

Variance Reduction with Model-Based Counterfactual Estimation

by

Kyu Min Shim

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Data Science

Waterloo, Ontario, Canada, 2024

© Kyu Min Shim 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Variance reduction is an important area of research in the realm of *online controlled experiments*, also known as *A/B tests*. Reducing outcome variability in an A/B test improves the test’s statistical power and improves the efficiency of the experimentation process. Many variance reduction techniques already exist, and they typically utilize data collected prior to the experiment (pre-experiment data) to reveal complex relationships between the outcome of interest and covariates. These insights are then applied to the data collected during the experiment (in-experiment data) to reduce the outcome variability in the A/B test. However, such methods are heavily reliant on the assumption that pre- and in-experiment data are highly correlated. This is questionable in online settings where trends change quickly due to heterogeneity in user behavior, the rapid development of technology, and the competitive landscape. In these settings, we cannot ignore that fluctuations in other factors may degrade the correlation between pre- and in-experiment data. We propose a two-stage framework for treatment effect estimation that adjusts for differences between pre- and in-experiment data, thereby producing treatment effect estimators with smaller variance than those associated with other variance reduction methods. Inference is conducted by modeling and estimating the counterfactual outcome of each unit and performing a pairwise comparison. This method of inference is shown to be asymptotically unbiased, with an asymptotic variance that scales with the model’s predictive accuracy. We compare the variance reduction capabilities of the proposed method with several alternatives through simulation studies using both simulated data and real-world data. In doing so, we demonstrate that the proposed method’s variance reduction capabilities are at least as good (and in some cases orders of magnitude better) than that of existing methods.

Acknowledgments

I would like to thank Professor Nathaniel Stevens for supervising my journey through this Master's degree. I first came across the field of A/B tests through his course on Experimental Design. I am very grateful for the many resources and guidance he provided me that enabled the exploration and testing of my curiosity.

I would also like to thank the other faculty members in the Department of Statistics and Actuarial Science who taught me everything I know about statistics since my undergraduate years here at the University of Waterloo.

Lastly, I would like to thank my mom, my dad, my brother, and my partner Arielle for being the pillars of my life. Their unconditional support makes even the toughest challenges appear trivial.

Dedication

This is dedicated to my family, especially to my late dog Coco. I knew one day we would run out of time, but I never thought that day was so near. Thank you for coming into my life, and thank you for all the happiness you brought to my family. I will be happy just to have known you. Until we meet again.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgments	iv
Dedication	v
List of Figures	viii
List of Tables	xi
1 Introduction	1
1.1 Online Controlled Experiments, Counterfactuals, and Variance Reduction .	1
1.2 Method of Control Variates	3
1.3 Time-Dependence of Online Activities	5
1.4 Outline of the Thesis	7
2 Methodology	8
2.1 Counterfactual Estimation with Constant Model	9
2.1.1 The Method	9
2.1.2 Exploration of Sampling Distributions	11

2.2	Counterfactual Estimation with Linear Regression	13
2.2.1	The Method	13
2.2.2	Demonstration of Variance Reduction	19
2.3	Counterfactual Estimation with Machine Learning	21
2.3.1	The Method	21
2.3.2	Demonstration of Variance Reduction	24
2.4	Counterfactual Estimation with Machine Learning and Secondary Linear Adjustment	25
2.4.1	The Method	25
2.4.2	Demonstration of Variance Reduction	28
3	Performance Investigation	33
3.1	Pageview Dataset	33
3.1.1	Overview	33
3.1.2	Simulating Experiments	34
3.2	Complexity of Secondary Adjustment	38
4	Conclusion	45
	References	47
	APPENDICES	49
A	Supplementary Materials	50
A.1	Proof of 0 covariance in conditional variance of $\tilde{\tau}_r$	50
A.2	Fitting a SARIMA Model	52

List of Figures

1.1	A simplified example of Google’s experiment. Two shades of blue are compared, with HTML color codes #2358C2 for the control group and #1A0DAB for the treatment group.	2
2.1	An instance of simulated experiment data and counterfactual estimates using the constant model in Equation (2.2).	12
2.2	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 2.2b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.2a.	13
2.3	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 2.3b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.3a.	14
2.4	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 2.4b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.4a.	14
2.5	Simulated experiment data ordered by covariate x . The dotted lines in 2.5b represent the fitted linear model from Equation (2.11). The open circles in 2.5b represent the estimated counterfactual outcome for each observed outcome (filled circles).	20
2.6	Visualization of pre- and in-experiment data. The grey line in Figure 2.6a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.6b represent treatment and control outcome estimates over covariate x	25

2.7	Visualization of pre- and in-experiment data. The grey line in Figure 2.7a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.7b represent treatment and control outcome estimates over covariate x with secondary linear adjustment.	29
2.8	Visualization of misaligned pre- and in-experiment data. The grey line in Figure 2.8a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.8b represent treatment and control outcome estimates over covariate x with secondary linear adjustment.	30
2.9	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 2.9b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.9a.	31
2.10	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 2.10b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.10a.	31
2.11	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 2.11b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.11a.	32
3.1	Time series of page visits. Figure 3.1a shows increasing variability over time, hence log transformation is applied to achieve constant variance as demonstrated in Figure 3.1b.	34
3.2	Simulating in-experiment data from test data using random sampling with replacement. The treatment units are applied a constant treatment effect of $\tau = 0.2$	35
3.3	Visualization of counterfactual estimators described in Chapter 2 on an instance of simulated experiment data, their treatment effect estimates $\hat{\tau}$ and standard errors $SE[\hat{\tau}]$	36
3.4	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 3.4b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.4a.	37
3.5	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 3.5b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.5a.	37

3.6	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 3.6b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.6a.	38
3.7	Estimating counterfactuals with degree 4 polynomial secondary adjustment. $\hat{\tau}_q = 0.2073, SE[\hat{\tau}_q] = 0.0230$	40
3.8	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 3.8b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.8a.	40
3.9	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 3.9b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.9a.	41
3.10	Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 3.10b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.10a.	41
A.1	ACF and PACF of page visits in training data.	52
A.2	ACF and PACF of seasonally differenced page visits in train data.	52
A.3	Optimal model fit with $p = 0, d = 1, q = 6, P = 1, D = 0, Q = 1, S = 7$. The residuals appear normally distributed and stationary.	54

List of Tables

2.1	Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$. .	21
2.2	Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$ when pre- and in-experiment data are misaligned.	32
3.1	Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$ in the Page Visits dataset.	39
3.2	Proportion of statistically significant experiments out of 10,000 with $n_1 =$ $n_0 = 10$	42
3.3	Proportion of statistically significant experiments out of 10,000 with $n_1 =$ $n_0 = 50$	42
3.4	Proportion of statistically significant experiments out of 10,000 with $n_1 =$ $n_0 = 100$	42

Chapter 1

Introduction

1.1 Online Controlled Experiments, Counterfactuals, and Variance Reduction

Online controlled experiments (OCEs), commonly known as A/B tests, are the gold standard method used to assess the causal impact of changes to websites, apps, and other internet-based products. Companies such as Google, LinkedIn, and Microsoft each run hundreds of online A/B tests on millions of users every day, quantifying the impact of changes to their platforms on key business metrics [1]. A/B tests involve randomly assigning experimental units (often the users) to one of two treatment variants: units in the treatment group (B) experience some change to the product or service, and the units in the control group (A) do not. By measuring the business metric of interest in each group and employing statistical analysis, A/B tests enable practitioners to draw causal conclusions about the relationship between the treatment and the metric. Such conclusions provide a data-driven means to decide whether to push the product or service change to the entire user base.

To provide some context for further discussion, consider Google’s now infamous “41 Shades of Blue” experiment [2]. Although this experiment has more than two treatment groups (such an experiment is referred to as an *A/B/n test*), we use it as an illustrative example because it allows us to describe many important components of OCEs. In this experiment, Google was interested in finding the optimal shade of blue for the links in its search results. Google randomly assigned users to different shades of blue and measured how often users clicked on the links. By choosing the shade of blue that had the highest

Control Group (A): www.google.com

Treatment Group (B): www.google.com

Figure 1.1: A simplified example of Google’s experiment. Two shades of blue are compared, with HTML color codes [#2358C2](#) for the control group and [#1A0DAB](#) for the treatment group.

click-through rate (CTR), Google was able to generate an estimated 200 million USD per year in additional revenue. The result of this experiment accounts for the particular shade of blue ([#1A0DAB](#) to be exact [3]) used on all Google links. On average, we unknowingly click these links more frequently than we would if a different shade of blue was chosen.

Consider a simplified version of the above experiment with only two shades of blue: [shade A](#) is the existing shade of blue ([#2358C2](#)), and [shade B](#) is a new shade of blue ([#1A0DAB](#)). See Figure 1.1 for a visual comparison of these shades of blue. In the context of online A/B tests, the counterfactual represents user behavior in the unobserved scenario. For a treatment user assigned to [shade B](#), the counterfactual outcome is the clicking behavior of that user had they been assigned to [shade A](#). For a control user assigned to [shade A](#), the counterfactual outcome is the clicking behavior had they been assigned to [shade B](#). In an ideal though hypothetical scenario, practitioners would be able to observe both the treatment and control outcomes of every user simultaneously. If this was possible, the true treatment effect (in this example, the difference in CTR between [shade A](#) and [shade B](#)) could be measured by averaging the pairwise differences between the two outcomes for each user. However, in reality, each user can only be assigned to a single treatment variant at a time, hence both treatment and control outcomes cannot be observed simultaneously. That is, a user cannot view a Google link in two different shades of blue at the same time. As such, the treatment effect is typically estimated at the group level by comparing the average outcome observed among users randomly assigned to the treatment and control groups. This method of estimating the treatment effect is called *difference in group means*. Causal conclusions made with difference in group means rely on the assumption that the randomization of individuals to each group renders the two groups homogeneous except for the treatment assignment, which allows any difference in outcomes between the two groups to be attributed to the treatment.

However, identifying a non-null treatment effect is difficult when the outcome and hence the metric of interest is highly variable. For instance, if the outcome has high variance, then the control and treatment groups may have different average outcomes purely by

chance even when there is no treatment effect. Similarly, even when there is a treatment effect, the control and treatment groups may result in similar outcomes due to their high variability, leading experimenters to incorrectly identify the absence of a treatment effect. High variability in the outcome is commonplace in A/B tests [4] and this negatively impacts the quality of their insights. As such, *variance reduction* has been described as one of the top challenges and an active research area in the context of A/B testing [4, 5]. Especially when interest lies in detecting a treatment effect that is small in magnitude, large variations in the outcome can compromise the validity of A/B tests even with large sample sizes. Further developments in the area of variance reduction are necessary to facilitate reliable decision-making with A/B tests.

The novel method introduced in this thesis synthetically generates a counterfactual outcome for each unit with which to perform pairwise comparisons. In doing so, the proposed method aims to reduce the variance of the treatment effect estimator. In the context of online experiments, pre-experiment data is readily accessible as online platforms collect user activity data through technologies such as cookies, device fingerprinting, and advertising identifiers, to provide personalized service, recommendations, and marketing analytics [6]. Like existing variance reduction methods [7, 8], the proposed technique exploits the availability of such pre-experiment data and leverages it for its variance reduction capability. However, existing variance reduction techniques fail to achieve substantial variance reduction if pre- and in-experiment data are not highly correlated. As will be detailed in Section 1.3, metrics in online settings are heavily dependent on measurement time. Hence, the pre- and in-experiment data may differ systematically if they are measured under different temporal conditions. The proposed method, on the other hand, makes appropriate adjustments in such cases to achieve variance reduction. This framework contributes to a more thorough understanding of the underlying mechanisms governing observed outcomes, thereby reducing the variance associated with treatment effect estimation.

1.2 Method of Control Variates

The *method of control variates* is a statistical technique used to reduce variance in the estimation of a parameter of interest [9]. By introducing an additional variable (referred to as the control variate) that is correlated with the primary variable of interest, the method creates a composite estimator that enhances the precision of parameter estimation without introducing bias. Let Y_i be the outcome observed for experimental unit (i.e., user) $i = 1, \dots, n$. Let X_i be some observed covariate of unit i such that $\mathbb{E}[X_i] = \mu_x$, which is used as the control variate. The covariates and the outcomes conditional on covariates are

assumed to be independently and identically distributed over all units. Define the sample means of the outcome and the covariate as $\bar{Y}$ and $\bar{X}$ respectively. Then, we can define an augmented metric $\bar{Y}^* = \bar{Y} - \theta(\bar{X} - \mu_x)$, where $\theta \in \mathbb{R}$ is some constant. The variance of the augmented metric is as follows:

$$\begin{aligned}\mathbb{V}[\bar{Y}^*] &= \mathbb{V}[\bar{Y} - \theta\bar{X}] \\ &= \mathbb{V}[Y_i - \theta X_i]/n \\ &= (\mathbb{V}[Y_i] + \theta^2\mathbb{V}[X_i] - 2\theta\text{Cov}[Y_i, X_i])/n.\end{aligned}$$

Note that $\mathbb{V}[Y_i - \theta X_i]$ is assumed to be constant with respect to the units $i = 1, \dots, n$. This is because we assume the outcome Y_i conditional on covariate X_i to be independently and identically distributed over the units. By taking the derivative of $\mathbb{V}[\bar{Y}^*]$ with respect to θ , it can be shown that $\theta = \text{Cov}[Y_i, X_i]/\mathbb{V}[X_i]$ minimizes $\mathbb{V}[\bar{Y}^*]$. With this optimal θ , the variance of the augmented metric simplifies to $\mathbb{V}[\bar{Y}^*] = (1 - \rho^2)\mathbb{V}[\bar{Y}] \leq \mathbb{V}[\bar{Y}]$ where $\rho = \text{Corr}[Y_i, X_i]$. Using subscripts 1 and 0 to denote treatment and control outcomes respectively, the variance of the traditional difference in means estimator $\bar{Y}_1 - \bar{Y}_0$ would be much higher than $\bar{Y}_1^* - \bar{Y}_0^*$ when $|\rho| \approx 1$.

The magnitude of variance reduction is highly dependent on the choice of the control variate, as the amount of variance reduction scales with the correlation between the outcome variable of interest and the control variate. An intuitive choice for the control variate is the historical value of the primary variable of interest. The application of control variates in online experiments has been made popular through a method known as Controlled-experiment Using Pre-Experiment Data (CUPED) [7], where we set $X_i = Y_i^{(pre)}$ (the pre-experiment outcome of unit i) hoping that $|\text{Corr}[Y_i, Y_i^{(pre)}]| \approx 1$. In today's world where massive amounts of online data are frequently logged and stored, the use of pre-experimental data is an efficient and effective way to reduce variation in online experiments. However, the effectiveness of CUPED relies on the existence of high-quality pre-experiment data. For scenarios where such pre-experimental data is not available, an alternative method referred to as Control Using Predictions as Covariates (CUPAC) has been suggested [10]. CUPAC uses covariates measured during the experiment to build a machine learning model of the outcome, then uses the predictions of this model as the control variate. A good predictive model would output a highly correlated synthetic control variate that can be used to reduce variance.

Poyarkov et al. [8] introduced a related method called Variance Reduction via Subtraction of Prediction. This method first fits a machine learning model $\hat{f}$ to predict the outcome variable Y based on a vector of covariates $\mathbf{X}$ from the pre-experiment data. Then, the model is used to predict the outcome of the in-experiment units based on their

covariates. Under the assumption that pre- and in-experiment data tend to be alike, the predictions $\hat{f}(\mathbf{X}_i)$ should well-model the in-experiment outcome Y_i related to the covariates $\mathbf{X}_i$. Instead of the observed outcomes Y_i , the difference $d_i = Y_i - \hat{f}(\mathbf{X}_i)$ is used to estimate the treatment effect. The variance of the new metric $\mathbb{V}[d_i]$ is expected to be lower than $\mathbb{V}[Y_i]$ as the variability in the outcome explained by the covariates has been removed by the prediction $\hat{f}(\mathbf{X}_i)$. Therefore, using d_i to construct a difference in group means estimator results in treatment effect estimation with lower variance. Since the pre-experiment data is observational, it is typically larger in size than the in-experiment data. Hence, it is possible to fit a flexible machine learning model to pre-experiment data with less concern for overfitting than if it were fit to the smaller in-experiment data.

The methods described here that utilize pre-experiment data ([7, 8]) assume a strong correlation between pre- and in-experiment data. If this is violated, variance reduction may be immaterial; the variance of the resulting treatment effect estimator can be as large as that of the usual difference-in-means estimator. Taking CUPED as an example, if $|\rho| = \text{Corr}[Y_i, Y_i^{(pre)}] \approx 0$, then $\mathbb{V}[\bar{Y}^*] \approx \mathbb{V}[\bar{Y}]$. This is a concern in online settings such as social media and e-commerce where, as will be discussed in Section 1.3, online usage and behavior can change rapidly, and so the assumption of high correlation between pre- and in-experiment data may be questionable. The method proposed in this thesis attempts to address this challenge by incorporating an additional linear adjustment of the predictions from a machine learning model fit to pre-experiment data to better align with the in-experiment data. In the scenario where pre- and in-experiment data differ, the secondary adjustment exploits the relationship between covariates and the outcome in both pre- and in-experiment data to achieve variance reduction. In the scenario where the pre- and in-experiment data are well-aligned, the secondary linear adjustment is not needed and the proposed method reduces to variance reduction via subtraction of prediction [8]. We present this methodology in Chapter 2.

1.3 Time-Dependence of Online Activities

User activities on online platforms such as e-commerce and social media are highly patterned over time. As such, it can be costly, both in terms of accuracy and precision of treatment effect estimation, to assume that outcomes related to user activity are independent of the time of measurement. Klarna, an e-commerce payment and shopping service technology company, performs quarterly consumer research. In the second quarter of 2023, Klarna analyzed the behaviors of approximately 20,000 consumers across 18 countries, coupled with shopping data from over 150 million consumers and 500,000 merchants on

its platform [11]. The study revealed dual peaks in online shopping activity during most days, typically around lunchtime (1 pm to 2 pm) and in the evening (8 pm to 9 pm). This insight aligns well with people’s typical work schedule of 9 am to 5 pm, and their tendency to use mobile devices and engage in mobile shopping during lunch breaks and leisure hours after work. The study further illuminates how online shopping behavior varies by day and country, highlighting distinct preferences across different regions. US consumers tend to shop much more on Fridays relative to other weekdays. Consumers in Sweden, Norway, and Finland tend to shop online on Sundays and Mondays, whereas consumers in Portugal and Spain sharply reduce online shopping activities over the weekends. A parallel investigation on online shopping hours in 2021 by SaleCycle, a conversion rate optimization company, corroborates these insights, emphasizing consistent peak online shopping times at 11 am and 8 pm [12]. SaleCycle’s study also found that Mondays are the busiest day of the week for online shopping, while Saturdays are the slowest day of the week for online shopping.

In terms of monthly patterns, the SaleCycle study noted peak online shopping activity towards the start and the end of the month, attributing this peak to the paydays of the general population. It is now common knowledge that online shopping activities also peak towards the end of the year due to the holidays, especially on days such as Black Friday and Cyber Monday which saw record-breaking US e-commerce spending in 2023 of \$9.8 billion and \$12.4 billion respectively, up 7.5% and 9.6% year-over-year [13]. The effect of holiday spending can be seen in Amazon’s quarterly sales revenue from 2015 to 2022, where Q4 sales in each year are 25% to 50% higher than the average quarterly sales of the first three quarters of the same year [14].

Similarly, a study of mobile activities of more than 27 million individuals in Beijing, China, revealed peak mobile phone usage at 10 am and 6 pm [15]. Another study of smartphone usage found that users tend to be the most active on Mondays, followed by slowly reduced activities throughout the week and increasing again on Sundays [16]. The study also showed that the day-of-the-week effect on usage varies depending on the type of application used such as social networking, finance, health, and web browsing. For example, social networking usage peaks on Mondays and slowly falls until Sunday, whereas web browsing shows a slow decrease in usage from Monday to Friday then sharply increases over the weekend.

As evidenced by these studies, users’ online usage and behavior are highly seasonal and follow different patterns at different increments of time (e.g., hourly, daily, weekly). Therefore, assuming a high correlation between pre- and in-experiment data in the presence of strong seasonality may not be appropriate. In the case that the correlation is low, variance reduction by methods that assume otherwise will be poor. To fully utilize pre-experiment data for variance reduction, there is a need for a flexible approach that can

adapt to possible differences between pre- and in-experiment data.

1.4 Outline of the Thesis

Chapter 2 describes the proposed methodology and hence the key contribution of this thesis. Throughout we draw connections to existing methods, beginning in Section 2.1 by considering the traditional difference in means estimator of treatment effect. This method is contrasted with those described in Sections 2.2, 2.3, and 2.4 which are treatment effect estimators based on progressively more flexible models of the outcome with lower variance. In particular, Section 2.4 introduces the novel method proposed in this thesis and considers a scenario in which pre- and in-experiment data are not well aligned, where the proposed method outperforms the existing methods described in Sections 2.1, 2.2, and 2.3. While the methodology is illustrated and justified via toy examples in Chapter 2, in Chapter 3 we conduct extensive simulation studies that explore the performance of the proposed variance reduction method on a real-world dataset. We conclude the thesis in Chapter 4 with a summary of the proposed method and a discussion of its limitations and extensions.

Chapter 2

Methodology

The primary motivation behind this methodology is variance reduction via clever exploitation of pre-experiment data. By fitting a prediction model to the pre-experiment data and making appropriate adjustments to the predictions with respect to the in-experiment data, we synthetically generate accurate counterfactual observations for each unit. Then, instead of comparing group means, we propose an analysis of the difference between the observed outcomes and the corresponding counterfactual estimates. The result is an augmentation of the traditional difference in means estimator with improved precision for treatment effect estimation.

In regression methods for completely randomized experiments, variance reduction by modeling the outcome as a linear function of covariates is common [17]. Such methods fit a linear model to in-experiment data to explain the relationship between the covariates and the outcome. If the covariates and the outcome are highly correlated, the regression model explains a large portion of the variation in the outcome. Then, the remaining variation in the outcomes is attributed to the treatment effect. In online experiment settings where large amounts of pre-experiment data may be available, we can fit a more flexible machine learning model to the pre-experiment data and use its prediction to account for potentially highly complex relationships between the covariates and the outcome observed during the experiment [8]. However, the performance of this method is heavily dependent on the correlation between the pre-experiment and in-experiment data. In the case where the relationship between covariates and the outcome can change over time, this method results in poor variance reduction. To solve this problem, we propose a two-staged modeling framework that exploits the relationships between the covariates and the outcome both before and during the experiment that can achieve significant variance reduction even when the relationships between covariates and the outcome change over time. We review

existing methods in Sections 2.1-2.3, priming the reader for the proposed method described in Section 2.4.

2.1 Counterfactual Estimation with Constant Model

2.1.1 The Method

We first consider the simplest case where no covariates are introduced. We use the model-based approach that assumes the potential outcomes are independently and identically distributed random variables from some joint distribution of a super-population, and randomness comes from inherent sampling variability. Below in Equation (2.1), we define notation for the mean and variance of potential outcomes in the super-population:

$$\alpha_W = \mathbb{E}[Y_i(W)], \quad \sigma_W^2 = \mathbb{V}[Y_i(W)]. \quad (2.1)$$

Above and throughout, $W \in \{0, 1\}$ denotes treatment assignment with $W = 0$ denoting control and $W = 1$ denoting treatment. The primary interest lies in estimating the super-population treatment effect: $\tau = \alpha_1 - \alpha_0$. We next consider the estimation of the potential outcomes, as this forms the basis of our counterfactual imputation methodology. We define the following model for the potential outcomes:

$$Y_i(W_i) = \alpha_1 W_i + \alpha_0 (1 - W_i) + \epsilon_i. \quad (2.2)$$

Here, $\epsilon_i \sim N(0, \sigma_{d_{W_i}}^2)$ are independently distributed error terms with finite variances $\sigma_{d_1}^2 = \sigma_1^2$ and $\sigma_{d_0}^2 = \sigma_0^2$. The subscript d specifies that the variance $\sigma_{d_{W_i}}^2$ is associated with the error terms of the model in Equation (2.2). In the following sections, new subscripts will be introduced to differentiate the variances of the error terms associated with different models.

We define the observed outcome as Y_i^{obs} , which corresponds to the treatment potential outcome for treatment units ($W_i = 1$) and control potential outcome for control units ($W_i = 0$). Given the observed outcomes from n units during the experiment (with n_1 treatment units and $n_0 = n - n_1$ control units), we define the estimators of the model coefficients with ordinary least squares (OLS):

$$\begin{aligned} (\tilde{\alpha}_1, \tilde{\alpha}_0) &:= \arg \min_{(\alpha_1, \alpha_0) \in \mathbb{R}^2} \rho(\alpha_1, \alpha_0) \\ &= \arg \min_{(\alpha_1, \alpha_0) \in \mathbb{R}^2} \sum_{i=1}^n (Y_i^{obs} - \alpha_{W_i})^2 \end{aligned} \quad (2.3)$$

We derive the estimates of α_1 and α_0 by solving the equations that arise by equating to 0 the partial derivatives of Equation (2.3) with respect to α_1 and α_0 .

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha_1} \rho(\alpha_1, \alpha_0) \\ \rightarrow \tilde{\alpha}_1 &= \frac{1}{n_1} \sum_{i:W_i=1} Y_i^{obs} \\ &= \bar{Y}_1, \end{aligned} \tag{2.4}$$

where $\bar{Y}_1$ is the mean of observations in the treatment group. Similarly,

$$\tilde{\alpha}_0 = \bar{Y}_0, \tag{2.5}$$

where $\bar{Y}_0$ is the mean of observations in the control group. The OLS coefficient estimators are unbiased:

$$\begin{aligned} \mathbb{E}[\tilde{\alpha}_W] &= \mathbb{E}[\bar{Y}_W] \\ &= E[\alpha_W + \bar{\epsilon}_W] \\ &= \alpha_W \end{aligned} \tag{2.6}$$

where $\bar{\epsilon}_W = \frac{1}{n_W} \sum_{i:W_i=W} \epsilon_i$. Using the estimators of the coefficients, estimators of the potential outcomes are constructed with the following model:

$$\tilde{Y}_i(W_i) = \tilde{\alpha}_1 W_i + \tilde{\alpha}_0 (1 - W_i). \tag{2.7}$$

This potential outcome estimator can be used to estimate the average treatment effect by averaging the pairwise differences between observed and estimated counterfactual outcomes for each unit. The average treatment effect estimator using the difference in group means is $\tilde{\tau}_d$.

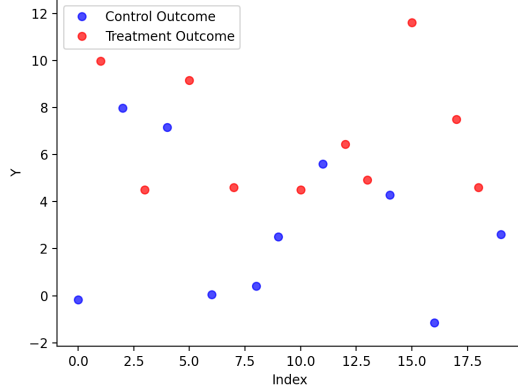
$$\begin{aligned} \tilde{\tau}_d &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{Y}_i(0)) + \sum_{i:W_i=0} (\tilde{Y}_i(1) - Y_i^{obs}) \right] \\ &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{\alpha}_0) + \sum_{i:W_i=0} (\tilde{\alpha}_1 - Y_i^{obs}) \right] \\ &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs}) - n_1 \tilde{\alpha}_0 - \sum_{i:W_i=0} (Y_i^{obs}) + n_0 \tilde{\alpha}_1 \right] \\ &= \frac{1}{n} [n_1 \tilde{\alpha}_1 - n_1 \tilde{\alpha}_0 - n_0 \tilde{\alpha}_0 + n_0 \tilde{\alpha}_1] \\ &= \tilde{\alpha}_1 - \tilde{\alpha}_0 \end{aligned} \tag{2.8}$$

Based on the results of Equations (2.4) and (2.5), we have that $\tilde{\tau}_d = \tilde{\alpha}_1 - \tilde{\alpha}_0$ is simply the difference in group means estimator. Therefore, the difference in group means method is equivalent to estimating the counterfactual outcome of each unit by the mean of the opposing group. The unbiasedness of $\tilde{\tau}_d$ and its variance follow from the familiar result concerning the difference in group means estimator:

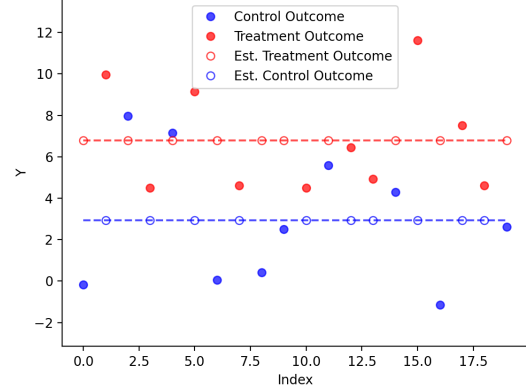
$$\begin{aligned}\mathbb{E}[\tilde{\tau}_d] &= \mathbb{E}[\bar{Y}_1 - \bar{Y}_0] \\ &= \alpha_1 - \alpha_0 \\ \mathbb{V}[\tilde{\tau}_d] &= \mathbb{V}[\bar{Y}_1 - \bar{Y}_0] \\ &= \frac{\sigma_{d_1}^2}{n_1} + \frac{\sigma_{d_0}^2}{n_0}\end{aligned}\tag{2.9}$$

2.1.2 Exploration of Sampling Distributions

The goal of this section is to verify the results in Section 2.1.1 regarding the bias and variance of $\tilde{\tau}_d$ using simulated experiment data. They serve as the benchmark against which the bias and variance of the methods introduced later in Chapter 2 are compared. Even though $\tilde{\tau}_d$ does not consider any covariates, we simulate the outcome variable Y as a function of a single covariate x for consistency of presentation as methods that are introduced later in this chapter rely on the relationship between the covariate x and the outcome Y . We simulate data assuming the covariates are distributed uniformly over the integers: $\{x \in \mathbb{Z} | 70 \leq x < 100\}$. The experiment data for covariate x is generated by taking two random samples of equal sizes $n_1 = n_0 \in \{10, 50, 100\}$ with replacement from its distribution. The sampling with replacement ensures independence between treatment assignment and the covariate, which is a key assumption that is made for methods in Sections 2.2, 2.3, and 2.4. If sampled without replacement, then assigning a particular value of x to the treatment group guarantees that it is not assigned to the control group which violates the independence assumption. The in-experiment outcomes are then generated by the following function: $Y_i(W_i) = \tau W_i + 2 \sin(2x_i) + 0.01(x_i - 70)^2 + 0.4\epsilon_i$, $\epsilon_i \sim N(0, 1)$ with the true treatment effect $\tau = 2$. The non-linearity of the function for generating the outcome Y is used to show the impact of model misspecification on the bias and variance of the estimators, which is important for the methods introduced in later sections. For each sample, the treatment effect estimate is computed as $\hat{\tau} = \hat{\alpha}_1 - \hat{\alpha}_0$ and the variance estimate is computed as $\hat{\mathbb{V}}[\hat{\tau}] = \hat{\sigma}_{d_1}^2/n_1 + \hat{\sigma}_{d_0}^2/n_0$, where $\hat{\sigma}_{d_w}^2$ is computed by taking the residual variance of the model for group W . In the case of the model in Equation (2.2), this is simply the sample variance of the outcomes in each group. The standard error is then $SE[\hat{\tau}] = \sqrt{\hat{\mathbb{V}}[\hat{\tau}]}$.



(a) Observed outcomes.

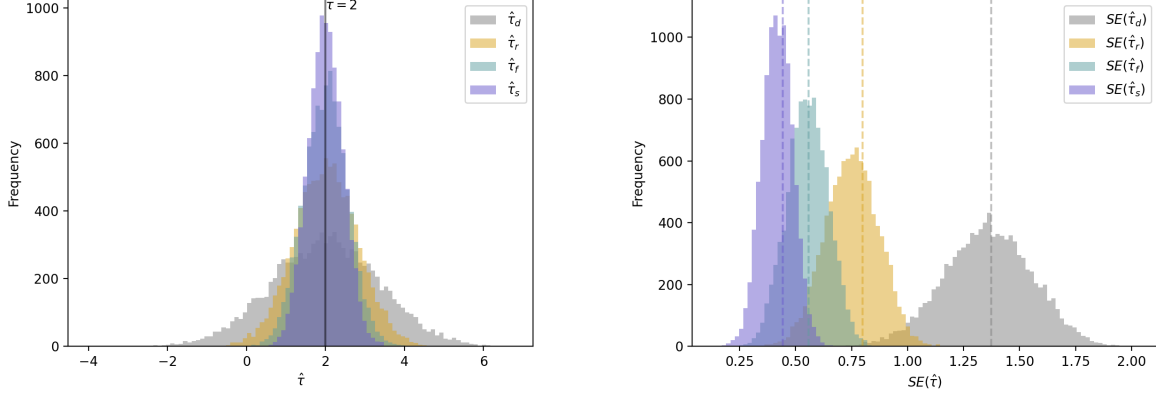


(b) Counterfactual estimates of outcomes. The dotted lines represent the mean outcome of each group. $\hat{\tau}_d = 3.86$, $SE[\hat{\tau}_d] = 1.32$.

Figure 2.1: An instance of simulated experiment data and counterfactual estimates using the constant model in Equation (2.2).

We acknowledge that the sample sizes used in the simulation ($\{10, 50, 100\}$) are very small by the standards of A/B tests, where it is common to have tens of thousands, hundreds of thousands, and even millions of users. We justify using such small samples as follows. First, given the variances underlying the data generation process, reasonably small samples are able to distinguish signal from noise. Moreover, the benefit of the variance reduction methods reviewed and proposed in this thesis is evident despite the small samples. Additionally, this toy scenario has just one covariate and as such is not overly complex. Relative to a more complicated high-dimensional covariate space, we do not require a large sample size to see the benefit of these methods. While we could have used larger sample sizes in these simulations, this would have imposed greater computational costs without providing additional insight into the efficacy of the methods. Thus, the sample sizes explored are sufficient for our purpose.

Figure 2.1 shows an instance of the simulated experiment data with a sample size of $n_1 = n_0 = 10$. For these data, the estimate $\hat{\tau}_d = 3.86$ overestimates the true treatment effect $\tau = 2$, with a large standard error of $SE[\hat{\tau}_d] = 1.32$. The estimate and standard error are poor since we are using a small sample size (which was chosen for clear visualization). However, this result is for a single instance of simulated experiment data. Repeating this process many times gives many treatment effect estimates and hence a distribution for the



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

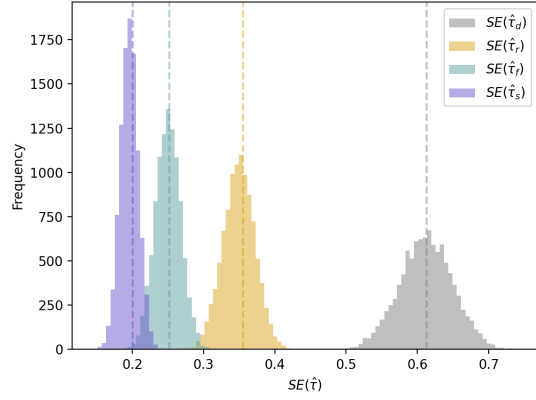
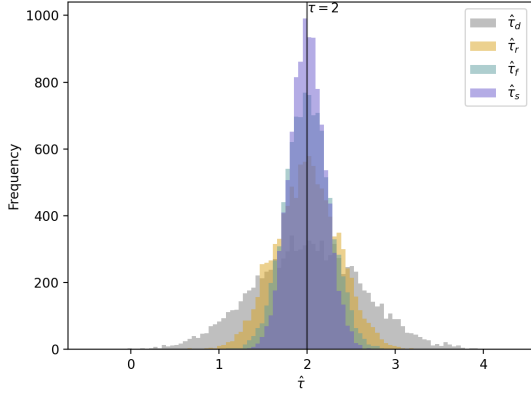
Figure 2.2: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 2.2b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.2a.

estimator. We can infer bias and variance by the location and the dispersion of these distributions. The simulated experiment is run 10,000 times to form the empirical distribution of $\hat{\tau}_d$ and $SE[\hat{\tau}_d]$ shown in Figures 2.2, 2.3, and 2.4 as grey histograms. The histograms for $\tilde{\tau}_r, \tilde{\tau}_f$, and $\tilde{\tau}_s$ will be introduced in subsequent sections, but their simulation results are combined to streamline the comparison of bias and variance between the estimators. Figures 2.2, 2.3, and 2.4 show the results for sample sizes 10, 50, and 100 respectively. The distributions of $\hat{\tau}_d$ in Figures 2.2a, 2.3a, and 2.4a are centered around the true treatment effect $\tau = 2$ at all sample sizes, verifying its unbiasedness. The distributions of standard error are also well centered around the standard deviation of the 10,000 treatment effect estimates represented by the vertical lines in Figures 2.2b, 2.3b, and 2.4b.

2.2 Counterfactual Estimation with Linear Regression

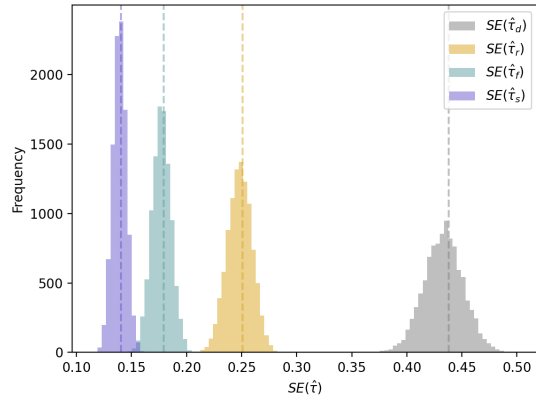
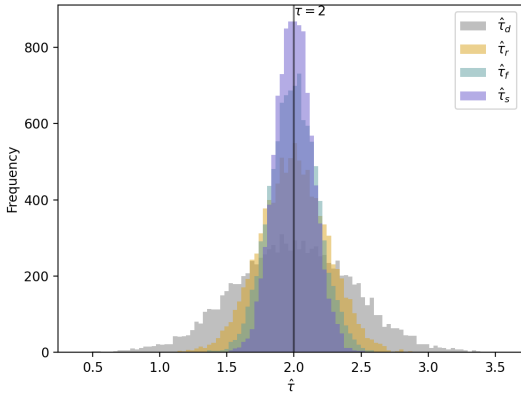
2.2.1 The Method

We move on to a more flexible model involving covariates that help reduce variance relative to the constant model introduced in Section 2.1. This method is a standard regression



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 2.3: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 2.3b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.3a.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 2.4: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 2.4b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.4a.

method for completely randomized experiments [17]. The regression method follows the same assumptions introduced in Section 2.1 but we now take into account a vector of covariates $\mathbf{x}$ which are also assumed to be independently and identically distributed from some super-population distribution. We define additional notation for the mean and variance of potential outcomes in the super-population conditional on covariates. We also define the mean vector and covariance matrix of the covariates. (2.10).

$$\begin{aligned}\alpha_W(\mathbf{x}) &= \mathbb{E}[Y_i(W)|\mathbf{X}_i = \mathbf{x}], \quad \sigma_W^2(\mathbf{x}) = \mathbb{V}[Y_i(W)|\mathbf{X}_i = \mathbf{x}], \\ \boldsymbol{\mu}_X &= \mathbb{E}[\mathbf{X}_i], \quad \Omega_X = \mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu}_X)^T(\mathbf{X}_i - \boldsymbol{\mu}_X)].\end{aligned}\tag{2.10}$$

The potential outcome is modeled as a linear function of covariates $\mathbf{x}_i$, the treatment indicator W_i , and a random error term (2.11).

$$Y_i(W_i) = \alpha_1 W_i + \alpha_0(1 - W_i) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i.\tag{2.11}$$

Here, $\mathbf{x}_i$ is an $m \times 1$ vector of covariates, $\boldsymbol{\beta}$ is an $m \times 1$ vector of coefficients, and $\epsilon_i \sim N(0, \sigma_{rW_i}^2)$ are independently distributed error terms with finite variances. The variance of error terms in this new model has subscript r to differentiate from that of the constant model, $\sigma_{dW_i}^2$, in Section 2.1. These variances are again assumed to be constant within each group for all $\mathbf{x}_i$: $\sigma_W^2(\mathbf{x}_i) = \sigma_{rW_i}^2$. Typically, more variability in the outcome will be explained by the addition of covariates $\mathbf{x}_i$ to the model, hence we expect $\sigma_{rW_i}^2 < \sigma_{dW_i}^2$. As is typical in regression modeling, we are interested in conditional expectations and variances of the outcome, treating W_i and $\mathbf{x}_i$ as fixed. The observed outcome is generated by randomly assigning each unit to treatment or control with equal probability, hence the treatment indicator is asymptotically uncorrelated with the covariates.

We again define the estimators of the model coefficients with ordinary least squares (OLS).

$$\begin{aligned}(\tilde{\alpha}_1, \tilde{\alpha}_0, \tilde{\boldsymbol{\beta}}) &:= \arg \min_{(\alpha_1, \alpha_0, \boldsymbol{\beta}) \in \mathbb{R}^{m+2}} \rho(\alpha_1, \alpha_0, \boldsymbol{\beta}) \\ &= \arg \min_{(\alpha_1, \alpha_0, \boldsymbol{\beta}) \in \mathbb{R}^{m+2}} \sum_{i=1}^n (Y_i^{obs} - (\alpha_{W_i} + \mathbf{x}_i^T \boldsymbol{\beta}))^2\end{aligned}\tag{2.12}$$

This yields group-specific estimates of α_1 and α_0 , and an estimate of $\boldsymbol{\beta}$ that is pooled across both groups. This is consistent with the constant treatment effect assumption. The

least squares estimators of α_1, α_0 are as follows:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha_1} \rho(\alpha_1, \alpha_0, \boldsymbol{\beta}), \\ \rightarrow \tilde{\alpha}_1 &= \frac{1}{n_1} \sum_{i:W_i=1} (Y_i^{obs} - \mathbf{x}_i^T \boldsymbol{\beta}) \\ &= \bar{Y}_1 - \bar{\mathbf{x}}_1^T \boldsymbol{\beta}, \end{aligned} \quad (2.13)$$

where $\bar{\mathbf{x}}_1$ is an $m \times 1$ vector with j^{th} entry the mean of covariate $j = 1, 2, \dots, m$ for treatment units. Similarly,

$$\tilde{\alpha}_0 = \bar{Y}_0 - \bar{\mathbf{x}}_0^T \boldsymbol{\beta}, \quad (2.14)$$

where $\bar{\mathbf{x}}_0$ is an $m \times 1$ vector with j^{th} entry the mean of covariate $j = 1, \dots, m$ for control units. The estimator of $\boldsymbol{\beta}$ is found by substituting the estimators from Equations (2.13) and (2.14) into Equation (2.12) and taking the derivative with respect to $\boldsymbol{\beta}$. We first substitute and then simplify:

$$\begin{aligned} \rho(\tilde{\alpha}_1, \tilde{\alpha}_0, \boldsymbol{\beta}) &= \sum_{i:W_i=1} (Y_i^{obs} - (\tilde{\alpha}_1 + \mathbf{x}_i^T \boldsymbol{\beta}))^2 + \sum_{i:W_i=0} (Y_i^{obs} - (\tilde{\alpha}_0 + \mathbf{x}_i^T \boldsymbol{\beta}))^2 \\ &= \sum_{i:W_i=1} (Y_i^{obs} - (\bar{Y}_1 - \bar{\mathbf{x}}_1^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\beta}))^2 + \sum_{i:W_i=0} (Y_i^{obs} - (\bar{Y}_0 - \bar{\mathbf{x}}_0^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\beta}))^2 \\ &= \sum_{i:W_i=1} (Y_i^c - (\mathbf{x}_i^c)^T \boldsymbol{\beta})^2 + \sum_{i:W_i=0} (Y_i^c - (\mathbf{x}_i^c)^T \boldsymbol{\beta})^2 \end{aligned} \quad (2.15)$$

where $Y_i^c = Y_i^{obs} - \bar{Y}_{W_i}$ and $\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}_{W_i}$ are centered outcome and covariate vectors for unit i . Equation (2.15) can be simplified using vector notation as follows.

$$\|\mathbf{Y}_1 - \mathbf{x}_1 \boldsymbol{\beta}\|^2 + \|\mathbf{Y}_0 - \mathbf{x}_0 \boldsymbol{\beta}\|^2, \quad (2.16)$$

where for treatment status W , $\mathbf{Y}_W$ is the $n_W \times 1$ vector of mean-centered outcomes Y_i^c , and $\mathbf{x}_W$ is the $n_W \times m$ matrix of mean-centered covariates with each row i equal to $(\mathbf{x}_i^c)^T$. Next, we take the derivative of Equation (2.16) with respect to the coefficients $\boldsymbol{\beta}$, and solve the system of equations arising by equating the sum to the zero vector $\mathbf{0}$.

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} [\|\mathbf{Y}_1 - \mathbf{x}_1 \boldsymbol{\beta}\|^2 + \|\mathbf{Y}_0 - \mathbf{x}_0 \boldsymbol{\beta}\|^2] \\ &= 2 [\mathbf{x}_1^T \mathbf{x}_1 \boldsymbol{\beta} - \mathbf{x}_1^T \mathbf{Y}_1] + 2 [\mathbf{x}_0^T \mathbf{x}_0 \boldsymbol{\beta} - \mathbf{x}_0^T \mathbf{Y}_0] \\ \rightarrow \tilde{\boldsymbol{\beta}} &= [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} [\mathbf{x}_1^T \mathbf{Y}_1 + \mathbf{x}_0^T \mathbf{Y}_0] \end{aligned} \quad (2.17)$$

An important special case of this is CUPED (introduced in Section 1.2). When there is just one covariate x and it corresponds to pre-experiment observations of the outcome variable Y , $\tilde{\boldsymbol{\beta}}$ here reduces to the optimal θ in CUPED. But we present the more general and multivariate case here.

The OLS coefficient estimators are unbiased as demonstrated below.

$$\begin{aligned}
\mathbb{E} [\tilde{\boldsymbol{\beta}}] &= \mathbb{E} \left[[\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} [\mathbf{x}_1^T \mathbf{Y}_1 + \mathbf{x}_0^T \mathbf{Y}_0] \right] \\
&= [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} \mathbb{E} [\mathbf{x}_1^T \mathbf{Y}_1 + \mathbf{x}_0^T \mathbf{Y}_0] \\
&= [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} \mathbb{E} [\mathbf{x}_1^T (\mathbf{x}_1 \boldsymbol{\beta} + \boldsymbol{\epsilon}_1) + \mathbf{x}_0^T (\mathbf{x}_0 \boldsymbol{\beta} + \boldsymbol{\epsilon}_0)] \\
&= [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0] \boldsymbol{\beta} \\
&= \boldsymbol{\beta} \\
\mathbb{E} [\tilde{\alpha}_W] &= \mathbb{E} [\bar{Y}_W - \bar{\mathbf{x}}_W^T \tilde{\boldsymbol{\beta}}] \\
&= \mathbb{E} [\alpha_W + \bar{\mathbf{x}}_W^T \boldsymbol{\beta} + \bar{\epsilon}_W] - \bar{\mathbf{x}}_W^T \boldsymbol{\beta} \\
&= \alpha_W
\end{aligned}$$

Here, $\boldsymbol{\epsilon}_W$ is a $n_W \times 1$ vector of error terms for group W . The potential outcomes estimators are constructed using the following estimated linear regression model:

$$\tilde{Y}_i(W_i) = \tilde{\alpha}_1 W_i + \tilde{\alpha}_0 (1 - W_i) + \mathbf{x}_i \tilde{\boldsymbol{\beta}}. \quad (2.18)$$

The average treatment effect estimator using regression is $\tilde{\tau}_r$:

$$\begin{aligned}
\tilde{\tau}_r &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{Y}_i(0)) + \sum_{i:W_i=0} (\tilde{Y}_i(1) - Y_i^{obs}) \right] \\
&= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{\alpha}_0 - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}) + \sum_{i:W_i=0} (\tilde{\alpha}_1 + \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} - Y_i^{obs}) \right] \\
&= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}) - n_1 \tilde{\alpha}_0 + \sum_{i:W_i=0} (\mathbf{x}_i^T \tilde{\boldsymbol{\beta}} - Y_i^{obs}) + n_0 \tilde{\alpha}_1 \right] \\
&= \frac{1}{n} [n_1 \tilde{\alpha}_1 - n_1 \tilde{\alpha}_0 - n_0 \tilde{\alpha}_0 + n_0 \tilde{\alpha}_1] \\
&= \tilde{\alpha}_1 - \tilde{\alpha}_0
\end{aligned} \quad (2.19)$$

The above derivation utilizes the result from Equations (2.13) and (2.14) to simplify the summations. Of interest is a comparison of $\tilde{\tau}_r$ relative to the classic difference in means

estimator $\tilde{\tau}_d = \bar{Y}_1 - \bar{Y}_0$. Setting $\beta = \mathbf{0}$ in Equation (2.11) yields $\tilde{\alpha}_W = \bar{Y}_W$. Therefore, $\tilde{\tau}_d$ is a special case of $\tilde{\tau}_r$ where the potential outcome for each group is modeled with a constant instead of a linear function of covariates. We first compare these estimators with respect to bias. We will concern ourselves with finite samples conditional on the matrix of observed covariates $\mathbf{x}^{obs}$, which consists of rows corresponding to treatment units ($\mathbf{x}_1^{obs}$) and rows corresponding to control units ($\mathbf{x}_0^{obs}$).

$$\begin{aligned}
\mathbb{E} [\tilde{\tau}_r | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{E} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{\mathbf{x}}_1^T \tilde{\beta} - \bar{Y}_0 + \bar{\mathbf{x}}_0^T \tilde{\beta} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] - (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{E} [\tilde{\beta} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \alpha_1(\mathbf{x}_1^{obs}) - \alpha_0(\mathbf{x}_0^{obs}) - (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \beta
\end{aligned} \tag{2.20}$$

where $\alpha_W(\mathbf{x}_W^{obs})$ is the conditional expectation of the outcome in group W given observed covariates $\mathbf{x}_W^{obs}$ as defined in Equation (2.10). We showed in Section 2.1.1 that $\tilde{\tau}_d$ is an unbiased estimator of the treatment effect $\tau = \alpha_1 - \alpha_0$, but we see here that $\tilde{\tau}_r$ is not. As noted in [17], we cannot guarantee unbiasedness with a regression-based estimator of the treatment effect in finite samples. Only when both treatment and control groups are identical in covariates will $\tilde{\tau}_r$ be unbiased, and this is difficult to achieve in small samples. However, as sample size increases, because treatment assignment is made independent of covariates, we expect the covariate distribution in both groups to approach that of the super-population. That is, $\bar{\mathbf{x}}_1^{obs}, \bar{\mathbf{x}}_0^{obs} \rightarrow \mu_{\mathbf{x}}$ and hence $\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs} \rightarrow \mathbf{0}$. Similarly, since the distribution of $\mathbf{x}_W^{obs}$ approaches the covariate distribution of the super-population, $\alpha_W(\mathbf{x}_W^{obs}) \rightarrow \alpha_W$ and hence $\alpha_1(\mathbf{x}_1^{obs}) - \alpha_0(\mathbf{x}_0^{obs}) \rightarrow \alpha_1 - \alpha_0$. Therefore, $\tilde{\tau}_r$ is an asymptotically unbiased estimator of the treatment effect $\alpha_1 - \alpha_0$. This asymptotic unbiasedness doesn't require the regression model to be correctly specified with respect to the covariates. Regardless of whether the potential outcomes truly are linear functions of the covariates, the treatment effect estimator will be unbiased in large samples as long as the treatment assignments are made randomly and hence independent of covariates.

Next, we compare the two estimators $\tilde{\tau}_d$ and $\tilde{\tau}_r$ with respect to their variances. First,

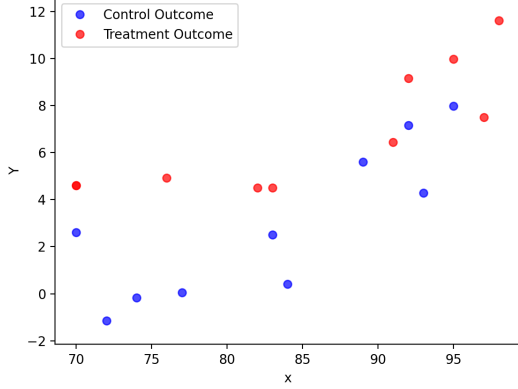
we derive the variance for $\tilde{\tau}_r$.

$$\begin{aligned}
\mathbb{V}[\tilde{\tau}_r|\mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{V}[\tilde{\alpha}_1 - \tilde{\alpha}_0|\mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{V}\left[\bar{Y}_1 - \bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}} - \bar{Y}_0 + \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}}|\mathbf{x} = \mathbf{x}^{obs}\right] \\
&= \mathbb{V}[\bar{Y}_1 - \bar{Y}_0|\mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V}\left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \tilde{\boldsymbol{\beta}}|\mathbf{x} = \mathbf{x}^{obs}\right] \\
&= \frac{\sigma_{r_1}^2}{n_1} + \frac{\sigma_{r_0}^2}{n_0} + (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V}[\tilde{\boldsymbol{\beta}}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs}).
\end{aligned} \tag{2.21}$$

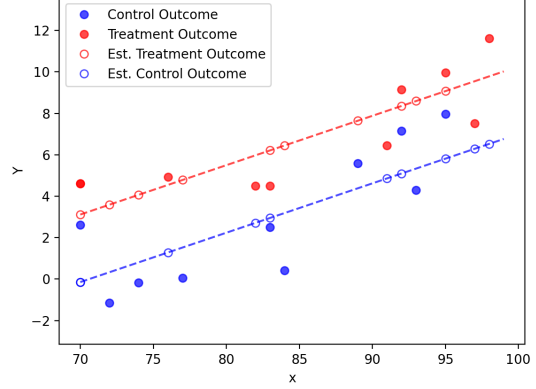
To see where this comes from, note that with the potential outcomes model in Equation (2.11), $\sigma_{r_{W_i}}^2$ is the variance of the error term ϵ_i and hence $Y_i(W_i)$ given covariates $\mathbf{x}_i$. Moreover, the covariance between $\bar{Y}_1 - \bar{Y}_0|\mathbf{x} = \mathbf{x}^{obs}$ and $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \tilde{\boldsymbol{\beta}}|\mathbf{x} = \mathbf{x}^{obs}$ is 0, which is why the result in Equation (2.21) omits a covariance term. This covariance is 0 because the estimator $\tilde{\boldsymbol{\beta}}$ is computed with mean-centered outcomes $\mathbf{Y}_W$ as specified in Equation (2.17). Therefore, $\tilde{\boldsymbol{\beta}}$ is independent of the mean outcomes, $\bar{Y}_1$ and $\bar{Y}_0$. A detailed proof of this can be found in Appendix A.1. As the sample size gets larger, the distribution of the covariates in each group approaches that of the super-population. This implies that $(\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V}[\tilde{\boldsymbol{\beta}}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})$ is approximately 0 and hence ignorable as the sample size increases. In this limiting case where $\bar{\mathbf{x}}_1^{obs} \approx \bar{\mathbf{x}}_0^{obs}$, the variability of the model coefficients $\mathbb{V}[\tilde{\boldsymbol{\beta}}]$ does not influence the variance of the treatment effect estimator. This will be evident in simulations in Section 2.2.2, where the approximate variance $\mathbb{V}[\tilde{\tau}_r|\mathbf{x} = \mathbf{x}^{obs}] \approx \sigma_{r_1}^2/n_1 + \sigma_{r_0}^2/n_0$ is well-aligned with the true variance of treatment effect estimates for even moderately sized n_1 and n_0 . Therefore, with large enough n_1 and n_0 , the task of minimizing $\mathbb{V}[\tilde{\tau}_r|\mathbf{x} = \mathbf{x}^{obs}]$ is achieved by minimizing the conditional variances $\sigma_{r_1}^2$ and $\sigma_{r_0}^2$ which can be achieved by increasing the flexibility of the model used to describe the relationship between the covariates and the outcome. As mentioned before, more variability in the outcome can be explained by the addition of covariates $\mathbf{x}_i$ to the model. This reduces the unexplained variability of the outcome or the variance of the error terms, hence $\sigma_{r_{W_i}}^2 \leq \sigma_{d_{W_i}}^2$. Then, the variance of $\mathbb{V}[\tilde{\tau}_r] \approx \sigma_{r_1}^2/n_1 + \sigma_{r_0}^2/n_0$ is asymptotically at most $\mathbb{V}[\tilde{\tau}_d] = \sigma_{d_1}^2/n_1 + \sigma_{d_0}^2/n_0$. This reduction in variance becomes even greater if the covariate is highly correlated with the outcome, hence further reducing $\sigma_{r_{W_i}}^2$.

2.2.2 Demonstration of Variance Reduction

The same simulation as Section 2.1.2 is utilized with uniformly distributed covariates $\{x \in \mathbb{Z}|70 \leq x < 100\}$ to generate the experiment data via the model $Y_i(W_i) = \tau W_i + 2\sin(2x_i) +$



(a) Experiment outcomes ordered by the value of covariate x .



(b) Linear model counterfactual imputation.
 $\hat{\tau}_r = 3.26$, $SE[\hat{\tau}_r] = 0.77$.

Figure 2.5: Simulated experiment data ordered by covariate x . The dotted lines in 2.5b represent the fitted linear model from Equation (2.11). The open circles in 2.5b represent the estimated counterfactual outcome for each observed outcome (filled circles).

$0.01(x_i - 70)^2 + 0.4\epsilon_i, \epsilon_i \sim N(0, 1)$. However, the covariate x is now used to model the outcome Y following Equation (2.11). For each sample, the treatment effect estimate is computed as $\hat{\tau}_r = \hat{\alpha}_1 - \hat{\alpha}_0$ and the variance estimate is computed as $\hat{V}[\tilde{\tau}_r] = \hat{\sigma}_{r_1}^2/n_1 + \hat{\sigma}_{r_0}^2/n_0$, where $\hat{\sigma}_{r_w}^2$ is computed by taking the residual variance in group W of the model. Figure 2.5 shows the same instance of the simulated experiment data as Figure 2.1 but plotted against x . Here, we observe a clear relationship between the outcome Y and the covariate x . This relationship can be exploited for purposes of variance reduction by fitting the linear model for Y as described in Equation (2.11). This model is then used to estimate counterfactuals. Similar to the result for $\hat{\tau}_d$ in Figure 2.1b, the estimate $\hat{\tau}_r = 3.26$ overestimates the true treatment effect. However, the standard error of $\hat{\tau}_r$ (0.77) is about half that of $\hat{\tau}_d$ (1.32) because the counterfactual estimates that compose $\hat{\tau}_r$ account for the relationship between Y and x whereas $\hat{\tau}_d$ does not.

The empirical distribution of 10,000 $\hat{\tau}_r$ and $SE[\hat{\tau}_r] = \sqrt{\hat{V}[\tilde{\tau}_r]}$ values from repeated simulations are represented by the yellow histograms in Figures 2.2, 2.3, and 2.4. The variance reduction of $\tilde{\tau}_r$ relative to $\tilde{\tau}_d$ is calculated as $1 - \mathbb{V}[\tilde{\tau}_r]/\mathbb{V}[\tilde{\tau}_d]$ where $\mathbb{V}[\tilde{\tau}_r]$ and $\mathbb{V}[\tilde{\tau}_d]$ are calculated as the variance of the respective 10,000 treatment effect estimates visualized in Figures 2.2a, 2.3a, and 2.4a. Note that $\mathbb{V}[\tilde{\tau}_d]$ is different from $\hat{V}[\tilde{\tau}_d]$. The former represents the true variance in treatment effect estimates $\hat{\tau}_d$ observed over 10,000

different experiments, whereas the latter represents an estimate of the former given a particular sample. The same idea extends to $\tilde{\tau}_r$ as well as $\tilde{\tau}_f$ and $\tilde{\tau}_s$ which will be discussed in Sections 2.3 and 2.4. Table 2.1 summarizes the variance reduction of each estimator at different sample sizes.

Sample Size	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$
10	66.08%	83.48%	89.59%
50	66.41%	83.12%	89.29%
100	67.28%	83.27%	89.71%

Table 2.1: Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$

Due to the strong relationship between the outcome Y and the covariate x , we observe variance reductions of over 66% across all sample sizes. When the sample size is small (Figure 2.2b), we see that the standard error of $\tilde{\tau}_r$ tends to underestimate the true standard deviation of treatment effects represented by the vertical line. From Equation (2.21), we see this is because the variance component of the treatment effect estimator related to the model, $(\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V}[\tilde{\beta}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})$, is not approximately 0 when sample sizes are small because the covariate distributions tend to be different from each other and hence $\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs} \not\rightarrow 0$. However, as the sample size increases, we see the distribution of standard errors of $\hat{\tau}_r$ center more closely to the true standard deviation of treatment effects $\sqrt{\mathbb{V}[\tilde{\tau}_r]}$ represented by the vertical line because $\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs} \rightarrow 0$ and so $(\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V}[\tilde{\beta}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs}) \rightarrow 0$. In Section 2.3, we introduce a more general approach where the linear model is replaced by some function of covariate $f(\mathbf{x})$. Increasing the flexibility of $f(\mathbf{x})$ can result in greater variance reduction while maintaining the asymptotic unbiasedness in the resulting treatment effect estimator.

2.3 Counterfactual Estimation with Machine Learning

2.3.1 The Method

From Equation (2.21), it was shown that reducing $\mathbb{V}[\tilde{\tau}_r | \mathbf{x} = \mathbf{x}^{obs}]$ can be achieved by reducing the conditional variances $\sigma_1^2(\mathbf{x}_i) = \sigma_{\tau_1}^2$ and $\sigma_0^2(\mathbf{x}_i) = \sigma_{\tau_0}^2$, provided that the sample

sizes of each group are large enough to ensure the covariates in each group are evenly distributed. This conditional variance can further be reduced by using a model more complex than a linear model to describe the relationship between the outcome and the covariates. Even though model misspecification does not impact the unbiasedness of the estimators, it does impact their variances by potentially increasing the conditional variance. The potential outcomes may be modeled with flexible machine learning (ML) models represented by some generic function of covariates, $f(\mathbf{x}_i)$:

$$Y_i(W_i) = \alpha_1 W_i + \alpha_0(1 - W_i) + f(\mathbf{x}_i) + \epsilon_i \quad (2.22)$$

where $\epsilon_i \sim N(0, \sigma_{f_{W_i}}^2)$ and hence $\mathbb{V}[Y_i(W_i)|\mathbf{x}_i = \mathbf{x}] = \sigma_{f_{W_i}}^2$. The least squares estimators of α_1, α_0 are analogous to those in the linear model (2.11):

$$\begin{aligned} \tilde{\alpha}_1 &= \bar{Y}_1 - \bar{f}(\mathbf{x}_1), \\ \tilde{\alpha}_0 &= \bar{Y}_0 - \bar{f}(\mathbf{x}_0), \end{aligned} \quad (2.23)$$

where $\bar{f}(\mathbf{x}_W) = \frac{1}{n_W} \sum_{i:W_i=W} f(\mathbf{x}_i)$. Suppose we estimate the function $f(\mathbf{x})$ with $\tilde{f}(\mathbf{x})$ using a different dataset independent from the experiment data, and this estimate is used in place of $f(\mathbf{x})$, thus redefining $\bar{f}(\mathbf{x}_W) = \frac{1}{n_W} \sum_{i:W_i=W} \tilde{f}(\mathbf{x}_i)$ in Equation (2.23). In this case, the estimation of $\tilde{\alpha}_1$ and $\tilde{\alpha}_0$ is impacted by $\tilde{f}(\mathbf{x}_i)$, but not the other way around. Then, the potential outcomes estimators are as follows:

$$\tilde{Y}_i(W_i) = \tilde{\alpha}_1 W_i + \tilde{\alpha}_0(1 - W_i) + \tilde{f}(\mathbf{x}_i). \quad (2.24)$$

Using this potential outcome model, we impute the counterfactual for every unit and define $\tilde{\tau}_f$ similar to $\tilde{\tau}_r$ in Equation (2.19), but with a generic model f instead of a linear model:

$$\begin{aligned} \tilde{\tau}_f &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{Y}_i(0)) + \sum_{i:W_i=0} (\tilde{Y}_i(1) - Y_i^{obs}) \right] \\ &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{\alpha}_0 - \tilde{f}(\mathbf{x}_i)) + \sum_{i:W_i=0} (\tilde{\alpha}_1 + \tilde{f}(\mathbf{x}_i) - Y_i^{obs}) \right] \\ &= \frac{1}{n} \left[\sum_{i:W_i=1} (Y_i^{obs} - \tilde{f}(\mathbf{x}_i)) - n_1 \tilde{\alpha}_0 + \sum_{i:W_i=0} (\tilde{f}(\mathbf{x}_i) - Y_i^{obs}) + n_0 \tilde{\alpha}_1 \right] \\ &= \tilde{\alpha}_1 - \tilde{\alpha}_0 \end{aligned} \quad (2.25)$$

The summations simplify to $\tilde{\alpha}_1 - \tilde{\alpha}_0$ using Equation (2.23). The bias and variance of $\tilde{\tau}_f$ are similar to $\tilde{\tau}_r$, as shown below.

$$\begin{aligned}
\mathbb{E} [\tilde{\tau}_f | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{E} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] - \mathbb{E} [\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \alpha_1(\mathbf{x}_1^{obs}) - \alpha_0(\mathbf{x}_0^{obs}) - \mathbb{E} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})]
\end{aligned} \tag{2.26}$$

$$\begin{aligned}
\mathbb{V} [\tilde{\tau}_f | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{V} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{V} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{V} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] + C \\
&= \frac{\sigma_{f_1}^2}{n_1} + \frac{\sigma_{f_0}^2}{n_0} + \mathbb{V} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] + C
\end{aligned} \tag{2.27}$$

We observe that $\tilde{\tau}_f$ is not unbiased in Equation (2.26). However, if we examine the definitions of $\bar{f}(\mathbf{x}_1)$ and $\bar{f}(\mathbf{x}_0)$, we see they are both averages of the same function $\tilde{f}(\mathbf{x}_i)$, averaged over the covariates in the treatment and control groups respectively. Therefore, as sample size increases and the distribution of the covariates in both groups approach that of the super-population, then $\mathbb{E} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})]$ approaches 0. Therefore, $\tilde{\tau}_f$ (like $\tilde{\tau}_r$) is an asymptotically unbiased estimator of τ . However, even in finite samples, the slight bias of the treatment effect estimator is often compensated by a large reduction in variance if covariates have a strong relationship with the outcome. Because we know that the covariate distributions of the two groups approach that of the super-population as we increase sample size, we can conclude that asymptotically $\mathbb{V} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] \rightarrow 0$ in Equation (2.27).

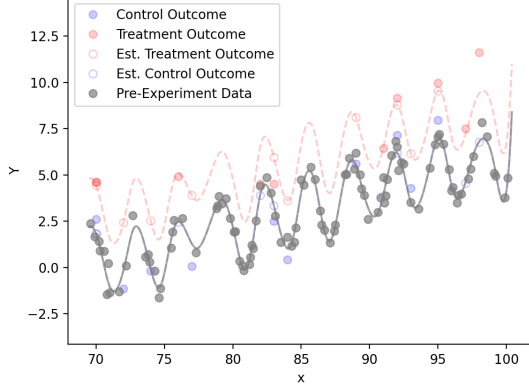
In Equation (2.27), C is equal to $Cov [\bar{Y}_1 - \bar{Y}_0, \bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}]$. There are several circumstances in which these terms are independent and hence $C = 0$. Previously, we had assumed that $\tilde{f}(\mathbf{x})$ is estimated using a different dataset independent of the experiment data. This could be a separate sample from the super-population that does not experience any treatment. Using an independent sample to train $\tilde{f}(\mathbf{x})$ ensures that its parameter estimates are independent of the experiment data. Therefore, conditional on the covariates $\mathbf{x}_i$, the outcome $Y_i(W_i)$ and its prediction $\tilde{f}(\mathbf{x}_i)$ are independent and thus $C = 0$. Because we cannot always take additional samples in experiments, it seems appropriate to utilize an existing sample of data that is large in size, purely observational, and hence does not experience any treatment. Especially in online settings where the outcome of interest may be a user activity metric that is measured and logged, utilizing a

large amount of pre-experiment data to fit the model $\tilde{f}(\mathbf{x})$ is attractive. Due to a large amount of pre-experiment data available in online settings, it is feasible to fit a complex ML model for $\tilde{f}(\mathbf{x})$ without significant overfitting. This method of using an ML model fit to pre-experiment data to reduce conditional variance corresponds to the “Variance Reduction by Subtraction of Prediction” method due to Poyarkov et al. [8] (see Section 1.2), where a boosted decision tree trained on pre-experiment data is taken as $\tilde{f}(\mathbf{x})$ and its predictions are used to reduce variance in treatment effect estimation. Instead of using an independent sample to fit $\tilde{f}(\mathbf{x})$, we can also achieve $C = 0$ by using mean-centered outcomes. If centered outcomes $Y_{iW} = Y_i^{obs} - \bar{Y}_W$ are used to fit the model $\tilde{f}(\mathbf{x})$ instead of observed outcomes Y_i^{obs} , the parameters of $\tilde{f}(\mathbf{x})$ are indeed independent of the group means $\bar{Y}_W$ and hence $C = 0$. The linear model introduced in Section 2.2 is an example of this case, where the coefficient vector β is estimated using centered outcomes, and its proof is given in Appendix A.1. However, given the limited size of in-experiment data, it may be difficult to fit a highly flexible model without overfitting to the particular sample.

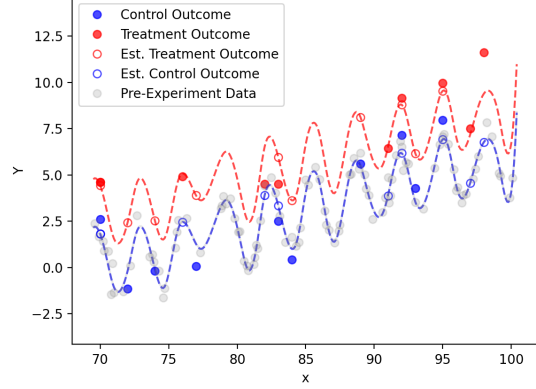
2.3.2 Demonstration of Variance Reduction

We perform the same simulation as in Section 2.2.2, but this time we choose a non-linear model and fit it to the pre-experiment data. A cubic spline with a smoothing parameter of 15 was chosen to capture the fluctuations produced by the true data-generating function specified in Section 2.2.2. In Figure 2.6a, using a flexible ML model (cubic splines) with pre-experiment data allows us to identify and exploit the non-linear, cyclical pattern that cannot be captured by a linear model fit to in-experiment data. The model is then used to generate counterfactual estimates of each observation in Figure 2.6b, and hence the resulting treatment effect estimate of $\hat{\tau}_f = 2.62$ and standard error of $SE[\hat{\tau}_f] = 0.45$. As in Sections 2.1.2 and 2.2.2, the standard error is calculated using the standard deviation of residuals: $SE[\hat{\tau}_f] = \sqrt{\widehat{V}[\tilde{\tau}_f]} = \sqrt{\hat{\sigma}_{f_1}^2/n_1 + \hat{\sigma}_{f_0}^2/n_0}$. Notice that $SE[\hat{\tau}_f] = 0.45 < 0.77 = SE[\hat{\tau}_r]$ from Figure 2.5b. Since the counterfactual estimates now reflect additional complexity in the relationship between the outcome and the covariate, we expect the variance associated with the treatment effect estimates to be smaller than that of a linear model. We perform this simulation 10,000 times and obtain a distribution of treatment effect estimates and their standard errors with sample sizes $n_1 = n_0 \in \{10, 50, 100\}$.

Figures 2.2, 2.3, and 2.4 show results for $\hat{\tau}_f$ in teal histograms. The flexible ML-based estimator, $\tilde{\tau}_f$, is asymptotically unbiased for the true treatment effect $\tau = 2$, with greater variance reduction over all sample sizes relative to $\tilde{\tau}_d$ and $\tilde{\tau}_r$. The variance reduction achieved by using $\tilde{\tau}_f$ instead of $\tilde{\tau}_d$ is roughly 83% across the three sample sizes (Table



(a) A cubic spline model fit to pre-experiment data with smoothing parameter of 15.



(b) Estimating counterfactuals with model predictions. $\hat{\tau}_f = 2.62$, $SE[\hat{\tau}_f] = 0.45$.

Figure 2.6: Visualization of pre- and in-experiment data. The grey line in Figure 2.6a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.6b represent treatment and control outcome estimates over covariate x .

2.1), which is greater than that achieved by using $\tilde{\tau}_r$ in Section 2.2.2. The distribution of standard errors in Figures 2.2b, 2.3b, and 2.4b also appear to be well-centered around the observed standard deviation $\sqrt{\mathbb{V}[\tilde{\tau}_f]}$, which verifies our argument concerning the negligibility of the covariance term C and the model variance term $\mathbb{V}[\tilde{f}(\mathbf{x}_1) - \tilde{f}(\mathbf{x}_0)]$ from Equation (2.27).

2.4 Counterfactual Estimation with Machine Learning and Secondary Linear Adjustment

2.4.1 The Method

High-quality predictions are necessary for variance reduction using a model-based approach (such as those described in Sections 2.2 and 2.3); the performance depends on the alignment of the predictions and the observed outcomes. Since the variance of $\tilde{\tau}_f$ is an increasing function of the variances of the error terms $\sigma_{f_w}^2$, poor predictions of the in-experiment outcomes from $\tilde{f}$ will hamper variance reduction. Hence, for the method in Section 2.3 to be

effective, it requires the in-experiment data to be highly correlated with the pre-experiment data so that the predictions of the in-experiment data based on the pre-experiment data are accurate. However, pre-experiment and in-experiment data may differ systematically. This is a large obstacle in the online experiment setting, where the observed outcomes are heavily dependent on the measurement time as demonstrated by [11, 12, 15, 16] and discussed in Section 1.3. Therefore, $\tilde{\tau}_f$ may not provide variance reduction if the pre-experiment and in-experiment data are not correlated. To maintain low conditional variance even when pre-experiment and in-experiment data are poorly aligned, we propose combining the linear model for potential outcomes with in-experiment data from Equation (2.11) and the flexible model for potential outcomes with pre-experiment data from Equation (2.22):

$$Y_i(W_i) = \tilde{f}(\mathbf{x}_i) + \alpha_1 W_i + \alpha_0(1 - W_i) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad (2.28)$$

with $\epsilon_i \sim N(0, \sigma_{s_{W_i}}^2)$ and hence $\mathbb{V}[Y_i(W_i)|\mathbf{x}_i = \mathbf{x}] = \sigma_{s_{W_i}}^2$ based on this new model of potential outcomes. By moving the prediction $\tilde{f}(\mathbf{x}_i)$ to the other side of the equation in (2.28), we can rewrite this model of the potential outcomes as a linear model of residual potential outcomes $R_i(W_i) = Y_i(W_i) - \tilde{f}(\mathbf{x}_i)$ based on pre-experiment data:

$$R_i(W_i) = Y_i(W_i) - \tilde{f}(\mathbf{x}_i) = \alpha_1 W_i + \alpha_0(1 - W_i) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i. \quad (2.29)$$

We assume $\tilde{f}(\mathbf{x}_i)$ is estimated using independent pre-experiment data while $\tilde{\alpha}_1, \tilde{\alpha}_0, \tilde{\boldsymbol{\beta}}$ are estimated using the in-experiment data with $R_i(W_i)$ in place of $Y_i(W_i)$. Consequently, the covariates $\mathbf{x}_i$ must be the same (though the units do not) in both pre-experiment and in-experiment data. We can then define the potential outcomes estimators as

$$\tilde{Y}_i(W_i) = \tilde{f}(\mathbf{x}_i) + \tilde{\alpha}_1 W_i + \tilde{\alpha}_0(1 - W_i) + \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}, \quad (2.30)$$

where

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= [\mathbf{x}_1^T \mathbf{x}_1 + \mathbf{x}_0^T \mathbf{x}_0]^{-1} [\mathbf{x}_1^T \mathbf{R}_1 + \mathbf{x}_0^T \mathbf{R}_0], \\ \tilde{\alpha}_1 &= \bar{R}_1 - \bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}}, \\ \tilde{\alpha}_0 &= \bar{R}_0 - \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}}. \end{aligned} \quad (2.31)$$

Similar to Equation (2.17), the covariate vector $\mathbf{x}_W$ and residual outcome vector $\mathbf{R}_W$ for estimating $\tilde{\boldsymbol{\beta}}$ are mean-centered. With this potential outcomes estimator, we can define a new treatment effect estimator $\tilde{\tau}_s$ that combines the predictive power of a flexible ML model $\tilde{f}(\mathbf{x}_i)$ and a secondary linear adjustment that accounts for possible differences between pre-

experiment and in-experiment data. The treatment effect estimator is:

$$\begin{aligned}
\tilde{\tau}_s &= \frac{1}{n} \left[\sum_{i:W_i=1} \left(Y_i^{obs} - \tilde{Y}_i(0) \right) + \sum_{i:W_i=0} \left(\tilde{Y}_i(1) - Y_i^{obs} \right) \right] \\
&= \frac{1}{n} \left[\sum_{i:W_i=1} \left(Y_i^{obs} - \tilde{f}(\mathbf{x}_i) - \tilde{\alpha}_0 - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} \right) + \sum_{i:W_i=0} \left(\tilde{f}(\mathbf{x}_i) + \tilde{\alpha}_1 + \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} - Y_i^{obs} \right) \right] \quad (2.32) \\
&= \frac{1}{n} \left[\sum_{i:W_i=1} \left(R_i(W_i) - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} \right) - n_1 \tilde{\alpha}_0 - \sum_{i:W_i=0} \left(R_i(W_i) - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} \right) + n_0 \tilde{\alpha}_1 \right] \\
&= \tilde{\alpha}_1 - \tilde{\alpha}_0,
\end{aligned}$$

and $\tilde{\tau}_s$ has the following expectation and variance.

$$\begin{aligned}
\mathbb{E} [\tilde{\tau}_s | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{E} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}} - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) + \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{E} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] - \mathbb{E} [\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] - \mathbb{E} [\bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}} - \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \alpha_1(\mathbf{x}_1^{obs}) - \alpha_0(\mathbf{x}_0^{obs}) - \mathbb{E} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] - (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \boldsymbol{\beta} \quad (2.33)
\end{aligned}$$

$$\begin{aligned}
\mathbb{V} [\tilde{\tau}_s | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{V} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{V} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}} - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) + \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \mathbb{V} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V} [\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V} [\bar{\mathbf{x}}_1^T \tilde{\boldsymbol{\beta}} - \bar{\mathbf{x}}_0^T \tilde{\boldsymbol{\beta}} | \mathbf{x} = \mathbf{x}^{obs}] \\
&= \frac{\sigma_{s_1}^2}{n_1} + \frac{\sigma_{s_0}^2}{n_0} + \mathbb{V} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] + (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V} [\boldsymbol{\beta}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs}) \quad (2.34)
\end{aligned}$$

Similar to $\tilde{\tau}_f$, we can see that $\tilde{\tau}_s$ is biased in a finite sample of experimental units. However, as the sample size grows, both $\mathbf{x}_1^{obs}$ and $\mathbf{x}_0^{obs}$ approach the covariate distribution of the super-population since treatment assignment is made randomly, thus $\mathbb{E} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})]$ and $(\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \boldsymbol{\beta}$ approach 0. The variance of $\tilde{\tau}_f$ is still mainly determined by the conditional variance of the outcome given the covariates. The covariance between $\bar{R}_1 - \bar{R}_0 = (\bar{Y}_1 - \bar{f}(\mathbf{x}_1)) - (\bar{Y}_0 - \bar{f}(\mathbf{x}_0))$ and $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \tilde{\boldsymbol{\beta}}$ conditional on the covariates $\mathbf{x} = \mathbf{x}^{obs}$ is 0 because $\tilde{\boldsymbol{\beta}}$ is not dependent on the mean of the residual outcome $\bar{R}_W$. With slight modifications, the proof of Appendix A.1 can be applied here to show that the covariance

is 0. Since we assume that $\tilde{f}(\mathbf{x})$ is fit using pre-experiment data, its model parameters are independent of the in-experiment outcomes hence the covariance between $\bar{Y}_1 - \bar{Y}_0$ and $\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0)$ conditional on covariates is also 0. Since the treatment assignment is made independently of the covariates, we again expect $\bar{\mathbf{x}}_1^{obs} \approx \bar{\mathbf{x}}_0^{obs}$ and consequently the model variances will be minimal, only leaving behind the conditional variance of the outcome.

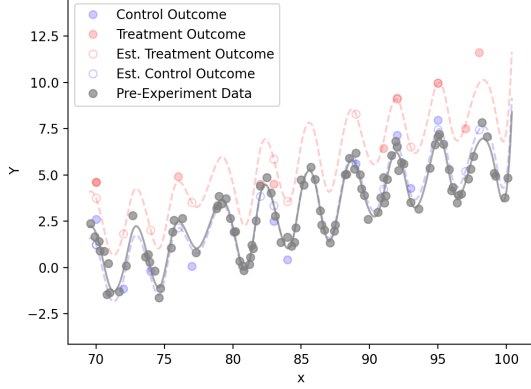
For each method described in Sections 2.1, 2.2, 2.3, and 2.4, the variance of the treatment effect estimator conditional on covariates follows the form $\mathbb{V}[\tilde{\tau}|\mathbf{x} = \mathbf{x}^{obs}] \approx \sigma_1^2/n_1 + \sigma_0^2/n_0$ where σ_W represents the conditional variance of the outcome based on the respective models. As the counterfactual models have gotten increasingly flexible, the conditional variances are reduced, leading to reduced-variance treatment effect estimates. Here, we expect the conditional variances to decrease even further as the model has become more flexible with the secondary linear adjustment. Thus the proposed method stands to provide larger variance reduction in treatment effect estimation. In principle, we can obtain even greater variance reduction by increasing the complexity of the secondary adjustment beyond a linear one, but we advocate for the linear adjustment for reasons to be discussed in Section 3.2.

2.4.2 Demonstration of Variance Reduction

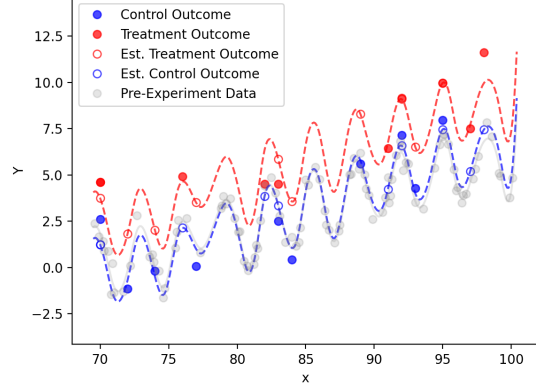
We repeat the same simulations as in Sections 2.2.2 and 2.3.2 with $\tilde{\tau}_s$. We can see that the counterfactual estimates in Figure 2.7 are linear transformations of the predictions from pre-experiment data, and they better align with the observed experiment data relative to only using linear regression (see Figure 2.5b) or only using predictions from the ML model fit to pre-experiment data (see Figure 2.6b). The treatment effect estimate $\hat{\tau}_s = 2.51$ in Figure 2.7b has a lower standard error of 0.40 relative to that of $\hat{\tau}_f$ (0.45) and $\hat{\tau}_r$ (0.77).

From analyzing Figures 2.2, 2.3, and 2.4, it is evident that $\hat{\tau}_s$ can estimate the treatment effect with even greater precision than $\hat{\tau}_f$. Notice that, especially with small samples, $SE[\hat{\tau}_s] = \sqrt{\hat{\sigma}_{s_1}^2/n_1 + \hat{\sigma}_{s_0}^2/n_0}$ tends to underestimate the true standard deviation of treatment effect estimates similar to $SE[\hat{\tau}_f]$. For $\tilde{\tau}_s$, there are two different models, $\tilde{f}(\mathbf{x})$ and $\tilde{\beta}$, that are contributing to the variance in treatment effect estimation (Equation (2.34)). With small sample sizes, the effect of model variance is not reduced to 0 hence $\mathbb{V}[\tilde{\tau}_s] > \sigma_{s_1}^2/n_1 + \sigma_{s_0}^2/n_0$. However, such underestimation dissipates quickly as sample size increases as can be seen through Figures 2.2b, 2.3b, and 2.4b.

The value of $\hat{\tau}_s$ is better demonstrated when the predictions from the pre-experiment data are poor. We rerun the simulation with an alternative model fit to different pre-experiment data and compare the performance of $\hat{\tau}_s$ to the other estimators. The new



(a) A cubic spline model fit to pre-experiment data.

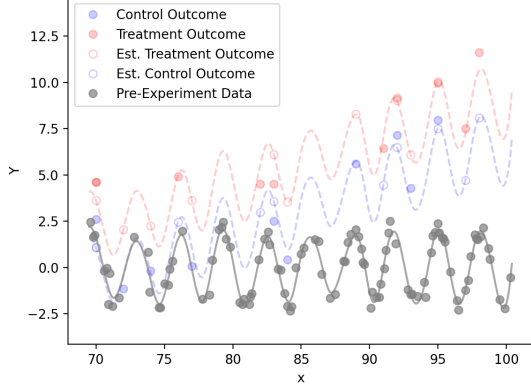


(b) Estimating counterfactuals with model predictions. $\hat{\tau}_s = 2.51$, $SE[\hat{\tau}_s] = 0.40$.

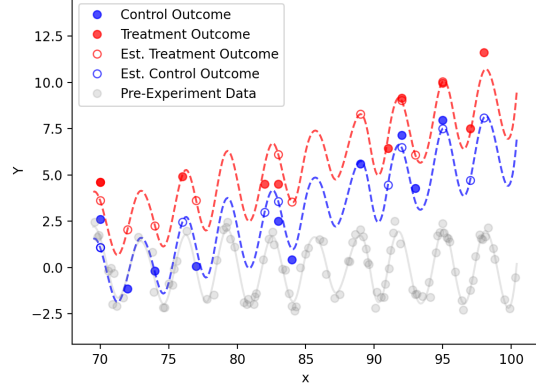
Figure 2.7: Visualization of pre- and in-experiment data. The grey line in Figure 2.7a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.7b represent treatment and control outcome estimates over covariate x with secondary linear adjustment.

pre-experiment outcomes are generated as follows: $Y_i = 2\sin(2x_i) + 0.5\epsilon_i$, $\epsilon_i \sim N(0, 1)$. Although this function shares the cyclical pattern in the outcome with the in-experiment data, it has a very different trend over the domain of x to ensure a high level of misalignment. Figure 2.8 is an example of a situation where pre-experiment data and in-experiment data have a misalignment. Even though the cyclical relationship between the outcome Y and covariate x still exists, the overall trend in the relationship before the experiment is different from that during the experiment. Utilizing predictions from the ML model $\tilde{f}(\mathbf{x})$ fit to misaligned pre-experiment data inflates the conditional variance of the model in Equation (2.22), which in turn increases the variance of $\tilde{\tau}_f$. However, $\tilde{\tau}_s$ can adapt to this change in trend as it makes a secondary linear adjustment to the experiment data. In Figure 2.8b, we observe the treatment effect estimate $\hat{\tau}_s = 2.54$ and its standard error of 0.39 is just as good as the previous scenario where the pre-experiment data was well-aligned with the in-experiment data in Figure 2.7b.

Figures 2.9, 2.10, and 2.11 show the distributions of treatment effect estimates and their standard errors over sample sizes $\{10, 50, 100\}$ respectively under this new scenario. Note, the misalignment between pre- and in-experiment data does not impact $\tilde{\tau}_d$ and $\tilde{\tau}_r$ which only use the in-experiment data. We observe that the distribution of $\hat{\tau}_f$ is much wider across



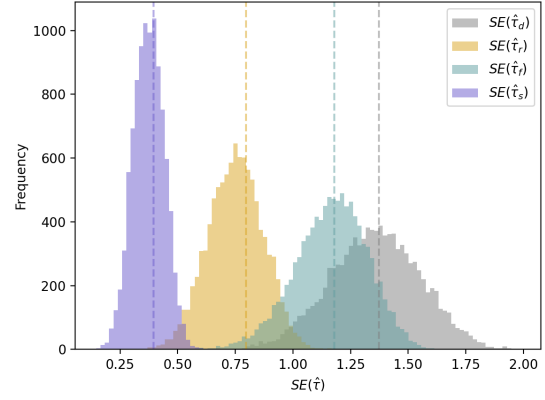
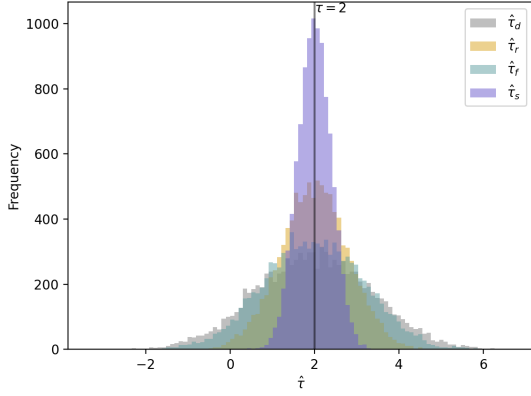
(a) A cubic spline model fit to pre-experiment data in the second scenario.



(b) Estimating counterfactuals with model predictions. $\hat{\tau}_s = 2.54$, $SE[\hat{\tau}_s] = 0.39$.

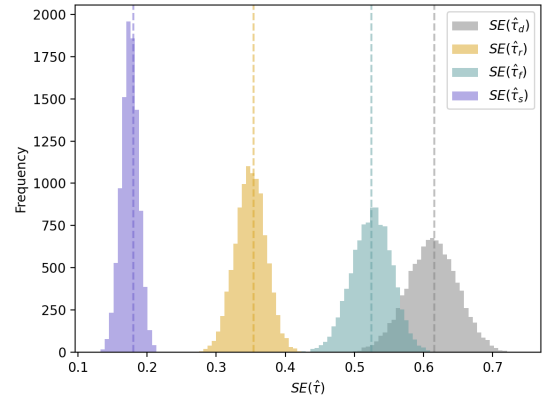
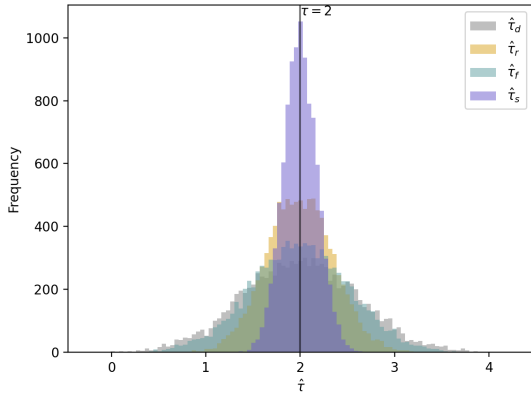
Figure 2.8: Visualization of misaligned pre- and in-experiment data. The grey line in Figure 2.8a represents ML model fit to pre-experiment data. The red and blue dotted lines in Figure 2.8b represent treatment and control outcome estimates over covariate x with secondary linear adjustment.

the sample sizes compared to those in Figures 2.2a, 2.3a, and 2.4a, and this is the result of misalignment between pre- and in-experiment data. With this level of misalignment, even the regression method estimates $\hat{\tau}_r$ shows greater precision than $\hat{\tau}_f$ in treatment effect estimation. From the standard error distributions in Figures 2.9b, 2.10b, and 2.11b, the distribution of $SE[\hat{\tau}_f]$ is comparable to the distribution of $SE[\hat{\tau}_d]$, the difference-in-means method which does not utilize any pre-experiment data or covariates. Therefore, in the event of misalignment between pre- and in-experiment data, utilizing the pre-experiment data without additional adjustment can lead to negligible variance reduction relative to difference-in-means. The summary of variance reduction in this new scenario is shown in Table 2.2. We see the variance reduction remains high for $\tilde{\tau}_s$ in this case, but the performance of $\tilde{\tau}_f$ has dramatically deteriorated. Thus, regardless of whether there is good alignment between pre- and in-experiment data, $\tilde{\tau}_s$ works well and outperforms the other estimators with respect to variance reduction in all scenarios.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 2.9: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 2.9b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.9a.

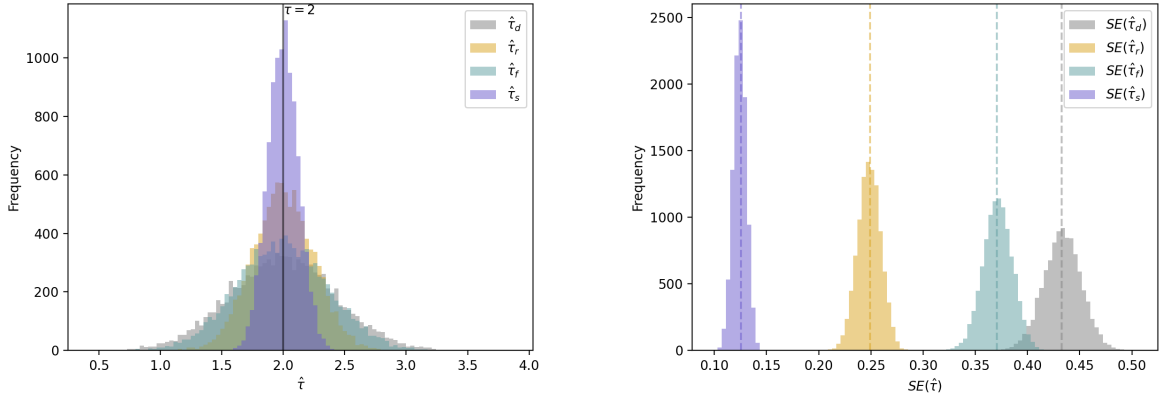


(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 2.10: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 2.10b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.10a.

Sample Size	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$
10	66.26%	26.18%	91.68%
50	66.87%	27.34%	91.40%
100	66.79%	26.72%	91.61%

Table 2.2: Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$ when pre- and in-experiment data are misaligned.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 2.11: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 2.11b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 2.11a.

Chapter 3

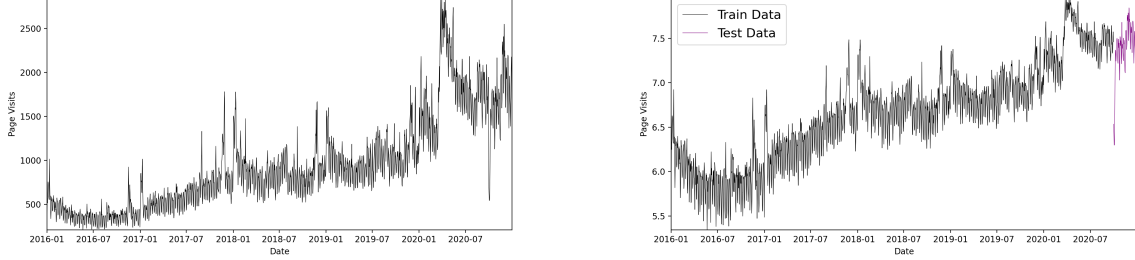
Performance Investigation

In Chapter 2, we introduced several potential outcomes models and hence treatment effect estimators and we demonstrated their relative performances on simulated data. From the simulations, the novel method with secondary linear adjustment (Section 2.4) outperformed the constant model (Section 2.1), the regression model (Section 2.2), and the ML model (Section 2.3) in terms of variance reduction. The purpose of this chapter is to validate the above results using real-world data that is noisier than simulated data and to showcase the novel method’s variance reduction ability in a realistic environment.

3.1 Pageview Dataset

3.1.1 Overview

The “Page Visits” dataset from Kaggle [18] is used to demonstrate the potential variance reduction that can be achieved by the four treatment effect estimators explored in Chapter 2. The data contains daily page visits to a website from January 1, 2016, to December 31, 2020, and the only covariate is the date. This is visualized in Figure 3.1a. As the page visits data shows a steady increase in variance, a log transformation is applied to the number of page visits to maintain homogeneity in variance over time. We will consider the first 95% of data as pre-experiment data, and use it as training data for the ML model $\tilde{f}(\mathbf{x})$. The remaining 5% of data is then used to generate synthetic experiment outcomes (Figure 3.1b) to which the predictions of page visits will be made using $\tilde{f}(\mathbf{x})$. The details of model fitting can be found in Appendix A.2, and the final model of choice is a SARIMA model based on the entire training data.



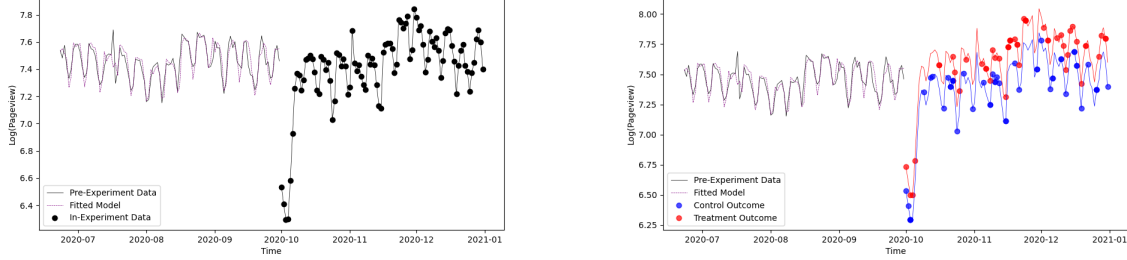
(a) Time series of all page visits from 2016 to 2020. (b) Log transformed time series of page visits.

Figure 3.1: Time series of page visits. Figure 3.1a shows increasing variability over time, hence log transformation is applied to achieve constant variance as demonstrated in Figure 3.1b.

3.1.2 Simulating Experiments

The experiment data is simulated using the Test Data portion in Figure 3.1b. Figure 3.2 shows a portion of the training (pre-experiment) data followed by the test (in-experiment) data, as well as the model predictions for the training data. Similar to Section 2.1.2, the test data is randomly sampled with replacement to generate an instance of control and treatment data, where the treatment data is applied a constant treatment effect of $\tau = 0.2$ in this case. We can fit the four different models described in Chapter 2 to the particular instance of experiment data shown in Figure 3.2b. The results are described in Figure 3.3. The standard error of $\hat{\tau}_r$ in Figure 3.3b is smaller than the standard error of $\hat{\tau}_f$ in Figure 3.3c, even though $\hat{\tau}_r$ does not utilize predictions based on pre-experiment data that can explain the cyclical behavior in page visits while $\hat{\tau}_f$ does. By observation of in-experiment data, we can identify a sudden drop in the outcome around October 2020. However, the raw predictions from the SARIMA model do not adjust for this (Figure 3.3c), and making a secondary linear adjustment more closely aligns the predictions to the in-experiment data (Figure 3.3d). Hence, this results in $SE[\hat{\tau}_f] > SE[\hat{\tau}_s]$.

We now repeat this experiment 10,000 times to create an empirical distribution of treatment effect estimates and their standard errors for each of the 4 estimators: $\tilde{\tau}_d$, $\tilde{\tau}_r$, $\tilde{\tau}_s$, and $\tilde{\tau}_f$. Sample sizes of $n_1 = n_0 \in \{10, 50, 100\}$ are utilized to demonstrate the asymptotic bias and variance of the estimators in Figures 3.4, 3.5, and 3.6. The distribution of treatment effect estimates in Figures 3.4a, 3.5a, and 3.6a demonstrates the unbiasedness of the estimators at all sample sizes. However, the distribution of treatment effect estimates of each

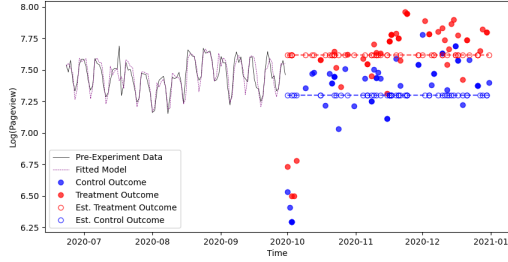


(a) Test data as the in-experiment data before applying random treatment. (b) In-experiment data generated with random application of treatment ($n_1 = n_0 = 50$).

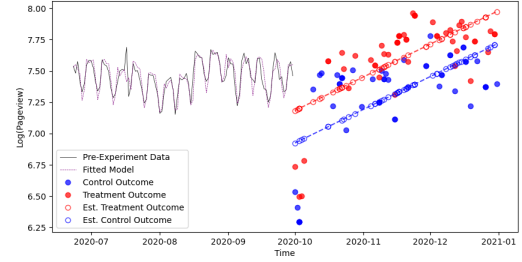
Figure 3.2: Simulating in-experiment data from test data using random sampling with replacement. The treatment units are applied a constant treatment effect of $\tau = 0.2$.

method has similar variability. As can be seen in Figure 3.3, the in-experiment outcomes exhibit a much more complex behavior than the simulated data considered in Chapter 2. Hence, even the more complex models shown in Figures 3.3c and 3.3d cannot completely capture this complexity, which results in larger residuals and thus less variance reduction. Also, the experiment data contains some extreme points near October 2020 that have much lower values than their neighbors (Figure 3.2). These extreme points, if included in the random sample used to simulate in-experiment data, significantly increase the standard error of the treatment estimate.

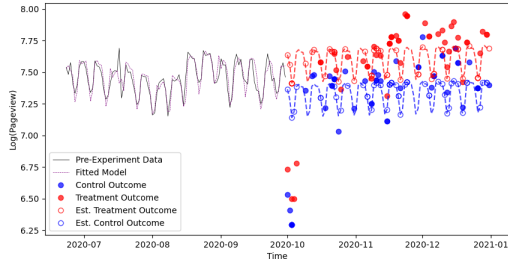
The effects of such extreme points on standard errors are demonstrated in Figure 3.4b. In Figure 3.4b, the distributions of standard errors are multimodal. This is because of the small sample size ($n_1 = n_0 = 10$), where the extreme points may or may not be included in the sample used to simulate an instance of in-experiment data. The inclusion of extreme points can significantly impact the standard error due to their large residuals, hence creating this multimodal distribution in standard errors. However, with a larger sample size, the likelihood of selecting a sample without any extreme point is very small, and the distribution of standard error becomes unimodal (Figures 3.5b and 3.6b). We observe that as sample size increases, the distribution of standard errors becomes well-centered around $\mathbb{V}[\tilde{\tau}]$ which is the true variance in treatment effect estimates observed over 10,000 simulations. This again validates that $\hat{\mathbb{V}}[\tilde{\tau}] = \hat{\sigma}_1^2/n_1 + \hat{\sigma}_0^2/n_0$ is an approximation for the variance of the treatment effect estimators, and the variance associated with the model such as $(\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})^T \mathbb{V}[\tilde{\beta}] (\bar{\mathbf{x}}_1^{obs} - \bar{\mathbf{x}}_0^{obs})$ and $\mathbb{V}[\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})]$ tend to 0 as sample size increases and the covariate distribution between the control and treatment groups become balanced.



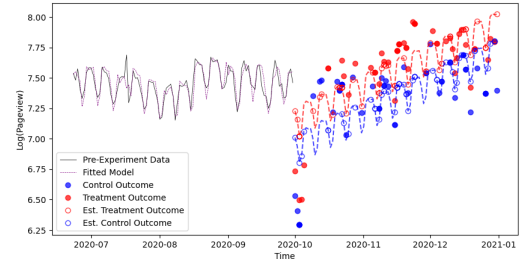
(a) Estimating counterfactuals with a constant model from Section 2.1.1.
 $\hat{\tau}_d = 0.3180, SE[\hat{\tau}_d] = 0.0717$.



(b) Estimating counterfactuals with a linear model from Section 2.2.1.
 $\hat{\tau}_r = 0.2593, SE[\hat{\tau}_r] = 0.0554$.

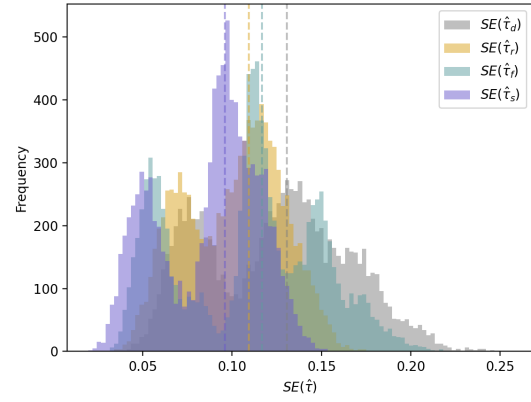
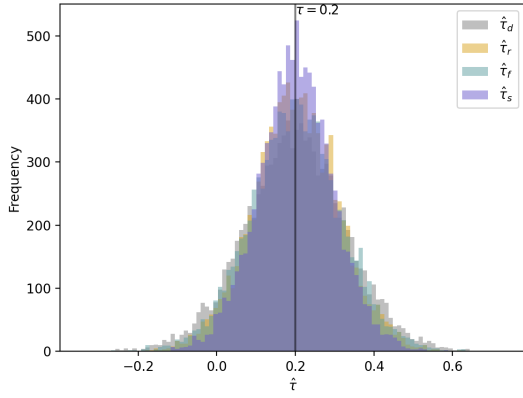


(c) Estimating counterfactuals with a flexible ML model from Section 2.3.1.
 $\hat{\tau}_f = 0.2736, SE[\hat{\tau}_f] = 0.0626$.



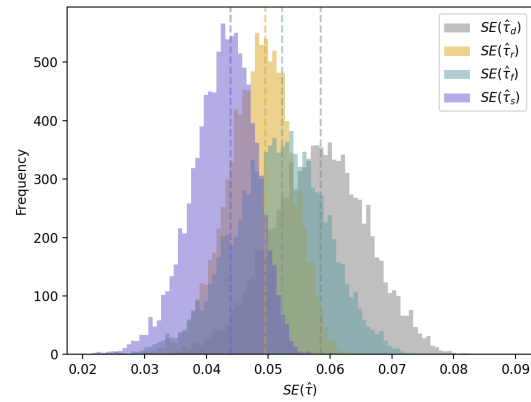
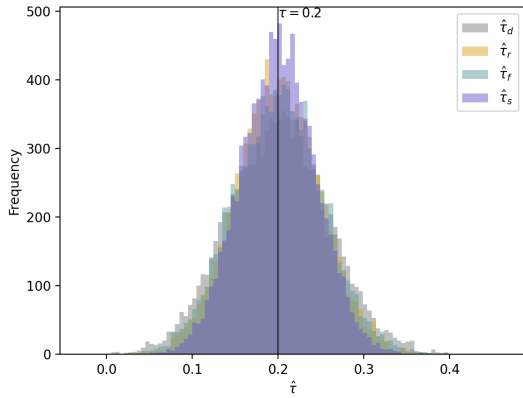
(d) Estimating counterfactuals with a two-stage model from Section 2.4.1.
 $\hat{\tau}_s = 0.2183, SE[\hat{\tau}_s] = 0.0457$.

Figure 3.3: Visualization of counterfactual estimators described in Chapter 2 on an instance of simulated experiment data, their treatment effect estimates $\hat{\tau}$ and standard errors $SE[\hat{\tau}]$.



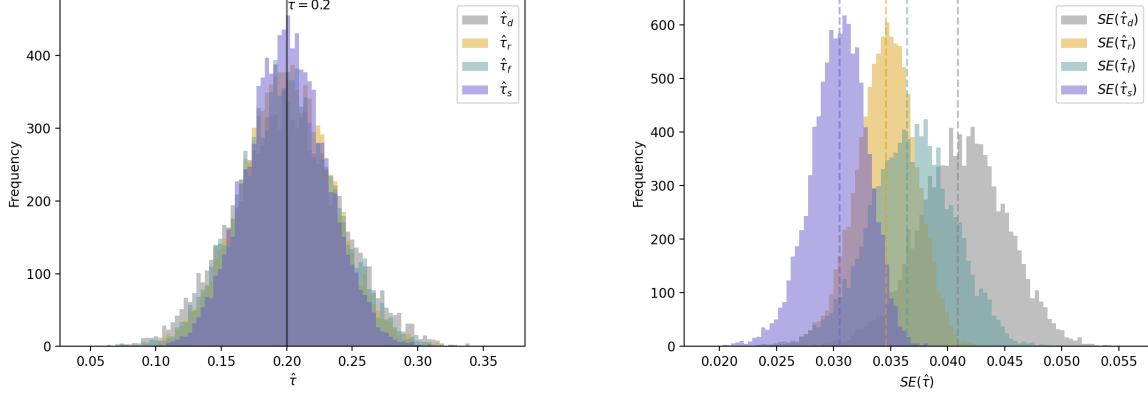
(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.4: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 3.4b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.4a.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.5: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 3.5b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.5a.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.6: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 3.6b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.6a.

The variance reduction of each estimator is summarized in Table 3.1. Even though the page visits data is much more complex and noisy than the simulated data considered in Chapter 2, we can observe that the two-stage framework continues to outperform $\tilde{\tau}_f$ and $\tilde{\tau}_r$ in terms of variance reduction. The predictions from the SARIMA model, although they capture the seasonality in page visits, were unable to align well with the in-experiment data. The presence of extreme points also deteriorated the alignment between predictions and the in-experiment data, which impacted the variance reduction of $\tilde{\tau}_f$. As $\tilde{\tau}_r$ makes linear adjustments within the in-experiment data, it would adjust its counterfactual estimates to the extreme points and achieve greater variance reduction than $\tilde{\tau}_f$. The two-stage framework maximizes the correlation between predictions and the in-experiment data (which may contain extreme observations) through a linear adjustment to achieve better alignment, resulting in greater variance reduction in treatment effect estimation.

3.2 Complexity of Secondary Adjustment

In Figure 3.3, we see that the predictions from pre-experiment data and even the linearly adjusted predictions do not closely model the in-experiment outcomes due to their complex behavior over time and the presence of extreme points. The natural question that arises

Sample Size	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$
10	29.88%	19.97%	46.02%
50	28.44%	20.16%	43.81%
100	28.45%	20.56%	44.28%

Table 3.1: Variance reduction of different treatment effect estimators relative to $\tilde{\tau}_d$ in the Page Visits dataset.

here is: *why stop at linear adjustment?* If the goal is to minimize the conditional variance, using a more complex secondary adjustment may be able to achieve this goal better than a simple linear adjustment. We repeat the above simulation using a degree 4 polynomial for the secondary adjustment and compare this result to our existing linear secondary adjustment. The potential outcomes estimators are defined as

$$\tilde{Y}_i(W_i) = \tilde{f}(x_i) + \tilde{\alpha}_1 W_i + \tilde{\alpha}_0(1 - W_i) + \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} \quad (3.1)$$

where x_i represents the time, $\mathbf{x}_i = [x_i, x_i^2, x_i^3, x_i^4]^T$ is a vector of powers of x_i , $\tilde{f}(x_i)$ is the page visits prediction of the SARIMA model fit to pre-experiment data, and $\tilde{\alpha}_1, \tilde{\alpha}_0 \in \mathbb{R}$ and $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^4$ are parameters estimated with OLS. We will use $\tilde{\tau}_q = \tilde{\alpha}_1 - \tilde{\alpha}_0$ to denote the corresponding treatment effect estimator. An example of these counterfactual estimates can be seen in Figure 3.7. Using a highly complex secondary adjustment seems to better fit our in-experiment data and result in an even lower standard error of 0.0230 for $\hat{\tau}_q$ compared to the standard error of 0.0457 of $\hat{\tau}_s$ in Figure 3.3d. However, the consequence of overfitting to the in-experiment data can be seen in Figures 3.8, 3.9, and 3.10. Only the treatment effect estimators with secondary adjustments, $\tilde{\tau}_s$ and $\tilde{\tau}_q$, are compared on these figures for visual clarity. The treatment effect estimates appear to be unbiased, yet the standard errors of $\hat{\tau}_q$ appear to underestimate the true standard deviation of the treatment effect estimates more than $\hat{\tau}_s$ across the different sample sizes. Even with sample sizes of 50 and 100, the distributions of standard errors are not well-centered around the true standard deviation of treatment effect estimates denoted by the dotted lines (Figures 3.9b and 3.10b respectively). This shows that the asymptotic approximation of the variance $\widehat{\mathbb{V}}[\tilde{\tau}_q] = \hat{\sigma}_{q_1}^2/n_1 + \hat{\sigma}_{q_0}^2/n_0$ is not achieved with the sample sizes of 50 or 100 due to the complexity of the secondary adjustment, and such underestimation can inflate Type I error when there is no treatment effect.

For a comprehensive comparison of type I and II error rates of all 5 estimators $\tilde{\tau}_d, \tilde{\tau}_r, \tilde{\tau}_f, \tilde{\tau}_s$, and $\tilde{\tau}_q$, we performed the same simulation across different treatment effect sizes, $\tau \in \{0, 0.1, 0.2, 0.3\}$. For each scenario with a particular estimator, sample size, and effect size,

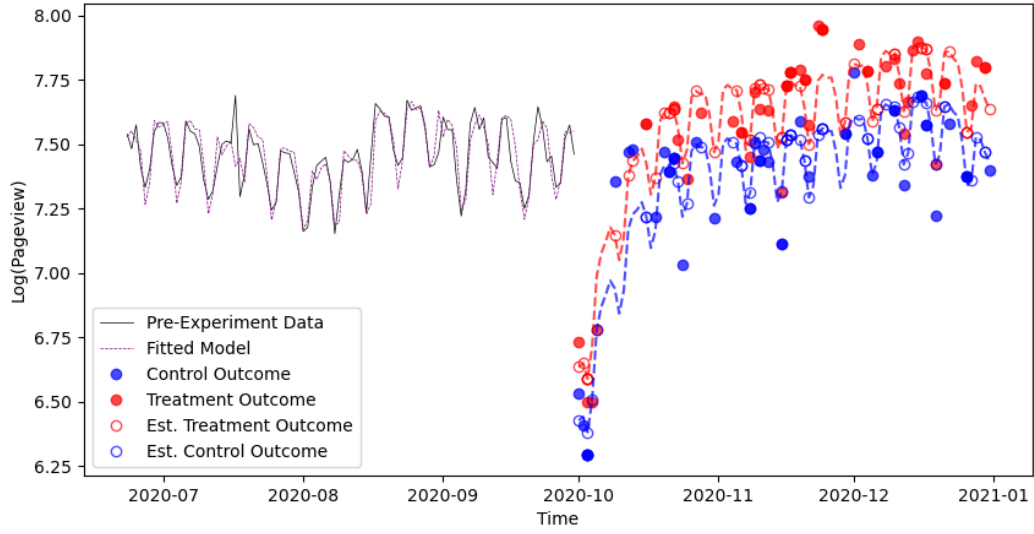
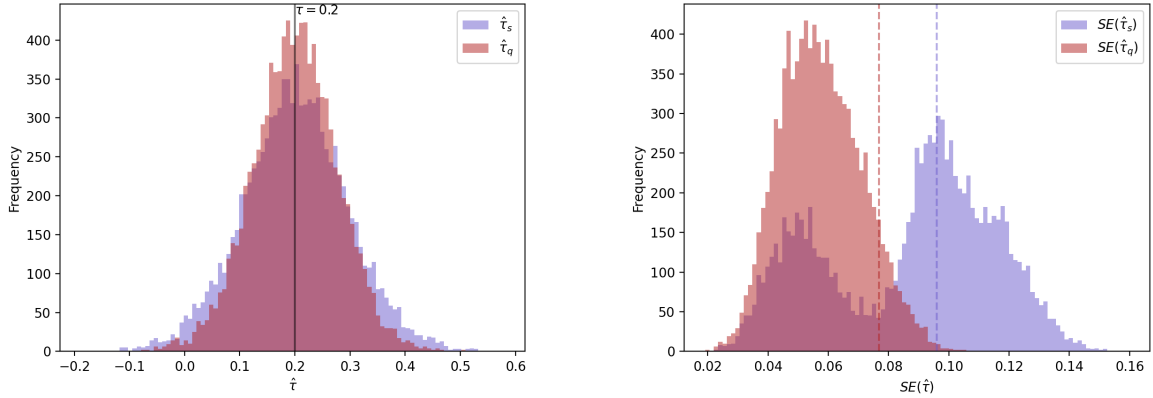
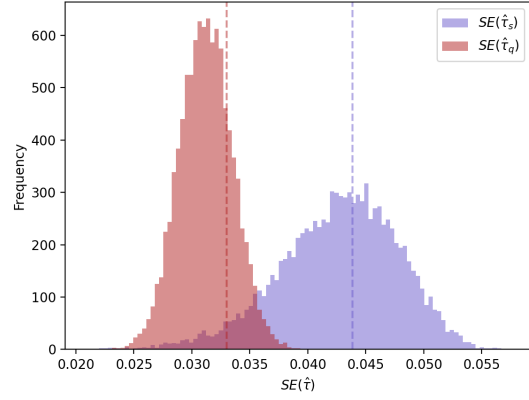
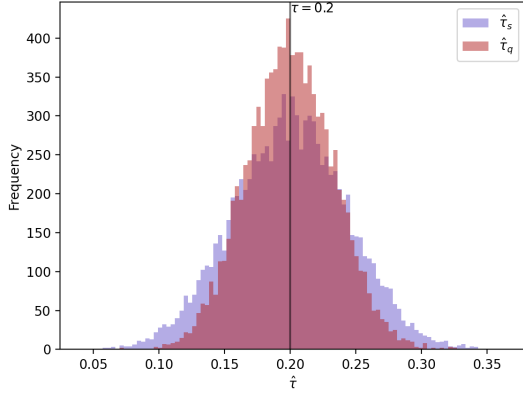


Figure 3.7: Estimating counterfactuals with degree 4 polynomial secondary adjustment. $\hat{\tau}_q = 0.2073$, $SE[\hat{\tau}_q] = 0.0230$.



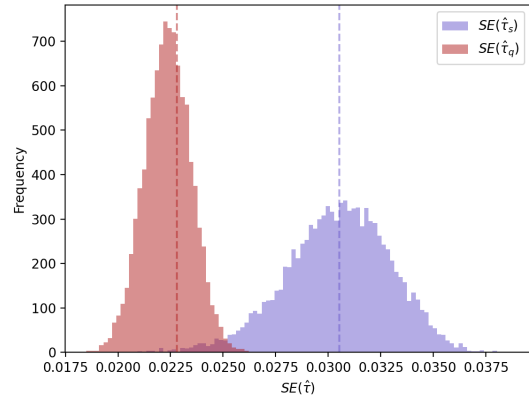
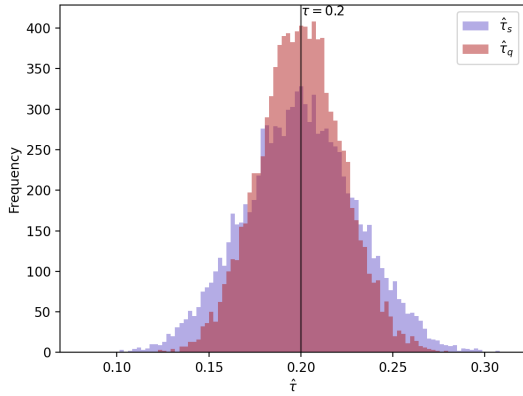
(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.8: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 10$. Vertical lines in Figure 3.8b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.8a.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.9: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 50$. Vertical lines in Figure 3.9b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.9a.



(a) Distribution of treatment effect estimates. (b) Distribution of standard error estimates.

Figure 3.10: Treatment effect results from 10,000 simulated experiments with $n_1 = n_0 = 100$. Vertical lines in Figure 3.10b represent the standard deviation of the 10,000 $\hat{\tau}$ in Figure 3.10a.

τ	$\tilde{\tau}_d$	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$	$\tilde{\tau}_q$
0	4.20%	5.64%	3.48%	5.31%	12.59%
0.1	13.90%	18.70%	19.70%	26.12%	41.27%
0.2	40.64%	50.18%	50.13%	59.35%	83.84%
0.3	62.82%	76.20%	68.03%	82.90%	98.15%

Table 3.2: Proportion of statistically significant experiments out of 10,000 with $n_1 = n_0 = 10$.

τ	$\tilde{\tau}_d$	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$	$\tilde{\tau}_q$
0	4.75%	5.48%	4.76%	5.33%	6.15%
0.1	41.31%	54.09%	49.89%	64.58%	87.51%
0.2	91.57%	97.88%	96.05%	99.48%	100.00%
0.3	99.78%	99.99%	99.96%	100.00%	100.00%

Table 3.3: Proportion of statistically significant experiments out of 10,000 with $n_1 = n_0 = 50$.

we compute the test statistic $|\hat{\tau}|/SE[\hat{\tau}]$ associated with $H_0 : \tau = 0$, and compare this with the 97.5% quantile of the t -distribution (assuming type I error rate of 5%) with degrees of freedom equal to $n_1 + n_0 - 2$. We then report the proportion of 10,000 simulations with statistically significant results, i.e., $|\hat{\tau}|/SE[\hat{\tau}] > t_{(0.975)}^{n_1+n_0-2}$. This yields empirical estimates of type I error rates and power (1 - type II error rate) when the treatment is zero and non-zero, respectively.

Tables 3.2, 3.3, and 3.4 summarize the proportion of statistically significant experiments for sample sizes 10, 50, and 100 respectively. First, we consider the type I error rate.

τ	$\tilde{\tau}_d$	$\tilde{\tau}_r$	$\tilde{\tau}_f$	$\tilde{\tau}_s$	$\tilde{\tau}_q$
0	4.58%	5.02%	4.63%	4.94%	5.44%
0.1	67.41%	81.94%	77.50%	90.05%	99.26%
0.2	99.75%	99.98%	99.95%	100.00%	100.00%
0.3	100.00%	100.00%	100.00%	100.00%	100.00%

Table 3.4: Proportion of statistically significant experiments out of 10,000 with $n_1 = n_0 = 100$.

Since we use a type I error rate of 5% for each test, we expect roughly 5% of the 10,000 experiments to be statistically significant when there is no treatment effect, i.e., $\tau = 0$. In Table 3.2, we observe the type I error rates for all methods are roughly around 5% except for $\tilde{\tau}_q$ with 12.59%, more than double what we want. Even with a sample size of 50, the result in Table 3.3 shows that $\tilde{\tau}_q$ has a type I error rate of 6.15%, which is still greater than the desired 5%. However, when $\tau > 0$, $\tilde{\tau}_q$ is associated with an increased proportion of statistically significant experiments relative to the other methods (across all sample sizes). This reflects its variance reduction capabilities. Especially for detecting small treatment effects ($\tau = 0.1$), $\tilde{\tau}_q$ achieves statistical significance more frequently than $\tilde{\tau}_s$. Therefore, it appears that increasing the complexity of the secondary adjustment generally increases the sensitivity of the test and hence both type I errors and power.

We can examine this in more detail by generalizing the two-stage framework introduced in Section 2.4.1 by replacing the secondary linear adjustment with a general function of covariates, $g(\mathbf{x}_i)$, whose parameters are estimated concurrently with α_1 and α_0 (Equation (3.2)).

$$\tilde{Y}_i(W_i) = \tilde{f}(\mathbf{x}_i) + \tilde{\alpha}_1 W_i + \tilde{\alpha}_0(1 - W_i) + \tilde{g}(\mathbf{x}_i) \quad (3.2)$$

The resulting treatment effect estimator, $\tilde{\tau}_g$, has the following bias and variance:

$$\begin{aligned} \mathbb{E} [\tilde{\tau}_g | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{E} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \mathbb{E} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{g}(\mathbf{x}_1) - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) + \bar{g}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \mathbb{E} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] - \mathbb{E} [\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] - \mathbb{E} [\bar{g}(\mathbf{x}_1) - \bar{g}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \alpha_1(\mathbf{x}_1^{obs}) - \alpha_0(\mathbf{x}_0^{obs}) - \mathbb{E} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] - \mathbb{E} [\bar{g}(\mathbf{x}_1^{obs}) - \bar{g}(\mathbf{x}_0^{obs})], \end{aligned} \quad (3.3)$$

$$\begin{aligned} \mathbb{V} [\tilde{\tau}_g | \mathbf{x} = \mathbf{x}^{obs}] &= \mathbb{V} [\tilde{\alpha}_1 - \tilde{\alpha}_0 | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \mathbb{V} [\bar{Y}_1 - \bar{f}(\mathbf{x}_1) - \bar{g}(\mathbf{x}_1) - \bar{Y}_0 + \bar{f}(\mathbf{x}_0) + \bar{g}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \mathbb{V} [\bar{Y}_1 - \bar{Y}_0 | \mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V} [\bar{f}(\mathbf{x}_1) - \bar{f}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] + \mathbb{V} [\bar{g}(\mathbf{x}_1) - \bar{g}(\mathbf{x}_0) | \mathbf{x} = \mathbf{x}^{obs}] \\ &= \frac{\sigma_{g_1}^2}{n_1} + \frac{\sigma_{g_0}^2}{n_0} + \mathbb{V} [\bar{f}(\mathbf{x}_1^{obs}) - \bar{f}(\mathbf{x}_0^{obs})] + \mathbb{V} [\bar{g}(\mathbf{x}_1^{obs}) - \bar{g}(\mathbf{x}_0^{obs})]. \end{aligned} \quad (3.4)$$

Here, $\bar{g}(\mathbf{x}_W) = \frac{1}{n_W} \sum_{i:W_i=W} \tilde{g}(\mathbf{x}_i)$. Similar to the results in Section 2.4.1, the covariance between $\bar{R}_1 - \bar{R}_0 = (\bar{Y}_1 - \bar{f}(\mathbf{x}_1)) - (\bar{Y}_0 - \bar{f}(\mathbf{x}_0))$ and $\bar{g}(\mathbf{x}_1) - \bar{g}(\mathbf{x}_0)$ conditional on the covariates $\mathbf{x} = \mathbf{x}^{obs}$ can be made 0 by using mean-centered outcomes and covariates when estimating the parameters of g .

However, using a highly complex ML model for the secondary adjustment implies that the function $\tilde{g}(\mathbf{x})$ varies widely over the different values of covariate $\mathbf{x}$. Therefore, a small difference in covariate distribution between the control and treatment group can result in a meaningful difference between $\bar{g}(\mathbf{x}_1)$ and $\bar{g}(\mathbf{x}_0)$ due to the high variability of $\tilde{g}$. Thus, a larger sample size is required to ensure that the covariate distribution between the two groups is minuscule and $\bar{g}(\mathbf{x}_1^{obs}) - \bar{g}(\mathbf{x}_0^{obs}) \rightarrow 0$, which is necessary for the asymptotic unbiasedness of $\tilde{\tau}_g$ and variance approximation with $\mathbb{V}[\tilde{\tau}_g | \mathbf{x} = \mathbf{x}^{obs}] \approx \sigma_{g_1}^2/n_1 + \sigma_{g_0}^2/n_0$. The curse of dimensionality also provides motivation to prefer a simpler secondary adjustment. Balancing the covariate distribution between the two finite treatment and control groups grows exponentially more difficult as the dimension of covariates increases. This wasn't a concern in the "Page Visits" data because x was a scalar, but using a highly flexible secondary adjustment with high-dimensional covariates is bound to result in a biased treatment effect estimate and underestimated variance even with a large sample size.

It is possible to closely examine the data to determine the optimal complexity of the secondary adjustment that would achieve greater variance reduction than $\tilde{\tau}_f$ without significant impact to bias and underestimation of variance. However, the search for the optimal complexity via a thorough model selection procedure significantly deteriorates the agility of experimentation. The model selection may require analysis of both graphical and numerical summaries of data, thus increasing the time it takes to complete an experiment and ultimately limiting the number of experiments that can be completed in a given day. We also expect the pre-experiment data to perform decently well in modeling the in-experiment data, and so the residual outcome $R_i = Y_i - \tilde{f}(\mathbf{x}_i)$ is not likely to exhibit highly complex behavior that necessitates the use of flexible secondary adjustment.

Given the points outlined above, we advocate for linear secondary adjustment over a more complex secondary adjustment. Especially in online settings where we can collect large amounts of high-quality pre-experiment data that can provide insights into the complex relationships between covariates and the outcome of interest, a complex secondary adjustment risks biased treatment effect estimates, underestimation of treatment effect variance, and a slowing of the experimentation process for a possibly negligible improvement in variance reduction.

Chapter 4

Conclusion

In this thesis, we proposed and explored a new variance reduction technique for online A/B tests. Variance reduction is an important problem in A/B tests as it directly improves their statistical power, hence enabling more accurate decisions with regard to product or service changes. Existing variance reduction methods utilize pre-experiment data under the assumption that pre- and in-experiment data are very similar to each other. When this is true, complex relationships between covariates and the outcome in the pre-experiment data can be exploited to model their relationship in the in-experiment data, thereby reducing the variance of treatment effect estimators. However, especially in online settings, changes in trends and environment over time can degrade the correlation between pre- and in-experiment data which impairs the variance reduction capabilities of existing methods that utilize pre-experiment data.

We propose a two-stage framework to model the potential outcomes which utilizes predictions from a machine learning model fit to pre-experiment data and makes linear adjustments to these predictions according to the in-experiment data. The two-stage framework can adjust for differences between pre- and in-experiment data, and can be used to construct a treatment effect estimator with a smaller variance than what is achievable with existing methods. This improvement over existing variance reduction techniques can significantly enhance the efficiency of A/B tests in online settings where large amounts of pre-experiment data are readily available. The performance of this method was analyzed first through simulated data in Chapter 2, then through real-world data in Chapter 3, and in both settings the proposed two-stage framework demonstrated superior variance reduction performance relative to existing methods.

An important consideration in the two-stage framework lies in determining the appro-

priate amount of pre-experiment data to incorporate in the variance reduction method. In a rapidly changing online environment, outdated pre-experiment data may share no similarities with the in-experiment data. Including outdated pre-experiment data to train a machine learning model may therefore result in inaccurate predictions of the in-experiment outcomes, hence resulting in poor variance reduction. Thus, a challenge remains in identifying what pre-experiment data to retain and which is considered “outdated” and hence should be discarded. The bias-variance trade-off should also be considered here, as fitting a complex machine learning model comes with the risk of over-fitting if the training data is too small. Another consideration to be made with the two-stage framework is determining the complexity of the secondary adjustment. In Section 3.2, we advocated against the use of a highly complex secondary adjustment because of the risk of biased treatment effect estimation, underestimation of treatment effect variance, and slowing of the experimentation process. However, in any context where the agility of experimentation is of less importance, it could be worthwhile to search for the optimal complexity of secondary adjustment (which can be more complex than linear) with careful model selection to achieve better variance reduction while mitigating the risks of biased treatment effect estimation and underestimation of treatment effect variance.

Finally, we describe an important extension of this methodology: sample size determination. Variance reduction brings two related benefits to A/B tests; all else equal, it allows a smaller treatment effect to be detected with the same sample size. In a similar vein, it allows a treatment effect of a given magnitude to be detected with a smaller sample size. However, it is difficult to take advantage of the second benefit as the two-stage framework is performed after the data is collected. If there was a way to estimate the potential variance reduction from using the two-stage framework before the data is collected, we could design the experiment with a smaller sample size with assurance that power and type I error rates will still be maintained for the effect size of interest. This is a promising avenue for future work.

References

- [1] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments*. Cambridge University Press, 2020.
- [2] Alex Hern. *Why Google has 200m reasons to put engineers over designers*. 2014. URL: <https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers> (visited on 07/15/2024).
- [3] Jakub Linowski. *Google has been A/B testing link colors (again) and this light blue didn't pass*. 2019. URL: <https://goodui.org/leaks/google-has-been-a-b-testing-link-colors-again-and-this-light-blue-didnt-pass/> (visited on 07/15/2024).
- [4] Nicholas Larsen et al. “Statistical challenges in online controlled experiments: A review of A/B testing methodology”. In: *The American Statistician* 78.2 (2024), pp. 135–149.
- [5] Somit Gupta et al. “Top challenges from the first practical online controlled experiments summit”. In: *ACM SIGKDD Explorations Newsletter* 21.1 (2019), pp. 20–35.
- [6] Gema de las Heras et al. *How websites and apps collect and use your information*. 2023. URL: <https://consumer.ftc.gov/articles/how-websites-and-apps-collect-and-use-your-information> (visited on 07/15/2024).
- [7] Alex Deng et al. “Improving the sensitivity of online controlled experiments by utilizing pre-experiment Data”. In: *Proceedings of the sixth ACM international conference on Web search and data mining* (2013). DOI: [10.1145/2433396.2433413](https://doi.org/10.1145/2433396.2433413).
- [8] Alexey Poyarkov et al. “Boosted decision tree regression adjustment for variance reduction in online controlled experiments”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 235–244.

- [9] Peter W Glynn and Roberto Szechtman. “Some new perspectives on the method of control variates”. In: *Monte Carlo and Quasi-Monte Carlo Methods 2000: Proceedings of a Conference held at Hong Kong Baptist University, Hong Kong SAR, China, November 27–December 1, 2000*. Springer. 2002, pp. 27–49.
- [10] Jeff Li, Yixin Tang, and Jared Bauman. *Improving experimental power through control using predictions as covariate (CUPAC)*. 2020. URL: <https://doordash.engineering/2020/06/08/improving-experimental-power-through-control-using-predictions-as-covariate-cupac/> (visited on 07/15/2024).
- [11] Klarna. *Shopping pulse: Online Shopping Habits*. 2023. URL: <https://insights.klarna.com/shopping-pulse/> (visited on 07/15/2024).
- [12] Brad Ward. *2022 Ecommerce Stats & Trends Report*. 2022. URL: <https://www.salecycle.com/1/22702/2022-02-15/7g3kt3> (visited on 07/15/2024).
- [13] Adobe Analytics. *Media Alert: Adobe: Cyber Monday Surges to \$12.4 Billion in Online Spending, Breaking E-Commerce Record*. 2023. URL: <https://news.adobe.com/news/news-details/2023/Media-Alert-Adobe-Cyber-Monday-Surges-to-12.4-Billion-in-Online-Spending-Breaking-E-Commerce-Record/default.aspx> (visited on 07/15/2024).
- [14] Marketplace Pulse. *Amazon Online Stores Sales 2015-2023*. URL: <https://www.marketplacepulse.com/stats/amazon-online-stores-sales> (visited on 07/15/2024).
- [15] Zhang Liu et al. “Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records”. In: *Transactions in GIS* 22.2 (2018), 494–513. DOI: [10.1111/tgis.12323](https://doi.org/10.1111/tgis.12323).
- [16] Tao Deng et al. “Measuring smartphone usage and task switching with log tracking and self-reports”. In: *Mobile Media & Communication* 7.1 (2019), pp. 3–23. DOI: [10.1177/2050157918761491](https://doi.org/10.1177/2050157918761491). eprint: <https://doi.org/10.1177/2050157918761491>. URL: <https://doi.org/10.1177/2050157918761491>.
- [17] Guido Imbens and Donald B. Rubin. *Causal inference for statistics, social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2022.
- [18] Vikram Amin. *Time series forecasting using prophet in R*. 2023. URL: <https://www.kaggle.com/datasets/vikramamin/time-series-forecasting-using-prophet-in-r/data> (visited on 07/15/2024).

APPENDICES

Appendix A

Supplementary Materials

A.1 Proof of 0 covariance in conditional variance of $\tilde{\tau}_r$

Consider a simple case where the covariate is one-dimensional. Hence, the individual unit's covariate is denoted as x_i , the average covariate value of group W is denoted as $\bar{x}_W$, and the linear coefficient is denoted as $\tilde{\beta}$ without bold. The following is Equation (2.21) that does not omit the covariance term:

$$\mathbb{V}[\tilde{\tau}_r | \mathbf{X} = \mathbf{X}^{obs}] = \frac{\sigma_{r1}^2}{n_1} + \frac{\sigma_{r0}^2}{n_0} + (\bar{x}_1^{obs} - \bar{x}_0^{obs})^2 \mathbb{V}[\tilde{\beta}] + Cov\left(\bar{Y}_1 - \bar{Y}_0, (\bar{x}_1^{obs} - \bar{x}_0^{obs}) \tilde{\beta}\right). \quad (\text{A.1})$$

Here, we algebraically show that the covariance term is 0. Conditional on the observed covariates x_1^{obs} and x_0^{obs} , $\tilde{\beta}$ is defined as

$$\tilde{\beta} = \frac{\sum_{i:W_i=1} x_i^c Y_i^c + \sum_{i:W_i=0} x_i^c Y_i^c}{\sum_{i:W_i=1} (x_i^c)^2 + \sum_{i:W_i=0} (x_i^c)^2}, \quad (\text{A.2})$$

where $Y_i^c = Y_i^{obs} - \bar{Y}_{W_i}$ and $x_i^c = x_i - \bar{x}_{W_i}$ are centered outcome and covariate for unit i . Then, the covariance term is

$$\begin{aligned} & Cov\left(\bar{Y}_1 - \bar{Y}_0, (\bar{x}_1^{obs} - \bar{x}_0^{obs}) \frac{\sum_{i:W_i=1} x_i^c Y_i^c + \sum_{i:W_i=0} x_i^c Y_i^c}{\sum_{i:W_i=1} (x_i^c)^2 + \sum_{i:W_i=0} (x_i^c)^2}\right) \\ &= \frac{(\bar{x}_1^{obs} - \bar{x}_0^{obs})}{\sum_{i:W_i=1} (x_i^c)^2 + \sum_{i:W_i=0} (x_i^c)^2} Cov\left(\bar{Y}_1 - \bar{Y}_0, \sum_{i:W_i=1} x_i^c Y_i^c + \sum_{i:W_i=0} x_i^c Y_i^c\right), \end{aligned} \quad (\text{A.3})$$

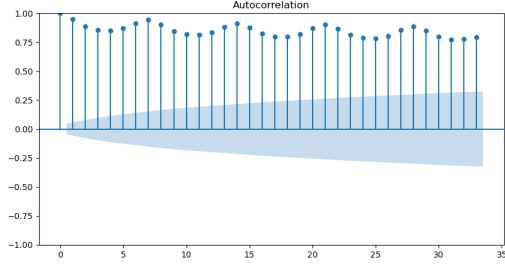
since we condition on the covariates hence they are considered as constants. We assume the units are independent of each other. Since a unit can only be assigned to either the control or the treatment group, $\bar{Y}_W = \frac{1}{n_W} \sum_{i:W_i=W} Y_i^{obs}$ does not share any common units with $\sum_{i:W_i=1-W} x_i^c Y_i^c$ and

$$Cov \left(\bar{Y}_1 - \bar{Y}_0, \sum_{i:W_i=1} x_i^c Y_i^c + \sum_{i:W_i=0} x_i^c Y_i^c \right) = Cov \left(\bar{Y}_1, \sum_{i:W_i=1} x_i^c Y_i^c \right) - Cov \left(\bar{Y}_0, \sum_{i:W_i=0} x_i^c Y_i^c \right). \quad (\text{A.4})$$

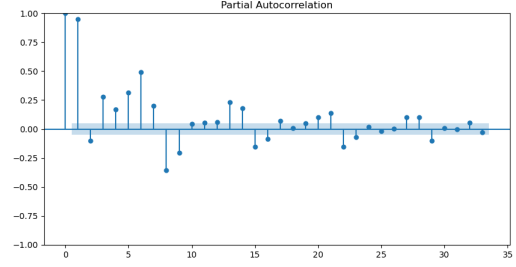
We consider the first covariance term in the above equation. Again, we use the assumption of independence between units and hence Y_i 's.

$$\begin{aligned} Cov \left(\bar{Y}_1, \sum_{i:W_i=1} x_i^c Y_i^c \right) &= Cov \left(\frac{1}{n_1} \sum_{i:W_i=1} Y_i^{obs}, \sum_{i:W_i=1} x_i^c Y_i^c \right) \\ &= \frac{1}{n_1} \sum_{i:W_i=1} x_i^c Cov(Y_i^{obs}, Y_i^c) \\ &= \frac{1}{n_1} \sum_{i:W_i=1} x_i^c Cov(Y_i^{obs}, Y_i^{obs} - \bar{Y}_1) \\ &= \frac{1}{n_1} \sum_{i:W_i=1} x_i^c (Cov(Y_i^{obs}, Y_i^{obs}) - Cov(Y_i^{obs}, \bar{Y}_1)) \\ &= \frac{1}{n_1} \sum_{i:W_i=1} x_i^c \left(\mathbb{V}[Y_i^{obs}] - \frac{1}{n_1} \mathbb{V}[Y_i^{obs}] \right) \\ &= \frac{1}{n_1} \sum_{i:W_i=1} x_i^c \left(\sigma_{r_1}^2 - \frac{1}{n_1} \sigma_{r_1}^2 \right) \\ &= \frac{(n_1 - 1) \sigma_{r_1}^2}{n_1^2} \sum_{i:W_i=1} x_i^c \end{aligned} \quad (\text{A.5})$$

By the definition of $x_i^c = x_i - \bar{x}_{W_i}$, we have that $\sum_{i:W_i=1} x_i^c = 0$ and Equation (A.5) reduces to 0. Similarly, the second covariance term in Equation (A.4) can be shown to equal to 0, and consequently Equation (A.3) is equal to 0. Therefore, the covariance term in (A.1) is equal to 0, and this idea extends to the case where the covariate is a vector of length greater than one. Note that we state $\mathbb{V}[Y_i^{obs}] = \sigma_{r_W}^2$ instead of σ_W^2 because we are conditioning on the covariates throughout our derivation.

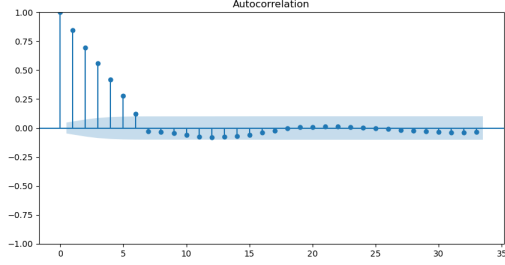


(a) ACF of page visits in training data.

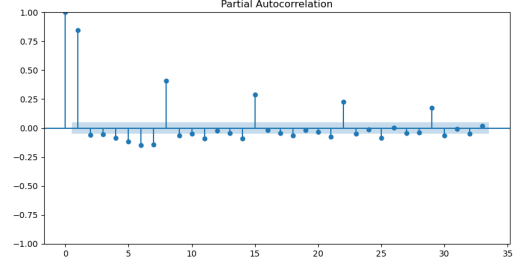


(b) PACF of page visits in training data.

Figure A.1: ACF and PACF of page visits in training data.



(a) ACF of seasonally differenced page visits in training data.



(b) PACF of seasonally differenced page visits in training data.

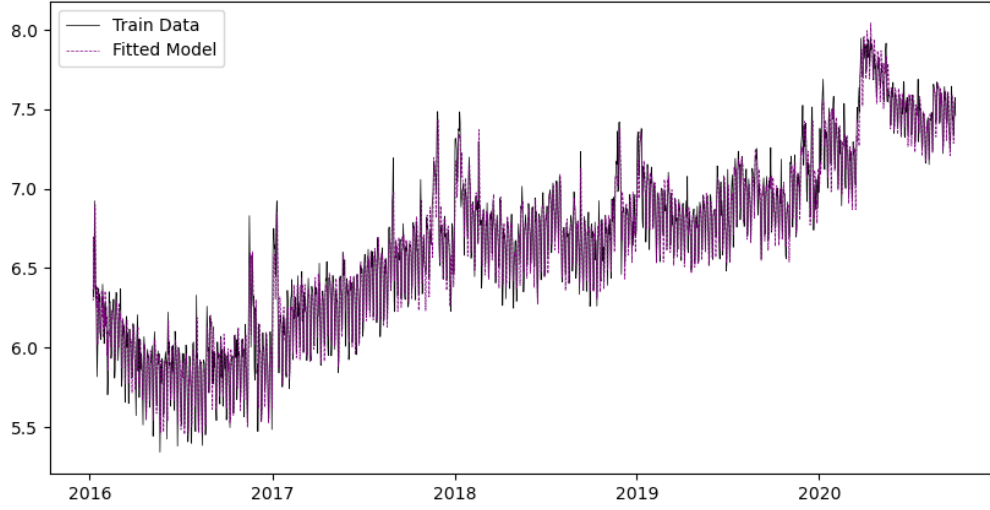
Figure A.2: ACF and PACF of seasonally differenced page visits in train data.

A.2 Fitting a SARIMA Model

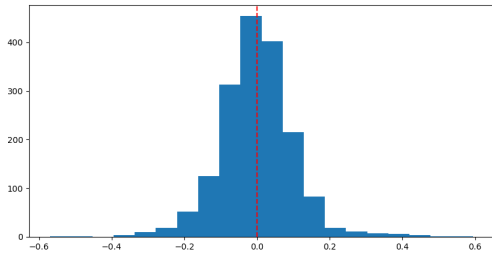
Given that this data is a time series, we will take the machine learning model $\tilde{f}$ to be a SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model. The SARIMA model is fitted by following the Box-Jenkins methodology. First, we analyze the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the pre-experiment data (training data) in Figure A.1. Seasonality appears to be present in the data with a period of $S = 7$. Figure A.1a shows a slightly decreasing trend, suggesting a need for a difference parameter $d = 1$.

Figure A.2 shows the ACF and PACF plots for seasonally differenced data. By analyzing the spikes and trends in these plots, we propose the following values for model parameters: $p \in \{0, 1\}$, $d \in \{1\}$, $q \in \{1, 6\}$, $P \in \{0, 1\}$, $D \in \{0, 1\}$, $Q \in \{0, 1\}$, and $S = 7$.

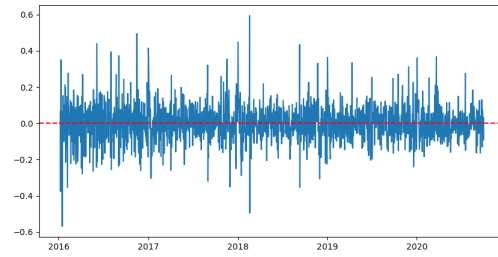
All possible combinations of these proposed values of parameters were considered, and the best model was chosen by comparing the Akaike information criterion (AIC). The optimal parameter values are as follows: $p = 0, d = 1, q = 6, P = 1, D = 0, Q = 1, S = 7$. This is surprising as seasonal differencing of $D = 1$ seemed intuitive. However, analyzing the residual distribution and the ACF/PACF in Figure A.3 seems to validate the optimality of this model. In Figures A.3b and A.3c, the residuals appear to be normally distributed and also randomly scattered around 0 with only minor fluctuations in variance over time. Hence, this is the model we use to predict the in-experiment outcomes.



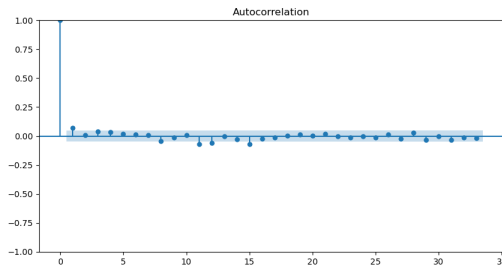
(a) Fitted values of the optimal model.



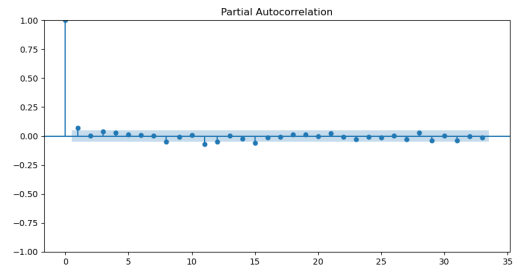
(b) Histogram of residuals.



(c) Residuals over time.



(d) ACF of residuals.



(e) PACF of residuals

Figure A.3: Optimal model fit with $p = 0, d = 1, q = 6, P = 1, D = 0, Q = 1, S = 7$. The residuals appear normally distributed and stationary.