An Asymptotically Powerful Test for the Average Treatment Effect

Emil Pitkin, Richard Berk, Lawrence Brown,
Andreas Buja, Edward George, Kai Zhang, Linda Zhao

Abstract

The average treatment effect (ATE) is a global measure of the effectiveness of an experimental treatment intervention. Regression based approaches are popular ways to estimate the ATE, especially in the context of randomized trials, because the inclusion of covariates can reduce the estimator's variance. Different frameworks lead to different estimators and associated standard errors, and we present a convenient regression derived ATE estimator for which covariates are treated as random not fixed, and minimal assumptions are placed on the respective response surfaces of the treatment and control groups. In particular, the regression is thought of only as a linear approximation to the response surface. We show that the estimator is asymptotically unbiased, and in a novel way derive its marginally valid asymptotic standard error. It is shown to underlie a more powerful test for the ATE than a classical estimator. Real and simulated data are used to illustrate typical gains from this estimator, as well as to shed light on the conditions when the gains are substantial.

1 Introduction

In the study of randomized controlled trials (RCTs), the average treatment effect (ATE) is a measure of an experimental intervention's global effect on a study population. For a treatment population response T and control population response C, the ATE is defined as $\tau = \mathbb{E}[T] - \mathbb{E}[C]$, where the measured responses can be continuous or categorical. While the simplest estimator is the difference of means $\bar{T} - \bar{C}$, there are many ways to estimate τ, depending on the sampling framework, choice of auxiliary information about the treated and control population, and target of inference. Even $\bar{T} - \bar{C}$ will have different standard errors depending on the sampling assumptions and scope of inference.

Our purpose in this paper is to clearly define an unrestrictive "assumption-lean" sampling and modeling framework, to define an ATE estimator within that framework, to explicitly derive its standard error, and to explicitly compare its asymptotic risk to the difference-in-means estimator. In order to properly position our work, we will briefly describe the principal strands of ATE estimation.

1.1 Brief development of ATE estimation

1.1.1 Fixed X, potential outcomes

The first strand of ATE study, first described by Neyman, focused on randomized experiments in fixed, finite populations [Splawa-Neyman et al., 1990]. The fixed subjects would be randomized either into treatments [1] or control. The only source of randomness came from the assignment of treatment condition, and inference extended only as far as to these subjects in the trial. Later developed by Rubin, this came to be known as the estimation of ATE within the "potential outcomes" framework, of which the $ATE_{interact}$ estimator from [Lin, 2013] is a recent example. This Neyman framework has since evolved to accommodate a superpopulation from which the experimental units can be thought to have been sampled [Imbens and Rubin, 2007].

1.1.2 Regression adjustments

More recently, ATE has been estimated via regression in order to improve its precision. Appealingly, yhe ATE can be explicitly written as a model parameter of the regression that describes the RCT, and the random discrepancies in treated and control covariate distributions are adjusted away. Within the regression framework, there are those who consider the covariates to be fixed, and others who consider them to be random; further, there is the choice of whether the statistical model as written down should be treated as true true. A fixed-X, true model approach was influentially critiqued by Freedman [Freedman, 2008], who showed that ATE estimators so derived can lead to reduced asymptotic precision, and can be beset by small-sample bias. In response, [Lin, 2013] while working in the potential outcome, fixed-X case, no longer assumes correct model specification, and shows that ATE estimators defined in this context have desirable properties. Another paper [Imbens and Wooldridge, 2008] analyzes ATEs under more flexible circumstances, allowing covariates to have a distribution and assuming heterogeneous effects. The authors present their useful results "assuming the linear regression model is correctly specified." More recently, [Samii and Aronow, 2012] compare the variances of the Neyman based and sandwich based estimators of the variance of the ATE, although the jump between fixed and random covariates is not obvious.

1.1.3 Assumption lean

More similarly to our framework, ATE has also been described in the context of semiparametric theory, as in [Yang and Tsiatis, 2001], [Tsiatis et al., 2008], [Zhang et al., 2008], [Rosenblum and van der Laan, 2010], where minimal assumptions are made about covariate distributions and correct model specification is not always assumed. Such works show the asymptotic optimality of particular ATE estimators in the case of model specification, though it is not clear if the derived standard errors fully account for the randomness of covariates. A separate and rich literature in statistics and econometrics, distinct however from our work, considers matching to improve the power of ATE estimates. For one good example of many, see [Abadie and Imbens, 2006] for a study of large sample properties of such estimators.

[1] "Treatments," because the number of treatments can exceed 1.

1.2 Our framework

With minimal assumptions, we derive efficient, explicit, asymptotically unbiased estimates for the unconditional average difference between the treatment and control groups. We show its difference from other estimators and improvement over the difference in means estimator, and therefore show that it is a more powerful test of the ATE.

We assume only that there exists a joint distribution between the covariates, the treatment indicator, and the response, and we do not assume that the model is correctly specified. Conveniently, the ATE can be estimated through least squares, and the targets of inference in the treatment and control regressions are the best linear approximations to their regression surfaces.

We position these results in the more realistic, assumption lean framework where the covariates are treated as random. This is appealing because the written statistical model rarely captures the data generating process, and subjects' covariates in an RCT should therefore be treated as random[2]. Moreover, for an analysis of an RCT to be useful, we want the scope of inference to extend outside of the sample in question.

In section 2 we define our framework and we define our ATE estimator. We then compare its performance to the difference in means estimator and show that it dominates, and consider further extensions. Section 3 illustrates the performance of the regression based estimator on real and simulated data, and compares it to two similar estimators. Section 4 concludes.

2 Framework

2.1 Assumption lean

In this section we make precise the mathematical framework in which we we define regressions for the treatment and control groups, and from which the ATE estimator is derived. It can be described as an "assumption-lean" framework, and the setup is borrowed from [Buja et al., 2015]. Specifically, we hold with Freedman, who writes that "randomization does not justify the assumptions behind the OLS model" [Freedman, 2008] and so relax assumptions to permit the subjects' covariates to be drawn from a distribution; and though we rely on OLS for estimation, we do not assume that linear relationships hold in the population. And, to reflect the possibly heterogeneous effects in the treatment and control groups, we include a full set of covariate-treatment indicator interactions.

Without loss of generality, consider the treated population. Let the population of subjects be described by the random variables $X_1, \ldots, X_p, Y$, with $\vec{\boldsymbol{X}} = (1, X_1, \ldots, X_p)'$ the random vector of predictor variables. The joint distribution between the predictors and the response $\mathbf{P} = \mathbf{P}(\mathbf{dx_1}, \ldots, \mathbf{dx_p}, \mathbf{dy})$ is assumed to have a full rank covariance matrix and four moments. We define the conditional mean of Y at $\vec{\boldsymbol{X}}$ by $\mu(\vec{\boldsymbol{X}}) = \mathbb{E}\left[Y|\vec{\boldsymbol{X}}\right]$. OLS assumptions are clearly relaxed: minimal assumptions are placed on errors, predictor variables can be omitted, and the true response surface is not assumed to be linear in the predictors (the operating assumption is that it is not).

[2]After patients have entered a clinical trial, nobody seriously presumes that other, putative patients in the target population have the same individual characteristics as the study subjects.

The conditional mean can be decomposed into a linear and a non-linear component. The linear component is the *best linear approximation* to the true conditional response surface; its partial slopes are defined by $\boldsymbol{\beta} = \left(\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right]\right)^{-1}\mathbb{E}\left[\boldsymbol{X}\mu(\boldsymbol{X})\right]$, where the expectation is over the joint distribution of the $\boldsymbol{X}$ and the Y. To foreshadow, the ATE can be represented by a partial slope thus defined. Finally, the difference between $\mu(\vec{\boldsymbol{X}})$ and $\boldsymbol{\beta}^T\vec{\boldsymbol{X}}$ is denoted by $\eta(\vec{\boldsymbol{X}})$, which is itself a random variable[3]. In equations:

$$\eta(\vec{\boldsymbol{X}}) = \mathbb{E}\left[Y|\vec{\boldsymbol{X}}\right] - (\beta^{(0)} + \vec{\boldsymbol{X}}'\boldsymbol{\beta}) \tag{1}$$

where $\beta^{(0)}$ is the first coordinate of $\boldsymbol{\beta}$ and represents the population intercept.

Estimation is straightforward: estimate $\boldsymbol{\beta}$ in the usual least squares fashion: $\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^TY$. With the population parameter $\boldsymbol{\beta}$ and its method of estimation thus defined, we present the results that we will appeal to in the paper:

1. $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a random variable with mean 0, where N is the sample size;
2. $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}$.

Admittedly, in finite samples, $\hat{\boldsymbol{\beta}}$ may be a biased estimator of $\boldsymbol{\beta}$.

2.2 Treatment and control populations

Recall that for treated response T and control response C, the target of estimation in our problem is $\tau = \mathbb{E}[T] - \mathbb{E}[C]$. In this section we define T and C as responses to separate regressions with random covariates. We note that the subjects of both the treatment and control groups are assumed to have been sampled at random from the same population, so that the treated and control covariate distributions are the same at the population level. With the notation from 2.1:

$$T_i = \beta_T^{(0)} + \vec{\boldsymbol{X}}'_{Ti}\boldsymbol{\beta}_T + \eta_T(\vec{\boldsymbol{X}})_i + \epsilon_{Ti} \tag{2}$$

and, analagously,

$$C_i = \beta_C^{(0)} + \vec{\boldsymbol{X}}'_{Ci}\boldsymbol{\beta}_C + \eta_C(\vec{\boldsymbol{X}})_i + \epsilon_{Ci} \tag{3}$$

We make three comments about the regression terms in the footnotes[4] [5] [6].

[3] Our operating assumption is that $\eta(\vec{\boldsymbol{X}})$ will not be identically equal to zero – that is, that non-linearity will be present in the population

[4] *Errors.* We place minimal demands on the errors: only that they should have zero mean and satisfy minimal moment conditions (there should be four moments). Because of iid sampling, they will be independent. Their distributional form is unspecified, and we do not assume normality of errors. We also allow their respective variances to differ, and denote the treated and control error variances, respectively, by σ_T^2 and σ_C^2.

[5] *Heterogeneity.* Note, also, that in the population, slopes are not assumed to be the same: we allow for heterogeneous effects. The random variables representing nonlinearity (η) are also allowed to differ between the treatment and control groups.

[6] *Population least squares* Because we no longer assume that the response is linear in the covariates, $\beta_T^{(0)} + \vec{\boldsymbol{X}}'_T\boldsymbol{\beta}_T$ should be thought of as the treated group's best linear approximation, at the population level, to $\mathbb{E}\left[T|\vec{\boldsymbol{X}}\right]$. So $\beta_T^{(0)}$ and $\boldsymbol{\beta}_T$ are population parameters derived from population least squares regression and minimize the expected squared distance between the linear surface and the true response surface.

3 The ATE estimator

3.1 ATE as regression parameter

In this section we show that the ATE (τ) is a parameter in the regression models and how it can be easily estimated. In the section to follow we compute its efficiency and compare it to the efficiency of the simple, difference-in-means estimator.

Subtracting (3) from (2) and taking expectations, we see that

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right) + \mathbb{E}\left[\vec{\boldsymbol{X}}_T\right]\boldsymbol{\beta}_T - \mathbb{E}\left[\vec{\boldsymbol{X}}_C\right]\boldsymbol{\beta}_C \tag{4}$$

By assumption, $\mathbb{E}\left[\vec{\boldsymbol{X}}_T\right] = \mathbb{E}\left[\vec{\boldsymbol{X}}_C\right] = \mathbb{E}\left[\vec{\boldsymbol{X}}\right]$, because the treated and control subjects are drawn from the same population, so (4) can be written as

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right) + \mathbb{E}\left[\vec{\boldsymbol{X}}\right](\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \tag{5}$$

Note that when $\mathbb{E}[X] = \mathbf{0}$, then $\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right)$, and the ATE is just the difference between the respective population intercepts. The question is how to estimate it optimally. If plug-in estimates are used in (4), with the empirical covariate means of the treatment and control groups as estimates for their population counterparts, the result is actually the simple difference-in-means estimator $\hat{\tau}_{\text{diff}} = \bar{T} - \bar{C}$. So controlling for covariates loses any advantage if no information is shared between the treatment and the control groups.

Instead, we estimate τ as in (5) by estimating $\mathbb{E}\left[\vec{\boldsymbol{X}}\right]$ by the complete set of pooled covariates: $\left(n_T\bar{\vec{\boldsymbol{X}}}_T + n_C\bar{\vec{\boldsymbol{X}}}_C\right)/N$. Substituting this single estimate into (5), we find that

$$\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right) + \frac{n_T\bar{\vec{\boldsymbol{X}}}_T + n_C\bar{\vec{\boldsymbol{X}}}_C}{N}\left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C\right)$$

When estimating, there is a way to preserve the convenient difference in intercepts interpretation of the ATE. The estimator $\hat{\tau}_{\text{regression}}$ is invariant to location, so it is convenient to mean center the covariates (with respect to the common, pooled mean), so that $\left(\vec{\boldsymbol{X}}_T\right)_i^* = \left(\vec{\boldsymbol{X}}_T\right)_i - \bar{\vec{\boldsymbol{X}}}$ and $\left(\vec{\boldsymbol{X}}_C\right)_i^*$ defined similarly. The ATE can therefore be estimated simply, via

$$\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{*(0)} - \hat{\beta}_C^{*(0)}\right) \tag{6}$$

which is just the difference of intercepts from the (mean centered) treatment and control regressions.

Theorem 3.1 prepares us for the following section where we compare the variances of different ATE estimators.

Theorem 3.1 *$\hat{\tau}_{regression}$ is an asymptotically unbiased estimator of τ.*

Corollary 3.2 *$\mathbb{E}[\hat{\tau}_{regression}] = \tau$ when*

1. *The population response is linear in the covariates, and all covariates have been included in the statistical model, or*

2. $\mathbb{E}[T|X] = \mathbb{E}[C|X] + k$, *and* $n_T = n_C$, *with* $k \in \mathbb{R}$.

The second condition says that if the treatment and control response functions are offset by a constant, then $\hat{\tau}_{\text{regression}}$ will be unbiased exactly, so long as the treatment and control sample sizes are equal. When they are unequal, unbiasedness can still be attained by inversely reweighting the observations. The proofs are deferred to the appendix.

3.1.1 Comments

Regression with interactions. The heterogeneous effects in the treatment and control regressions can be equivalently modeled in a single regression with an interaction term; the ATE is then an even more convenient parameter. Letting

$$I_T = \begin{cases} 1 & \text{Treatment is administered} \\ 0 & \text{Control is administered} \end{cases}$$

the response can be written as

$$Y_i = \beta^{(0)} + \beta^{(T)} I_T + \vec{\boldsymbol{X}}_i' \boldsymbol{\beta} + \vec{\boldsymbol{X}}_i' \boldsymbol{\beta}^{(Int)} I_T + \eta(\vec{\boldsymbol{X}})_i + I_T g(\vec{\boldsymbol{X}})_i + \epsilon_i \tag{7}$$

where $g(\vec{\boldsymbol{X}})$ is the difference in the treatment and control non-linearity functions, and I_T is the treatment indicator at the population level. Similarly to the discussion above, $\boldsymbol{\beta}^{(T)}$ is the ATE when the covariate expectation is zero. Accordingly, $\hat{\beta}^{(T)}$ is the estimated ATE when the empirical covariates are first mean-centered[7].

Arbitrary response. The analysis is not altered if the T_i, C_i are assumed to be count data, or to take on values $0, 1$. When the response is binary, for example, the target of estimation is still $\mathbb{E}[T] - \mathbb{E}[C]$, but these terms can be rewritten as $P(T) - P(C)$, where $P(T)$ represents the proportion of treatment outcomes in the population that take on the value 1.

If one estimates $\hat{\tau}$ by $\bar{T} - \bar{C}$, the estimate $\hat{P}(T) - \hat{P}(C)$ will fall inside $[-1, 1]$ But since $\hat{\tau}_{\text{regression}}$ estimates the response Y not at the respective sample means of the covariates $\vec{\boldsymbol{X}}_T$ and $\vec{\boldsymbol{X}}_C$ but at the weighted average $\frac{n_T \bar{\vec{\boldsymbol{X}}}_T + n_C \bar{\vec{\boldsymbol{X}}}_C}{N}$, $\hat{P}(T) - \hat{P}(C)$ is not guaranteed with probability 1 to be restricted to $[-1, 1]$. The problem arises in the unlikely case that there is limited overlap between the observed treatment and control covariates and the slope coefficients differ significantly between the two groups.

3.2 Relative performance of ATE estimators

In this section we explicitly write down the asymptotic variance of $\hat{\tau}_{\text{regression}}$ and write down the asymptotic variance of $\hat{\tau}_{\text{diff}}$ in an unconventional way, and show that $\hat{\tau}_{\text{regression}}$ dominates the $\hat{\tau}_{\text{diff}}$ in asymptotic risk. By explicitly writing out the variance expressions, we give practitioners the tools to execute this more powerful test for the ATE. The expressions are presented one after the other for ease of comparison:

[7]The *predicted* response, represented by a single regression with interactions, looks like $\hat{Y}_i = \hat{\beta}^{(0)} + \hat{\beta}^{(T)} I_T + \vec{\boldsymbol{X}} \hat{\boldsymbol{\beta}} + \vec{\boldsymbol{X}} \hat{\boldsymbol{\beta}}^{(Int)} I_T$

Lemma 3.3

$$Var\left(\hat{\tau}_{diff}\right) = \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T}\left[\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\right] \quad (8)$$

Lemma 3.4

$$Var(\hat{\tau}_{regression}) = \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + O(N^{-2}) + \frac{1}{N}(\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \Sigma_X (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \quad (9)$$

where Σ_X is the variance-covariance matrix of the predictors. The proofs are deferred to the appendix.

The more familiar expression for $\mathrm{Var}\left[\hat{\tau}_{\mathrm{diff}}\right]$ is not 8 but rather $\mathrm{Var}\left[T\right]/n_T + \mathrm{Var}\left[C\right]/n_C$. We chose to express the variance as above for the purpose of term by term comparison to $\mathrm{Var}\left[\hat{\tau}_{\mathrm{regression}}\right]$. It was found by first conditioning on the regression covariates, and then marginalizing over their distribution.

The respective standard deviations of $\hat{\tau}_{\mathrm{diff}}$ and $\hat{\tau}_{\mathrm{regression}}$ can be estimated, in an asymptotically unbiased fashion, by

$$\widehat{SE}(\hat{\tau}_{\mathrm{diff}}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{n_T}\left(\hat{\boldsymbol{\beta}}_T \hat{\Sigma}_X^{(T)} \hat{\boldsymbol{\beta}}_T\right) + \frac{1}{n_C}\left(\hat{\boldsymbol{\beta}}_C \hat{\Sigma}_X^{(C)} \hat{\boldsymbol{\beta}}_C\right)} \quad (10)$$

$$\widehat{SE}(\hat{\tau}_{\mathrm{regression}}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{N}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)' \hat{\Sigma}_X (\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)} \quad (11)$$

In the above estimates, MSE_T is the mean square error computed in the treatment regression, defined as usual by $MSE_T = \left(\sum_{i=1}^{n}(T_i - \hat{T}_i)^2\right)/(N - p - 1)$, $\hat{\Sigma}_X$ is the empirical variance-covariance matrix of the complete collection of covariates, and $\hat{\Sigma}_X^{(T)}$ of the treatment covariates[8].

The main claim now follows: the asymptotic variance of the regression-based estimator dominates the variance of the naive estimator:

Theorem 3.5

$$AVar(\hat{\tau}_{diff}) \geq AVar(\hat{\tau}_{regression}) \quad (12)$$

The proof is found in the appendix.

A comparison of just the asymptotic variances suffices because both estimators are unbiased: $\hat{\tau}_{\mathrm{diff}}$ trivially, and $\hat{\tau}_{\mathrm{regression}}$ asymptotically according to 3.1. It therefore follows that $\hat{\tau}_{\mathrm{regression}}$ has a lower asymptotic risk than $\hat{\tau}_{\mathrm{diff}}$ and is therefore more asymptotically efficient.

[8]Remark: the mean squared error is a scaled estimate of the total variability in the response that is not captured by the linear approximation. So MSE_T estimates two components: the variability in the structural errors $\epsilon_T = \sigma_T^2$, together with the variability of $\eta(\vec{\boldsymbol{X}})$, the random variable measuring the non-linearity in the conditional mean.

3.2.1 Additional remarks

Variance estimates. Theorem 3.5 concerns the true variance of the respective estimators, rather than to their estimated variances[9]. The theorem could analogously have been written, and should be seen here for clarity, as

$$\mathbb{E}\left[\widehat{Var}(\hat{\tau}_{\text{diff}})\right] \geq \mathbb{E}\left[\widehat{Var}(\hat{\tau}_{\text{regression}})\right]$$

Equality of variances. The inequality in theorem 3.5 is not strict. Equality between the asymptotic variances can is attained iff $\boldsymbol{\beta}_C = -\frac{n_C}{n_T}\boldsymbol{\beta}_T$ (see appendix). When the treatment and control sample sizes are equal, then equality is attained when $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_T$.

3.2.2 Conditional versus marginal estimation

The standard errors we have derived are valid for marginal inference for the ATE. This brief section is dedicated to making explicit the difference between conditional and marginal inference in ATE estimation. It is needed because some authors describe a framework with random covariates where marginal inference is warranted, but give expressions for standard errors that ignore the variability of the covariates.

Familiarly, the variance of $\hat{\tau}_{\text{diff}}$ is a marginal variance, measuring the variability of the estimator over all possible replications of the experiment, irrespective of any measured or unmeasured covariates:

$$\text{Var}\left[\hat{\tau}_{\text{diff}}\right] = \frac{Var[T]}{n_T} + \frac{Var[C]}{n_C}. \tag{13}$$

and is estimated, unbiasedly, by ${s^2}_T/n_T + {s^2}_C/n_C$.

We will now compute the variance of $\bar{T} - \bar{C}$ in a conditional manner and highlight the difference from the marginal result. Define the treatment and control regression as in (2) and (3) and estimate the treatment and response, respectively, by $\hat{T} = \hat{\beta}_T^{(0)} + \vec{\boldsymbol{X}}_T\hat{\boldsymbol{\beta}}_T$ and $\hat{C} = \hat{\beta}_C^{(0)} + \vec{\boldsymbol{X}}_C\hat{\boldsymbol{\beta}}_C$. Now estimate the treatment response at the mean of the treatment predictors, and likewise for control. Then, from elementary regression, $\hat{T}_i\Big|_{\vec{\boldsymbol{X}}_T=\vec{\boldsymbol{X}}_{\boldsymbol{T}}} = \bar{T}$, and $\hat{C}_i\Big|_{\vec{\boldsymbol{X}}_C=\vec{\boldsymbol{X}}_{\boldsymbol{C}}} = \bar{C}$.

Classical regression theory applied to this estimator will give a different variance from (13) because it considers the response conditional on the observed covariates. WLOG consider $\bar{T}$, which was arrived at by estimating the response at $\vec{\boldsymbol{X}}_T = \vec{\boldsymbol{X}}_{\boldsymbol{T}}$. To fix the idea, consider simple regression where the prediction variance is given by

$$\hat{Var}[\hat{y}|X = x_p] = MSE\left[1 + \frac{1}{n_T} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n_T}(x_i - \bar{x})^2}\right] \tag{14}$$

At the covariate mean,

$$\hat{Var}[\hat{T}|\vec{\boldsymbol{X}} = \vec{\boldsymbol{X}}_T] = MSE\left[1 + \frac{1}{n_T}\right] \tag{15}$$

[9]Which means that in a given sample, $\widehat{SE}(\hat{\tau}_{\text{regression}})$ may exceed $\widehat{SE}(\hat{\tau}_{\text{diff}})$

which does not uniformly equal $\frac{s^2_T}{n_T}$[10].

The variance estimated in (15) ignores the fact that the covariate mean, at which the response is estimated, is a random quantity, since inference was done conditional on the covariates. With that in mind, when we computed the variance of $\hat{\tau}_{\text{regression}}$, we incorporated the variability of the covariates, even though the estimator can be thought of as a coefficient in a regression model. Our results, then, are valid marginally.

3.3 Alternative Conditions

3.3.1 Distribution of $\boldsymbol{X}$ known

Throughout the discussion and analysis, we have assumed that the underlying distribution of $\vec{\boldsymbol{X}}$ is unknown. That distribution, however, might be known: in practice, covariates like age, weight and income, or for which measurements might exist for the whole population, can be adjusted for in the study. In such a case, the component of the variance due to estimating $\mathbb{E}[\boldsymbol{X}]$ is removed, and only the regression slopes remain to be estimated. The standard error estimate of the ATE diminishes correspondingly. The appendix quantifies the precise degree to which the standard error diminishes when the mean of $\vec{\boldsymbol{X}}$ is known.

3.3.2 Treatment correlated with covariates

In the preceding discussion, we had assumed that the assignment of treatment (the treatment indicator) was independent of the covariates. Any correlation between treatment indicator and covariates would only appear in data and be the result of random sampling. In practice, however, the decision to administer treatment might depend on the covariates: perhaps, by design and because of cost constraints in the study, a researcher will wish to offer treatment which is expensive to subjects whose covariates suggest they will require it for a shorter duration.

To mathematically define this scenario, suppose that the regression is written as in 7, except that $I_T = H(\vec{\boldsymbol{X}})$, either deterministically or stochastically, as when $I_T \sim Bern\left(H\left(\vec{\boldsymbol{X}}\right)\right)$. The treatment indicator becomes a function of the covariates so the assignment mechanism is different across different strata. In the simplest case, the functional form of $H(\cdot)$ is known, so that $\pi_i = P\left(I_T = 1|\vec{\boldsymbol{X}}\right)$ does not need to be estimated.

With the goal of estimating the ATE, an inverse probability weighting scheme is natural because it can reduce the bias that would result from the differing sampling regimes across strata. Accordingly, reweight the observed response y_i according to

$$y_i^{(T)*} = \frac{y_i^{(T)}}{\pi_i}$$

with π_i defined as above for the treated units, and

[10]The two estimated variances will be equal only when the R^2 from the regression equals $\frac{p+2}{n_T+1}$, where p is the number of covariates[11]. When R^2 is larger, then the regression based estimated variance will be smaller than that of the marginal, conventional estimated variance.

$$y_i^{(C)*} = \frac{y_i^{(C)}}{1-\pi_i}$$

Such a reweighting has been considered by, for example, [Freedman and Berk, 2008], except the functional relationship between the confounders and the treatment indicator was unknown and was consequently estimated via propensity scores. Our future work will extend to cases when this functional relationship needs to be estimated. One proceeds with the analysis as before, running the two separate treated and control regressions, estimating the (weighted response) at the pooled mean of the covariates, and taking the difference.

Another estimate of the ATE would be

$$\frac{1}{n_T}\sum_{i=1}^{n} y_i^{(T)*} - \frac{1}{n_C}\sum_{i=1}^{n} y_i^{(C)*} \tag{16}$$

what [Freedman and Berk, 2008] call a weighted contrast, and is the weighted variant of the difference in means estimator considered earlier. The latter is a Horvitz-Thompson type estimator (the formal H-T estimator assumes a finite population from which one samples). The derivations and analysis relating to the weighted scheme are beyond the scope of the current paper, and will be considered in depth in a forthcoming work.

4 Illustrations on data

4.1 Simulated data

We have shown that $\hat{\tau}_{\text{regression}}$ dominates $\hat{\tau}_{\text{diff}}$ asymptotically. We have found that in small samples, however, R^2 should be larger than about 0.2 in order for our estimator to be more efficient[12]. In this brief section we observe the relative efficiencies of the two estimators as a function of R^2. The model chosen vividly illustrates the relationship. Define the treated and control groups respectively by

$$\begin{aligned} T &= 3X_1 - X_2 + Z_T \\ C &= X_1 + X_2 + Z_C \end{aligned}$$

In accord with the framework adopted by this paper, the variables are random not fixed. We allowed

$$\begin{aligned} X_1 &\sim Lognormal(0,1) \\ X_2 &\sim Gamma(3,4) \\ Z_T, Z_C &\overset{iid}{\sim} N(0, V_Z) \end{aligned}$$

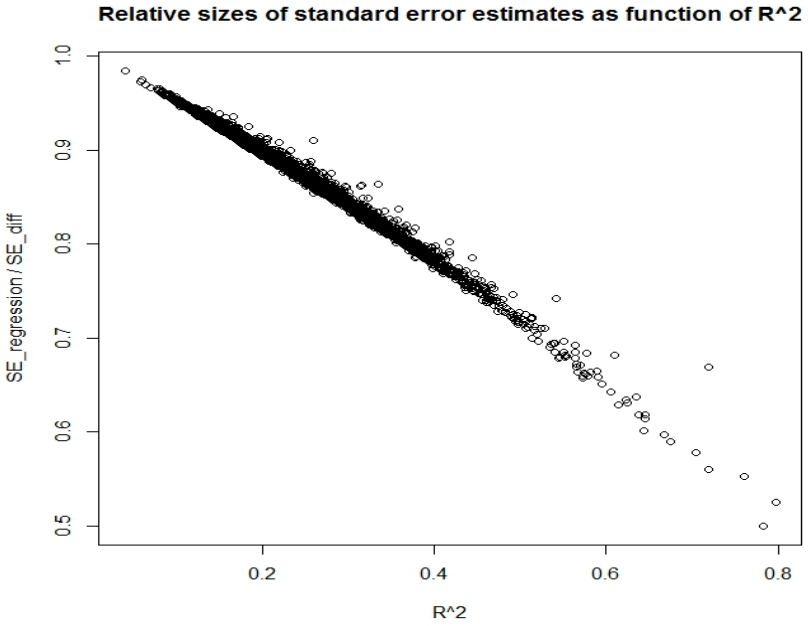
As stated previously, $\hat{\tau}_{\text{diff}}$ depends only on the response, so that estimate does not depend on the quality of the regression fit. $\hat{\tau}_{\text{regression}}$, however, does, so by manipulating the R^2 of

[12] In practice, RCTs usually don't have an R^2 that exceeds 0.25

the model, we can see the effect on the relative efficiencies of the estimators. The R^2 was manipulated by varying V_Z.

We ran 10,000 simulations with 250 treated and 250 control units in each. Sequentially, V_Z was dialed from 1 to 100 (100 simulations for each value of the variance), with attendant decreases in the R^2 on account of the increased noise. In each simulation, the ratio of the standard errors $\hat{SE}\left(\hat{\tau}_{\text{regression}}\right)/\hat{SE}\left(\hat{\tau}_{\text{diff}}\right)$ was computed, as well as the R^2 of the combined model with interactions. The plot below gives a sense of the magnitude of the efficiency gains as a function of R^2. As R^2 decreases, the estimated standard errors converge as expected. For high R^2, for example above 0.5, the $\hat{\tau}_{\text{regression}}$ enjoys a dramatically lower standard error.

Figure 1: R^2 plotted against $\frac{\widehat{SE}\left(\hat{\tau}_{\text{regression}}\right)}{\widehat{SE}\left(\hat{\tau}_{\text{diff}}\right)}$



A numeric illustration fixes the point. In the model above, the target of estimation $ATE = \mathbb{E}[T] - \mathbb{E}[C] = 2e^{1/2} - 3/2 = 1.797$. Again we simulated 10,000 draws, with $V_Z = 3$, balanced groups of 250 treated and 250 control units in each simulation. We recorded the R^2 of the combined regression, as well as the ATE and SE estimates for both estimators considered in this paper.

The average R^2 in the 10,000 simulation was 0.75. With this good fit, the average $\hat{SE}\left(\hat{\tau}_{\text{diff}}\right) = 0.452$ (with simulation SE = 0.0011), more than 80% larger than while the average $\hat{SE}\left(\hat{\tau}_{\text{regression}}\right) = 0.249$ (with simulation SE = 0.0002). Both estimators were unbiased (up to simulation granularity), with difference-in-mean and regression-based average ATEs equal to 1.793 and 1.795, respectively. Using $\Phi^{-1}(0.975)$ as the multiplier, coverage of the true ATE (1.797) was equal to 0.9473 and 0.949, respectively.

There was nothing particular about the model chosen; similar phenomena are observed for other choices of underlying distribution. The lesson is once again that the regression based estimate leads to a more powerful test.

4.1.1 Relation to other estimators

The last decade has seen many alternate estimators of the ATE, each with its own desirable properties. In this section we compare and contrast $\hat{\tau}_{\text{regression}}$ and its precision to its counterparts.

The $\widehat{ATE}_{OLS}$ estimator from [Bloniarz et al., 2015], corresponding to the $\widehat{ATE}_{interact}$ from [Lin, 2013], is written down like $\hat{\tau}_{\text{regression}}$ (in the sense that it can be viewed as a difference of intercepts from two regressions with heterogeneous effects) but these estimators above have numerically different standard errors than $\hat{\tau}_{\text{regression}}$ because the sources of randomness are different. In 2013, [Lin, 2013] reexamined Freedman's critique in the context of the Neyman model of randomization inference, which is the finite population, potential outcomes formulation. The randomness comes from which of the potential outcomes is observed. The target of estimation is the average difference across this population between the two potential responses.

The analog to the difference between the [Lin, 2013] estimator and $\hat{\tau}_{\text{regression}}$ is the difference between sample-based estimates and super-population estimates in sampling theory. As a simple illustration of the difference, the same model as in 4.1 was used to generate one sample of 500 cases. In each of 10,000 replications, 250 were randomly assigned treatment, and the rest were assigned control. The empirical standard deviation of the $\widehat{ATE}_{interact}$ estimator was 0.41, as compared to 0.25 for $\hat{\tau}_{\text{regression}}$. In other circumstances with lower R^2 the empirical standard deviation of $\widehat{ATE}_{interact}$ should be lower. In any case, it is different.

Another group of authors [Rosenblum and van der Laan, 2010] arrive at an estimator equivalent to $\hat{\tau}_{\text{regression}}$ through targeted maximum likelihood and appeals to general semiparametric theory (see their section 3). Impressively, they derive $\hat{\tau}_{\text{regression}}$ as a special case of ATE estimators for generalized linear models regression though they assume that the response covariate is bounded. They show that when the working model is correctly specified, the estimator is optimal in terms of asymptotic mean square error, but do not show optimality in the case of misspecification. In contrast, our variance expressions are explicitly written down (and remain valid under misspecification), so that, extending the work of [Rosenblum and van der Laan, 2010], it is easy for us to show by how much the improved estimator is better than the baseline difference in means estimator.

We are also indebted to the work of Tsiatis et al. [Tsiatis et al., 2008], whose $ANCOVA_2$ estimator (first defined in [Yang and Tsiatis, 2001]), like $\hat{\tau}_{\text{regression}}$, is based on the regression model with interactions (they consider it as a case of an augmentation estimator) and is notationally equivalent. However, their practical implementation of computing the ATE estimator's variance estimate (see their section 4) differs from ours. Their method, among other things, appeals to the sandwich estimator. For example, again using the model in 4.1, the standard error estimate using the formula from [Tsiatis et al., 2008] is 0.33, compared to our estimate of 0.25. In other models, it may be smaller. In any case, it is different. In a subsequent paper, [Zhang et al., 2008] impressively introduce a general class of estimators for which $ANCOVA_2$ is a special case. We amplify this work by making explicit the comparison to the difference in means estimator and by deriving an explicit, marginally valid formula for the standard errors and their estimates.

4.2 Illustration on real data

We have found that in typical RCT examples, the improvement enjoyed by $\hat{\tau}_{\text{regression}}$ over $\hat{\tau}_{\text{diff}}$ is small but real. We observe the comparison on data furnished from a classic study discussed in [LaLonde, 1986] and reanalyzed in [Dehejia and Wahba, 1999]. The data come from the National Support Work (NSW) Demonstration, where a pool of adults with economic and social problems was randomized into two groups. The treated group was offered job training while the control group was not. The response measured was earnings in 1978, after the job training had concluded. The covariates that were adjusted for included: age, education (number of years), an indicator for Black race, indicator for Hispanic race, indicator for marital status, indicator for attainment of high school degree, and earnings in 1974.

The intent of the work in [LaLonde, 1986] was to compare ATE estimates from experiments to those from observational studies. He compared the unbiased estimate of the ATE from NSW groups to an estimate drawn by comparing the treated adults to a batch of controls collected from separate comparison groups (PSID-1 and CPS-1 in his paper). Dehejia and Wahba [Dehejia and Wahba, 1999] apply matching techniques for this comparison; relevant for our work are the 297 treated and 260 control male subjects they analyze, and which are available from the original NSW experiment.

In this experimental context we compare the ATE estimates and standard errors. Here, $\hat{\tau}_{\text{diff}} = \$4709.4$, while the point estimate from our estimator is $\hat{\tau}_{\text{regression}} = \$4435.2$. The respective standard errors are \$443.5 and \$431.9. The reduction in SE amounts to 3.1%, with the R^2 of the regression of salary in 1978 on covariates and their interaction with the treatment indicator equal to 0.24. A gain of this magnitude is typical for an RCT with an R^2 of this size. As illustrated before, a higher R^2 would have resulted in higher SE gains.

5 Conclusion

We lay the foundation for conducting principled and efficient asymptotic inference on ATEs. In an infinite population, random design, regression based estimation, where the response surface needn't be linear, we showed how the ATE is equivalent to a regression parameter and wrote down how to estimate it. We subsequently derived explicit standard errors for $\hat{\tau}_{\text{regression}}$, the regression adjusted estimator, and directly showed how it dominates $\hat{\tau}_{\text{diff}}$, the difference in means estimator, and therefor underlies a more powerful test fo the ATE. Interestingly, despite the added source of variability from the randomness of the covariates, the derived standard error, which also adjusts for covariates, is in expectation actually lower than its conventional counterpart.

Bootstrapped confidence intervals can easily be generated and inference conducted for the population ATE. Moreover, the paired bootstrap, mimicking as it does the random X framework, is the natural technique for such intervals. Future work will focus on weighting schemes when the treatment is correlated with covariates, as it would be, for example, in observational studies. In this work we estimated with linear models. We hope to extend the work, including explicitly written standard errors, to GLMs.

6 Technical appendix

Proof of 3.1

After mean centering, $\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right)$. Direct application of proposition 5.4 in [Buja et al., 2015] shows that the difference of the independent quantities $\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}$ is an unbiased estimate of $\beta_T^{(0)} - \beta_C^{(0)}$, which is equal to τ when $\boldsymbol{\mu} = \mathbf{0}$.

Proof of 3.2

(a) When the regression model is correctly specified, then it is an introductory result that the LS estimates are unbiased: $\mathbb{E}\left[\hat{\beta}_T^{(0)}\right] = \beta_T^{(0)}$ and that $\mathbb{E}\left[\hat{\beta}_C^{(0)}\right] = \beta_C^{(0)}$, so $\mathbb{E}\left[\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right] = \beta_T^{(0)} - \beta_C^{(0)} = \tau$.

(b) Suppose that the treatment and response surfaces have a constant offset: $n_T = n_C$ and $\mathbb{E}[T|X] = \mathbb{E}[C|X] + k$. In the decomposition of $\hat{\tau}_{\text{regression}} - \tau$ in the proof of 3.4, the only term which does not generally have expectation $\mathbf{0}$ is the term denoted by R_2, and equal to $\left[\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right]\left[p_C(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T) + p_T(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C)\right]$. It will have expectation 0 when the two bracketed terms are uncorrelated. Exploiting the independence between the treated and control groups, the bracketed terms will be uncorrelated iff

$$p_C Cov\left(\bar{\boldsymbol{X}}_T, \hat{\boldsymbol{\beta}}_T\right) = p_T Cov\left(\bar{\boldsymbol{X}}_C, \hat{\boldsymbol{\beta}}_C\right) \tag{17}$$

Inversely weight the observations, giving weight $\frac{1}{n_T}$ to the control observations, and $\frac{1}{n_C}$ to the treatment, so that (17) will hold true when $Cov\left(\bar{\boldsymbol{X}}_T, \hat{\boldsymbol{\beta}}_T\right) = Cov\left(\bar{\boldsymbol{X}}_C, \hat{\boldsymbol{\beta}}_C\right)$ When $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$, then, since the $\bar{\boldsymbol{X}}_T$ and $\bar{\boldsymbol{X}}_C$ are identically distributed, the above equality will hold. $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$ when there is a constant offset.

Proof of 3.3

The conventional estimator of the ATE is $\hat{\tau}_{\text{diff}} = \bar{T} - \bar{C}$. Assume the covariates have zero mean; then its difference from the true ATE equals

$$\begin{aligned} \hat{\tau}_{\text{diff}} - \tau &= \bar{T} - \bar{C} - \left(\beta_T^0 - \beta_C^0\right) \\ &= \left[\bar{T} - \left(\beta_T^0 + \bar{X}_T\boldsymbol{\beta}_T\right)\right] - \left[\bar{C} - \left(\beta_C^0 + \bar{X}_C\boldsymbol{\beta}_C\right)\right] \\ &+ \bar{X}_T\boldsymbol{\beta}_T - \bar{X}_C\boldsymbol{\beta}_C \end{aligned} \tag{18}$$

The two terms – the former the residual means, and the latter a function of the covariates – are independent. Hence

$$
\begin{aligned}
Var\left(\hat{\tau}_{\text{diff}}\right) &= Var\left\{\left[\bar{T}-\left(\beta_T^0+\bar{X}_T\boldsymbol{\beta}_T\right)\right]-\left[\bar{C}-\left(\beta_C^0+\bar{X}_C\boldsymbol{\beta}_C\right)\right]\right\} \\
&+ Var\left\{\bar{X}_T\boldsymbol{\beta}_T-\bar{X}_C\boldsymbol{\beta}_C\right\} \\
&= \left[\frac{\sigma_T^2+Var[\eta_T]}{n_T}+\frac{\sigma_C^2+Var[\eta_C]}{n_C}\right]+\frac{1}{n_T}\left[\boldsymbol{\beta}_T'\Sigma_{X_T}\boldsymbol{\beta}_T\right]+\frac{1}{n_C}\left[\boldsymbol{\beta}_C'\Sigma_{X_C}\boldsymbol{\beta}_C\right] \\
&= \left[\frac{\sigma_T^2+Var[\eta_T]}{n_T}+\frac{\sigma_C^2+Var[\eta_C]}{n_C}\right]+\frac{1}{n_T}\left[\boldsymbol{\beta}_T'\Sigma_X\boldsymbol{\beta}_T\right]+\frac{1}{n_C}\left[\boldsymbol{\beta}_C'\Sigma_X\boldsymbol{\beta}_C\right]
\end{aligned}
$$

as the covariance matrices of the treatment and control distributions are equal, since the covariates are drawn from the same distribution.

Proof of 3.4

As before, we allow for unequal randomization, so that n_T cases receive treatment, and n_C cases receive control; denote the proportions p_T and p_C, respectively, and suppose that $\mathbb{E}[X]=\boldsymbol{\mu}$ and $Var[X]=\Sigma$. Denote the ATE in the population by τ. It equals $\mathbb{E}[T]-\mathbb{E}[C]=(\beta_T^0-\beta_C^0)+\boldsymbol{\mu}(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C)$. So

$$
\begin{aligned}
\hat{\tau}_{regression} &= \hat{\beta}_T^0-\hat{\beta}_C^0+\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\beta}}_T-\hat{\boldsymbol{\beta}}_C) \\
\hat{\tau}_{regression} &= \hat{\beta}_T^0-\hat{\beta}_C^0+\left[p_T\bar{\boldsymbol{X}}_T+p_C\bar{\boldsymbol{X}}_C\right](\hat{\boldsymbol{\beta}}_T-\hat{\boldsymbol{\beta}}_C) \\
&= \bar{T}-\bar{\boldsymbol{X}}_T\hat{\boldsymbol{\beta}}_T-(\bar{C}-\bar{\boldsymbol{X}}_C\hat{\boldsymbol{\beta}}_C)+\left[p_T\bar{\boldsymbol{X}}_T+p_C\bar{\boldsymbol{X}}_C\right](\hat{\boldsymbol{\beta}}_T-\hat{\boldsymbol{\beta}}_C) \\
&= \bar{T}-\bar{C}-\left(\bar{\boldsymbol{X}}_T-\bar{\boldsymbol{X}}_C\right)\left(p_C\hat{\boldsymbol{\beta}}_T+p_T\hat{\boldsymbol{\beta}}_C\right)
\end{aligned}
$$

The multivariate mean can be taken to equal $\mathbf{0_p}$ WLOG since the problem is one of scale, rather than location. So

$$
\begin{aligned}
\hat{\tau}_{regression}-\tau &= \bar{T}-\bar{C}-\left(\bar{\boldsymbol{X}}_T-\bar{\boldsymbol{X}}_C\right)\left(p_C\hat{\boldsymbol{\beta}}_T+p_T\hat{\boldsymbol{\beta}}_C\right)-\beta_T^0+\beta_C^0 \\
&= \left[\bar{T}-(\beta_T^0+\bar{\boldsymbol{X}}_T\boldsymbol{\beta}_T)\right]-\left[\bar{C}-(\beta_C^0+\bar{\boldsymbol{X}}_C\boldsymbol{\beta}_C)\right] \\
&- \left(\bar{\boldsymbol{X}}_T-\bar{\boldsymbol{X}}_C\right)\left[p_C(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_T)+p_T(\hat{\boldsymbol{\beta}}_C-\boldsymbol{\beta}_C)\right] \\
&+ (p_T\bar{\boldsymbol{X}}_T+p_C\bar{\boldsymbol{X}}_C)(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C) \\
&= R_1+R_2+R_3 \qquad (19)
\end{aligned}
$$

R_1, R_2, and R_3 are uncorrelated: R_1 is a function of the errors, which are independent of the covariates, whereas R_2 and R_3 lie in the column space of the covariates. That R_2 is uncorrelated with R_3 is verified algebraically.

Each of the R_i has expectation $\mathbf{0_p}$: the first, R_1, is a difference of average errors, equal to $\left(\bar{\epsilon}_T+\bar{f}_T\right)-\left(\bar{\epsilon}_C+\bar{f}_C\right)$. The ϵ have expectation 0 by assumption, and the f by construction. R_2 is asymptotically equal to $\mathbf{0}$, for the following reason: the treatment and controls are uncorrelated, and $\mathbb{E}\left[\bar{\boldsymbol{X}}\right]=\mathbf{0}$, so the only component of R_2 not equal for all n to $\mathbf{0}$ in expectation is $p_C\bar{\boldsymbol{X}}_T\hat{\boldsymbol{\beta}}_T-p_T\bar{\boldsymbol{X}}_C\hat{\boldsymbol{\beta}}_C$. We'll now show that $\mathbb{E}\left[\bar{\boldsymbol{X}}_T\hat{\boldsymbol{\beta}}_T\right]\to\mathbf{0}$:

$$
\begin{aligned}
\mathbb{E}\left[\bar{\boldsymbol{X}}_T\hat{\beta}\right] &= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\mathbb{E}\left[\hat{\beta}|\boldsymbol{X}_T\right]\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\left(\boldsymbol{X}'_T\boldsymbol{X}_T\right)^{-1}\boldsymbol{X}'_T\mathbb{E}\left[Y|\boldsymbol{X}_T\right]\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\left(\boldsymbol{X}'_T\boldsymbol{X}_T\right)^{-1}\boldsymbol{X}'_T\left(\boldsymbol{X}_T\boldsymbol{\beta}_T+\eta_T(\boldsymbol{X}_T)\right)\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\left(\boldsymbol{X}'_T\boldsymbol{X}_T\right)^{-1}\boldsymbol{X}'_T\boldsymbol{X}_T\boldsymbol{\beta}_T+\bar{\boldsymbol{X}}_T\left(\boldsymbol{X}'_T\boldsymbol{X}_T\right)^{-1}\eta_T(\boldsymbol{X}_T)\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\boldsymbol{\beta}_T\right]+\mathbb{E}\left[\bar{\boldsymbol{X}}_T\left(\boldsymbol{X}'_T\boldsymbol{X}_T\right)^{-1}\eta_T(\boldsymbol{X}_T)\right]
\end{aligned}
$$

The first term is equal to $\mathbf{0}$ because $\mathbb{E}[\boldsymbol{X}]=\mathbf{0}$ by assumption. The second term is equal to $\mathbf{0}$ because $\eta_T(\boldsymbol{X}_T)$ is uncorrelated with the covariates and itself has expectation zero. And $\mathbb{E}[R_3]=0$ because $\mathbb{E}[\boldsymbol{X}]=\mathbf{0}$. Since the R_i are uncorrelated and have expectation zero,

$$
\begin{aligned}
Var(\hat{\tau}_{\text{regression}}) &= \mathbb{E}\left[R_1^2\right]+\mathbb{E}\left[R_2^2\right]+\mathbb{E}\left[R_3^2\right] \\
&= \left\{\left(\mathbb{E}\left[\bar{\epsilon}_T^2\right]+\mathbb{E}\left[\bar{f}_T^2\right]\right)+\left(\mathbb{E}\left[\bar{\epsilon}_C^2\right]+\mathbb{E}\left[\bar{f}_C^2\right]\right)\right\}+O(N^{-2}) \\
&+ (\boldsymbol{\beta}_T-\boldsymbol{\beta}_C)'\left(p_T^2\frac{\Sigma_{X_T}}{n_T}+p_C^2\frac{\Sigma_{X_C}}{n_C}\right)(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C) \\
&= \left(\frac{\sigma_T^2}{n_T}+\frac{Var[\eta_T]}{n_T}\right)+\left(\frac{\sigma_C^2}{n_C}+\frac{Var[\eta_C]}{n_C}\right)+O(N^{-2}) \\
&+ (\boldsymbol{\beta}_T-\boldsymbol{\beta}_C)'\left(p_T\frac{\Sigma_{X_T}}{N}+p_C\frac{\Sigma_{X_C}}{N}\right)(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C) \\
&= \left[\frac{\sigma_T^2+Var[\eta_T]}{n_T}+\frac{\sigma_C^2+Var[\eta_C]}{n_C}\right]+O(N^{-2})+(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C)'\left(\frac{\Sigma_X}{N}\right)(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C)
\end{aligned}
$$

The last line follows since $\Sigma_{X_T}=\Sigma_{X_C}=\Sigma_X$, since they are all variances of a common distribution. ■

Proof of 3.3.1 Suppose now that the distribution of $\boldsymbol{X}$ is known. Its mean can be assumed to be $\mathbf{0}$ WLOG. Then $\tau=\beta_T^0-\beta_C^0$ and $\hat{\tau}_{regression}=\hat{\beta}_T^0-\hat{\beta}_C^0$, so that, using a similar rearrangement as before,

$$
\begin{aligned}
\hat{\tau}_{regression}-\tau &= \left(\bar{T}-\hat{\boldsymbol{\beta}}_T\bar{\boldsymbol{X}}_T\right)-\left(\bar{C}-\hat{\boldsymbol{\beta}}_C\bar{\boldsymbol{X}}_C\right)-(\boldsymbol{\beta}_T-\boldsymbol{\beta}_C) \\
&= \left[\bar{T}-\left(\beta_T^0+\bar{\boldsymbol{X}}_T\boldsymbol{\beta}_T\right)\right]-\left[\bar{C}-\left(\beta_C^0+\bar{\boldsymbol{X}}_C\boldsymbol{\beta}_C\right)\right] \\
&+ \bar{\boldsymbol{X}}_T\left(\boldsymbol{\beta}_T-\hat{\boldsymbol{\beta}}_T\right)-\bar{\boldsymbol{X}}_C\left(\boldsymbol{\beta}_C-\hat{\boldsymbol{\beta}}_C\right) \\
&= R_1+R_2^* \qquad (20)
\end{aligned}
$$

Direct comparison of 20 with 19 will show that the estimated ATE is also asymptotically unbiased, and that its asymptotic variance is decreased by the value of R_3, and some of R_2. With R_3 omitted, the standard error of the regression can just be estimated by $\sqrt{\frac{MSE_T}{n_T}+\frac{MSE_C}{n_C}}$

Proof of 3.5

We now verify that the standard error of the proposed estimator dominates the standard error estimator of the conventional ATE. We compare, therefore,

$$\left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)$$

to

$$\left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T}\left[\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\right]$$

We easily show that the asymptotic variance of the conventional estimator is higher than that of the regression estimator by comparing the variance components that differ among the two equations, noting that the $O(N^{-2})$ term vanishes.

$$\begin{aligned}
\left(\sqrt{\frac{n_C}{n_T}}\boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}}\boldsymbol{\beta}_C\right)' \Sigma_X \left(\sqrt{\frac{n_C}{n_T}}\boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}}\boldsymbol{\beta}_C\right) &\geq 0 \qquad (21)\\
\frac{n_C}{n_T}\left(\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T\right) + 2\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_C + \frac{n_T}{n_C}\left(\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\right) &\geq 0\\
\frac{N}{n_T}\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T + \frac{N}{n_C}\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C &\geq \boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T - 2\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_C + \boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\\
\frac{1}{n_T}\left[\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\right] &\geq (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \blacksquare
\end{aligned}$$

The only non-algebraic step is in the first line, which is true because the LHS is a quadratic form. Equality is attained iff $\boldsymbol{\beta}_C = -\frac{n_C}{n_T}\boldsymbol{\beta}_T$, which can be verified by direct substitution into (21).

Proof of remark on R^2 following equation (15):

$Var(\bar{T}) = \frac{SST}{n_T}$, whereas the regression based variance at the covariate mean is estimated by $MSE_T[1 + \frac{1}{n_T}]$, which can be rewritten as $\frac{SST - SSR}{n_T - p - 1} \times \left(\frac{n_T + 1}{n_T}\right)$. Dividing both expressions by SST leads us to compare $\frac{1}{n_T}$ to $\frac{1 - R^2}{n_T - p - 1} \times \left(\frac{n_T + 1}{n_T}\right)$. Equality is attained when R^2 is equal to $\frac{p+2}{n_T+1}$.

References

[Abadie and Imbens, 2006] Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

[Bloniarz et al., 2015] Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J., and Yu, B. (2015). Lasso adjustments of treatment effect estimates in randomized experiments. *arXiv preprint arXiv:1507.03652*.

[Buja et al., 2015] Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2015). Models as approximations – a conspiracy of random regressors and model deviations against classical inference in regression. Submitted to *Statistical Science*.

[Dehejia and Wahba, 1999] Dehejia, R. H. and Wahba, S. (1999). Causal effects in non-experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

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

[Freedman and Berk, 2008] Freedman, D. A. and Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409.

[Imbens and Rubin, 2007] Imbens, G. and Rubin, D. (2007). Causal inference: Statistical methods for estimating causal effects in biomedical, social, and behavioral sciences.

[Imbens and Wooldridge, 2008] Imbens, G. M. and Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. Technical report, National Bureau of Economic Research.

[LaLonde, 1986] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620.

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

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

[Samii and Aronow, 2012] Samii, C. and Aronow, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, 82(2):365–370.

[Splawa-Neyman et al., 1990] Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

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

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

[Zhang et al., 2008] Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715.