NBER WORKING PAPER SERIES

# USING MACHINE LEARNING FOR EFFICIENT FLEXIBLE REGRESSION ADJUSTMENT IN ECONOMIC EXPERIMENTS

John A. List
Ian Muir
Gregory K. Sun

Working Paper 30756
http://www.nber.org/papers/w30756

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022

We thank Lyft Inc. for providing a large portion of the data used in this project. We additionally thank Adeline Sutton for her help in accessing and interpreting the CHECC data as well as Brent Hickman, Michael Cuna, Atom Vayalinkal, and participants at the Advances in Field Experiments conference for helpful comments which have improved the paper. Documentation of our procedures and our Stata and R code can be found here: https://github.com/gsun593/FlexibleRA The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w30756.ack

Using Machine Learning for Efficient Flexible Regression Adjustment in Economic Experiments
John A. List, Ian Muir, and Gregory K. Sun
NBER Working Paper No. 30756
December 2022
JEL No. C9,C90,C91,C93


## ABSTRACT

This study investigates how to use regression adjustment to reduce variance in experimental data. We show that the estimators recommended in the literature satisfy an orthogonality property with respect to the parameters of the adjustment. This observation greatly simplifies the derivation of the asymptotic variance of these estimators and allows us to solve for the efficient regression adjustment in a large class of adjustments. Our efficiency results generalize a number of previous results known in the literature. We then discuss how this efficient regression adjustment can be feasibly implemented. We show the practical relevance of our theory in two ways. First, we use our efficiency results to improve common practices currently employed in field experiments. Second, we show how our theory allows researchers to robustly incorporate machine learning techniques into their experimental estimators to minimize variance.



John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

Ian Muir
Lyft
muir.ian.m@gmail.com

Gregory K. Sun
University of Chicago
gsun1@uchicago.edu

# 1 Introduction

Since the pioneering work of Fisher (1935), it has been well established in the scientific community that experimentation through randomized controlled trials is the gold standard through which it is possible to learn about cause and effect. However, the large subsequent literature on the nuances of experimental methodology illustrates that randomization is merely the beginning, not the end, of the experimental process. After leveraging advantages of controlling the assignment mechanism, many decisions remain: the population of people and situations to examine, how many units to include, what outcomes and observables to collect, and how to best examine and model the collected data, to name a few.

In this paper, we focus on the problem of how to make use of pre-exposure covariates to reduce variance optimally when estimating parameters in an i.i.d. sampled experimental dataset. In particular, we derive the efficient regression adjustment over a large class of estimators in this setting. This class is sufficiently broad to subsume all common variance reduction strategies of which we are aware.[1] We show that this efficient regression adjustment can be feasibly estimated in a data driven way using non-parametric regression, including by making use of machine learning methods. Moreover, we find that it is possible to obtain asymptotic standard errors equal to what would be obtained if the optimal adjustment was ex-ante known, even though in general, the adjustment must be estimated. This property of our proposed estimators also makes the construction of asymptotically correct standard errors robust and computationally straightforward. We demonstrate that this asymptotic result is a good approximation in finite samples by conducting a number of simulation exercises and additionally demonstrate the application of our proposed estimator in four real world examples. Our results can also be used to critique certain practices in the experimental literature and show how they can cause researchers to conduct inefficient experiments.

The methods proposed in this paper extend and unify a number of literatures. Most directly, our paper expands upon the recent work of Negi and Wooldridge (2020) and Negi and Wooldridge (2021) (henceforth NW), who derive a number of results about better and worse ways to use linear regression to reduce variance in experimental settings. Our efficiency results also subsume those of Frison and Pocock (1992), who show the efficiency of regression adjustment over difference-in-differences (DiD) in settings where the researcher has access to pre-intervention outcomes. Indeed, the approach taken in our study unifies these previous developments under a single framework and allows us to show what NW denote "full regression adjustment" is efficient in a certain class of estimators which includes

[1] More precisely, this statement ignores attempts to reduce variance by changing the randomization approach, instead focusing on strategies to reduce variance via the choice of estimator given a fixed data generating process.

all of the estimators considered in the above three papers.[2]

Our paper also contributes to a broader literature on using covariates as a tool for variance reduction in settings with randomization. Here, we briefly review a number of recent papers which address settings different than ours. Roth and Sant'Anna (2021) (henceforth RS) consider a setting where researchers have access to a panel dataset and treatment timing is assigned at random. While we consider the case of i.i.d. sampling from an infinite population, the asymptotics in RS are with respect to a sequence of finite populations. Aside from this distinction, our recommendation to use pre-treatment observations as covariates rather than estimating a two-way fixed effects can be thought of as a special case of results in RS (for their example when there is only a single treatment timing group in addition to the control group). Our framework, however, is flexible enough to allow for the presence of other covariates. A related recent paper is due to Cohen and Fogarty (2021), which like RS considers asymptotics with respect to a sequence of finite populations. Their proposed estimator can be shown to be contained within the class considered in this paper, so at least under an i.i.d. sampling assumption, is no more efficient than ours.

Relative to the above two papers, an additional important difference is that our results apply to *any* function of potential outcome means, and therefore continue to apply even if the researcher is interested in estimating parameters other than average treatment effects. Finally, Carneiro et al. (2020) considers the optimal tradeoff between spending a limited budget on collecting more covariates compared to collecting more observations. As in their paper, we find that the expected benefits from regression adjustment can be approximately measured by the mean squared error in using the covariates to predict the outcome of interest.

Our results are also related to a recent literature describing the approaches to variance reduction in "big data" settings. In particular, Deng et al. (2013) propose an estimator for the treatment effect which is asymptotically equivalent to the "full regression adjustment" of NW.[3] Another approach to regression adjustment in big data settings is given by Poyarkov et al. (2016). The authors consider fitting a Machine Learning (ML) model of the conditional expectation function, but explicitly advocate against sample splitting. Our simulation results show that not sample splitting when the regression adjustment function is complex can produce a sizable bias *against* finding a treatment effect, even in fairly large samples.

Our advice is therefore in the spirit of Wager et al. (2016), who propose using ML methods and sample splitting to reduce variance in estimating average treatment effects.

[2]More specifically, NW also study the asymptotic properties of an estimator which is not contained in our class. However, this estimator is infeasible, relying on a priori knowledge of population quantities that the researcher is unlikely to know, and is therefore of little practical interest for our purposes.

[3]Specifically, the Deng et al. (2013) proposal is to do optimal regression adjustment with a single covariate that is easy to compute in the typical setting faced by a some firms.

Yet, it is important to point out that our results generalize their non-parametric results along two dimensions. First, following Negi and Wooldridge (2020), we compute the asymptotic standard errors from estimating the vector of treatment group mean outcomes, not just the average treatment effect. Via the delta method, this allows us to characterize the asymptotic sampling distribution when estimating any differentiable function of treatment group means, which includes the average treatment effect as a special case. Relatedly, while the claimed efficiency of regression adjustment in Wager et al. (2016) is based on the efficiency results of Hahn (1998) in the context of average treatment effects, we show that this efficiency extends to the joint estimation of the entire vector of potential outcome means, and hence to any function of these means. This generalization is especially valuable to economists who plan to use the experimental variation to learn about quantities beyond average treatment effects. Examples include Goldszmidt et al. (2021), who use treatment assignment as an instrument for measuring sensitivities to more fundamental quantities like price and time, DellaVigna et al. (2012), who use experimental variation to estimate parameters in a structural model, and Cotton et al. (2020), who use a structurally motivated field experiment to estimate a model of human capital production in adolescents.

An aspect of regression adjustment which to our knowledge has not been previously emphasized, but which we view as the key point underlying why regression adjustment works with relatively few assumptions, is that it satisfies an orthogonality property with respect to the adjustment term. This observation allows us to connect regression adjustment to a large literature in semiparametric statistics on estimating parameters in the presence of a high dimensional, but orthogonal nuisance parameter (e.g. Andrews (1994); Newey (1994); Chernozhukov et al. (2018)). This perspective allows us to give considerably shorter and conceptually more transparent proofs of key results about regression adjustments.

The plan for the rest of this paper is as follows. In Section 2, we develop our theory of regression adjustment. In Section 3, we use a number of empirical evaluations leveraging data from the ridesharing firm, Lyft, to show that the asymptotic theory provides a good guide for conducting inference and to show that the optimal choice of estimator can make a difference in practical settings. Section 4 concludes.

# 2 Theory

In this section, we develop the theory for our generalized regression adjustment. We take as our point of departure the potential outcomes model. Specifically, we assume that our data are generated from an experiment with treatment groups $\{1, \ldots, G\}$. Let the potential outcomes in group $g$ be denoted by $Y(g)$ so that the objects of interest are the average

potential outcome within each group:

$$\mu_g = \mathbb{E}[Y(g)]$$

We assume that treatment is assigned via simple random sampling and define $W_g$ to be an indicator that equals 1 if and only if an individual was randomized into group $g$ and 0 otherwise. Let bolded versions of letters be the vector obtained by stacking the versions with $g$ subscripts so that, for example, $\mathbf{W} = (W_1, \ldots, W_G)'$ is the vector of treatments, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)'$ is the vector of mean potential outcomes. We additionally assume that the researcher has access to a vector of pre-treatment covariates, $X$, which are potentially informative about the outcome $Y$. Our goal is to characterize the extent to which $X$ can be used to reduce variance when estimating $\boldsymbol{\mu}$. Because we are working with an experiment, we assume the usual orthogonality condition arising from statistical independence:

$$\{Y(g)\}_{g=1}^{G}, X \perp \mathbf{W}$$

We model the data as comprising an i.i.d. sample of size $n$ from the experiment, denoted $\{Y_i, X_i, \mathbf{W}_i\}_{i=1}^{n}$, where $Y_i = Y_i(g_i)$ for the unique $g_i$ satisfying $W_{i,g_i} = 1$. Finally, let $\rho_g = \Pr(W_g = 1)$ be the probability that an individual is in treatment group $g$, thus $\boldsymbol{\rho}$ is the vector of treatment probabilities. Because we are discussing asymptotic variances, we maintain the following mild regularity condition throughout.

**Assumption 1.** $\mathbb{E}[Y(g)^2] < \infty$ for all $g = 1, \ldots, G$.

With the notation in hand, we can describe our theory in 4 sections. In Section 2.1, we highlight the key intuition behind our approach by simplifying the asymptotic variance derivation relative to NW for the case of linear regression adjustment. The ideas used to obtain this simplification reveal how their efficiency results might be generalized. In Section 2.2, we show how to embed the linear regression adjustment into a more general class of regression adjustment estimators and derive the optimal adjustment within this class. In Section 2.3, we discuss how to implement feasibly the efficient estimator implied by Section 2.2 and derive the asymptotic distribution of this estimator, which we refer to as "flexible regression adjustment" (FRA). Finally, Section 2.4 provides practical guidance for how a researcher should do power calculations when they use FRA.

## 2.1 Revisiting Linear Regression Adjustment

In this section, we simplify the derivation of the asymptotic variance of the optimal linear regression adjustment (LRA) discussed in NW. Their derivations involve working through

a number of tedious matrix computations which our simplification obviates. In doing so, we provide an alternative perspective on why regression adjustment works, which in turn motivates our subsequent generalizations. NW write their LRA as

$$\hat{\mu}_g = \hat{\alpha}_g + \hat{\beta}_g' \bar{X}, \tag{1}$$

where $\hat{\alpha}_g, \hat{\beta}_g$ are the OLS coefficients from the regression, $Y = \alpha_g + \beta_g' X + \varepsilon$, fit using only the observations from treatment group $g$, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the mean value of $X$ across the entire sample. Let $n_g$ be the number of observations in treatment group $g$ and define $\bar{Y}_g = \frac{1}{n_g}\sum_{i:W_{i,g}=1} Y_i$ and $\bar{X}_g = \frac{1}{n_g}\sum_{i:W_{i,g}=1} X_i$, so respectively, they are the mean values of $Y$ and $X$ among the observations in treatment group $g$. A basic fact about OLS is that $\bar{Y}_g = \hat{\alpha}_g + \hat{\beta}_g' \bar{X}_g$. Using this fact, we can rearrange the LRA as

$$\hat{\mu}_g = \bar{Y}_g + \hat{\beta}_g' \left(\bar{X} - \bar{X}_g\right).$$

The consistency of the full regression adjustment follows from several observations. First, randomization implies that $\mathbb{E}[\bar{X} - \bar{X}_g] = 0$, while $\hat{\beta}_g$ converges in probability to a fixed vector, so the law of large numbers and the continuous mapping theorem imply that the second term in $\hat{\mu}_g$ converges in probability to 0. The first term, on the other hand, is the sample mean of the outcome in group $g$ and converges to $\mu_g$ by the law of large numbers.

Second, the fact that $\mathbb{E}[\bar{X} - \bar{X}_g] = 0$ also greatly simplifies the calculation of the asymptotic variance of the estimator. To see why, consider the family of estimators $\hat{\mu}_g^{\beta} = \bar{Y}_g + \beta'(\bar{X} - \bar{X}_g)$ which fixes the slope of the regression adjustment at some $\beta$. Then

$$\frac{\partial}{\partial \beta}\mathbb{E}[\hat{\mu}_g^{\beta}] = \frac{\partial}{\partial \beta}\mathbb{E}[\bar{Y}_g + \beta'(\bar{X} - \bar{X}_g)] = \mathbb{E}[\bar{X} - \bar{X}_g] = 0.$$

In other words, the class of estimators $\hat{\mu}_g^{\beta}$ is orthogonal to the adjustment $\beta$, which implies that the asymptotic variance of $\hat{\mu}_g$ is equivalent to the asymptotic variance of $\hat{\mu}_g^{\beta_g}$ where $\beta_g = \text{plim}_{n\to\infty}\hat{\beta}_g$ is the limiting value of $\hat{\beta}_g$.[4] Similarly, letting $\hat{\rho}_g$ be the proportion of individuals in treatment $g$, we can write

$$\hat{\mu}_g = \bar{Y}_g + \hat{\beta}_g'(\hat{\rho}_g \bar{X}_g + (1 - \hat{\rho}_g)(\bar{X}_{-g} - \bar{X}_g)) = \bar{Y}_g + \hat{\beta}_g'(1 - \hat{\rho}_g)(\bar{X}_{-g} - \bar{X}_g),$$

and $\hat{\mu}_g$ is also orthogonal to $\rho_g$, so replacing $\hat{\rho}_g$ with $\rho_g$, the probability that a given individual ends up in treatment $g$, will also not affect asymptotic variance computations. As a result,

[4] See, for example, Theorem 6.1 of Newey and McFadden (1994)

we can easily compute that

$$\lim_{n\to\infty} n\text{Var}(\hat{\mu}_g) = \frac{\text{Var}(Y(g))}{\rho_g} + \frac{1-\rho_g}{\rho_g}\left(\beta_g'\text{Var}(X)\beta_g - 2\beta_g'\text{Cov}(Y(g),X)\right). \tag{2}$$

The choice of $\beta_g$ only affects the limiting variance of $\hat{\mu}_g$ via the second term, and its effect on the asymptotic variance is given by a quadratic form, which can thus be minimized by setting its first derivative equal to 0, i.e.

$$\beta_g^* = \text{Var}(X)^{-1}\text{Cov}(Y(g),X).$$

This is precisely the population OLS slope for observations in treatment group $g$, provided $\hat{\beta}_g \xrightarrow{p} \beta_g^*$.[5] In this case, the second term of Equation (2) will be minimized and becomes

$$-\frac{1-\rho_g}{\rho_g}(\beta_g^*)'\text{Var}[X]\beta_g^*.$$

Moreover, if we let $U(g)$ be the OLS error in treatment group $g$, then we have $\text{Var}(Y(g)) = \text{Var}(U(g) + (\beta_g^*)'X) = \text{Var}(U(g)) + (\beta_g^*)'\text{Var}(X)\beta_g^*$. Note that by construction, $U(g)$ is uncorrelated with $X$. Putting everything together, we have that

$$\lim_{n\to\infty} n\text{Var}(\hat{\mu}_g) = \text{Var}(U(g))/\rho_g + (\beta_g^*)'\text{Var}(X)\beta_g^*.$$

This takes on the same form as in NW, and extending the above argument to also account for correlations across different groups $g$ allows us to replicate their asymptotic variance computations. Rather than doing this, in the next subsection, we present a more general argument that allows us to arrive at our main efficiency results.

## 2.2 Efficiency Theory

The regression adjusted estimator for $\mu_g$ is composed of a number of sample means (whose asymptotic variances are easy to deal with) and the OLS slope $\hat{\beta}_g$ (whose asymptotic variance is more difficult to deal with). The key simplification made in the above derivation is to notice that orthogonality with respect to $\beta_g$ allowed us to avoid discussing the asymptotic influence that the estimation of $\beta_g$ had on the final estimator. We now generalize this idea

[5] Which it does under our i.i.d. sampling assumption and Assumption 1

by relaxing the restriction that our regression adjustment is linear in $X$. To do so, we write

$$\beta_g' \bar{X} = \frac{1}{n}\sum_{i=1}^{n} \beta_g' X_i = \frac{1}{n}\sum_{i=1}^{n} h_g(X_i) \equiv \bar{h}_{g,all},$$
$$\beta_g' \bar{X}_g = \frac{1}{n_g}\sum_{i:W_{g,i}=1} \beta_g' X_i = \frac{1}{n_g}\sum_{i:W_{g,i}=1} h_g(X_i) \equiv \bar{h}_{g,g}.$$

In other words, we have that these two quantities can be written as sample means of some function $h_g$ of $X_i$ where $h_g$ is implicitly parameterized by $\beta_g$. Systematizing the above notational convention, in what follows, for any function $f_g(x)$, we introduce the following notation

$$\bar{f}_{g,all} = \frac{1}{n}\sum_{i=1}^{n} f_g(X_i), \quad \bar{f}_{g,g} = \frac{1}{n_g}\sum_{i:W_{i,g}=1} f_g(X_i),$$

so that the first subscript indicates the $g$ in the subscript in $f_g$ while the second subscript indicates the dataset for which the sample mean is taken. Given this notation, we now re-write the form of the regression adjustment estimators we consider as

$$\hat{\mu}_g^{h_g} = \bar{Y}_g + \bar{h}_{g,all} - \bar{h}_{g,g}. \tag{3}$$

For any fixed, square-integrable $h_g$, randomization implies that the second two terms in $\hat{\mu}_g^{h_g}$ as defined in Equation (3) cancel one another, i.e. $\mathbb{E}[\bar{h}_{g,all} - \bar{h}_{g,g}] = 0$. Orthogonality of the estimator with respect to $\beta$ in the linear case then generalizes to the following

$$\frac{\partial}{\partial r}\mathbb{E}\left[\hat{\mu}_g^{h_g + r h_g'}\right] = 0, \ \forall h_g, h_g' \in L^2(X).$$

Standard results about the asymptotic distribution of semi-parametric estimators (Newey (1994); Chernozhukov et al. (2018)) suggest that even if we must estimate $h_g$ using data, this estimation will have asymptotically negligible effects on the standard errors of the resulting estimator. Before formalizing this statement, we first pin down the optimal value of $h_g$ that minimizes asymptotic variance. To do so, define

$$m_g(x) \equiv \mathbb{E}[Y|X = x, G = g] \quad \text{so that} \quad Y_i(g) = m_g(X_i) + \varepsilon_i(g), \ \mathbb{E}[\varepsilon_i(g)|X_i] = 0. \tag{4}$$

Then denoting $d_g(x) = h_g(x) - m_g(x)$, we can rearrange $\hat{\mu}_g^{h_g}$ as

$$\hat{\mu}_g^{h_g} = \bar{\varepsilon}_g + \bar{m}_{g,g} + \bar{h}_{g,all} - \bar{h}_{g,g} = \underbrace{\bar{\varepsilon}_g}_{A_g} + \underbrace{\bar{m}_{g,all}}_{B_g} + \underbrace{(\bar{d}_{g,all} - \bar{d}_{g,g})}_{C_g^{\mathbf{h}}}, \tag{5}$$

where $\bar{\varepsilon}_g = \frac{1}{n_g}\sum_{i:W_{i,g}=1}\varepsilon_i(g)$ is the mean value of $\varepsilon_i(g)$ among individuals in treatment $g$. Since $\varepsilon_i(g)$ is mean independent of $X_i$, $A_g$ is uncorrelated with $B_{g'}$ and $C_{g'}^{\boldsymbol{h}}$ for any $g, g'$ and for all choices of $d_g$. To see why, note that mean independence implies that $\text{Cov}(\varepsilon_i(g), f(X_i)) = 0$ for any function $f$ of $X_i$ and for any $i \neq j$, our i.i.d. sampling assumption implies that $\text{Cov}(\varepsilon_i(g), f(X_j)) = 0$. But because the summands comprising $B_g$ and $C_g^{\boldsymbol{h}}$ are all functions of $X_i$, this shows that all summands comprising $A_g$ are uncorrelated with all summands comprising $B_g$ and $C_g^{\boldsymbol{h}}$.

Additionally, $B_g$ is uncorrelated with $C_{g'}^{\boldsymbol{h}}$, again for all $g, g'$ and for any fixed choice of $\mathbf{h}$. To see why, note that because $\mathbb{E}[\bar{d}_{g',all} - \bar{d}_{g',g'}] = 0$,

$$\begin{aligned}
\text{Cov}\left(B_g, C_{g'}^{\boldsymbol{h}}\right) &= \mathbb{E}[\bar{m}_{g,all}(\bar{d}_{g',all} - \bar{d}_{g',g'})] \\
&= \frac{1}{n^2}\left(\underbrace{\sum_{i=1}^{n}\mathbb{E}[m_g(X_i)d_{g'}(X_i)]}_{S_1} + \underbrace{\sum_{i\neq j}\mathbb{E}[m_g(X_j)d_{g'}(X_i)]}_{S_2}\right) \\
&\quad - \frac{1}{nn_g}\left(\underbrace{\sum_{i:W_{i,g}=1}\mathbb{E}[m_g(X_i)d_{g'}(X_i)}_{S_3} + \underbrace{\sum_{i:W_{i,g}=1}\sum_{j\neq i}\mathbb{E}[m_g(X_j)d_{g'}(X_i)]}_{S_4}\right).
\end{aligned}$$

As is clear, the summands in $S_1$ and $S_3$ are identical, as are the summands of $S_2$ and $S_4$. Moreover, $S_1$ has $n$ elements, $S_2$ has $n(n-1)$ elements, $S_3$ has $n_g$ elements, and $S_4$ has $n_g(n-1)$ elements. It is then straightforward to see that everything in the above expression cancels, so the covariance vanishes, as desired.

This reveals that the three terms in (5) are all uncorrelated with each other. Stacking Equation (5) for all of the $g$ into a single equation and using the fact that the variance of the sum of uncorrelated random vectors is the sum of the variances, we obtain the following variance decomposition

$$\text{Var}\left(\hat{\boldsymbol{\mu}}^{\mathbf{h}}\right) = \text{Var}\left(\mathbf{A}\right) + \text{Var}\left(\mathbf{B}\right) + \text{Var}\left(\mathbf{C}^{\mathbf{h}}\right). \tag{6}$$

Note that $\mathbf{h}$ does not affect $\mathbf{A}$ or $\mathbf{B}$ and therefore only affects the asymptotic variance of the regression adjusted estimator by affecting $\mathbf{C}^{\mathbf{h}}$. Moreover, it must contribute a positive semi-definite matrix, which is minimized if we can choose a value of $\mathbf{h}$ making $\text{Var}(\mathbf{C}^{\mathbf{h}}) = 0$. This is exactly what happens when we set $h_g = m_g$.[6] We thus arrive at the intuitive conclusion that

[6]Such a choice makes $\mathbf{C}^{\mathbf{h}}$ deterministically 0.

the conditional expectation function is the optimal function to use for regression adjustments. Summarizing, we have just shown the following result

**Proposition 1.** *Consider the class of regression adjustment estimators for* $\boldsymbol{\mu}$*, where* $\mu_g = \mathbb{E}[Y(g)]$*, of the form*

$$\hat{\boldsymbol{\mu}}^{\mathbf{h}} = \bar{\mathbf{Y}}_{\mathbf{g}} + \bar{\mathbf{h}}_{all} - \bar{\mathbf{h}}_{\mathbf{g}}, \tag{7}$$

*where* $\bar{\mathbf{Y}}_{\mathbf{g}}$ *is the vector of sample means of* $Y$ *partitioned by* $g$ *and* $\bar{\mathbf{h}}_{\mathbf{g}}$ *is the vector of sample means of* $h_g(X)$ *partitioned by* $g$*. Then, for any fixed* $\boldsymbol{h}$*,* $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$ *is a consistent estimator for* $\boldsymbol{\mu}$*. Moreover the (asymptotic) variance minimizing choice of* $\mathbf{h}$ *is given by* $h_g = m_g$*.*[7]

*Remark* 1. The argument leading to the variance decomposition (6) continues to hold if $Y$ is vector valued, so in that case, the variance minimizing choice of $\mathbf{h}$ would be the conditional expectation of each component of $Y$ for each group $g$.

*Remark* 2. Another way to think about our result is to consider what moment conditions are implied by the fact that we have an experiment. In particular, we note that randomization implies an infinite family of moment conditions of the form

$$\mathbb{E}\left[\frac{\mathbf{W}(Y - \mathbf{h}(X))}{\boldsymbol{\rho}} + \mathbf{h}(X) - \boldsymbol{\mu}\right] = 0 \tag{8}$$

for fixed function $\mathbf{h}$ making the above expectation well defined. In the above, both the multiplication and division of vectors in the first term are pointwise. This family of moment conditions is closed under linear transformations, so WLOG, we can build a Generalized Method of Moments (GMM) estimator by picking any single value of $\mathbf{h}$. Because we allow $\boldsymbol{h}$ to be any function in a linear space $L^2(X)$, it actually suffices to pick just a single optimal choice, which we just showed is $\boldsymbol{m}$. Viewed in this light, the results of Proposition 1 show that the conditional expectation functions of the outcome conditional on $X$ in each treatment group $g$ are the efficient choice of $\mathbf{h}$ within the GMM efficiency framework of Newey and McFadden (1994). Specializing to an average treatment effect, $\mu_g - \mu_{g'}$, (6) yields the efficiency bounds of Hahn (1998).

We can modify the argument leading to Proposition 1 to show that NW's optimal LRA is in fact optimal with respect to the family of linear in $X$ regression adjustments. The idea is to derive a variance decomposition analogous to Equation (6), but with minor modifications to account for the linearity in $X$ restriction. First, we define the *best linear predictor* of $Y$

[7]Where due to the multi-dimensional nature of the object being estimated, "variance minimizing" is with respect to the matrix partial order defined by positive semi-definiteness.

given $X$ in treatment group $g$ as

$$\text{BLP}_g(Y|x) = \alpha^* + (\beta^*)'x, \quad \alpha^*, \beta^* = \underset{\alpha,\beta}{\text{argmin}}\, \mathbb{E}\left[(Y(g) - \alpha - \beta' X)^2\right].$$

The error term $\delta_i(g) \equiv Y_i(g) - \text{BLP}_g(Y|X_i)$ is uncorrelated with $X$. Denoting $\ell_g(x) \equiv \text{BLP}_g(Y|x)$, we have a decomposition similar to (5)

$$\hat{\mu}_g^{\beta_g} = \bar{\varepsilon}_g + \bar{\ell}_{g,g} + \bar{h}_{g,all} - \bar{h}_{g,g} \equiv \underbrace{\bar{\delta}_g}_{A_g} + \underbrace{\bar{\ell}_{g,all}}_{B_g} + \underbrace{(\bar{d}_{g,all} - \bar{d}_{g,g})}_{C_g^{\boldsymbol{\beta}}}, \tag{9}$$

where now, we define $d_g(x) = \beta_g' x - \ell_g(x)$. It is still the case that $A_g$ is uncorrelated with $B_{g'}$ and $C_{g'}^{\boldsymbol{\beta}}$ for any choice of $\boldsymbol{\beta}$ and any choice of $g, g'$.[8] Similarly, the argument for why $B_g$ is uncorrelated with $C_g^{\boldsymbol{\beta}}$ remains unchanged as well. We thus arrive at the following analogue of Equation (6)

$$\text{Var}\left(\hat{\boldsymbol{\mu}}^{\boldsymbol{\beta}}\right) = \text{Var}\left(\mathbf{A}\right) + \text{Var}\left(\mathbf{B}\right) + \text{Var}\left(\mathbf{C}^{\boldsymbol{\beta}}\right). \tag{10}$$

Again, the choice of regression adjustment only affects the third term, so the variance of the above expression can be minimized if the third term can be set to 0, which is exactly what happens when $\beta_g$ is set as the OLS slope within group $g$. This analysis yields a next proposition

**Proposition 2.** *Consider the class of regression adjustment estimators for $\boldsymbol{\mu}$, where $\mu_g = \mathbb{E}[Y(g)]$, of the form*

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{\beta}} = \bar{\mathbf{Y}}_{\mathbf{g}} + \bar{\boldsymbol{\ell}}_{all}^{\boldsymbol{\beta}} - \bar{\boldsymbol{\ell}}_{g}^{\boldsymbol{\beta}}, \tag{11}$$

*where $\bar{\mathbf{Y}}_{\mathbf{g}}$ is the vector of sample means of $Y$ partitioned by $g$, $\bar{\boldsymbol{\ell}}_{all}^{\boldsymbol{\beta}}$ is the vector whose $g^{th}$ component is $\frac{1}{n}\sum_{i=1}^{n} \beta_g' X_i$, and $\bar{\boldsymbol{\ell}}_{\mathbf{g}}^{\boldsymbol{\beta}}$ is the vector whose $g^{th}$ component is $\frac{1}{n_g}\sum_{i:W_{i,g}=1} \beta_g' X_i$. Then for any fixed $\boldsymbol{\beta}$, $\hat{\boldsymbol{\mu}}^{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\mu}$. Moreover the variance minimizing choice of $\boldsymbol{\beta}$ is given by the group $g$ specific OLS slope $\beta_g = \text{Var}(X)^{-1}\text{Cov}(Y(g), X)$.*

*Remark* 3. This proposition generalizes NW's efficiency results. Specifically, their subsample means estimator is simply linear regression adjustment taking $\boldsymbol{\beta} = \mathbf{0}$ and their pooled regression adjustment is asymptotically also equivalent to a linear regression adjustment.

*Remark* 4. The above proposition also expands the scope of NW's conclusions beyond simply comparing estimators that are traditionally understood to be regression adjustments. Consider, for example, a field experiment where the researcher has access to pre-treatment

[8] Note a subtle difference in the justification for this fact. In this case, $A_g$ is uncorrelated only with linear functions of $X$, but because we are restricting ourselves to the class of linear in $X$ regression adjustments, the summands of $B_g$ and $C_g^{\boldsymbol{\beta}}$ are all restricted to be linear as well

analogues of the outcome of interest. This setting appears to be quite common in applied work (Brandon et al. (2021); Fowlie et al. (2020); Gosnell et al. (2020); Kaplan et al. (2013); Todd and Wolpin (2006)). All of these papers use a two-way fixed effect (TWFE) panel estimator at least once in estimating treatment effects. Why is the use of such an estimator so common in these experimental settings? If randomization is done properly, simply comparing the raw means post-treatment would suffice to unbiasedly estimate the treatment effect.

One important explanation comes from Burlig et al. (2019), who frame the question in terms of the potential of using the panel structure of the data to reduce the variance of treatment effect estimates and hence to improve power. Our efficiency results imply that focusing on asymptotic variances, this strategy is always (weakly) dominated by a regression adjustment. To see why, consider a field experiment setting where $T$ periods of pre-exposure data are recorded, indexed by $t = 0, \ldots, T-1$, and in period $T$, some proportion $p$ of individuals are assigned to a treatment group with the remainder being assigned to control. The TWFE estimator of the average treatment effect would then be the $\beta$ from the OLS regression corresponding to the model

$$Y_{it} = \beta D_{it} + \alpha_i + \gamma_t + \varepsilon_{it}, \quad t = 0, \ldots, T,$$

where $D_{it}$ is an indicator for whether unit $i$ is treated in period $t$ (so $D_{it} = 1$ if and only if $i$ is treated and $t = T$), and $\alpha_i, \gamma_t$ are respectively individual and time fixed effects. A special case of Theorem 1 of Goodman-Bacon (2021) implies that $\hat{\beta} = \bar{Y}_{T,1} - \bar{Y}_{T,0} - (\bar{Y}_{-T,1} - \bar{Y}_{-T,0})$ where $\bar{Y}_{T,1}, \bar{Y}_{T,0}, \bar{Y}_{-T,1}, \bar{Y}_{-T,0}$ are respectively the average outcomes in treatment post exposure, control post exposure, treatment pre exposure, and control pre exposure.

Yet, if we consider our panel dataset as a cross-sectional dataset and view the pre-exposure values of $Y$ as the covariates instead of outcomes from separate observations, the above expression for $\hat{\beta}$ is of the form $\hat{\mu}_1^{\mathbf{1}/T} - \hat{\mu}_0^{\mathbf{1}/T}$ (here, we are using the notation of Equation (9) and $\mathbf{1}$ is a vector of 1s). This shows that the TWFE estimator belongs to the class of linear regression adjustments and in particular uses a vector of values $1/T$ as the slope instead of an OLS fit. But, then Proposition 2 implies that the TWFE estimator is always (weakly) dominated by instead treating the pre-experiment outcomes as covariates and running a regression adjustment.

## 2.3 Sample Splitting and Inference

In this subsection, we discuss how to estimate the optimal regression adjustment derived above. In particular, we show that it is possible to tractably conduct inference while using

a flexible non-parametric model for fitting the conditional expectation functions $m_g(X)$. In particular, one could even make use of machine learning methods, which have demonstrated a remarkable ability to solve prediction problems in real world datasets. A potential pitfall of estimating $m_g(X)$ too flexibly, which we deal with, is that it is possible that this flexibility can introduce substantial finite-sample bias in the estimates for $\boldsymbol{\mu}$. To motivate how one should deal with this concern, we first consider why this is often thought to not be a concern in the linear case. Recall that LRA is given by

$$\hat{\mu}_g = \bar{Y}_g + \hat{\beta}_g'(\bar{X} - \bar{X}_g).$$

The first term in LRA is an unbiased estimate for $\mu_g$. When the conditional expectation function $\mathbb{E}[Y|X = x, G = g]$ is not linear in $X$, however, the second term is not necessarily unbiased. The intuition is that a given observation $i$ in group $g$ affects the second term both through its influence on the $\bar{X} - \bar{X}_g$ and through its influence on $\hat{\beta}_g$. This leads to a form of "overfitting" bias. In the classical linear regression adjustment, we typically think of $\hat{\beta}_g$ as being a fixed length parameter; thus, as the sample size increases, this overfitting becomes negligible, and hence can be ignored in asymptotic analysis.[9] Nonetheless, once this source of bias is identified, it is straightforward to handle. In particular, we can use a "jackknife" idea similar in spirit to the jackknife instrumental variables approach of Angrist et al. (1999). Specifically, define

$$\hat{\mu}_g^{JK} = \bar{Y}_g + \sum_{i=1}^{n} \hat{\beta}_g^{(-i)'} \left( \frac{1}{n} X_i - \frac{W_{i,g}}{\hat{\rho}_g} X_i \right), \tag{12}$$

where $\hat{\beta}_g^{(-i)}$ is the OLS estimate of the regression $Y_g = \alpha_g + \beta_g X_g + \varepsilon$ using all observations in group $g$ except for the $i^{th}$ observation. It is clear that as the sample size goes to infinity, we have $\hat{\mu}_g^{JK} \xrightarrow{p} \hat{\mu}_g$, so this modification has the same asymptotic properties as the usual estimator.

Despite the fact that it does not make a difference asymptotically, the jackknife approach is useful at removing much of the finite sample bias. In particular, $\hat{\beta}_g^{(-i)}$ is uncorrelated with $X_i$. Unfortunately, as pointed out in Wager et al. (2016), if $\hat{\beta}_g^{(-i)}$ is not an unbiased estimator for $\beta_g$ in finite samples, which will in general not be the case if $\mathbb{E}[Y|X = x, G = g]$ is not linear in $X$, then $\hat{\beta}_g^{(-i)}$ will be estimated using $n_g$ observations if $W_{i,g} = 0$ and $n_g - 1$ observations otherwise. In general, the bias will be different for these two sample sizes, which causes a small amount of correlation between $\hat{\beta}_g^{(-i)}$ and $W_{i,g}$. This introduces a small amount of bias to the estimator in Equation (12), but the bias will generally be of lower order than the overfitting bias.

[9] Recent literature relaxes this assumption; see, for instance, Belloni et al. (2013).

Returning to our initial suggestion to estimate $m_g(X)$ non-parametrically, we now show how the same de-biasing idea plays an important role in making regression adjustment work in finite samples. An issue with the above jackknife approach is that it may be prohibitively computationally expensive to perform as $n$ grows large, as it requires fitting one ML model *per observation*. Fortunately, recent advances in semi-parametric estimation as summarized, for instance, in Chernozhukov et al. (2018) (henceforth CCDDHNR) provides an approach to avoid this issue. Specifically, the results from CCDDHNR imply that an estimator with a much coarser sample splitting scheme, which the authors call "$k$-fold cross fitting," maintains the same favorable asymptotic properties as the jackknife estimator. We therefore use this form of sample splitting throughout our empirical examples.

To apply the insights in CCDDHNR to our setting, we note that the influence function corresponding to $\hat{\mu}_g^{\mathbf{m}}$ has the property that its derivatives to all orders with respect to $\mathbf{m}$ vanish and in particular, all second order Gateaux derivatives vanish. Then, checking Assumptions 1 and 2 and appealing to Theorems 3.1 and 3.2 of CCDDHNR provides our third Proposition,

**Proposition 3.** *Suppose that $\mathbb{E}[|Y|^q] < \infty$ for some $q > 2$. Assume we have some procedure for generating $\hat{\mathbf{h}}$ which is consistent in the sense that $||\hat{m}_g - m(\cdot; g)||_2 \xrightarrow{p} 0$. Suppose that we estimate $\hat{\mathbf{m}}$ using a cross fitting procedure as described in CCDDHNR and plug in the fitted values of $\hat{\mathbf{h}}$ into our estimates $\hat{\boldsymbol{\mu}}^{\mathbf{m}}$. Then we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\mu}}^{\hat{\mathbf{h}}} - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}(0, V), \quad V = \mathrm{Var}(\mathbf{A}) + \mathrm{Var}(\mathbf{B}).^{10} \tag{13}$$

*Moreover, $V$ can be consistently estimated using sample variances and covariances as described below.*

It is worth emphasizing that the conditions imposed in Proposition 3 are fairly weak. In particular, because of the higher order orthogonality of the class of regression adjustments to $\mathbf{m}$, we merely require that $\hat{\mathbf{m}}$ is consistent and do not require any conditions on how quickly the convergence occurs.[11] Many non-parametric estimators are known to be strongly consistent for fairly general classes of $\mathbf{m}$ (see, for instance, Györfi et al. (2002)). We, therefore, find that asymptotically, not much is lost by switching from using linear regression adjustment to a more flexible non-parametric regression adjustment with appropriate sample splitting.[12]

[10] Where here, $\mathbf{A}$ and $\mathbf{B}$ are as in the previous section.

[11] As we will see in our simulations, we still prefer $\hat{\mathbf{m}}$ to be high quality, as the ability of $\hat{\mathbf{m}}$ to fit the data affects the sampling variability of the resulting estimator.

[12] This point can be overstated. Nonparametric estimators typically suffer from slower rates of convergence than parametric estimators, so in a finite sample, one may still prefer linear regression adjustment. The results from our empirical examples suggest that, in general, one should pick the method that produces the highest quality out-of-sample predictions of the outcome as measured by mean squared error.

For completeness, we provide an explicit algorithm using Proposition 3 to estimate efficiently a regression adjustment and to obtain correct asymptotic standard errors. We use the sample splitting procedure suggested by CCDDHNR and randomly split the data into $K$ roughly equal size folds.[13] We then fit $\hat{\mathbf{m}}$ in the following way:

1. For each fold $k \in 1, \ldots, K$, and for each $g$, fit $\hat{m}_g^{(-k)}(X)$ using a non-parametric method for estimating a conditional expectation function on data not in fold $k$, but in treatment group $g$.

2. For each index $i$ in fold $k$, let $\hat{m}_{g,i} = \hat{m}_g^{(-k)}(X_i)$

Finally, we form the point estimate

$$\hat{\mu}_g^{FRA} = \frac{1}{n_g} \sum_{i:W_{i,g}=1} \underbrace{(Y_i - \hat{m}_{g,i})}_{a_i} + \frac{1}{n} \sum_{i=1}^{n} \underbrace{\hat{m}_{g,i}}_{b_{g,i}}, \tag{14}$$

where $n_g$ is the number of observations in group $g$. We have now written each individual treatment group mean estimator in terms of two sample averages: one in the treatment sample and one in the full sample. Additionally, from the efficiency proof, we have $\text{Cov}(a_{g,i}, b_{g',j}) = 0$ for both $i = j$ and $i \neq j$. Computing the full covariance matrix between all of the groups is now simply an accounting exercise, where only covariances between terms from the same group need to be accounted for in the computation. Separately doing this for diagonal ($V_{g,g}$) and off diagonal ($V_{g,g'}$) terms gives us:

$$\begin{aligned} \hat{V}_{g,g} &= \frac{1}{n_g} \widehat{\text{Var}}(a_g) + \frac{1}{n} \widehat{\text{Var}}(b_g) \\ \hat{V}_{g,g'} &= \frac{1}{n} \widehat{\text{Cov}}(b_g, b_{g'}), \end{aligned} \tag{15}$$

where $\widehat{\text{Var}}$, $\widehat{\text{Cov}}$ are respectively the sample variance and sample covariance. If multiple means are estimated per treatment group, then $\hat{V}_{g,g}$ and $\hat{V}_{g,g'}$ will be $k \times k$ matrices instead, but the basic form of (15) remains the same. For parameters that depend on a large number of subsample means, the formula given in (15) may be cumbersome to work with, so a bootstrap approach for standard errors is potentially easier computationally.

[13] Note that if the sample size is not sufficiently large, some care should be taken to ensure that each fold gets observations from each of the treatment groups $g$.

## 2.4 Power Calculations

We conclude our theoretical discussion by using our results to provide practical advice for power analyses when using a flexible regression adjustment, as described in this paper. We assume that the researcher has access to pre-exposure analogues of the $X$'s and $Y$'s. In particular, we focus our attention on Equation (5). Recall that for the optimal choice of $\mathbf{h} = \mathbf{m}$, the third term vanishes. Consider now, an estimator of the average treatment effect between two groups $g, g'$. The variance of the average treatment effect constructed using FRA is given by

$$\frac{1}{n_g}\text{Var}(a_g) + \frac{1}{n_{g'}}\text{Var}(a_{g'}) + \frac{1}{n}\text{Var}(m_g - m_{g'}). \tag{16}$$

Since power calculations are typically conducted to ensure the ability to detect small treatment effects, we make the assumption that treatment effects are "negligible" in the sense that $m_g - m_{g'} \approx 0$ and $\text{Var}(a_g) \approx \text{Var}(a_{g'}) = \text{Var}(a)$. We can therefore take the third term in Equation (16) to be approximately 0. Under these conditions, Equation (16) simplifies to approximately $\left(\frac{1}{n_g} + \frac{1}{n_{g'}}\right)\text{Var}(a)$.

Assume that we have a sample containing pre-exposure analogues of the $Y$'s and $X$'s available to perform power calculations. We now show that we can obtain reasonable estimates of $\text{Var}(a)$ using these data. In particular, we observe that by definition

$$\text{Var}(a) = \mathbb{E}\left[\text{Var}(Y|X)\right] = \mathbb{E}\left[(Y - m(X))^2\right], \quad m(x) = \mathbb{E}[Y|X = x].$$

We can then take $\text{Var}(a)$ to be the low dimensional parameter in a semi-parametric estimation problem with nuisance parameter $m$. We note, furthermore, that set up in this way, this problem satisfies Neyman's orthogonality.

$$\frac{\partial}{\partial r}\mathbb{E}\left[(Y - (m(X) + r(h(X) - m(X))))^2\right] = 0$$

Given this observation, we can estimate $\text{Var}(a)$ using a cross-fitting procedure. Examining the estimator, we can see that the cross-fitting procedure is equivalent to computing the mean squared error (MSE) of the non-parametric model evaluated on a test set. Orthogonality implies that even if $m$ is difficult to estimate, we expect that the test set MSE converges to the $\text{Var}(a)$ at about twice the rate as $m$ converges to its true value. Thus, for the purposes of performing a power calculation, we can plug in the test set MSE when predicting $Y$ using $X$ on the available data where we would ordinarily plug in the variance of $Y$ to obtain an approximately valid estimate of power. An analogue of this result in the linear case is that we can take $1 - R^2$ of an OLS regression of $Y$ on $X$ to estimate the factor by

which a regression adjustment can be expected to reduce variance under the assumption of homogeneous treatment effects.

# 3 Empirical Examples

We now take the theory for flexible regression adjustment derived in the previous section to both simulated and naturally-occurring data. We begin with a number of simulation exercises. First, we construct a synthetic dataset by adding a simulated treatment effect on top of the naturally-occurring data. We then use these simulations to show that the asymptotic theory in Proposition 3 is reliable, but only if proper sample-splitting procedures are followed. This exercise shows that our proposed estimator has good asymptotic properties, but it does not provide much guidance for when one might wish to use a linear regression adjustment over a more sophisticated ML-based approach. We thus turn to a number of theory-driven simulation exercises that showcase the salient features of the data that cause ML methods to substantially outperform linear methods.

After showcasing our methods using synthetic datasets, we show that our estimators perform well in naturally-occurring datasets as well. We first turn our attention back to Lyft data and analyze a natural field experiment (see Harrison and List (2004) we conducted at Lyft to show how using regression adjustment reduces variance in a non-negligible manner. In the Lyft setting, we find that although regression adjustments in general make a large difference, the additional flexibility from using an ML model does not substantially improve precision. To explore if ML-based models can yield statistical improvements, we analyze three additional field experimental datasets where we do find improvements from using a more flexible form of regression adjustment.[14] The four real world applications we consider have substantially varying sample sizes and demonstrate that the flexibility from ML-based approaches can be helpful for various populations of people and situations.

## 3.1 Simulation Study: Augmented Lyft Data

As a baseline, we first check that the asymptotic distribution implied by Proposition 3 approximates the actual sampling distribution of FRA estimators well. To do this, we take a dataset containing one row per registered Lyft passenger. Throughout the exercises in this subsection, we only report the distribution of estimated z-scores using the theory derived in the previous section. Because the z-scores are normalized, they reveal no information about

[14] R Code implementing our flexible regression adjustment along with the analyses of the three non-Lyft settings can be found at the following link: https://github.com/gsun593/FlexibleRA. We have also included a copy of the code in Appendix A

Lyft's underlying data. However, by comparing their distributions to a standard normal distribution, we can verify that the asymptotic theory we derived in the previous section works well in finite samples.

For each passenger in the dataset, we record as our outcome variable the count of rides that that passenger consumed in a fixed two month window. As covariates, we use a number of summary statistics of past behavior computed on the day before our two month window begins. As one would expect, past behavior tends to be predictive of future behavior, so using these summary statistics as covariates for regression adjustment is a reasonable approach to reduce variance. We split this dataset randomly into 10,000 smaller datasets and construct "placebo" experiments on these smaller datasets by randomly assigning each individual into "treatment" and "control" with 50/50 probability. For each experiment, we use the sample splitting procedure (with five folds) described in Section 2.3 to estimate the mean potential outcomes in treatment and control. We consider both a "null" case where there is uniformly no treatment effect and an alternative case where we synthetically induce a treatment effect.

To simulate treatment, for each observation $i$ assigned to "treatment," we add a random number of rides drawn from Poisson$(0.1 \cdot r_i)$, where $r_i$ is the number of rides actually attributed to observation $i$. By construction, the average treatment effect (ATE) from this data generating process is a 10% increase in the number of rides. We construct a point estimate of the ATE by taking the difference of the FRA-adjusted group means. We then use $\hat{V}$ as defined in Equation (15) to compute a standard error estimate. The asymptotics stated in Proposition 3 imply that subtracting the actual treatment effect (0 in the "null" case and $0.1 * mean(\text{rides})$ in the "alternative" case) from the point estimate of the treatment and dividing this difference by the estimated standard error gives us a random variable distributed approximately according to $\mathcal{N}(0, 1)$, so we call these values "z-scores".

In Panels A and B of Figure 1, we plot a histogram of the 10,000 z-scores as well as a qq-plot comparing the empirical quantiles of the z-scores to the normal theoretical quantiles for the dataset with a treatment effect.[15] In Panels A and B of Figure 2, we do the same for the null dataset. In Table 1, we report the mean and variance of the z-scores along with the coverage of 95% and 99% confidence intervals. We find that the z-scores fit the standard normal distribution remarkably well and the coverage of the resulting confidence intervals are statistically indistinguishable from their theoretical values.

We next show the importance of sample splitting. In Panels C and D of Figures 1 and 2, we replicate Panels A and B respectively, except we do not use sample splitting when

[15]Specifically, the $x$ axis in these qq-plots is defined by the theoretical quantiles of a standard normal distribution while the $y$ axis corresponds to the empirical quantiles. If the asymptotic theory is correct, the points in these plots should lie close to the 45 degree line, and deviations from this prediction allow us to more precisely visualize deviations from asymptotic normality.

Table 1: Summmary Statistics of Normalized Treatment Effect Estimates

| Simulation | Mean | Std. Dev. | 95% CI Coverage | 99% CI Coverage |
|---|---|---|---|---|
| TE, Moderate, Cross | -0.012 | 1.00 | 0.9521 | 0.9887 |
| TE, Moderate, No Cross | -1.33 | 1.06 | 0.8792 | 0.7256 |
| Null, Moderate, Cross | 0.005 | 1.00 | 0.9509 | 0.9904 |
| Null, Moderate, No Cross | -0.001 | 0.99 | 0.9507 | 0.9906 |
| TE, Large, Cross | 0.017 | 1.02 | 0.946 | 0.989 |
| TE, Large, No Cross | -1.16 | 1.07 | 0.764 | 0.894 |
| Null, Large, Cross | -0.04 | 0.96 | 0.959 | 0.990 |
| Null, Large, No Cross | -0.02 | 1.03 | 0.947 | 0.989 |
| TE, Small, Cross | -0.019 | 1.00 | 0.95191 | 0.99181 |
| TE, Small, No Cross | -1.06 | 1.10 | 0.91515 | 0.78837 |
| Null, Small, Cross | 0.004 | 1.00 | 0.95264 | 0.99242 |
| Null, Small, No Cross | -0.001 | 1.01 | 0.9495 | 0.9885 |
| Ratio, Moderate, Cross | -0.01 | 1.00 | 0.95264 | 0.99242 |
| Ratio, Moderate, No Cross | -0.005 | 1.20 | 0.8983 | 0.9675 |

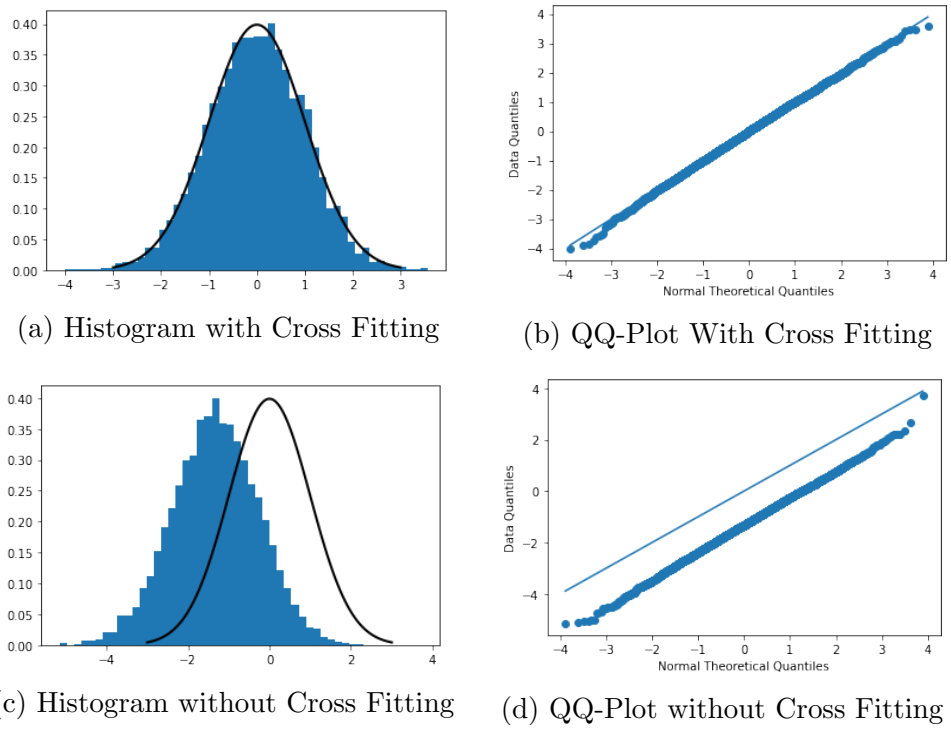
**Notes:** This table contains summary statistics for the normalized treatment value estimates. If the asymptotic theory is valid, these normalized estimates should be mean zero with a standard deviation of 1, and the 95% and 99% CIs should cover 95% and 99% of the values respectively. Each simulation is indexed by (TE, Sample Size, Cross) where TE and Cross denote respectively whether or not the dataset has a treatment effect and whether or not cross fitting was used. We refer to the simulations obtained by splitting the original dataset into 1,000, 10,000, and 100,000 pieces respectively as "Large", "Moderate", and "Small" samples. The last two rows contain results from doing inference for a ratio metric.

constructing our estimators. Summary statistics are again in Table 1. Interestingly, when the treatment has a null effect, we find that even without sample splitting, our standard errors provide a reliable approximation to the true sampling distribution of the estimator. However, the situation changes dramatically in the presence of a treatment effect.

For example, examining Figure 1, we see that without sample splitting, the estimates for the average treatment effect with regression adjustment are biased downwards considerably. Intuition about the role of regression adjustment provides a key reason why this bias arises. Recall that FRA estimates treatment group means by adding $\bar{h}_{g,all} - \bar{h}_{g,g}$ to the raw group mean. When $\bar{h}_g$ is estimated in sample, the value of $Y_i$ influences both $\bar{h}_{g,all}$ and $\bar{h}_{g,g}$, but its weight in $\bar{h}_{g,g}$ will be larger. The influence of $Y_i$ on $\bar{h}_g$ tends to push in the opposite direction as its influence in computing $\bar{Y}_g$. This will tend to bias the estimated group means $\hat{\mu}_g$ towards homogeneity, which in turn biases the average treatment effect estimates towards zero. The negative bias we observe in our simulations thus reflects the fact that we constructed a positive treatment effect, so that a bias towards 0 is a negative bias.

Thus far, we have found that not using sample splitting leads to biased estimators, where the bias tends to make group means more similar. One might wonder to what extent a larger

Figure 1: Distribution of Normalized Estimates (TE, Moderate)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

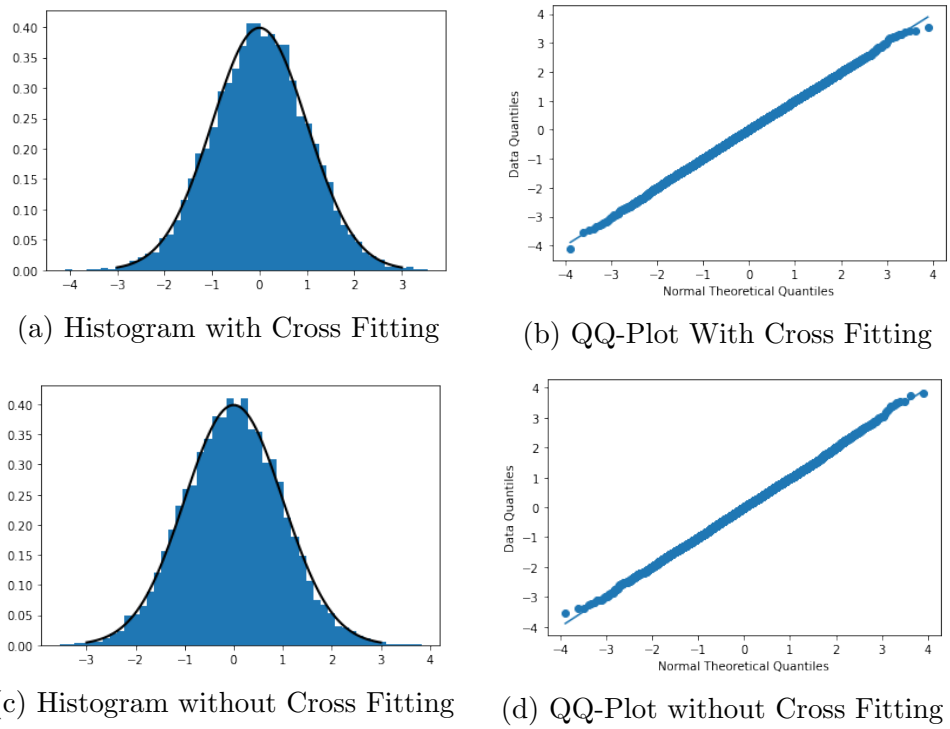
sample mitigates this issue. To investigate this question, we repeat the exercises above, but split the data evenly into 1,000 evenly sized datasets, thus increasing the sample size by a factor of 10. We replicate the results of Figures 1 2 in Figures 3 and 4. Summary statistics can again be found in Table 1.

These results make it evident that a large bias continues to exist, even on these larger datasets. Nonetheless, increasing the sample size does appear to attenuate the bias somewhat: while in the moderate sized sample the bias is -1.33 standard errors, in the large sized sample the bias decreases to -1.16.[16] This result suggests that asymptotic unbiasedness might hold, even for the non sample-split estimator, but the asymptote may not be a good approximation, even for large sample sizes. Since cross-fitting does not negatively affect asymptotic efficiency and is usually computationally simple,[17] the evidence presented

[16]This reduction is not just due to noise: the difference would be statistically significant if subjected to formal hypothesis testing.

[17]If the non-parametric method being used has algorithmic complexity growing faster than linearly in dataset size (which is common), two-fold cross-fitting would be even faster than not using a split sample for sufficiently large datasets.

Figure 2: Distribution of Normalized Estimates (Null, Moderate)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

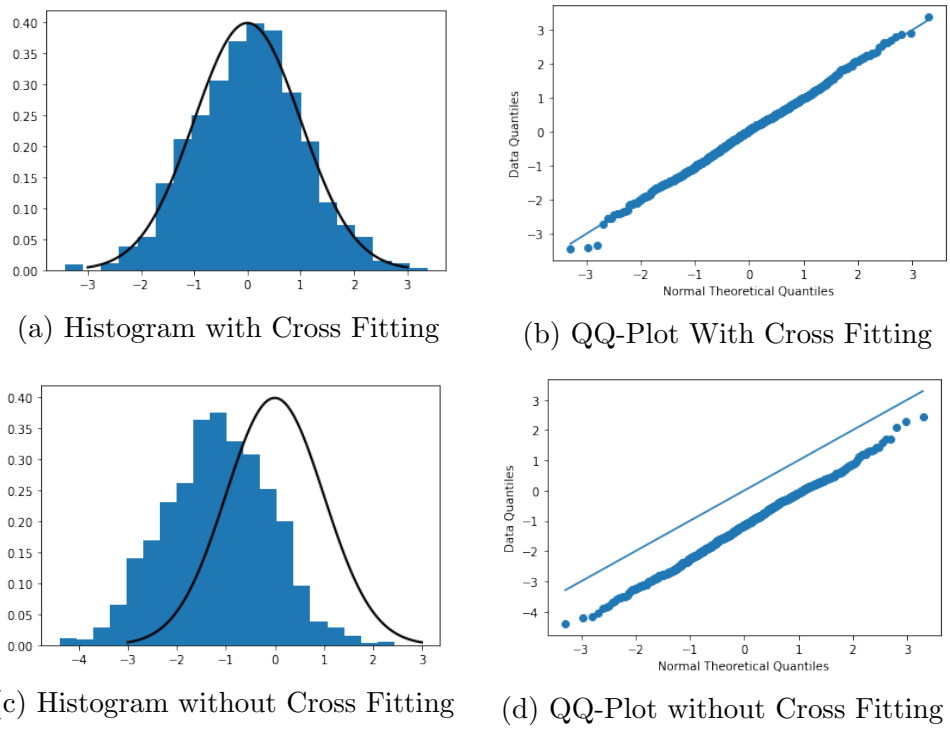
(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. The dataset being used here by construction has no treatment effects. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

here supports the conclusion that a practitioner using a flexible regression adjustment should leverage sample splitting for reliable inference.

Proceeding in the opposite direction in terms of sample sizes, one might wonder if flexible regression adjustment can be fruitfully applied on relatively small datasets, which may be the relevant case for many experimental researchers. To address this question, we split our dataset into roughly 100,000 evenly sized smaller datasets, which therefore only have hundreds of observations. We again construct a set of "z-scores" and compare them to the standard normal distribution. We plot the results of this exercise in Figure 5 for the non-null case and 6 for the null case and report summary statistics in Table 1. In this small data regime, a number of additional interesting features are present. First, even the estimator with sample splitting displays a slight amount of bias, although the size of this bias is only about 2% of a standard error. As a result, 95% and 99% confidence intervals still have excellent coverage properties. Second, examining the null case, both estimators, but especially the non-sample split estimator, display slight deviations from normality, even

Figure 3: Distribution of Normalized Estimates (TE, Large)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. Compared to Figure 1, we use sample sizes that are 10 times as large. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

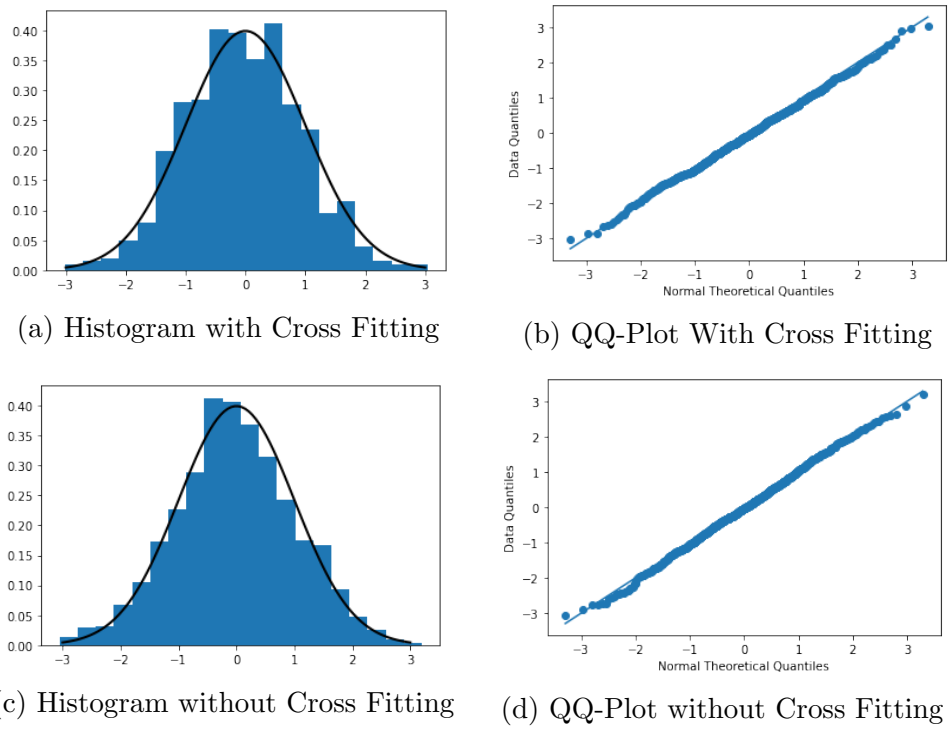
while the first and second moments appear to closely match their theoretical values. In particular, both estimators display slightly thinned tails (as evidenced by the *over*-coverage of the 95% and 99% confidence intervals), and slightly less mass around values very close to 0 relative to a Gaussian distribution. The thinned tails are likely due to a similar mechanism to the one driving bias in the non-null case: the overfitting from not sample splitting creates a tendency towards mean-reversion, which may be especially effective at correcting extreme cases of imbalance arising due to sampling variability.

Finally, we consider an example where we estimate a quantity that is not an average treatment effect. Consider, for instance, metrics of the form

$$\mathbb{E}[Y_1|W_{i,g} = 1]/\mathbb{E}[Y_2|W_{i,g} = 1].$$

Two examples of this metric type in a rideshare context are "conversion" (i.e. the probability of taking a ride conditional on opening the app and receiving a price quote and time estimate)

Figure 4: Distribution of Normalized Estimates (Null, Large)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

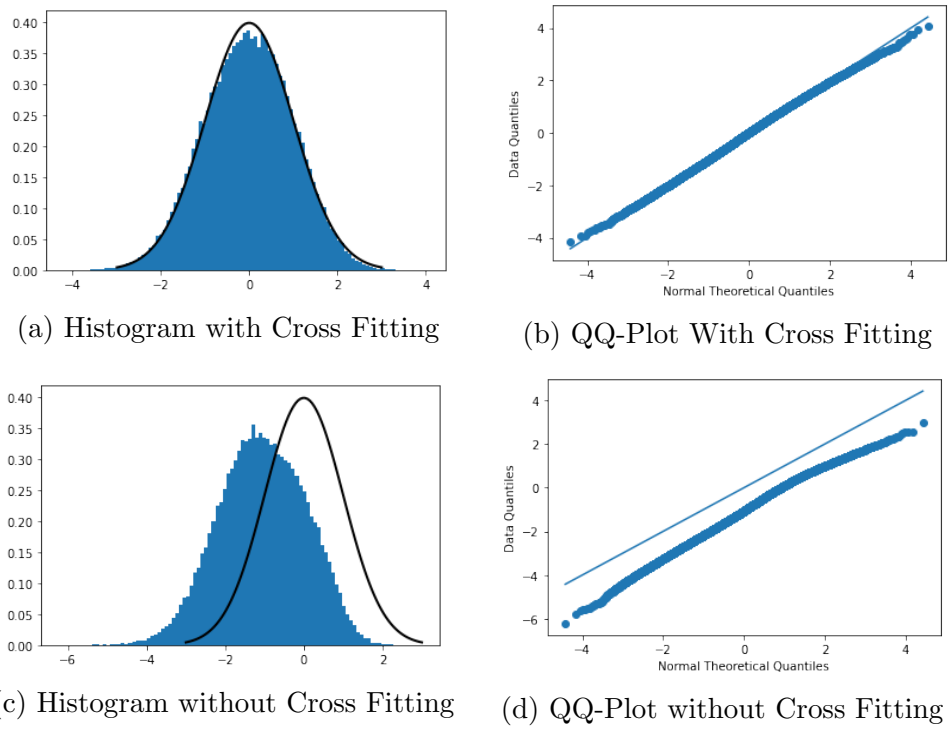
(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. The dataset being used here by construction has no treatment effects. Compared to Figure 1, we use sample sizes that are 10 times as large. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

and intensive margin labor supply outcomes (i.e. the number of hours a Lyft driver works conditional on working within a given time period). Firms are often interested in learning how this ratio varies in response to different interventions: $\mathbb{E}[Y_1|W_{i,g} = 1]/\mathbb{E}[Y_2|W_{i,g} = 1] - \mathbb{E}[Y_1|W_{i,g'} = 1]/\mathbb{E}[Y_2|W_{i,g'} = 1]$. We therefore replicate the exercise in Figure 1 in the context of a ratio metric. Again, we only report z-scores here, which allows us to test the validity of the asymptotic theory while normalizing so as to obscure any identifiable information about Lyft's underlying (confidential) data.

In what follows, for each individual, in addition to examining the number of rides an individual consumes, we also include data on the number of times passengers checked the app, which we call "sessions". The probability of taking a ride conditional on opening the app is therefore given by the average number of rides divided by the average number of sessions, which takes the form of a ratio metric. We add a treatment effect to this dataset by first adding sessions to each treated individual according to Poisson($s_i$), where $s_i$ is 0.05 times the number of sessions taken by individual $i$. For each added session, we add a ride for

Figure 5: Distribution of Normalized Estimates (TE, Small)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

(d) QQ-Plot without Cross Fitting

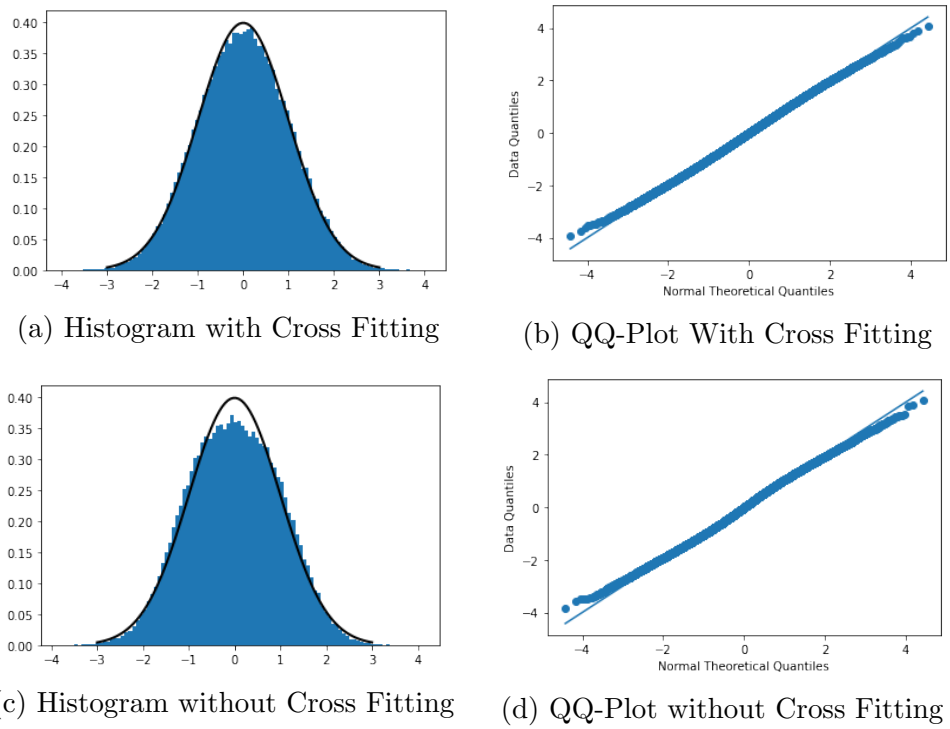
**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. Compared to Figure 1, we use sample sizes that are 10 times as small. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

that session according to Bernoulli($p_i$), where $p_i$ is the proportion of sessions for individual $i$ that resulted in a ride. Finally, for all sessions still not associated with a ride, we add additional rides according to Bernoulli(0.02). The change in conversion in the simulation is therefore $0.02(1-\bar{p})$, where $\bar{p}$ is the population level of conversion in control. We construct point estimates for this quantity by plugging in the regression adjusted means in place of the population means in the formula defining conversion and compute standard errors using Equation (15) and the delta method.

Empirical results are found in Figure 7, and summary statistics can be found in Table 1. For this particular simulation, we are unable to detect bias in our estimator, but the standard errors are misleadingly small when we do not use sample splitting. As before, the results suggest that sample splitting allows us to do valid inference.

Before closing this subsection, we report the average estimated standard errors from the various simulations performed in Table 2 as a proportion of the standard errors from taking the raw difference in means. There are a number of interesting patterns in this table. In the

Figure 6: Distribution of Normalized Estimates (Null, Small)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

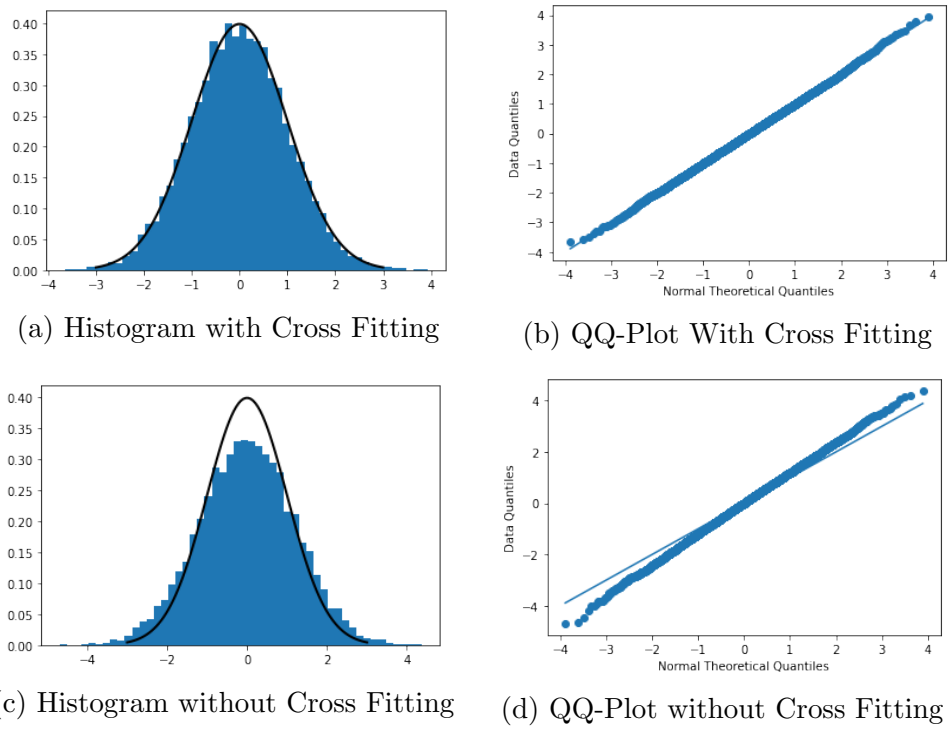
(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the treatment effect, which subtract off the population average treatment effect and divide by the sample standard deviation. The dataset being used here by construction has no treatment effects. Compared to Figure 2, we use sample sizes that are 10 times as small. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

non sample-split estimators, we find that the within-sample standard errors are considerably smaller than the sample split estimators and become *larger* as a fraction of the difference in means standard error as sample size increases. This fact on its own is unsurprising: for small samples, there is more overfitting while in larger samples, there is less. What is surprising is that the sampling distribution of the "z-scores" suggest that these smaller standard errors end up still being reasonable estimates of the true sampling variability of the estimator. As seen in the non-null data, however, when a treatment effect is present, this reduced sampling variability comes at the cost of large amounts of bias against finding a treatment effect.

A second notable fact is that flexible regression adjustment absorbs a larger proportion of the variability in the data with larger samples. This is a reflection of the fact that non-parametric estimators typically need at least a moderate amount of data to deliver valid results. Interestingly, while we find that in small samples, the non-parametric estimator can deviate considerably from the asymptotic efficiency bound, the plotted distributions of normalized estimates suggest that inference based on such an estimator remains valid. In small

Figure 7: Distribution of Normalized Estimates (Ratio)




(a) Histogram with Cross Fitting

(b) QQ-Plot With Cross Fitting

(c) Histogram without Cross Fitting

(d) QQ-Plot without Cross Fitting

**Notes:** This figure plots the distribution of normalized estimates of the difference in ratios, which subtract off the population average difference and divide by the sample standard deviation. Panels (a) and (b) show the distribution from a procedure where sample splitting is used while Panels (c) and (d) show the distribution from a procedure where sample splitting is not used.

data settings, we may therefore prefer to use linear regression adjustment, provided we are confident that a linear functional form is a decent approximation to the conditional expectation function. As discussed in Section 2.4, when data are available ex-ante for doing power calculations, comparing the $R^2$ from a linear regression to the $R^2$ from a non-parametric model is a practical way to decide if there is enough data to use a non-parametric method. Moreover, this can be completed prior to examining the experimental results themselves, thus mitigating concerns about specification search or p-hacking.

## 3.2 Simulation Example: When Does ML Outperform OLS?

While the theoretical results in this paper show that FRA asymptotically can do no worse (and often does better) than LRA, in practice, one may still not always wish to use ML techniques when performing regression adjustment. For example, one reason to prefer LRA is that ML methods can be unstable in small samples and thus deliver less variance reduction than OLS in practice. Another reason to prefer LRA is that ML methods are computationally

Table 2: Shrinkage Factor for Confidence Intervals

| Simulation | Ratio |
|---|---|
| Small, Cross | 0.85 |
| Small, No Cross | 0.24 |
| Moderate, Cross | 0.75 |
| Moderate, No Cross | 0.27 |
| Large, Cross | 0.67 |
| Large, No Cross | 0.38 |
| Ratio, Cross | 0.96 |
| Ratio, No Cross | 0.80 |

**Notes:** This table shows the factor by which the estimated confidence intervals shrink depending on whether or not sample splitting is used, what data is being used, and whether an average treatment effect or a difference in ratios is being estimated.

expensive to use. Fortunately, by examining the proof of the efficiency of FRA, we can gain an intuition for when it is likely to deliver substantial statistical gains relative to LRA.

Specifically, the main use of ML techniques in the FRA procedure is in estimating the conditional expectation function (CEF) of the outcome given the covariates. The LRA, on the other hand, takes the same form as the FRA *except* that the CEF is approximated to be linear in covariates. Thus, FRA outperforms LRA when the linear approximation is of low quality. At a high level, we expect a linear model to poorly approximate the CEF either when the CEF is highly non-linear with respect to each covariate individually or when the CEF contains important interaction effects.

In this section, we conduct a simple family of simulation studies to demonstrate the importance of these two features. Specifically, we simulate treatments $W \sim \text{Bernoulli}(0.5)$ as well as three latent variables $L_1, L_2, L_3$ with $L_1, L_2, L_3 \sim \text{Unif}(0, 1)$ which are unobserved to the econometrician. We also simulate an unobserved "error term" $U \sim \mathcal{N}(0, 1)$. Our outcome $Y$ is then given by

$$Y = W + L_1 + L_2 + L_3 + U$$

In three of our simulations, we derive our covariates $X_1, X_2, X_3$ from $L_1, L_2, L_3$ according to $X_i = L_i^p$ for $p = 1, 5, 10$. In addition, we consider three additional specifications where $X_1 = (U_1U_2)^p$, $X_2 = (U_2U_3)^p$ and $X_3 = (U_3U_1)^p$ for $p = 1, 5, 10$ to generate non-trivial interaction terms. For each simulation, we draw $B = 100$ samples of size $N = 1,000$.

We specified our simulations so that the true signal/noise ratio is identical across specifications, but as $p$ increases further from 1, the true CEF becomes increasingly non-linear. In Table 3, we report the ratio of the standard errors from FRA compared to LRA in the

6 specifications (3 values of $p$ times 2 specifications varying whether or not there are interactions amongst covariates). As expected, as $p$ increases away from 1 or as we add an interaction amongst covariates, the gains from FRA over LRA increase. Note however, that when there is no interaction and $p=1$, FRA performs slightly worse than LRA. Asymptotically, because the CEF is linear, we expect FRA and LRA to attain the same sized standard errors. However, in finite sample sizes, we expect FRA in this case to perform worse due to the fact that machine learning techniques tend to require more data to achieve a good fit. This shows that when the true CEF is known to be linear (or approximately linear), LRA may still be preferable over FRA in practice, even ignoring computational considerations.

Table 3: Variance Reduction for Average Treatment Effects

| | $p=1$ | $p=5$ | $p=10$ |
|---|---|---|---|
| Interaction | 1.03 | 0.94 | 0.72 |
| No Interaction | 0.93 | 0.81 | 0.68 |

**Notes:** This table displays the ratio between the standard deviation of the FRA estimator relative to the standard deviation of the LRA estimator. The columns vary non-linearity, as parameterized by $p$ while the rows vary whether or not there are interactions between covariates.

## 3.3 Application I: Natural Field Experiment on Lyft Cancellations

Having shown that cross-fitting allows us to robustly quantify uncertainty in a simulation setting, we apply the FRA to a natural field experiment. The premise behind our field experiment is that, ceteris paribus, Lyft prefers to minimize the number of cases wherein a driver agrees to pick up a passenger, but then cancels the ride before pickup occurs. Formally, Lyft would like to design a policy to minimize the cancellation rate, $\frac{\frac{1}{n}\sum_{i=1}^{n} \#\text{ rides canceled}}{\frac{1}{n}\sum_{i=1}^{n} \#\text{ rides accepted}}$. When considering a new intervention, Lyft would therefore like to track how this metric varies across different treatment conditions.

Here, we focus on a field experiment we helped to conduct at Lyft in 2018. To ensure that drivers do not waste time waiting for passengers who never show up, Lyft allows drivers the option to mark a passenger as a "no show" if sufficient time elapses after the driver arrives at the pickup location (this is considered a special type of cancellation). Passengers who are marked as "no-show" are charged a fee which is passed on to the driver to compensate for lost passenger time. When Lyft introduced its Shared rides (a product whereby a passenger receives a discount in return for allowing Lyft to match them with another passenger taking a similar route at the same time), it had to rethink its original no-show policy. Specifically, because of the fact that multiple passengers potentially shared the same driver, a passenger not promptly arriving to their pickup location would impose a negative externality on

passengers already in the car. As a result, Lyft decided that the window of time passengers received before the driver was allowed to mark them as a "no show" for Shared rides should be less generous. This led to the rate of no-show cancels to be higher on Shared rides relative to Standard rides.

Before our field experiment, the status quo policy was that *all* Shared rides had a shorter no-show window. However, such a uniform policy was irrational if a passenger requesting a Shared ride was matched to a driver without other passengers already in the car, since the negative externality is not present. This reasoning represents the genesis of our experiment. In our field experiment, drivers were assigned to treatment or control. Control drivers received the status quo policy whereas if a passenger was matched to a treated driver and was the first passenger in the car, they would receive the more generous no-show window that Standard passengers received.

Our two key outcome metrics are the cancellation rate, as defined above, and the no show rate, defined as $\frac{\frac{1}{n}\sum_{i=1}^{n} \#\text{ rides canceled because no show}}{\frac{1}{n}\sum_{i=1}^{n} \#\text{ rides accepted}}$. We fit three models. First, we consider simply plugging the subsample means (SM) within each variant into the formulae defining the cancellation and no show rates. Second, we consider plugging in the linear regression adjusted estimates (LRA).[18] Finally, we apply the fully flexible regression adjustment (FRA). Standard errors are constructed using the delta method. For each metric, we report the point estimate of the effect as a percent of the baseline in control and the standard errors.

A summary of our empirical results are reported in Table 4. A number of notable facts stand out. First, the experiment was a success when considering the no-show rate, which decreased by roughly 4.5% across estimators; the null hypothesis of no effect is easily rejected. Second, despite this, when examining the treatment effect on the overall cancellation rate, we have difficulty detecting a significant effect with the non-regression adjusted estimate. Yet, reducing variance using a regression adjustment (either LRA or FRA) makes the effect considerably easier to detect. Third, while it is true that, as the theory predicts, a fully flexible regression adjustment delivers slight efficiency gains over a linear adjustment, these gains appear to be modest in practice.

In Table 5, we investigate this further by reporting the $R^2$ (defined as one minus the out-of-sample mean squared prediction error divided by total variance) of the linear model compared to the non-parametric model in explaining variation in the number of accepts, number of cancels, and number of no shows. Indeed, for most of the outcomes of interest studied here, the non-parametric models appear to explain only slightly more of the variation than a linear model, and even provides a slightly worse fit for predicting the number of

[18] Specifically, we implemented our point estimates according to 14 and our standard errors according to 15, but using an OLS fit for $\hat{m}_{g,i}$ in place of a fitted machine learning model.

cancels. The limited additional gains from using the flexible regression adjustment in our setting likely implies that the conditional expectation functions for the outcomes we are examining are reasonably well approximated by linear functions.

Table 4: Variance Reduction for Average Treatment Effects

| | SM | LRA | FRA |
|---|---|---|---|
| Cancel Rate | -0.60% | -1.26% | -1.32% |
| | (0.52) | (0.35) | (0.34) |
| No Show Rate | -4.4% | -4.2% | -4.3% |
| | (0.51) | (0.40) | (0.40) |

**Notes:** This table shows point estimates and standard errors in parentheses for the percent difference in cancel rate and no show rate between treatment and control. The first column looks at the estimator obtained by plugging in the subsample means. The second column considers a linear regression adjustment. The third column uses a nonparamteric regression adjustment. We are unable to report sample sizes for this analysis.

Table 5: $R^2$ in predicting outcomes

| | # Accepts | # Cancels | # No Shows |
|---|---|---|---|
| Linear $R^2$ | 0.687 | 0.587 | 0.514 |
| Flexible $R^2$ | 0.712 | 0.580 | 0.525 |

**Notes:** This table shows the $R^2$ in predicting various metrics necessary to compute cancellation and no show rates. The first row considers $R^2$ from fitting an OLS model while the second row considers the $R^2$ from fitting a flexible non-parametric model.

We conclude this section by also comparing regression adjustments to the TWFE model. For each individual, in addition to the number of acceptances, cancels, and no shows during the experimental period, we also obtain data about the number of acceptances, cancels, and no shows in the period before the experiment started. In Table 6, we report the degree to which differences-in-differences reduces variance compared to a regression adjustment estimator that simply takes these pre-experimental outcomes as covariates.

Depending on the outcome we are examining, we find that regression adjustment results in a 1-10% reduction in the size of the standard errors. Moreover, the square of the ratios reported in Table 6 represent how much smaller the sample size needs to be holding statistical power fixed if one uses regression adjustment instead of differences-in-differences. Across our three outcomes, we find that regression adjustment allows an experimenter to garner the same power for a sample with only 84% to 97% of the number of observations, suggesting that the current practice among experimental economists of estimating TWFE models may be causing researchers to "overpay" for their experiments by a non-trivial amount, leading to a greater number of Type 2 errors. This result suggests that experimental economists commitment

to focusing solely on sample size or sample allocation across cells when considering power is unduly restrictive, and indeed quite inefficient.

Table 6: Reduction in Standard Errors

| | # Accepts | # Cancels | # No Shows |
|---|---|---|---|
| Diff-in-diff | 0.622 | 0.667 | 0.786 |
| Regression Adjustment | 0.596 | 0.658 | 0.721 |

**Notes:** This table shows the reduction in standard errors from using different variance reduction techniques. The columns index the outcome measure while the rows index the estimator considered. For a given outcome $Y$ and for a given estimator $\hat{\beta}$, each entry in the table displays the ratio between the standard errors from fitting $\hat{\beta}$ on outcome $Y$ relative to the standard errors from taking a simple difference in means.

## 3.4 Application II: Oregon Health Insurance Experiment

We next turn our attention to an analysis of the data from the Oregon Health Insurance Experiment (OHIE). We focus in particular on replicating the results of Finkelstein et al. (2016), which measures the impact of Medicaid on emergency room visits. We take our covariates to be gender, age, prior health, and education along with a detailed vector of counts for various types of ER visits prior to randomization. Our outcome of interest is whether or not an individual visited an emergency department during the experiment. We additionally estimate the impact of treatment status on medicaid take-up, and by dividing the reduced form effect of treatment on outcome by the effect of treatment on take-up, we can estimate the LATE of Medicaid take-up on ER visits.

Empirical results of this exercise are summarized in Table 7, which has a similar form as Table 4 and compares the subsample means estimator which does not use any covariates to the linear and flexible adjustments. Across specifications, we find that the flexible regression improves standard errors by about 2-3% relative to the next best alternative. While these gains are modest fixing sample size, they imply that for a similar level of statistical power, researchers could reduce sample sizes by about 5-6%, thus reducing variable experimental costs by a similar quantity. We suspect that with a richer covariate set the gains would have been even greater.

## 3.5 Application III: Water Conservation Nudges

We next re-analyze data from a natural field experiment conducted by Ferraro and Price (2013), which studies the effect of a number of nudges on water conservation. The intervention was designed to reduce water consumption during the summer months of 2007. In

Table 7: Variance Reduction for OHIE

| | SM | LRA | FRA |
|---|---|---|---|
| ER Visits | 0.0132 | 0.0143 | 0.0139 |
| | (0.0085) | (0.0079) | (0.0077) |
| Medicaid Take-Up | 0.172 | 0.159 | 0.150 |
| | (0.0063) | (0.0062) | (0.0062) |
| LATE | 0.0892 | 0.0902 | 0.0870 |
| | (0.0496) | (0.0498) | (0.0482) |

**Notes:** This table shows point estimates and standard errors in parentheses for a number of causal parameters from the OHIE across a number of regression adjustment specifications. In the first row, we measure the reduced form impact of treatment assignment on ER visits. In the second row, we measure the first stage impact of treatment assignment on Medicaid take-up. In the third row, we divide the first row by the second row to obtain an estimate of the LATE of Medicaid uptake. The sample size is $N = 13,051$.

addition to collecting data on summer consumption, the authors also collect month-by-month water consumption for each individual in the experiment in the year prior to experimentation.

For simplicity, we only consider the effect of their strongest nudge treatment relative to the control group. We consider three sets of analyses, with results displayed in Table 8. First, we replicate the basic specification of Ferraro and Price (2013), with the small difference that we further disaggregate their pre-intervention measures of water use and include a separate covariate for each month. Our outcome $Y$, measures levels of water consumption in June, July, August, and September of 2007. With this specification, ML delivers similarly levels of precision improvements as in the OHIE, reducing standard errors by 3% and implying that the sample size could be reduced by 6% while holding statistical power fixed. When replicating their results, we noticed that the distribution of outcomes is fairly skewed, so we next considered a specification where we instead took our outcome to be $\log(Y + 1)$. In this specification, we found substantial gains to using an ML technique. Relative to the linear specification, ML reduced standard errors by 13%, which equivalently can be thought of as implying that a sample size reduction of 24% would leave statistical power unchanged.

An important reason for this discrepancy is that while outcomes are measured in logs, the covariates in the second specification continued to be expressed as levels. We thus consider a third specification where we measure covariates $X$ in logs as well, $\log(X + 1)$. Under this specification, the gains to using an ML technique once again look modest relative to the linear specification. Standard errors are smaller by roughly 2%, or equivalently, sample size could be reduced by roughly 4% holding power fixed. We view this example as demonstrating an important methodological point. If the researcher has good intuition about the functional relation between outcomes and covariates, there are limited gains to using ML techniques over a well-specified linear regression. In this particular case, it is fairly intuitive that if outcomes

are logged, then corresponding covariates should be measured in logs as well. However, our example shows that that ML-based regression adjustments are considerably more robust to pre-analysis transformations that the researcher might make to covariates and thus may be especially helpful when the researcher does not have strong prior information about which functional form specifications are most likely to be accurate.

Table 8: Variance Reduction for Water Conservation

| | SM | LRA | FRA |
|---|---|---|---|
| Un-Logged Outcomes and Covariates | -1.44 | -1.84 | -1.85 |
| | (0.155) | (0.160) | (0.354) |
| Logged Outcomes, Unlogged Covariates | -0.0293 | -0.0365 | -0.0368 |
| | (0.00860) | (0.00463) | (0.00402) |
| Logged Outcomes and Covariates | -0.0293 | -0.0374 | -0.0377 |
| | (0.00860) | (0.00415) | (0.00406) |

**Notes:** This table shows point estimates and standard errors in parentheses for a number of causal parameters from the OHIE across a number of regression adjustment specifications. In the first row, we measure raw outcomes and take raw pre-exposure outcomes as covariates. In the second row, we measure log outcomes. In the third row, we also apply the log transformation to covariates. The sample size is $N = 100,026$

### 3.6 Application IV: CogX, An Early Education Program

Our final empirical example is the evaluation of data from the CogX program described in Fryer et al. (2020). The experiment studies the effect of an early childhood intervention, CogX, on cognitive and non-cognitive test scores. Following the authors, we focus in particular on the effect of CogX on an index of cognitive test scores. For controls, we include a number of demographic variables (birth weight, mother's education, mother's age, household income, race, and gender) as well as pre-intervention test scores.

We include a summary of empirical results in Table 9. We find that ML techniques are able to reduce standard errors by roughly 4% in this setting, which alternatively implies that sample size could have been reduced by roughly 8% while maintaining statistical power. In this setting, this would have amounted to hundreds of thousands of dollars.

## 4 Conclusion

In this paper, we synthesize and generalize a number of approaches to reducing variance when analyzing experimental data. We expand on prior theory along a number of dimensions. We consider a broad class of regression adjustment estimators and identify the conditional expectation function as the minimum variance function to use for the adjustment. We

Table 9: Variance Reduction for CogX

| | SM | LRA | FRA |
|---|---|---|---|
| Cog Test Scores | 7.13 | 10.75 | 8.97 |
| | (2.69) | (2.59) | (2.49) |

**Notes:** This table shows point estimates and standard errors in parentheses for the effect of the CogX program on cognitive test scores across a number of regression adjustment specifications. The sample size is $N = 395$.

then show that regression adjustment estimators can be written in a way that satisfies an *orthogonality* property, which in turn makes it possible to use non-parametric/machine learning estimators to implement feasibly the optimal adjustment under mild regularity assumptions.

These efficiency and feasibility results have important implications for researchers designing and analyzing experiments. They suggest that provided a good approximation to the conditional expectation function is being used for variance reduction, there are limited gains to using additional clever econometrics to enhance precision in experimental data. Importantly, regardless of the parameter being estimated, our results suggest that a researcher seeking a greater level of experimental power should focus more on finding better covariates than on finding better estimators.

We also showcase the practical implications of our theoretical results in a number of synthetic and naturally-occurring datasets. We begin by performing a number of simulation studies by augmenting naturally-occurring Lyft data on outcomes and covariates with synthetic treatments. We show that across a range of sample sizes from hundreds to hundreds of thousands, our asymptotic results provide a reliable guide to inference. We then construct a set of additional simulations to show when researchers may wish to use ML vs linear regression adjustment techniques. After running these simulation studies, we turn towards the analysis of a number of real-world datasets where we are able to quantify the performance various regression adjustment techniques.

Our empirical examples provide a number of key takeaways. First, flexible regression adjustment is not solely a technique for the big data world. Even with relatively small datasets with only hundreds of observations, FRA can improve precision meaningfully. Second, our simulations suggest that sample splitting is crucial for ensuring that the standard errors from the regression adjustment are valid, but once sample splitting is used, inference is reliable and tractable. Third, our example using Lyft data gives a real-world example where a currently popular applied practice of analyzing field experiments with pre-treatment outcomes using two-way fixed effects estimators can lead to substantial losses in statistical precision

relative to regression adjustment. This empirically supports our theoretical result that the two-way fixed effects estimators are statistically inefficient. They should therefore typically be replaced with some form of regression adjustment, linear or otherwise. Fourth, in a number of non-Lyft datasets, we find that the additional gains to using ML techniques over linear regression adjustment can allow researchers to attain similar levels of statistical power with 4-8% fewer observations.

While the results we present are able to generalize the existing literature in a number of ways, they are suggestive of some avenues for future work, which we briefly discuss. First, our efficiency results only apply to the case of i.i.d. sampling. Future work should explore whether the efficiency of our proposed estimator is robust to alternative experiment designs such as blocking or stratification. Second, our results only apply to the cases where the parameters of interest can be expressed as functions of a finite set of sample means. As aforementioned, in some settings, researchers are interested in using experimental variation to estimate a structural parameter. When the estimator of the structural parameter of interest is obtained by optimizing a criterion function, our results may not directly apply. We are currently exploring the possibility of generalizing the ideas behind this present work to that setting.

# References


Andrews, D. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72.

Angrist, J., Imbens, G., and Krueger, A. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2).

Brandon, A., Clapp, C., List, J., Metcalfe, R., and Price, M. (2021). Smart tech, dumb humans: The perils of scaling household technologies. *Working Paper*, 64(4).

Burlig, F., Preonas, L., and Woerman, M. (2019). Panel data and experimental design. *Journal of Development Economics*, 144:102458.

Carneiro, P., Lee, S., and Wilhelm, D. (2020). Optimal data collection for randomized control trials. *The Econometrics Journal*, 23:1–31.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cohen, P. L. and Fogarty, C. B. (2021). No-harm calibration for generalized oaxaca-blinder estimators. *Working Paper*.

Cotton, C., Hickman, B., List, J., Price, J., and Roy, S. (2020). Productivity versus motivation: Combining field experiments with structural econometrics to study adolescent human capital production. *Working paper*.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127:1–56.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132.

Ferraro, P. J. and Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

Finkelstein, A. N., Taubman, S. L., Allen, H. L., Wright, B. J., and Baicker, K. (2016). Effect of medicaid coverage on ed use—further evidence from oregon's experiment. *New England Journal of Medicine (NEJM/MMS)*.

Fisher, R. (1935). *The Design of Experiments*. Oliver and Boyd.

Fowlie, M., Wolfram, C., Spurlock, C. A., Todd, A., Baylis, P., and Cappers, P. (2020). Default effects and follow-on behavior: Evidence from an electricity pricing program. *Working Paper*.

Frison, L. and Pocock, S. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, 11:1685–1704.

Fryer, Jr., R. G., Levitt, S. D., List, J. A., and Samek, A. (2020). Introducing cogx: A new preschool education program combining parent and child interventions. Technical report, National Bureau of Economic Research.

Goldszmidt, A., List, J., Metcalfe, R., Muir, I., Smith, V. K., and Wang, J. (2021). The value of time in the united states: Estimates from nationwide natural field experiments. *Working Paper*.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Gosnell, G., List, J., and Metcalfe, R. (2020). The impact of management practices on employee productivity: A field experiment with airline captains. *Journal of Political Economy*, 128(4):1195–1233.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.

Kaplan, S., Moskowitz, T., and Sensoy, B. (2013). The effects of stock lending on security prices: An experiment. *Journal of Finance*, 68(5):1891–1936.

Negi, A. and Wooldridge, J. (2020). Robust and efficient estimation of potential outcome means under random assignment. *Working Paper*.

Negi, A. and Wooldridge, J. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Poyarkov, A., Drutsa, A., Khalyavin, A., Gusev, G., and Serdyukov, P. (2016). Boosted decision tree regression adjustment for variance reduction in online controlled experiments. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 235–244.

Roth, J. and Sant'Anna, P. H. (2021). Efficient estimation for staggered rollout designs. *Working Paper*.

Todd, P. and Wolpin, K. (2006). Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review*, 96(5):1384–1417.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

# A Code for Non-Lyft Analyses

```
library(dplyr)
library(gbm)
library(randomForest)
library(numDeriv)
library(ggplot2)


# Perform Flexible Regression Adjustment Pre-Processing
# FRA(dat, outcome_cols, treat_col, covariate_cols, n_folds, method)
# Inputs:
#   dat: data frame with outcomes, treatments, and covariates
#   outcome_cols: column names for outcomes of interest
#   treat_col: column name of treatment
#   covariate_cols: column names of covariates
#   n_folds: number of folds for sample splitting
#   method: regression method used for regression adjustment
#   ML_func: Custom ML model supplied by user. Should be of the form ML_func(formula, data).
#            Output should have a predict function.
#
# Output:
#   dat_with_FRA: original dataframe with extra columns of the form
#       'm_{otcuome name}_{treatment name}': fitted value of conditional expectation,
#         E[outcome | X, treatment] for the outcome and treatment named
#       'u_{outcome name}_{treatment name}': "influence function" for mean potential outcome,
#         E[outcome(treatment)]. Mean of this column is the regression adjusted estimator for
#         E[outcome(treatment)] and variance-covariance matrix of these columns is asymptotically
#         valid estimator of covariance matrix of the regression adjusted point estimates
######
FRA <- function(dat, outcome_cols = c('Y'),
                     treat_col = 'W',
                     covariate_cols = c('X1', 'X2', 'X3'),
                     n_folds = 2,
                     method = '',
                     ML_func = NULL, num_trees = 300) {
  # Split sample to ensure balance in treatment status across samples
  dat <- dat %>% as.data.frame
  dat$order <- sample(1:nrow(dat), nrow(dat))
  dat <- dat %>% arrange(!!sym(treat_col), order)
  fold_col <- rep(1:n_folds, ceiling(nrow(dat) / n_folds))
  fold_col <- fold_col[1:nrow(dat)]
  dat$fold <- fold_col

  # Get unique treatment levels
  treat_levels <- unique(dat[,treat_col]) %>% as.vector


  # Perform Crossfitting
  # Split out by method
  # For each outcome/treatment pair, create column called 'm_{outcome name}_{treatment name}'
  # which is the best predictor of outcome given covariates within treatment group
  if (method == 'linear') {
    for (y in outcome_cols) {
      for (treat in treat_levels) {
        # Create new column for m_{outcome name}_{treatment name}
        dat[,paste('m_', y, '_', treat, sep = '')] <- 0
        for (f in 1:n_folds) {
          # Fit OLS model using data from folds except current fold
          lmod <- lm(formula(paste(y, '~', paste(covariate_cols, collapse = '+'))),
                     dat %>% filter(f != fold, !!sym(treat_col) == treat))
          # Project fitted values based on covariates of current fold
```

```
        dat[dat$fold == f,paste('m_', y, '_', treat, sep = '')] <- predict(lmod, dat %>%
                                                                           filter(fold == f))
      }
    }
  }
}
else if (method == 'rf') {
  for (y in outcome_cols) {
    for (treat in treat_levels) {
      # Create new column for m_{outcome name}_{treatment name}
      dat[,paste('m_', y, '_', treat, sep = '')] <- 0
      for (f in 1:n_folds) {
        # Fit random forest model using data from folds except current fold
        rfMod <- randomForest(formula(paste(y, '~', paste(covariate_cols, collapse = '+'))),
                   dat %>% filter(f != fold, !!sym(treat_col) == treat))
        # Project fitted values based on covariates of current fold
        dat[dat$fold == f,paste('m_', y, '_', treat, sep = '')] <- predict(rfMod, dat %>%
                                                                           filter(fold == f))
      }
    }
  }
}
else if (method == 'gbm') {
  for (y in outcome_cols) {
    for (treat in treat_levels) {
      # Create new column for m_{outcome name}_{treatment name}
      dat[,paste('m_', y, '_', treat, sep = '')] <- 0
      for (f in 1:n_folds) {
        # Fit gradient boosting machine model using data from folds except current fold
        gbmMod <- gbm(formula(paste(y, '~', paste(covariate_cols, collapse = '+'))),
                              dat %>% filter(f != fold, !!sym(treat_col) == treat),
                      interaction.depth = 2, n.trees = num_trees, shrinkage = 0.05,
                      distribution = 'gaussian', verbose = F)
        # Project fitted values based on covariates of current fold
        dat[dat$fold == f,paste('m_', y, '_', treat, sep = '')] <- predict(gbmMod, dat %>%
                                                                           filter(fold == f))
      }
    }
  }
}
else if (!is.null(ML_func)) {
  for (y in outcome_cols) {
    for (treat in treat_levels) {
      # Create new column for m_{outcome name}_{treatment name}
      dat[,paste('m_', y, '_', treat, sep = '')] <- 0
      for (f in 1:n_folds) {
        # Fit OLS model using data from folds except current fold
        ML_mod <- ML_func(formula(paste(y, '~', paste(covariate_cols, collapse = '+'))),
                   dat %>% filter(f != fold, !!sym(treat_col) == treat))
        # Project fitted values based on covariates of current fold
        dat[dat$fold == f,paste('m_', y, '_', treat, sep = '')] <- predict(ML_mod, dat %>%
                                                                           filter(fold == f))
      }
    }
  }
}
else {
  stop("Method most be in c('linear', 'rf', 'gbm') or custom method must be supplied")
}

# For each outcome/treatment pair, create column for influence function of the form
# 1 / prob(treatment) * (Y - E[Y|X,treatment]) * 1{treatment} + E[Y|X,treatment]
for (treat in treat_levels) {
  prop_treat <- mean(dat[,treat_col] == treat)
  for (y in outcome_cols) {
    dat <- dat %>% mutate(
      !!sym(paste('u_', y, '_', treat, sep = '')) :=
        case_when(!!sym(treat_col) == treat ~ 1/prop_treat *
                    (!!sym(y) - !!sym(paste('m_', y, '_', treat, sep = ''))),
                  TRUE ~ 0) + !!sym(paste('m_', y, '_', treat, sep = ''))
    )
  }
}
dat_with_FRA <- dat
```

```
  dat_with_FRA
}
#####


# Estimate Average Treatment Effect after Full Regression Adjustment Pre-processing
# FRA_ATE(dat_with_FRA, outcome_col, treat_lvl, ctrl_lvl)
# Inputs:
#   dat_with_FRA: dataframe with regression adjusted columns
#   outcome_col: name of outcome whose ATE is being estimated
#   treat_lvl: value of W corresponding to "treatment"
#   ctrl_lvl: value of W corresponding to "control"
#
# Output:
#   Vector with point estimate and standard error
#####
FRA_ATE <- function(dat_with_FRA, outcome_col = 'Y', treat_lvl, ctrl_lvl) {
  tmp <- dat_with_FRA %>%
    mutate(u = !!sym(paste('u_', outcome_col, '_', treat_lvl, sep = '')) -
             !!sym(paste('u_', outcome_col, '_', ctrl_lvl, sep = '')))

  c(tmp %>% .$u %>% mean, (tmp %>% .$u %>% sd) / sqrt(nrow(tmp)))
}
#####


# Estimate local average treatment effect when experiment assignment W is instrument for treatment
# FRA_LATE(dat_with_FRA, outcome_col, endog_col, treat_lvl, ctrl_lvl)
# using regression-adjusted Wald-style estimator
# Inputs:
#   dat_with_FRA: dataframe with regression adjusted columns
#   outcome_col: name of outcome whose LATE is being estimated
#   endog_col: treatment, which experiment assignment instruments for
#   treat_lvl: value of W corresponding to "treatment"
#   ctrl_lvl: value of W corresponding to "control"
#
# Output:
#   Vector with point estimate and standard error
#####
FRA_LATE <- function(dat_with_FRA, outcome_col = 'Y', endog_col = 'D', treat_lvl, ctrl_lvl) {
  tmp <- dat_with_FRA %>%
    mutate(u_num = !!sym(paste('u_', outcome_col, '_', treat_lvl, sep = '')) -
             !!sym(paste('u_', outcome_col, '_', ctrl_lvl, sep = '')),
           u_denom = !!sym(paste('u_', endog_col, '_', treat_lvl, sep = '')) -
             !!sym(paste('u_', endog_col, '_', ctrl_lvl, sep = '')))

  pe <- mean(tmp$u_num) / mean(tmp$u_denom)
  VCV <- 1/nrow(dat_with_FRA) * matrix(c(var(tmp$u_num), cov(tmp$u_num, tmp$u_denom),
                                         cov(tmp$u_num, tmp$u_denom), var(tmp$u_denom)), nrow = 2)
  D <- c(1 / mean(tmp$u_denom), - mean(tmp$u_num) / mean(tmp$u_denom)^2)

  c(pe, sqrt(D %*% VCV %*% D))
}
#####


# Estimate function of potential outcome means after regression adjustment
# FRA_theta(para_func, dat_with_FRA, outcome_treats)
# Inputs:
#   param_func: function of potential outcome means being estimated
#   dat_with_FRA: dataframe with regression adjusted columns
#   outcome_treats: vector of strings of the form '{outcome name}_{treatment name}' which
#   are the inputs into param_func
# Output:
#   Vector with point estimate and standard error
#####
FRA_theta <- function(param_func, dat_with_FRA, outcome_treats) {
  input_cols = sapply(outcome_treats, function(x) paste('u_', x, sep = ''))
  VCV = matrix(sapply(input_cols, function(x) sapply(input_cols, function(y)
    cov(dat_with_FRA[,x], dat_with_FRA[,y]))),
    nrow = length(outcome_treats))
  m = as.vector(sapply(input_cols, function(x) mean(dat_with_FRA[,x])))

  D <- grad(param_func, m)
```

```
  pe = param_func(m)
  se = sqrt(1/nrow(dat_with_FRA) * D %*% VCV %*% D)
  c(pe, se)
}
#####



# GOTV
#####
data(GerberGreenImai)
dat <- GerberGreenImai
rm(GerberGreenImai)
dat <- dat %>% mutate(Y = VOTED98, W = APPEAL) %>%
  select(Y,W,WARD, AGE, MAJORPTY, VOTE96.0, VOTE96.1, NEW)
dat$WARD <- as.factor(dat$WARD)

set.seed(6124)
covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 2)
dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'rf', n_folds = 10)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)


FRA_ATE(dat_with_FRA, treat_lvl = 3, ctrl_lvl = 1)
FRA_ATE(dat_with_LRA, treat_lvl = 3, ctrl_lvl = 1)

dat %>% group_by(W) %>% summarise(m = mean(Y), v = var(Y) / n()) %>%
  summarise(pe = mean(m[W==3]) - mean(m[W==1]), se = sqrt(mean(v[W==3]) + mean(v[W==1])))
#####

# Ferraro Price
#####
set.seed(326)

# Unlogged Everything
dat <- read_csv('dat_for_RA_ferraroprice.csv') %>% na.omit %>% filter(Y < 200)
dat$Y %>% hist
hist(dat$Y)
covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 2)

dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'gbm', n_folds = 3,
                    num_trees = 600)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)

FRA_ATE(dat_with_FRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)
FRA_ATE(dat_with_LRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)

dat %>% summarise(
  pe = mean(Y[W==3]) - mean(Y[W==4]),
  se = sqrt(var(Y[W==3]) / sum(W==3) + var(Y[W==3])/sum(W==3)))


# Logged Outcome Only
dat <- read_csv('dat_for_RA_ferraroprice.csv') %>% na.omit %>% filter(Y < 200)
dat$Y %>% hist
dat$Y <- log(dat$Y + 1)
hist(dat$Y)
covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 2)

dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'gbm', n_folds = 3,
                    num_trees = 600)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)

FRA_ATE(dat_with_FRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)
FRA_ATE(dat_with_LRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)

dat %>% summarise(
  pe = mean(Y[W==3]) - mean(Y[W==4]),
```

```
  se = sqrt(var(Y[W==3]) / sum(W==3) + var(Y[W==3])/sum(W==3)))


# Logged Everything
dat <- read_csv('dat_for_RA_ferraroprice_logged.csv') %>% na.omit %>% filter(Y < 200)
dat$Y %>% hist
dat$Y <- log(dat$Y + 1)
hist(dat$Y)
covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 2)

dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'gbm', n_folds = 3,
                    num_trees = 600)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)

FRA_ATE(dat_with_FRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)
FRA_ATE(dat_with_LRA, outcome_col = 'Y', treat_lvl = 3, ctrl_lvl = 4)

dat %>% summarise(
  pe = mean(Y[W==3]) - mean(Y[W==4]),
  se = sqrt(var(Y[W==3]) / sum(W==3) + var(Y[W==3])/sum(W==3)))
######

# CHECC
######
dat <- read_csv('dat_for_RA_CHECC.csv')
dat$hl <- as.factor(dat$hl)

set.seed(161)
covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 2)
dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'rf', n_folds = 10)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)


dat_with_FRA %>% filter(W == 0) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_0)^2)))
dat_with_FRA %>% filter(W == 1) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_1)^2)))

dat_with_LRA %>% filter(W == 0) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_0)^2)))
dat_with_LRA %>% filter(W == 1) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_1)^2)))


FRA_ATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)
FRA_ATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)

dat %>% group_by(W) %>% summarise(m = mean(Y), v = var(Y) / n()) %>%
  summarise(pe = mean(m[W==1]) - mean(m[W==0]),
            se = sqrt(mean(v[W==1]) + mean(v[W==0])))
######

# OHIE
######
dat <- read_csv('dat_for_RA_OHIE.csv') %>% na.omit


covariate_cols <- dat %>% colnames %>% tail(ncol(dat) - 3)
set.seed(623)
dat_with_FRA <- FRA(dat, outcome_cols = c('Y','D'),
                    covariate_cols = covariate_cols, method = 'rf', n_folds = 3)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y','D'),
                    covariate_cols = covariate_cols, method = 'linear', n_folds = 10)


dat_with_FRA %>% filter(W == 0) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_0)^2)))
dat_with_FRA %>% filter(W == 1) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_1)^2)))
```

```
dat_with_LRA %>% filter(W == 0) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_0)^2)))
dat_with_LRA %>% filter(W == 1) %>%
  summarise(Y_sd_0 = sd(Y), rmse_0 = sqrt(mean((Y-m_Y_1)^2)))



FRA_ATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)
FRA_ATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)

(FRA_ATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)[2]/
    FRA_ATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)[2])^2


FRA_ATE(dat_with_FRA, outcome_col = 'D', treat_lvl = 1, ctrl_lvl = 0)
FRA_ATE(dat_with_LRA, outcome_col = 'D', treat_lvl = 1, ctrl_lvl = 0)


dat %>% summarise(
  pe_rf = mean(Y[W==1]) - mean(Y[W==0]),
  se_rf = sqrt(var(Y[W==1]) / sum(W==1) + var(Y[W==0])/sum(W==0)),
  pe_fs = mean(D[W==1]) - mean(D[W==0]),
  se_fs = sqrt(var(D[W==1]) / sum(W==1) + var(D[W==0])/sum(W==0)))


FRA_LATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)
FRA_LATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)
dat %>% felm(Y~1|0|(D~W), data = .) %>%
  summary


(FRA_LATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)[2]/
FRA_LATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)[2])^2

dat %>%
  felm(formula(paste('Y~',paste(covariate_cols,collapse='+'),'|0|(D~W)')),
       data = .) %>%
  summary
#####




# Simulation example
#####
# Latent variables L1, L2, L3


get_pe <- function(p, interaction,N = 1000, method = 'rf') {
  W <- sample(c(0,1), N, replace = T)
  L1 <- runif(N, 0, 1)
  L2 <- runif(N, 0, 1)
  L3 <- runif(N, 0, 1)
  if(interaction == 1) {
    X1 <- (L1 * L2)^p
    X2 <- (L2 * L3)^p
    X3 <- (L3 * L1)^p
  } else {
    X1 <- L1^p
    X2 <- L2^p
    X3 <- L3^p
  }
  U <- rnorm(N, 0, 0.5)

  Y <- L1 + L2 + L3 + W + U

  dat <- data.frame(W = W, X1 = X1, X2 = X2, X3 = X3, Y = Y)



  # Apply regression adjustment pre-processing
  dat_with_FRA <- FRA(dat, outcome_cols = c('Y'), method = method, n_folds = 5)
```

```
  # Compare FRA_theta with FRA_ATE estimates of average effect
  FRA_ATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)[1]
}


set.seed(216)
for (p in c(1, 5,10)) {
  for (interaction in c(0,1)) {
    fits_ml <- sapply(1:100, function(x) get_pe(p,interaction, method = 'rf'))
    fits_linear <- sapply(1:100, function(x) get_pe(p,interaction, method = 'linear'))
    print(c(p,interaction, sd(fits_ml) / sd(fits_linear)))
  }
}


N <- 1000
p <- 1
interaction = 1

W <- sample(c(0,1), N, replace = T)
L1 <- runif(N, 0, 1)
L2 <- runif(N, 0, 1)
L3 <- runif(N, 0, 1)
if(interaction == 1) {
  X1 <- (L1 * L2)^p
  X2 <- (L2 * L3)^p
  X3 <- (L3 * L1)^p
} else {
  X1 <- L1^p
  X2 <- L2^p
  X3 <- L3^p
}
U <- rnorm(N, 0, 0.5)

Y <- L1 + L2 + L3 + W + U

dat <- data.frame(W = W, X1 = X1, X2 = X2, X3 = X3, L1 = L1, L2 = L2, L3 = L3, Y = Y)

# Apply regression adjustment pre-processing
dat_with_FRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = c('X1', 'X2', 'X3'), method = 'rf',n_folds = 5)
dat_with_LRA <- FRA(dat, outcome_cols = c('Y'),
                    covariate_cols = c('X1', 'X2', 'X3'), method = 'linear',n_folds=5)


# Produce plots to show model fit
dat_with_FRA %>% filter(W == 0) %>% mutate(truth = L1 + L2 + L3) %>%  ggplot +
  geom_point(aes(x=truth,y=m_Y_0, col = 'CEF')) +
  geom_point(aes(x=truth, y = Y, col = 'Actual Data')) +
  geom_abline(aes(slope = 1,intercept=0)) + labs(x = 'E[Y|X]', y = 'Fit', title = 'Flexible RA') +
  scale_color_discrete(name = 'Type')

dat_with_LRA %>% filter(W == 0) %>% mutate(truth = L1 + L2 + L3) %>%  ggplot +
  geom_point(aes(x=truth,y=m_Y_0, col = 'CEF')) +
  geom_point(aes(x=truth, y = Y, col = 'Actual')) +
  geom_abline(aes(slope = 1,intercept=0)) + labs(x = 'E[Y|X]', y = 'Fit', title = 'Linear RA') +
  scale_color_discrete(name = 'Type')


# Compare FRA_theta with FRA_ATE estimates of average effect
FRA_ATE(dat_with_FRA, treat_lvl = 1, ctrl_lvl = 0)
FRA_ATE(dat_with_LRA, treat_lvl = 1, ctrl_lvl = 0)
#####
```