{pmatrix}^{-1} \left[N^{-1/2} \sum_{i=1}^{N} \mathbf{W}_i' U_i\right] + o_p(1)$$

$$\begin{aligned} \mathbf{W}_i' U_i &= \begin{pmatrix} W_{i1} \\ W_{i2} \\ \vdots \\ W_{iG} \end{pmatrix} \left[\sum_{h=1}^{G} (W_{ih} - \rho_g) \dot{\mathbf{X}}_i \boldsymbol{\beta}_g + \sum_{h=1}^{G} W_{ih} U_i(h)\right] \\ &= \begin{pmatrix} W_{i1} \sum_{h=1}^{G} (W_{ih} - \rho_h) \dot{\mathbf{X}}_i \boldsymbol{\beta}_h \\ W_{i2} \sum_{h=1}^{G} (W_{ih} - \rho_h) \dot{\mathbf{X}}_i \boldsymbol{\beta}_h \\ \vdots \\ W_{iG} \sum_{h=1}^{G} (W_{ih} - \rho_h) \dot{\mathbf{X}}_i \boldsymbol{\beta}_h \end{pmatrix} + \begin{pmatrix} W_{i1} U_i(1) \\ W_{i2} U_i(2) \\ \vdots \\ W_{iG} U_i(G) \end{pmatrix} \end{aligned}$$

and so

$$\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)=N^{-1/2}\sum_{i=1}^{N}\left[\begin{pmatrix}\dot{\mathbf{X}}_i\boldsymbol{\beta}\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}\\ \vdots\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}\end{pmatrix}+\begin{pmatrix}\sum_{h=1}^{G}W_{i1}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\\ \sum_{h=1}^{G}W_{i2}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\\ \vdots\\ \sum_{h=1}^{G}W_{iG}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\end{pmatrix}+\begin{pmatrix}W_{i1}U_i(1)/\rho_1\\ W_{i2}U_i(2)/\rho_2\\ \vdots\\ W_{iG}U_i(G)/\rho_G\end{pmatrix}\right] \quad ((k4))$$

For each g, we can write the second term in bracNow combine the first and second parts and simplify using the expression for $\boldsymbol{\beta}$. For example,

$$\begin{aligned}\sum_{g=1}^{G}W_{i1}(W_{ig}-\rho_g)\dot{\mathbf{X}}_i\boldsymbol{\beta}_g/\rho_1 &= \rho_1^{-1}\left[W_{i1}(W_{i1}-\rho_1)\dot{\mathbf{X}}_i\boldsymbol{\beta}_1-W_{i1}\rho_2\dot{\mathbf{X}}_i\boldsymbol{\beta}_2-\cdots-W_{i1}\rho_G\dot{\mathbf{X}}_i\boldsymbol{\beta}_G\right]\\ &= \rho_1^{-1}\dot{\mathbf{X}}_i\left[W_{i1}(1-\rho_1)\boldsymbol{\beta}_1-W_{i1}\rho_2\boldsymbol{\beta}_2-\cdots-W_{i1}\rho_G\boldsymbol{\beta}_G\right]\\ &= \rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left[\boldsymbol{\beta}_1-(\rho_1\boldsymbol{\beta}_1+\rho_2\boldsymbol{\beta}_2+\cdots+\rho_G\boldsymbol{\beta}_G)\right]\\ &= \rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}\right).\end{aligned}$$

Using (k4) and adding $\dot{\mathbf{X}}_i\boldsymbol{\beta}$ and rearranging, we obtain the following theorem,

Theorem 5 (Asymptotic variance of Pooled regression adjustment estimator of PO means)**.** *Under assumptions (1) and (2), along with finite second moments*

$$\begin{aligned}\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right) &= \begin{pmatrix}N^{-1/2}\sum_{i=1}^{N}\left[\rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{i1}U_i(1)/\rho_1\right]\\ N^{-1/2}\sum_{i=1}^{N}\left[\rho_2^{-1}W_{i2}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_2-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{i2}U_i(2)/\rho_2\right]\\ \vdots\\ N^{-1/2}\sum_{i=1}^{N}\left[\rho_G^{-1}W_{iG}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_G-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{iG}U_i(G)/\rho_G\right]\end{pmatrix}+o_p(1)\\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{F}_i+\mathbf{K}_i+\mathbf{Q}_i\right)+o_p(1)\end{aligned}$$

where $\mathbf{K}_i$ *and* $\mathbf{Q}_i$ *are as before and, with* $\boldsymbol{\delta}_g=\boldsymbol{\beta}_g-\boldsymbol{\beta}$*,*

$$\mathbf{F}_i=\begin{pmatrix}\rho_1^{-1}\left(W_{i1}-\rho_1\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_1\\ \rho_2^{-1}\left(W_{i2}-\rho_2\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_2\\ \\ \rho_G^{-1}\left(W_{iG}-\rho_G\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_G\end{pmatrix} \tag{5}$$

Notice that, again by random assignment and the linear projection property,

$$\mathbb{E}\left(\mathbf{F}_i\mathbf{K}_i'\right) = \mathbb{E}\left(\mathbf{F}_i\mathbf{Q}_i'\right) = \mathbf{0}$$

4 Comparing the Asymptotic Variances

We now take the representations derived in Section 3 and use them to compare the asymptotic variances of the three estimators. For notational clarity, it is helpful summarize the conclusions reached in Section 3:

$$\begin{aligned}
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM} - \boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\boldsymbol{L}_i + \mathbf{Q}_i\right) + o_p(1) \\
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA} - \boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{K}_i + \mathbf{Q}_i\right) + o_p(1) \\
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{PRA} - \boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{F}_i + \mathbf{K}_i + \mathbf{Q}_i\right) + o_p(1),
\end{aligned}$$

where $\mathbf{L}_i$, $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{F}_i$ are defined in 2, 3, 4 and 5 respectively.

4.1 Comparing FRA to Subsample Means

We now show that, asymptotically, $\hat{\boldsymbol{\mu}}_{FRA}$ is no worse than $\hat{\boldsymbol{\mu}}_{SM}$. From (m1), (m2), $\mathbb{E}\left(\mathbf{L}_i\mathbf{Q}_i'\right) = \mathbf{0}$, and $\mathbb{E}\left(\mathbf{K}_i\mathbf{Q}_i'\right) = \mathbf{0}$, it follows that

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM} - \boldsymbol{\mu}\right)\right] = \boldsymbol{\Omega}_{\mathbf{L}} + \boldsymbol{\Omega}_{\mathbf{Q}}$$

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA} - \mu\right)\right] = \boldsymbol{\Omega}_{\mathbf{K}} + \boldsymbol{\Omega}_{\mathbf{Q}}$$

where $\boldsymbol{\Omega}_{\mathbf{L}} = \mathbb{E}\left(\mathbf{L}_i\mathbf{L}_i'\right)$ and so on. Therefore, to show that Avar $\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA} - \boldsymbol{\mu}\right)\right]$ is smaller (in the matrix sense), we must show

$$\boldsymbol{\Omega}_{\mathbf{L}} - \boldsymbol{\Omega}_{\mathbf{K}}$$

is PSD, where

$$\mathbf{K}_i = \begin{pmatrix} \dot{\mathbf{X}}_i\boldsymbol{\beta}_1 \\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_2 \\ \vdots \\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_G \end{pmatrix} \text{ and } \mathbf{L}_i = \begin{pmatrix} W_{i1}\dot{\mathbf{X}}_i\boldsymbol{\beta}_1/\rho_1 \\ W_{i2}\dot{\mathbf{X}}_i\boldsymbol{\beta}_2/\rho_2 \\ \vdots \\ W_{iG}\dot{\mathbf{X}}_i\boldsymbol{\beta}_G/\rho_G \end{pmatrix}$$

The elements of $\mathbf{L}_i$ are uncorrelated because $W_{ig}W_{ih} = 0$ for $g \neq h$. The variance of the g^{th} element is

$$\mathbb{E}\left[\left(W_{ig}\dot{\mathbf{X}}_i\boldsymbol{\beta}_g/\rho_g\right)^2\right] = \mathbb{E}\left(W_{ig}\right)\rho_g^{-2}E\left[\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right)^2\right] = \rho_g^{-1}\mathbb{E}\left[\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right)^2\right] = \rho_g^{-1}\boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_g.$$

Therefore,

$$\begin{aligned} \mathbb{E}\left(\mathbf{L}_i\mathbf{L}_i'\right) &= \begin{pmatrix} \boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\beta}_2'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_2/\rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\beta}_G'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_G/\rho_G \end{pmatrix} \\ &= \mathbf{B}'\begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{X}}/\rho_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Omega}_{\mathbf{X}}/\rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\Omega}_{\mathbf{X}}/\rho_G \end{pmatrix}\mathbf{B} = \mathbf{B}'\left[\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} \otimes \boldsymbol{\Omega}_{\mathbf{X}}\right]\mathbf{B} \end{aligned}$$

where

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\beta}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\beta}_G \end{pmatrix}$$

For the variance matrix of $\mathbf{K}_i$,

$$\begin{aligned} \mathbb{V}\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right) &= \boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_g \\ \mathbb{C}(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g, \dot{\mathbf{X}}_i\boldsymbol{\beta}_h) &= \boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_h \end{aligned}$$

Therefore,

$$\mathbb{E}\left(\mathbf{K}_i\mathbf{K}_i'\right) = \mathbf{B}'\begin{pmatrix} \mathbf{\Omega_X} & \mathbf{\Omega_X} & \cdots & \mathbf{\Omega_X} \\ \mathbf{\Omega_X} & \mathbf{\Omega_X} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{\Omega_X} \\ \mathbf{\Omega_X} & \cdots & \mathbf{\Omega_X} & \mathbf{\Omega_X} \end{pmatrix}\mathbf{B} = \mathbf{B}'\left[\left(\mathbf{j}_G\mathbf{j}_G'\right)\otimes\mathbf{\Omega_X}\right]\mathbf{B}$$

where $\mathbf{j}_G' = (1, 1, ..., 1)$. Therefore, the comparison we need to make is

$$\begin{pmatrix} \rho_1^{-1} & 0 & & 0 \\ 0 & \rho_2^{-1} & & \\ & & & 0 \\ 0 & & 0 & \rho_G^{-1} \end{pmatrix}\otimes\mathbf{\Omega_X} \text{ versus } \left(\mathbf{j}_G\mathbf{j}_G'\right)\otimes\mathbf{\Omega_X}$$

That is, we need to show

$$\left[\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} - \left(\mathbf{j}_G\mathbf{j}_G'\right)\right]\otimes\mathbf{\Omega_X}$$

is PSD. The Kronecker product of two PSD matrices is also PSD, so it suffices to show

$$\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} - \left(\mathbf{j}_G\mathbf{j}_G'\right)$$

is PSD when the ρ_g add to unity. Let $\mathbf{a}$ be any $G \times 1$ vector. Then

$$\mathbf{a}' \begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} \mathbf{a} = \sum_{g=1}^{G} a_g^2/\rho_g$$

$$\mathbf{a}' \left(\mathbf{j}_G \mathbf{j}_G'\right) \mathbf{a} = \left(\mathbf{a}'\mathbf{j}_G\right)^2 = \left(\sum_{g=1}^{G} a_g\right)^2$$

So we have to show

$$\sum_{g=1}^{G} a_g^2/\rho_g \geq \left(\sum_{g=1}^{G} a_g\right)^2.$$

Define vectors $\mathbf{b} = \left(a_1/\sqrt{\rho_1}, a_2/\sqrt{\rho_2}, ..., a_G/\sqrt{\rho_G}\right)'$ and $\mathbf{c} = \left(\sqrt{\rho_1}, \sqrt{\rho_2}, ..., \sqrt{\rho_G}\right)'$ and apply the Cauchy-Schwarz inequality:

$$\begin{aligned} \left(\sum_{g=1}^{G} a_g\right)^2 &= \left(\mathbf{b}'\mathbf{c}\right)^2 \leq \left(\mathbf{b}'\mathbf{b}\right)\left(\mathbf{c}'\mathbf{c}\right) = \left(\sum_{g=1}^{G} a_g^2/\rho_g\right)\left(\sum_{g=1}^{G} \rho_g\right) \\ &= \left(\sum_{g=1}^{G} a_g^2/\rho_g\right) \end{aligned}$$

because $\sum_{g=1}^{G} \rho_g = 1$. This completes the derivation.

We summarize with a theorem.

Theorem 6. *Under assumptions of theorems 3 and 4,*

$$Avar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM} - \boldsymbol{\mu}\right)\right] - Avar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA} - \boldsymbol{\mu}\right)\right] = \boldsymbol{\Omega_L} - \boldsymbol{\Omega_K} \tag{6}$$

is PSD.

The one case where there is no gain in asymptotic efficiency in using FRA is when $\boldsymbol{\beta}_g = \mathbf{0}$, $g = 1, ..., G$, in which case $\mathbf{X}$ does not help predict any of the potential outcomes. Importantly, there is no gain in asymptotic efficiency in imposing $\boldsymbol{\beta}_g = \mathbf{0}$, which is what the subsample means estimator does. From an asymptotic perspective, it is harmless to separately estimate the $\boldsymbol{\beta}_g$ even when they are zero. When they are not all zero, estimating them leads to asymptotic efficiency

gains.

Theorem 6 implies that any smooth nonlinear function of $\boldsymbol{\mu}$ is estimated more efficiently using $\hat{\boldsymbol{\mu}}_{FRA}$. For example, in estimating a percentage difference in means, we would be interested in μ_2/μ_1, and using the FRA estimators is asymptotically more efficient than using the SM estimators.

4.2 Full RA versus Pooled RA

The comparision between FRA and PRA is simple given the expressions in (m2) and (m3) because, as stated earlier, $\mathbf{F}_i$, $\mathbf{K}_i$, and $\mathbf{Q}_i$ are pairwise uncorrelated.

Theorem 7. *Under the assumptions of theorem 4 and 5,*

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{PRA}-\boldsymbol{\mu}\right)\right]-\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]=\boldsymbol{\Omega}_{\boldsymbol{F}}$$

which is PSD.

Therefore, $\hat{\boldsymbol{\mu}}_{FRA}$ is never less asymptotically efficient than $\hat{\boldsymbol{\mu}}_{PRA}$. There are some special cases where the estimators achieve the same asymptotic variance, the most obvious being when the slopes in the linear projections are homogeneous:

$$\boldsymbol{\beta}_1=\boldsymbol{\beta}_2=\cdots=\boldsymbol{\beta}_G$$

As with comparing FRA with with subsample means, there is no gain in efficiency from imposing this restriction when it is true. This is another fact that makes FRA attractive if the sample size is not small.

Other situations where there is no asymptotic efficiency gain in using FRA are more subtle. In general, suppose we are interested in linear combinations $\tau=\mathbf{a}'\boldsymbol{\mu}$ for a given $G\times 1$ vector $\mathbf{a}$. If

$$\mathbf{a}'\boldsymbol{\Omega}_{\mathbf{F}}\mathbf{a}=0$$

then $\mathbf{a}'\hat{\boldsymbol{\mu}}_{PRA}$ is asymptotically as efficient as $\mathbf{a}'\hat{\boldsymbol{\mu}}_{FRA}$ for estimating τ. Generally, the diagonal

elements of

$$\mathbf{\Omega}_F = \mathrm{E}\left(\mathbf{F}_i\mathbf{F}_i'\right)$$

are

$$\frac{(1-\rho_g)}{\rho_g}\boldsymbol{\delta}_g'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_g$$

because $E\left[(W_{ig}-\rho_g)^2\right] = \rho_g(1-\rho_g)$. The off diagonal terms of $\mathbf{\Omega}_\mathbf{F}$ are

$$-\boldsymbol{\delta}_g'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_h$$

because $E\left[(W_{ig}-\rho_g)(W_{ih}-\rho_h)\right] = -\rho_g\rho_h$. Now consider the case covered in Negi and Wooldridge (2019), where $G = 2$ and $\mathbf{a}' = (-1, 1)$, so the parameter of interest is $\tau = \mu_2 - \mu_1$ (the average treatment effect). If $\rho_1 = \rho_2 = 1/2$ then

$$\mathbf{\Omega}_\mathbf{F} = \begin{pmatrix} \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & -\boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_2 \\ -\boldsymbol{\delta}_2'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & \boldsymbol{\delta}_2'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_2 \end{pmatrix}.$$

Now $\boldsymbol{\delta}_2 = -\boldsymbol{\delta}_1$ because $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1 - (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)/2 = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)/2 = -\boldsymbol{\delta}_2$. Therefore,

$$\mathbf{\Omega}_\mathbf{F} = \begin{pmatrix} \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 \end{pmatrix}$$

and

$$\begin{pmatrix} -1 & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 & \boldsymbol{\delta}_1'\mathbf{\Omega}_\mathbf{X}\boldsymbol{\delta}_1 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = 0.$$

This finding does not extend to the $G \geq 3$ case when Interestingly, it is not true that for estimating each mean separately that PRA is asymptotically equivalent to FRA. So, for example, with lower bound WTP, it might require that bid values have the same frequency. But it is not clear that even that is sufficient.

What about general G with $\rho_g = 1/G$ for all g? Then

$$1 - \rho_g = 1 - \frac{1}{G} = \frac{(G-1)}{G}$$

and so

$$\frac{1-\rho_g}{\rho_g} = G - 1.$$

Note that

$$\boldsymbol{\delta}_g = \boldsymbol{\beta}_g - (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 + \cdots + \boldsymbol{\beta}_G)/G$$

and it is less clear when there is a degeneracy. Seems very likely for estimating pairwise differences. In fact, seems like that almost must be true because can always do each two-by-two case, right? What about the diff-in-diffs parameter, or lower bound WTP?

Do some $G = 3$ simulations. Are the pairwise differences efficiently estimated using PRA under equal assignment? xx simulations show that in the $G \geq 3$ case that it is still better to do separate, I think.

5 Nonlinear Regression Adjustment

We now discuss a class of nonlinear regression adjustment methods that preserve consistency without adding additional assumptions (other than weak regularity conditions). In particular, we extend the setup in Negi and Wooldridge (2019) to allow for more than two treatment levels.

We show that both separate and pooled methods are consistent provided we choose the mean functions and objective functions appropriately. Not surprisingly, using a canonical link function in the context of quasi-maximum likelihood in the linear exponential family plays a key role.

Unlike in the linear case, we can only show that full RA improves over the subsample means estimator when the conditional mean is correctly specified. Whether one can prove efficiency more general is an interesting topic for future research.

5.1 Full Regression Adjustment

We model the conditional means, $E[Y(g)|\mathbf{X}]$, for each $g = 1, 2, ..., G$. Importantly, we will not assume that the means are correctly specified. As it turns out, to ensure consistency, the mean should have the index form common in the generalized linear models literature. In particular, we use mean functions

$$m(\alpha_g + \mathbf{x}\boldsymbol{\beta_g}),$$

where $m(\cdot)$ is a smooth function defined on $\mathbb{R}$. The range of $m(\cdot)$ is chosen to reflect the nature of $Y(g)$. Given that the nature of $Y(g)$ does not change across g, we choose a common function $m(\cdot)$ across all g. Also, as usual, the vector $\mathbf{X}$ can include nonlinear functions (typically squares, interactions, and so on) of underlying covariates.

As discussed in Negi and Wooldridge (2019) in the binary treatment case, the function $m(\cdot)$ is tied to a specific quasi-log-likelihood function in the linear exponential family (LEF). Table 1 gives the pairs of mean function and quasi-log-likelihood function that ensure consistent estimation. Consistent estimation follows from the results on doubly-robust estimation in the context of missing data in Wooldridge (2007). Each quasi-LLF is tied to the mean function associated with the canonical link function.

Table 1: Combinations of Means and QLLFs to Ensure Consistency

Support Restrictions	**Mean Function**	**Quasi-LLF**
None	Linear	Gaussian (Normal)
$Y(g) \in [0, 1]$ (binary, fractional)	Logistic	Bernoulli
$Y(g) \in [0, B]$ (count, corners)	Logistic	Binomial
$Y(g) \geq 0$ (count, continuous, corner)	Exponential	Poisson
$Y_j(g) \geq 0$, $\sum_{j=0}^{J} Y_j(g) = 1$	Logistic	Multinomial

The binomial QMLE is rarely applied, but is a good choice for counts with a known upper bound, even if it is individual-specific (so B_i is a positive integer for each i). It can also be applied to corner solution outcomes in the interval $[0, B_i]$ where the outcome is continuous on $(0, B_i)$ but perhaps has mass at zero or B_i. The leading case is $B_i = B$ for all i. Note that we do not recommend a

Tobit model in such cases because Tobit is not generally robust to distributional or mean failure. Combining the multinomial QLL and the logistic mean functions is attractive when the outcome is either a multinomial response or more than two shares that necessarily sum to unity.

As discussed in Wooldridge (2007), the key feature of the single outcome combinations in Table 1 is that it is always true that

$$E\left[Y(g)\right] = E\left[m(\alpha_g^* + \mathbf{X}\boldsymbol{\beta}_g^*)\right],$$

where α_g^* and $\boldsymbol{\beta}_g^*$ are the probability limits of the QMLEs whether or not the conditional mean function is correctly specified. The analog also holds for the multinomial logit objective function.

Applying nonlinear RA with multiple treatment levels is straightforward. For treatment level g, after obtaining $\hat{\alpha}_g$, $\hat{\boldsymbol{\beta}}_g$ by quasi-MLE using only units from treatment level g, the mean, μ_g, is estimated as

$$\hat{\mu}_g = N^{-1}\sum_{i=1}^{N} m(\hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_g),$$

which includes linear RA as a special case. This estimator is consistent by a standard application of the uniform law of large numbers; see, for example, Wooldridge (2010) (xx 12.xx).

As in the linear case, and of the mean/QLL combinations in Table 1 allow us to write the subsample average as

$$\bar{Y}_g = N_g^{-1}\sum_{i=1}^{N} W_{ig}m(\hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_g).$$

It seems that $\hat{\mu}_g$ should be asymptotically more efficient than $\bar{Y}_g$ because $\hat{\mu}_g$ averages across all of the data rather than just the units at treatment level g. Unfortunately, the proof used in the linear case does not go through in the nonlinear case. At this point, we must be satisfied with consistent estimators of the POs that impose the logical restrictions on $E\left[Y(g)|\mathbf{X}\right]$. In the binary treatment case, Negi and Wooldridge (2019) find nontrivial efficiency gains in using logit, fractional logit, and Poisson regression even compared with full linear RA.

5.2 Pooled Regression Adjustment

In cases where N is not especially large, one might, just as in the linear case, resort to pooled RA. Provided the mean/QLL combinations are chosen as in Table 1, the pooled RA estimator is still consistent under arbitrary misspecification of the mean function. To see why, write the mean function, without an intercept in the index, as

$$m(\gamma_1 w_1 + \gamma_2 w_2 + \cdots + \gamma_G w_G + \boldsymbol{x}\boldsymbol{\beta}).$$

The first-order conditions of the pooled QMLE include the G conditions

$$N^{-1}\sum_{i=1}^{N} W_{ig}\left[Y_i - m(\hat{\gamma}_1 W_{i1} + \hat{\gamma}_2 W_{i2} + \cdots + \hat{\gamma}_G W_{iG} + \mathbf{X}_i\hat{\boldsymbol{\beta}})\right] = 0,\ g = 1, ..., G.$$

Therefore, assuming no degeneracies, the probability limits of the estimators, denoted with a $*$, solve the population analogs:

$$\mathrm{E}\left(W_g Y\right) = \mathrm{E}\left[W_g Y\left(g\right)\right] = \mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right],$$

where $\mathbf{W} = (W_1, W_2, ..., W_G)$. By random assignment, $\mathrm{E}\left[W_g Y\left(g\right)\right] = \rho_g \mu_g$. By iterated expectations and random assignment,

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)\right] = \mathrm{E}\left\{\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)|\mathbf{X}\right]\right\}$$

and

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)|\mathbf{X}\right] = \mathrm{P}(W_g = 1|\mathbf{X}) m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*) = \rho_g m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*).$$

Therefore,

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)\right] = \rho_g \mathrm{E}\left[m(\gamma_g^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right]$$

and, using $\rho_g > 0$, we have shown

$$\mu_g = \mathrm{E}\left[m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*)\right]$$

By definition, $\hat{\gamma}_g$ is consistent for γ_g^* and $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$. Therefore, after the pooled QMLE estimation, we obtain the estimated means as

$$\check{\mu}_g = N^{-1} \sum_{i=1}^{N} m(\hat{\gamma}_g + \mathbf{X}_i \hat{\boldsymbol{\beta}}),$$

and these are consistent by application of the uniform law of large numbers.

As in the case of comparing full nonlinear RA to the subsample averages, we have no general asymptotic efficiency results comparing full nonlinear RA to pooled nonlinear RA. As shown in Section 4.2, in the linear case it is never worse, asymptotically, to use full RA.

6 Applications

6.1 Treatment Effects with Multiple Treatment Levels

The most direct application of the previous results is in the context of a randomized intervention with more than two treatment levels. Regression adjustment can be used for any kind of response variable. With a reasonable sample size per treatment level, full regression adjustment is preferred to pooled regression adjustment.

If the outcome $Y(g)$ is restricted in some substantive way, a nonlinear RA method of the kind described in Section 5 can be used to exploit the logical restrictions on $\mathrm{E}\left[Y(g)|\mathbf{X}\right]$. While we cannot show this guarantees efficiency gains compared with using subsample averages, the simulation findings in Negi and Wooldridge (2019) suggest the gains can be nontrivial – even compared with full linear RA.

6.2 Difference-in-Differences Designs

Difference-in-differences applications can be viewed as a special case of multiple treatment levels. For illustration, consider the standard setting where there is a single before period and a single

post treatment period. Let C be the control group and T the treatment group. Label B the before period and A the after period. The standard DID treatment effect is a particular linear combination of the means from the four groups:

$$\tau = (\mu_{TA} - \mu_{TB}) - (\mu_{CA} - \mu_{CB})$$

Estimating the means by separate regression adjustment is generally better than not controlling for covariates, or putting them in additively.

6.3 Estimating Lower Bound Mean Willingness-to-Pay

In the context of contingent valuation, individuals are randomly presented with the price of a new good or tax for a new project. They are asked whether they would purchase the good at the given price, or be in favor of the project at the given tax. Generally, the price or tax is called the "bid value." The outcome for each individual is a binary "vote" (yes = 1, no = 0).

A common approach in CV studies is to estimate a lower bound on the mean willingness-to-pay (WTP). The common estimators are based on the area under the WTP survival function:

$$\mathrm{E}(WTP) = \int_0^\infty S(a)da$$

When a population of individuals is presented with a small number of bid values, it is not possible to identify $\mathrm{E}(WTP)$, but only a lower bound. Specifically, let b_1, b_2, ..., b_G be G bid values and define the binary potential outcomes as

$$Y(g) = 1[WTP > b_g],\ g = 1, ..., G.$$

In other words, if a person is presented with bid value b_g, $Y(g)$ is the binary response, which is assumed to be unity if WTP exceeds the bid value. The connection with the survivor function is

$$\mu_g \equiv \mathrm{E}\left[Y(g)\right] = P(WTP > b_g) = S(b_g)$$

Notice that μ_g is the proportion of people in the population who have a WTP exceeding b_g. This fits into the potential outcomes setting because each person is presented with only one bid value. Standard consumer theory implies that $\mu_{g+1} \leq \mu_g$, which simply means the demand curve is weakly declining in price.

It can be shown that, with $b_0 \equiv 0$ for notational ease,

$$\tau \equiv \sum_{g=1}^{G} (b_g - b_{g-1})\mu_g \leq \mathrm{E}(WTP),$$

and it is this particular linear combination of $\{\mu_g : g = 1, 2, ..., G\}$ that we are interested in estimating. The so-called ABERS (1955) estimator introduced by Ayer et al. (1955), without a downward sloping survival function imposed, replaces μ_g with its sample analog:

$$\hat{\tau}_{ABERS} = \sum_{g=1}^{G} (b_g - b_{g-1})\bar{Y}_g$$

where

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} Y_i 1[B_i = b_g]$$

is the fraction of yes votes at bid value b_g. Of course, the $\bar{Y}_g$ can also be obtained as the coefficients from the regression

$$Y_i \text{ on } Bid1_i, Bid2_i, ..., BidG_i, i = 1, ..., N$$

Lewbel (2000) and Watanabe (2010) allows for covariates in order to see how WTP changes with individual or family characteristics and attitudes, but here we are interested in estimating τ.

We can apply the previous results on efficiency because τ is a linear combination of the μ_g. Therefore, using separate linear RA to estimate each μ_g, and then forming

$$\hat{\tau}_{FRA} = \sum_{g=1}^{G} (b_g - b_{g-1})\hat{\mu}_g$$

is generally asymptotically more efficient than $\hat{\tau}_{ABERS}$. Moreover, because Y is a binary outcome, we might improve efficiency further by using logit models at each bid value to obtain the $\hat{\mu}_g$.

6.4 Application to California Oil Data

This section applies the linear RA estimators discussed in section 3 to survey data from the California Oil Spill study from Carson et al. (2004). The study implemented a CV survey to assess the value of damages to natural resources from future oil spills along California's Central Coast. This was achieved by estimating a lower bound mean WTP measure of the cost of such spills to California's residents. The survey provided respondents with the choice of voting for or against a governmental program that would prevent natural resource injuries to shorelines and wildlife along California's central coast over the next decade. In return, the public would be asked to pay a one time lump sum income tax surcharge for setting up the program.

The main sample survey which was used to elicit the yes or no votes was conducted by Westat, Inc. The data was a random sample of 1,085 interviews conducted with English speaking Californian households where the respondent was 18 years or older, and lived in private residences that were either owned or rented. To address issues of non-representativeness of the interviewed sample from the total initially chosen sample, weights were used. Each respondent was randomly assigned one of five tax amounts: \$5, \$25, \$65, \$120, or \$220 and the binary choice of "yes" or "no" for the oil spill prevention program was recorded at the randomly assigned tax amount.

Apart from the choice at different bid amounts, data was also collected on demographics for the respondent and the respondent's household such as total income, prior knowledge of the spill site, distance to the site, environmental attitudes, attitudes towards big businesses, understanding of the program and the task of voting, beliefs about the oil spill scenario etc.

Table 1 provides a summary of yes votes at the different bid or tax amounts presented to the respondents. Table 2 provides estimates for the PO means as well as the lower bound mean WTP estimate. We see that the FRA estimator delivers more precise estimates for both the PO means and the WTP estimate than the ABERS estimator.

Table 1: Summary of yes votes at different bid amounts

Bid	**Yes-votes**	**%**
\$5	219	20
\$25	216	20
\$65	241	22
\$120	181	17
\$220	228	21
Total	**1085**	**100**

Table 2: Lower bound mean willingness to pay estimate using ABERS and FRA estimators

	PO means	
Bids	**SM**	**FRA**
\$5	0.689	0.685
	(0.0313)	(0.0288)
\$25	0.569	0.597
	(0.0338)	(0.0307)
\$65	0.485	0.489
	(0.0323)	(0.0294)
\$120	0.403	0.378
	(0.0365)	(0.0332)
\$220	0.289	0.290
	(0.0301)	(0.0286)
	ABERS	**FRA**
$\hat{\tau}$	85.39	84.67
	(3.905)	(3.792)
Obs	**1085**	**1085**

7 Concluding Remarks

In this paper, we build on the work of Negi and Wooldridge (2019) to study efficiency improvements in linear regression adjustment estimators when there are more than two treatment levels. In particular, we consider any arbitrary 'G' number of treatments when these treatments have been randomly assigned. We show that jointly estimating the vector of potential outcome means using linear RA that allows for separate slopes for the different assignment levels is asymptotically never worse than just using subsample averages. One case when there is no gain in asymptotic efficiency from using FRA is when the slopes are all zero. In other words, when the covariates are not predictive of the potential outcomes, then using separate slopes does not produce more precise estimates compared to just estimating the subsample averages. We also show that separate slopes RA is generally more efficient compared to pooled RA, unless the true linear projection slopes are homogeneous. In this case using FRA to estimate the vector of PO means is harmless. In other words, using FRA in this case does not hurt.

In addition, this paper also extends the discussion around nonlinear regression adjustment made in Negi and Wooldridge (2019) to more than two treatment levels. In particular, we show that pooled and separate nonlinear RA estimators in the quasi maximum likelihood family are consistent if one chooses the mean and objective functions appropriately from the linear exponential family of distributions.

As an illustration of these efficiency arguments, we apply the different linear RA estimators for estimating the lower bound mean WTP using data from a contingent valuation study undertaken to provide an ex-ante measure of damages to natural resources from future oil spills along California's central coast. We find that the lower bound mean WTP is estimated more efficiently when we allow the slopes on the different bid values to be estimated separately as opposed to the ABERS estimator, which uses subsample averages for the PO means.

References

Angrist, J., Bettinger, E., and Kremer, M. Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American economic review*, 96(3): 847–862, 2006.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955.

Calónico, S. and Smith, J. The women of the national supported work demonstration. *Journal of Labor Economics*, 35(S1):S65–S97, 2017.

Carson, R. T., Conaway, M. B., Hanemann, W. M., Krosnick, J. A., Mitchell, R. C., and Presser, S. Valuing oil spill prevention, 2004.

Hirano, K. and Imbens, G. W. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278, 2001.

Lewbel, A. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177, 2000.

Negi, A. and Wooldridge, J. M. Regression adjustment in experiments with heterogeneous treatment effects. *Working paper*, 2019.

Watanabe, M. Nonparametric estimation of mean willingness to pay from discrete response valuation data. *American Journal of Agricultural Economics*, 92(4):1114–1135, 2010.

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.