

Improved Precision in Estimating Average Treatment Effects

Emil Pitkin, Richard Berk, Larry Brown,
Andreas Buja, Ed George, Kai Zhang, Linda Zhao

October 31, 2013

Abstract

The Average Treatment Effect (ATE) is a global measure of the effectiveness of an experimental treatment intervention. Classical methods of its estimation either ignore relevant covariates or do not fully exploit them. Moreover, past work has considered covariates as fixed. We present a method for improving the precision of the ATE estimate: the treatment and control responses are estimated via a regression, and information is pooled between the groups to produce an asymptotically unbiased estimate; we subsequently justify the random X paradigm underlying the result. Standard errors are derived, and the estimator’s performance is compared to the traditional estimator. Conditions under which the regression-based estimator is preferable are detailed, and a demonstration on real data is presented.

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 T and control population C , the ATE is defined as $\tau = \mathbb{E}[T] - \mathbb{E}[C]$ for some measured response that can be continuous or categorical. τ can be estimated in a multitude of ways, each estimator depending on the sampling framework and model specification. Such disparate estimators’ definitions have practical significance for the researcher, who must understand the population for which his analysis holds – that is, he must understand the scope of inference. What’s more, depending on the particular situation, the interpretation of and inference for the ATE parameter will be different.

Past work has followed two principal strands. The first, earliest investigations of randomized experiments centered around finite, fixed populations, all of whose members would be randomized into either treatment(s) (the number of treatments could exceed one) or control groups; the random assignment furnished the randomness, and inference extended only as far as to these subjects in the trial. The foundation was thereby laid by Neyman, and subsequently developed by Rubin, for the notion of “potential outcomes,” whose unbiased estimation represented the first attempts to estimate some ATE [Neyman 1923]. Neyman considered a series of plots in a field, on each of which one of several varieties of fertilizer was

applied; he wished to estimate the true average yield of the aggregated plots, even though the individual plots were fertilized with only one variety. In this, earliest exploration of the ATE, the scope of inference was the collection of plots examined in the study only.

More recent literature has aimed to improve the precision of the ATE estimates via regression; whenever signal exists, the conditional variance of the response is reduced, with attendant gains in efficiency. Some authors [3] assume the framework in which a true, generating model exists, which could be correctly and completely specified via a regression equation. The estimating regression model in practice, however, is often misspecified, and in this case covariance adjustment can lead to undesirable consequences: in an influential critique, Freedman demonstrates how regression-based ATE estimators can lead to reduced asymptotic precision, and how they can be beset by small-sample bias. In this paper we will step aside from the Neyman framework within which Freedman offers his analysis.

The conventional philosophy behind regression adjustments in RCTs is appealing: not only does the ATE become a parameter of the model, but the random discrepancies in empirical covariate distributions between the treatment and control groups are adjusted away, and the essential difference between treatment and control groups is retained. In this regression analysis based RCT framework, several sources of randomness may exist. When there is a larger, target population of interest, then variation is driven by the choice of sampling units. When inference is restricted to the sample at hand, which may not be generalizable, then randomness stems from the units' randomization to treated or control status. Playing to the same tune as classical statistics, however, such regression analysis still presumes a fixed-X design. Elsewhere, also in the name of improving precision of the ATE estimate, knowledge of the population mean of the covariate distribution is assumed [14].

We argue for an analysis of RCTs that places minimal assumptions on the population from which data are generated. Fixed X is rarely reasonable in the context of RCTs: 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. When random X is hinted at, the population mean of the covariate distribution is rarely known. For these reasons a random-X analysis, which more realistically models experimental trials, and upon which minimal assumptions about the covariate distributions are imposed, offers the most convincing analysis. Such an approach, with minimal assumptions placed on the data generating mechanism, echoes the work of [18] and [19]. We will assume trials with random design, no knowledge of the covariate distributions (besides moment conditions), and will derive an efficient ATE estimator.

2 Neyman framework

Most pithily, the heart of Neyman's paradigm can be described as a "repeated-sampling randomization-based" method [17]. Of N subjects $\{Y_i\}_{1:N}$, fixed once and for all, n_T are assigned to the treatment group, and the remaining $n_C = N - n_T$ are exposed to the control condition. In subsequent hypothetical realizations of the experiment, another n_T subjects out of the original N are exposed to the treatment, and the remainder to the control. Each of the $\binom{N}{n_T}$ subsets has an equal probability of being the "treated block" in any given experiment. Note that in the thought experiment, the same, fixed n_T number of units are assigned

treatment, rather than each of the n subjects being assigned treatment as a Bernoulli trial with probability $\frac{n_T}{n}$.

To each subject are associated two hypothetical states, one of which is observed in practice¹. These are called “potential outcomes,” and they refer to the (deterministic) response of the subject, had he been subjected to the treatment (or control) condition. Let $Y_i(0)$ be the i th patient’s response under the control, and let $Y_i(1)$ be the corresponding response under treatment. The i th patient’s unobserved treatment effect is defined as $Y_i(1) - Y_i(0)$. The sample-ATE, known as SATE, is defined as

$$\tau^S = \frac{1}{N} \sum_{i=1}^N [Y_i(1) - Y_i(0)] \quad (1)$$

and is estimated (WLOG let Y_1, \dots, Y_{n_T} be treated) by

$$\hat{\tau}^S = \frac{1}{n_T} \sum_{i=1}^{n_T} Y_i(1) - \frac{1}{n_C} \sum_{i=1}^{n_C} Y_i(0) \quad (2)$$

$\hat{\tau}^S$ is an unbiased estimate of τ^S . To emphasize the point, the *only* source of randomness in subsequent realizations of the experiment is the subset of the original N patients to whom the treatment will be assigned. Their potential outcomes are immutable, and all that has the potential to change is which of the potential outcomes are observed.

In the literature a complementary parameter exists, called the population average treatment effect (PATE). The parameter, if the potential outcomes were known, would be computed similarly to the SATE, except the summation in (1) would be taken not over the sample in question but over all subjects in the population. In RCTs, where the desired scope of inference extends beyond the sample in question, the PATE is the more logical parameter to estimate. The estimate will be more variable: “sample selection error,” defined by $\Delta_S = PATE - SATE$, adds to the uncertainty of the ATE estimate [10].

3 Fixed X

The attractiveness of the Neyman framework lay in its simplicity: at its core the estimator is just a difference of means. In the name of simplicity, however, potentially useful subject specific characteristics are sacrificed. So instead, rather than working exclusively with treatment and control responses (and taking the difference in their averages, etc.), it is possible also to estimate the ATE by way of regression. The intention behind this approach is to make more precise the estimate of the ATE parameter by adjusting for the treated and control units’ covariates. Freedman [3], responding to its pervasiveness as an estimation tool, specifically considers OLS. He calls the ATE parameter b_{ITT} , where ITT is the acronym for “intention to treat.”² b_{ITT} can be estimated via regression in several ways. In the first, most simple and slightly contrived way, one regresses the response on the treatment indicator only, and

¹Of course, with multiple treatments, multiple states will be associated with each subject

²“Intention to treat” is described as “the effect of assigning everybody to treatment, minus the effect of assigning them to control.”

takes note of the indicator’s coefficient. This is akin to measuring the difference of treated and control means. For testing the equality of b_{ITT} to some value, usually 0, one employs the usual t-tests ³

One may then proceed to introduce covariates into the regression; the new coefficient of the treatment indicator, \hat{b}_{ITT} , is now the estimator of b_{ITT} . Freedman demonstrates that while augmenting the design with covariates can improve the performance of the estimator, it can worsen it as well (standard error is either increased or decreased, depending on the data). What’s worse, the nominal standard error of \hat{b}_{ITT} , in addition to the estimator itself, can be severely biased. The counterintuitive result arises because, as Freedman writes: “randomization does not justify the assumptions behind the OLS model.” That is, the demands the Neyman paradigm places on the nature of the data are not nearly as stringent as those imposed by OLS, what with its requirements of homoscedasticity, linearity, and fixed design.

We, however, opt for a parallel framework, one which is not hidebound by the assumptions behind OLS. We consider regression of a sort different from the one that Freedman critiques so compellingly. First, he focuses in his discussion on regressions without interactions – that is, the treatment and control groups are assumed to share slope coefficients. We will relax the assumption of homogeneous effects, and allow the treatment and control group covariates to impact to different degrees the response. Second, in the critiqued paradigm the population of subjects is finite, and their covariates are fixed too. The only source of randomness comes from the random assignment of treatment and control conditions. We will permit the subjects themselves (more to the point, their covariates) to be drawn from a distribution. We will describe our formulation more fully in section 4.

A recent and interesting paper by [14] reacts to Freedman’s critique, and reports the conditions under which, even in the Neyman paradigm, regression adjustment can give asymptotically valid coverage. His most trenchant point is that, by including a full set of covariate-treatment indicator interactions in the regression model, OLS adjustment cannot worsen asymptotic precision. Another recent paper [11] analyzes ATEs under more flexible circumstances, no longer working under the Neyman paradigm. The authors present their useful results “assuming the linear regression model is correctly specified.” We come to similar conclusions, but under the relaxed assumptions of proper specification.

4 Random X formulation

In this formulation, in contrast with the preceding discussion, nearly all quantities are random. Whereas in the Neyman framework, only the assignment of the n_T treated units is random – but not the subject pool (hence not the covariates), nor the potential responses – now all that will remain fixed is the number of units assigned to treatment, and the number to control.⁴ Subjects are not assigned treatment with probability $\frac{n_T}{N}$. Mathematically, subjects are sampled independently from an infinite population; which subjects are chosen will

³Interestingly, the usual t-tests assume the units to have been randomly sampled, but conclusions are little affected when the assumption does not hold for a difference in means.[8]

⁴As before, the thought experiment requires, in the next realization of the experiment, for the same n_T number of subjects to be assigned treatment, and the remaining n_C – control.

vary from sample to sample, as will the observed covariates.

The following discussion follows from the exposition in [1]. Let the population of subjects be described by the random variables X_1, \dots, X_p, Y . Their joint distribution $\mathbf{P} = \mathbf{P}(dx_1, \dots, dx_p, dy)$ has a full rank covariance matrix and four moments. $\vec{\mathbf{X}} = (1, X_1, \dots, X_p)'$ is the random vector of the predictor variables. Finally, let $\mu(\vec{\mathbf{X}})$ be the conditional mean of Y at $\vec{\mathbf{X}}$: $\mu(\vec{\mathbf{X}}) = \mathbb{E}[Y|\vec{\mathbf{X}}]$. We relax OLS assumptions, permitting, for examples, predictor variables to be omitted, and do not require – indeed, the operating assumption is that it is not – we do not require the true response surface to be linear in the predictors. Instead, we work with a conditional mean that can be decomposed into linear and non-linear components. Indeed, the linear component is thought of as the best linear approximation to the true conditional response surface; its partial slopes are defined by $\boldsymbol{\beta} = (\mathbb{E}[\mathbf{X}\mathbf{X}^T])^{-1} \mathbb{E}[\mathbf{X}Y]$, where the expectation is over the joint distribution of the \mathbf{X} and the Y . The difference between $\mu(\vec{\mathbf{X}})$ and $\boldsymbol{\beta}^T \vec{\mathbf{X}}$ is denoted by $f(\vec{\mathbf{X}})$, which is itself a random variable. A typical decomposition of a response would look like $Y = \boldsymbol{\beta}^T \vec{\mathbf{X}} + f(\vec{\mathbf{X}}) + \epsilon$, where the ϵ is the difference $Y - \mu(\vec{\mathbf{X}})$. Our operating assumption is that $f(\vec{\mathbf{X}})$ will not be uniformly equal to zero – that is, that non-linearity will be present in the population.

The additional results relevant to this paper are the following:

1. $\boldsymbol{\beta}$ should be estimated in the usual least squares fashion: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$
2. $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a random variable with mean 0; $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}$.
3. In finite samples, $\hat{\boldsymbol{\beta}}$ may be a biased estimator of $\boldsymbol{\beta}$.

With the background presented, we paint in more detail the particulars of our random \mathbf{X} formulation of how responses might be adjusted for covariates. The subjects of both the treatment and control groups are all assumed to have been sampled at random from the same population – that is, at the population level, the covariate distributions are the same for the two groups, and assignment of treatment is independent of covariates. The formulation is more general than in [18], which considers a baseline measurement of Y (as well as a treatment indicator) as the sole covariates. The treatment and control responses, respectively, can be denoted in the population by

$$T_i = \beta_T^{(0)} + \vec{\mathbf{X}}'_{Ti} \boldsymbol{\beta}_T + f_T(\vec{\mathbf{X}})_i + \epsilon_{Ti} \quad (3)$$

and, analogously,

$$C_i = \beta_C^{(0)} + \vec{\mathbf{X}}'_{Ci} \boldsymbol{\beta}_C + f_C(\vec{\mathbf{X}})_i + \epsilon_{Ci} \quad (4)$$

The $\beta^{(0)}$ are the respective intercepts at the population level, and the $\boldsymbol{\beta}$ are the respective vectors of population partial slopes. $\vec{\mathbf{X}}'_T$ is a random vector of treated units' covariates. Again, because we no longer assume that the response is linear in the covariates, $\beta_T^{(0)} + \vec{\mathbf{X}}'_T \boldsymbol{\beta}_T$ should be thought of as the treated group's best linear approximation, at the population level, to $\mathbb{E}[T|\vec{\mathbf{X}}]$. $\beta_T^{(0)}$ and $\boldsymbol{\beta}_T$, then, are population parameters derived from population least squares regression, and minimize the expected squared distance between the linear surface

and the true response surface. $f_T(\vec{\mathbf{X}})$ is a random variable that represents the difference between the true conditional mean of T and its best linear approximation in the population. In equations:

$$f_T(\vec{\mathbf{X}}) = \mathbb{E} [T|\vec{\mathbf{X}}] - (\beta_T^{(0)} + \vec{\mathbf{X}}'_T \beta_T) \quad (5)$$

Certain other assumptions and comments are warranted here.

1. *Errors.* We place minimal demands on the errors: they should have zero mean, and be uncorrelated, conditional on the predictors. Their distributional form is unspecified, and we do not assume normality of errors. Their variances, however, we allow to differ: denote the treated and control error variances, respectively, by σ_T^2 and σ_C^2 .
2. *Heterogeneity* Note, also, that in the population slopes are not assumed to be the same; we allow for heterogeneous effects. The non-linear random variables, too, are allowed to differ between the treatment and control groups.
3. *Equation (5):* Moreover, $\mathbb{E} [f_T \vec{\mathbf{X}}] = \mathbf{0}$. The non-linear component is orthogonal to the covariates, since it is the residual from the population least squares regression. Finally, f_T should also be well-behaved, so that its variances can be assumed to exist (for example, it should be bounded, or else defined on a compact set). The same conditions apply, of course, to the control group as well.

As detailed in [1], the target of estimation – the intercept and slopes – should be estimated, even in the random X setting, by the classical least squares estimators, and we shall do the same.

4.1 ATE definition through regression

We are going to re-express the ATE parameter through regression, thereby foreshadowing the proposed estimator. As mentioned in the introduction, and using the notation developed above, the ATE is the difference between the population average of the treated subjects and their control counterparts:

$$\tau = \mathbb{E} [T] - \mathbb{E} [C] \quad (6)$$

Subtracting (4) from (3) and taking expectations, we see that

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)} \right) + \mathbb{E} \left[\vec{\mathbf{X}}_T \right] \beta_T - \mathbb{E} \left[\vec{\mathbf{X}}_C \right] \beta_C \quad (7)$$

Note that the non-linear components $f_T(\vec{\mathbf{X}})$ and $f_C(\vec{\mathbf{X}})$ from (4) and (3) do not appear in the equation above. Simply, they are both equal to zero in expectation over the joint distribution of $\vec{\mathbf{X}}$ and Y .⁵ It deserves mentioning that the β in preceding equations are de-

⁵This is an interesting point, whose derivation is not central to the discussion. In brief, that $\mathbb{E} [f_T(\vec{\mathbf{X}})] = 0$ follows from $\mathbb{E} [f_T \vec{\mathbf{X}}] = 0$. $\vec{\mathbf{X}}$, as defined, contains an intercept; and since the expectation of the dot product of f_T with a vector of ones must be zero, then $\mathbb{E} [f_T \vec{\mathbf{X}}] = 0$ is equivalent to saying that $\mathbb{E} [f_T] = 0$

rived from the best linear approximations to the response surface, and may differ appreciably therefrom.

The careful reader will remark that we did not simply fully, as $\mathbb{E}[\vec{\mathbf{X}}_T] = \mathbb{E}[\vec{\mathbf{X}}_C] = \mathbb{E}[\vec{\mathbf{X}}]$, since, according to our assumptions, the treated and control subjects are drawn from the same population. And, indeed, (7) can be written as

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right) + \mathbb{E}[\vec{\mathbf{X}}] (\beta_T - \beta_C) \quad (8)$$

We consciously write these two true statement separately. In (7), one is tempted to estimate the respective expected values separately, by the respective covariate means of treatment and control groups. In (8), as single estimate will do, perhaps through a mean of all observed covariates, both treated and control. There will be a difference, in practice, and we wished to emphasize it now.

One more remark: when $\mathbb{E}[\vec{\mathbf{X}}] = \mathbf{0}$, then $\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right)$, and the ATE is just the difference between the respective population intercepts. This formulation hints at how we may wish to estimate the ATE from sample regressions.

All the while, we have represented the treatment and control regressions separately, if only to emphasize that the two functional relations of covariates to the responses need bear no relation to one another in order for an ATE to be properly defined, and, later, estimated. A single regression formulation, with interactions, may be more familiar. The response can be written as:

$$Y_i = \beta^{(0)} + \beta^{(T)} I_T + \beta' \vec{\mathbf{X}}_i + \beta^{(Int)} I_T \vec{\mathbf{X}}_i + f(\vec{\mathbf{X}})_i + I_T g(\vec{\mathbf{X}})_i + \epsilon_i \quad (9)$$

where $g(\vec{\mathbf{X}})$ is the difference in the treatment and control non-linearity functions. I_T is the treatment indicator at the population level; $\beta^{(Int)}$ is the vector in which are collected the differences in coefficients found in the treatment and control regressions respectively. The linear approximation being the target of estimation, we will restrict our attention to estimating the β . In equation (9) above, $\beta^{(T)}$ is precisely the ATE parameter when the covariate expectation is equal to 0.

4.2 ATE estimation

In this section we define two ATE estimators that can be derived from a random-X regression. The first reduces to the most familiar difference in means estimator, while the second borrows information across the treated and control groups. For both estimators, we write the regression-derived expression that is equivalent to the ATE, and then appeal to plug-in MLE estimates for the associated estimator.

1. Difference in means estimator.

Recall this fact of elementary statistics: that there is one point through which the least squares regression line must pass, and that that point the mean of the predictors and the mean of the response: $\hat{y}|_{x=\bar{x}} = \bar{y}$. So if we substitute $\vec{\mathbf{X}}_T$ into the treatment regression, the estimated conditional response will be \bar{T} , an unbiased estimate of $\mathbb{E}[T]$. In the same way we can find an unbiased estimate of $\mathbb{E}[C]$, and, as a result, of the

ATE. One must be very careful when estimating the standard error of this quantity $\left[\hat{\beta}_T^{(0)} + \bar{\mathbf{X}}_T \hat{\beta}_T\right] - \left[\hat{\beta}_T^{(0)} + \bar{\mathbf{X}}_T \hat{\beta}_T\right]$, as we do in section 4.3.

What we have done, in effect, by substituting the respective covariate means into the separate regressions, is estimate $\mathbb{E}[\bar{\mathbf{X}}]$ separately in the treatment and the control regression, which is congruent with the decomposition in (7). But the winding path leads back to response sample means – to compute them no regressions need to have been run, no covariates measured. The lesson here is that for our purposes, controlling for covariates loses its appeal and effectiveness if no information is shared between the treatment and the control groups.

2. A strictly regression derived estimator.

Alternatively, $\mathbb{E}[\bar{\mathbf{X}}]$ can – and in most cases should – be estimated not separately as above, twice, but rather once, by the complete set of the pooled covariates. It should be estimated at the mean of *all* covariates, $\frac{n_T \bar{\mathbf{X}}_T + n_C \bar{\mathbf{X}}_C}{N}$. The efficiency gains will be seen in section 4.2. This approach is more congruent with (8), so that, substituting the single estimate into (7), we find that

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

The estimator is invariant to location – a shift of the empirical covariate distribution does not change the value of $\hat{\tau}_{\text{regression}}$, so for the sake of appealing interpretability, we mean center the covariates. Note that we mean-center with respect to the common, pooled mean, so that $\left(\bar{\mathbf{X}}_T\right)_i^* = \left(\bar{\mathbf{X}}_T\right)_i - \bar{\mathbf{X}}$, with $\left(\bar{\mathbf{X}}_T\right)_i^*$ defined similarly. We thereby estimate the ATE for a covariate distribution with expectation equal to 0. From this we learn that the ATE can be estimated simply, via

$$\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{*(0)} - \hat{\beta}_C^{*(0)}\right) \tag{10}$$

Theorem 4.1 $\hat{\tau}_{\text{regression}}$ is an asymptotically unbiased estimate of τ .

Corollary 4.2 $\mathbb{E}[\hat{\tau}_{\text{regression}}] = \tau$ when

- (a) The population response is linear in the covariates, and all covariates have been included in the statistical model, or
- (b) $\mathbb{E}[T|X] = \mathbb{E}[C|X] + k$, and $n_T = n_C$, where $k \in \mathbb{R}$.

That is, 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, the result continues to hold when the units are inversely reweighted. The proofs are deferred to the appendix.

The difference in intercepts (from a mean centered regression) enriches our understanding of the relationship between a single regression with interaction terms, and one without. In a single regression with no interactions, the ATE can be estimated via the least squares regression coefficient of the treatment indicator, which represents the constant gap between the treatment and control response surfaces. It is the difference of intercepts (that is, at $\vec{\mathbf{X}} = \mathbf{0}$), but it is also the difference in responses at any arbitrary $\vec{\mathbf{X}}$ value, the difference being constant. In a single regression with interaction, the gap between the response surfaces is allowed to vary, and depends on the location of those covariates interacting with the treatment indicator. What then, is the estimated ATE in the regression with interactions? It, too, is the coefficient of the treatment indicator. But how else can the treatment indicator be represented and understood? It, too, is equal to the estimated difference in intercepts. Why intercepts? Intercepts are what are left when the regression is evaluated at 0; and since we are evaluating at the average of the (pooled) mean-centered covariates, we are evaluating at $\mathbf{0}$.

When

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

then in equations, the predicted response, when represented by a single regression with interactions, looks like

$$\hat{Y}_i = \hat{\beta}^{(0)} + \hat{\beta}^{(T)} I_T + \vec{\mathbf{X}} \hat{\beta} + \vec{\mathbf{X}} \hat{\beta}^{(Int)} I_T \quad (11)$$

With the covariates mean centered, substituting in the mean of the mean-centered covariates results in

$$\hat{Y}_i \Big|_{\vec{\mathbf{X}} = \vec{\mathbf{x}}^*} = \hat{\beta}^{(0)} + \hat{\beta}^{(T)} I_T \quad (12)$$

for which, as described, $\hat{\beta}^{(T)}$ represents the difference in intercepts. Here, the coefficient of the treatment indicator is precisely equal to $\hat{\tau}_{\text{regression}}$.

Nowhere in the definition of the model were any assumptions made about the nature of the response variables. While a continuous response may have been implicitly assumed, 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, the target of estimation is still $\mathbb{E}[T] - \mathbb{E}[C]$, but these terms can now 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.

One hopes that the estimate $\hat{P}(T) - \hat{P}(C)$ should fall inside $[-1, 1]$. If one estimates $\hat{\tau}$ by the difference in means estimator, then such a desirable outcome is assured. However, $\hat{\tau}_{\text{regression}}$, since it estimates the response Y not at the respective sample means of the covariates $\vec{\mathbf{X}}_{i_T}$ and $\vec{\mathbf{X}}_{i_C}$ but at the weighted average $\frac{n_T \vec{\mathbf{X}}_{i_T} + n_C \vec{\mathbf{X}}_{i_C}}{N}$, $\hat{P}(T) - \hat{P}(C)$ is not guaranteed with probability one to be restricted to $[-1, 1]$. The problem arises if there is limited

overlap between the observed treatment and control covariates, and the slope coefficients differ appreciably between the two groups. The probability associated with this possibility is small.

4.3 Relative performance of ATE estimators

We present in this section the expression for the variances of the difference-in-means and our regression based estimator, as well as for the standard error estimates, and compare the sizes of the variances.

The most familiar expression for $Var[\hat{\tau}_{diff}]$, of course, is $\frac{Var[T]}{n_T} + \frac{Var[C]}{n_C}$. For the purposes of comparison to $Var[\hat{\tau}_{regression}]$, the variance can be re-expressed by conditioning on covariates, and then marginalizing over their distribution, so that

Lemma 4.3

$$Var(\hat{\tau}_{diff}) = \left[\frac{\sigma_T^2 + Var[f_T]}{n_T} + \frac{\sigma_C^2 + Var[f_C]}{n_C} \right] + \frac{1}{n_T} [\boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_T] + \frac{1}{n_C} [\boldsymbol{\beta}'_C \Sigma_X \boldsymbol{\beta}_C] \quad (13)$$

The proof is found in the appendix. The standard deviation of $\hat{\tau}_{diff}$ should be estimated by

$$\hat{SE}(\hat{\tau}_{diff}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{n_T} \left(\hat{\boldsymbol{\beta}}_T \hat{\Sigma}_X \hat{\boldsymbol{\beta}}_T \right) + \frac{1}{n_C} \left(\hat{\boldsymbol{\beta}}_C \hat{\Sigma}_X \hat{\boldsymbol{\beta}}_C \right)} \quad (14)$$

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

The mean squared error is a scaled estimate of all the variability in the response that is not captured by the linear approximation. So the MSE is composed of two components: the estimate of the variability in the structural errors ϵ , together with the variability of $f(\vec{X})$, the random variable measuring the non-linearity in the conditional mean.

$\hat{\tau}_{regression}$ also admits a clean variance expression:

Lemma 4.4

$$Var(\hat{\tau}_{regression}) = \left[\frac{\sigma_T^2 + Var[f_T]}{n_T} + \frac{\sigma_C^2 + Var[f_C]}{n_C} \right] + O(N^{-2}) + \frac{1}{N} (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \Sigma_X (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \quad (15)$$

The proof is deferred to the appendix.

The standard deviation of $\hat{\tau}_{regression}$ should be estimated by

$$SE(\hat{\tau}_{regression}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{N} (\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)' \hat{\Sigma}_X (\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)} \quad (16)$$

The more interesting claim follows: the asymptotic variance of the regression-based estimator dominates the variance of the naive estimator.

Theorem 4.5

$$AVar(\hat{\tau}_{diff}) \geq AVar(\hat{\tau}_{regression}) \quad (17)$$

[Larry – would the following form be preferred: $\lim_{n \rightarrow \infty} Var(\hat{\tau}_{diff}) - Var(\hat{\tau}_{regression}) \geq 0$] The proof is found in the appendix.

To compare the relative asymptotic efficiencies of $\hat{\tau}_{diff}$ and $\hat{\tau}_{regression}$, only their respective variances need be compared because $\hat{\tau}_{diff}$ is a trivially unbiased estimate of τ , and, according to 4.1, $\hat{\tau}_{regression}$ is an asymptotically unbiased estimator of the ATE.

[19] also show that the estimator based on the model with interactions – they call it the *ANCOVA*₂ model – is efficient, and compare it with a large class of augmentation estimators. The aim here is to describe the nature of the interaction model’s efficiency and demonstrate which terms contribute to it. The inequality in 4.5 is not strict; and equality between the asymptotic variances can be attained, and is attained iff $\beta_C = -\frac{n_C}{n_T}\beta_T$. When the treatment and control sample sizes are equal, for example, then equality is attained when $\beta_C = -\beta_T$. In this case, when the treatment and control slopes are negative inverses of each other, the regression-based estimate of the ATE is maximally variable. This makes sense: sample estimates of the difference in intercepts are just as likely to be positive as to be negative, with equal probabilities of linearly increasing magnitudes of difference.

Theorem 1 refers, however, to the true variance of the respective estimators, rather than to their estimated variances⁶. The theorem could analogously have been written, and should be seen here for clarity, as

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

A remark on the seemingly different estimators of $\hat{\tau}_{diff}$. Every introductory statistics textbook will teach that

$$Var[\bar{T} - \bar{C}] = \frac{Var[T]}{n_T} + \frac{Var[C]}{n_C} \quad (18)$$

and that it is estimated unbiasedly – for example, for the purpose of hypothesis testing – by

$$\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C} \quad (19)$$

In our paper, we wrote different expressions for the variance and standard error estimates of $\hat{\tau}_{diff}$. This was done for ease of comparison. In fact, (4.3) and (18) are equal, as are (14) and (19), which are unbiased estimates thereof.

4.4 Conditional and marginal estimation

We pause to make explicit the essential difference between conditional and marginal inference in our problem, and to emphasize the role of covariates that are here random. The variance of the difference-in-means estimator is a marginal variance: over all conceivable repetitions of

⁶Which means that in a given sample, $SE(\hat{\tau}_{regression})$ may exceed $SE(\hat{\tau}_{diff})$

the experiment, as new subjects are sampled and assigned a treatment or a control condition, irrespective of any other measured or unmeasured covariates,

$$\text{Var}[\bar{T} - \bar{C}] = \frac{\text{Var}[T]}{n_T} + \frac{\text{Var}[C]}{n_C}. \quad (20)$$

It is estimated, unbiasedly, by $\frac{s^2_T}{n_T} + \frac{s^2_C}{n_C}$.

Now, as in our problem, measure covariates, and run two separate regressions, so that $\hat{T} = \hat{\beta}_T^{(0)} + \vec{\mathbf{X}}_T \hat{\beta}_T$, and $\hat{C} = \hat{\beta}_C^{(0)} + \vec{\mathbf{X}}_C \hat{\beta}_C$. From elementary regression, if we estimate the response at the mean of the predictors, then $\hat{T}_i \Big|_{\vec{\mathbf{X}}_T = \vec{\bar{\mathbf{X}}}_T} = \bar{T}$, and $\hat{C}_i \Big|_{\vec{\mathbf{X}}_C = \vec{\bar{\mathbf{X}}}_C} = \bar{C}$. Apparently, in estimating the ATE, $\bar{T} - \bar{C} = \hat{T}_i \Big|_{\vec{\mathbf{X}}_T} - \hat{C}_i \Big|_{\vec{\mathbf{X}}_C}$, so the variance should depend on the the observed covariates! What, then, is the proper variance of $\bar{T} - \bar{C}$? Is it the same as that reported in (20)?

It will not be equal, for the simple reason that the classical variance is considered conditional on the observed covariates. To compute, note that \bar{T} is independent of \bar{C} , so let us for the moment consider just $\text{Var}[\bar{T}]$. \bar{T} was estimated in a regression at a specific covariate value. For ease of exposition, recall the prediction variance from simple regression, where

$$\hat{\text{Var}}[\hat{y}|X = x_p] = \text{MSE} \left[1 + \frac{1}{n_T} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n_T} (x_i - \bar{x})^2} \right] \quad (21)$$

That is to say, at the covariate mean,

$$\hat{\text{Var}}[\hat{T}|\vec{\mathbf{X}} = \vec{\bar{\mathbf{X}}}_T] = \text{MSE} \left[1 + \frac{1}{n_T} \right] \quad (22)$$

which, of course, does not uniformly equal $\frac{s^2_T}{n_T}$. As a matter of fact, the two estimated variances will be equal only when the R^2 from the regression exceeds $\frac{p+2}{n_T+1}$, where p is the number of covariates; then the regression based estimated variance will be smaller than that of the marginal, conventional estimated variance. The reason for this discrepancy, for how the relative variances of ostensibly the same statistic depend on the quality of the fit, is simple.

The variance estimated in (22) relies on classical regression theory, where the predictors are assumed to be fixed from one realization of the data to the next. Inference is therefore conditional on the covariates; the estimate of the variance of \bar{T} in (22) is *conditional* on being estimated at the (here, fixed) mean of the covariates. It is saying: when the mean of the covariates is equal exactly to the mean of the covariates in this sample, what is the variability of the average response? What is unaccounted for is that that selfsame covariate mean is a random quantity, and its variability will contribute to the variability in the average response. This naive regression based estimate (22), therefore, artificially deflates the true variance of the response mean. In our analysis we compare two marginal variances, from which an inequality follows that holds for all fits.

4.5 Alternative Conditions

4.5.1 Distribution of \mathbf{X} known

Throughout the discussion and analysis, we have assumed that the underlying distribution of \mathbf{X} is unknown. The alternative may present in practice where, for example, covariates like age, weight and income, for which measurements exist in the whole population, are used in the study. In such a case, the variability inherent in estimating $\mathbb{E}[\mathbf{X}]$ is removed (only the regression slopes remain to be estimated), with a corresponding diminution of the standard error of the ATE. The precise degree to which the standard error diminishes can be found in the appendix.

4.5.2 Treatment Correlated with Covariates

In the preceding discussion, we had assumed that the assignment of treatment (the treatment indicator) was independent of the covariates, with correlation among them presenting itself only in samples. It is conceivable and natural, however, that the decision to administer treatment should depend on the covariates: perhaps, by design and because of cost constraints in the study, the researcher wishes to offer expensive treatment to a higher proportion of those suspected to require it for a shorter duration.

Precisely, suppose that the regression is written as in 9, except that $I_T = H(\vec{\mathbf{X}})$, either deterministically or stochastically, as when $I_T \sim \text{Bern}\left(H\left(\vec{\mathbf{X}}\right)\right)$. The treatment indicator is a function of the covariates so the assignment mechanism is different across different strata. In this case, the functional form of $H(\cdot)$ is known, so that $\pi_i = P\left(I_T = 1|\vec{\mathbf{X}}\right)$ 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

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

Such a reweighting has been considered by, for example, [6], 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 $\frac{1}{n_T} \sum_{i=1}^n y_i^{(T)*} - \frac{1}{n_C} \sum_{i=1}^n y_i^{(C)*}$, what [6] 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 article.

4.5.3 Stratification

The results described in the preceding sections make no assumptions about the nature of the covariates, which may be discrete, continuous, or both. An interesting special case arises when, besides the treatment indicator, the other covariates represent stratum assignment, and interactions are permitted between the treatment indicators and assignment indicators. For example, subjects may be classified by treatment/control, and highest degree of educational attainment (no high school, high school, college, etc.) The result of this pre-stratification is a two-way ANOVA layout, with interactions. In the familiar ANOVA form, the regression model may be described by

$$Y_{ijk} = \mu + s_i + \tau_j + (s\tau)_{ij} + \epsilon_{ijk} \quad (23)$$

s_i is the i th stratum, $i = 1, \dots, I$, τ_j is the treatment effect, $j = 0, 1$ (WLOG, let $j = 1$ when treatment is administered), and $(s\tau)_{ij}$ is the interaction effect. Denote the number of patients in stratum i receiving regime j by K_{ij} .

The difference-in-means estimator is written simply as

$$\bar{\mu} = \bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 0}. \quad (24)$$

and is unbiased, since $\mathbb{E}[\bar{\mu}] = \mathbb{E}[Y_{\cdot 1}] - \mathbb{E}[Y_{\cdot 0}]$.

Now define the local ATEs, which represent the respective within-stratum ATEs by

$$ATE_i \equiv \theta_i = \mathbb{E}[Y_{i1} - Y_{i0}].$$

The second estimator weights the per-stratum difference-in-means by the proportion of the sample found in each stratum:

$$\tilde{\mu} = \sum_{i=1}^I (\bar{Y}_{i1} - \bar{Y}_{i0}) * \hat{p}_i \quad (25)$$

where \hat{p}_i is the sample proportion of all subjects in stratum i ; it is equivalently written as $\frac{K_{i+}}{K_{++}}$. $\mathbb{E}[\tilde{\mu}] = \sum_{i=1}^I p_i \theta_i = \theta$, so it is also unbiased. The estimator is unbiased under randomized assignment and under blocking since in both instances, the proportion of treated cases in a stratum is independent of the mean, and in both cases, $\mathbb{E}[\hat{p}_i] = \theta_i$. As in [15], which gives an impressive treatment of *post*-stratification in the Neyman framework, the ATE estimate here is assumed to be well-defined – that is, the estimator is computed conditional on the event that each stratum is populated by at least one treated and one control unit. This second estimator just described is precisely $\hat{\tau}_{\text{regression}}$. Our results, in particular Lemma 4.4 and Theorem 4.5 continue to hold. Under slightly modified conditions, [15] and [12] show, for example, that its variance is less than that of the difference-in-means estimator, and is higher than the variance resulting from blocking (or pre-stratification) on an order of $O(N^{-2})$.

4.6 Illustration on real data

We present a typical application of our regression based ATE estimator on real data. We illustrate the performance of the estimator on data furnished from a classic study discussed in [13] and reanalyzed in [7]. The data in question come from the National Support Work (NSW) Demonstration. 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 intent of the work in [13] 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). [7] apply matching techniques for this comparison; relevant for our work are the 185 treated and 260 control male subjects they analyze, and which are available from the original NSW experiment.

The following covariates were adjusted for: age, education (number of years), an indicator for black, indicator for hispanic, indicator for marital status, indicator for high school degree, and earnings in 1974. The response measured was earnings in 1978, after the job training had concluded.

In this experimental context the difference in means is equal to 4709.4 dollars, with a standard error equal to 443.5. The regression based method yields a point estimate of $\hat{\tau} = 4435.2$ dollars, with an SE estimate of 431.9. The gain in SE amounts to 3.1%, this when the R^2 of the regression of reservation price on covariates and their interaction with the treatment indicator was 0.24. A gain of this magnitude is typical for an R^2 of this size. Higher R^2 results in higher SE gains, which is vividly demonstrated in the following section.

4.7 Illustration on simulated data

The datasets on RCTs we have encountered have come with an R^2 that doesn't far exceed 0.2. To more vividly illustrate the results obtained in this paper, we considered the following model. The treated and control groups were defined, respectively, by

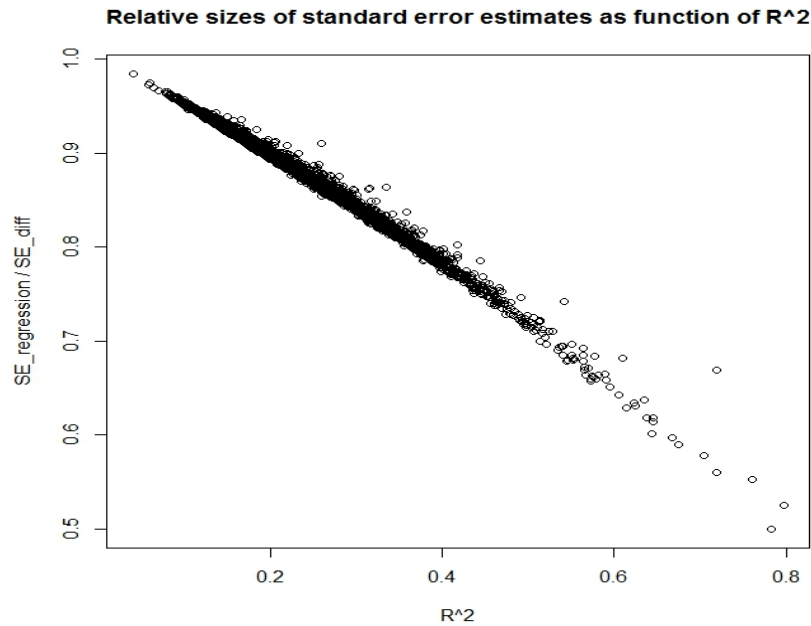
$$T = 2X_1 + 3X_2 + Z_T \tag{26}$$

$$C = X_1 + X_2 + Z_C \tag{27}$$

where $X_1 \sim \text{Lognormal}(0, 1)$, $X_2 \sim \text{Gamma}(3, 4)$, and $Z_T, Z_C \stackrel{iid}{\sim} N(0, 3)$. Under these conditions, $\mathbb{E}[T] - \mathbb{E}[C] = 2e^{1/2} - 3/2 = 1.797$. We simulated 10,000 times, with 250 treated and 250 control units in each simulation, and recorded the R^2 of the combined regression, as well as the ATE and SE estimates for both the difference-in-means, and for the regression based estimator considered in this paper. The average R^2 in the 10,000 simulation was 0.75. Accordingly, the average $\hat{SE}(\hat{\tau}_{\text{diff}}) = 0.676$ (with simulation SE = 0.0011), while the average $\hat{SE}(\hat{\tau}_{\text{regression}}) = 0.332$ (with simulation SE = 0.0002). Both estimators were unbiased (up to simulation granularity), with difference-in-mean and regression-based average ATEs equal to 1.798 and 1.796, respectively. Coverage of the true ATE was equal to 0.9473 and 0.949, respectively, when using $\Phi^{-1}(0.975)$ as the multiplier. The regression based estimate naturally leads to a more powerful test. There was nothing particular about the model chosen; similar phenomena are observed for other choices of underlying distribution.

As a final illustration, we show the relationship between the R^2 from the combined model and the respective standard error estimates. $\hat{\tau}_{\text{diff}}$, depending only on the response, does not depend on the quality of the regression fit. $\hat{\tau}_{\text{regression}}$, however, does. 10,000 simulations were again run, except the variance of Z_T, Z_C was dialed from 1 to 100, with attendant decreases in the R^2 . The plot of R^2 against $\frac{\hat{SE}(\hat{\tau}_{\text{regression}})}{\hat{SE}(\hat{\tau}_{\text{diff}})}$ is shown. As R^2 decreases, the estimated standard errors converge. For high R^2 , the $\hat{\tau}_{\text{regression}}$ enjoys a dramatically lower standard error.

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



5 Conclusion

This paper lays the foundation for conducting principled and efficient asymptotic inference on ATEs. After acknowledging the aesthetics but also limitations of the Neyman paradigm, and the unreality of fixed X, we restricted our focus to an infinite population, random design, regression based estimation, where the response surface needn't be linear. Since the regression covariates are seen as random, generated from a distribution, the formulation is a more realistic representation of the practice of random sampling: randomness arises not only from the random assignment of treatment and control to subjects, but also from these subjects' (random) characteristics as well. Despite the added source of variability, the derived standard error, which takes into account these sources of randomness but 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 to GLMs.

6 Technical appendix

Proof of 4.1

After mean centering, $\hat{\tau}_{\text{regression}} = (\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)})$. Direct application of the proposition on page 11 in [1] 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 4.2

- (a) When the regression model is correctly specified, then it is an introductory result that the LS estimates are unbiased: $\mathbb{E}[\hat{\beta}_T^{(0)}] = \beta_T^{(0)}$ and that $\mathbb{E}[\hat{\beta}_C^{(0)}] = \beta_C^{(0)}$, so $\mathbb{E}[\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}] = \beta_T^{(0)} - \beta_C^{(0)} = \tau$.
- (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 4.4, the only term which does not generally have expectation $\mathbf{0}$ is the term denoted by R_2 , and equal to $[\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C] [p_C(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T) + p_T(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C)]$. 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

$$p_C \text{Cov}(\bar{\mathbf{X}}_T, \hat{\boldsymbol{\beta}}_T) = p_T \text{Cov}(\bar{\mathbf{X}}_C, \hat{\boldsymbol{\beta}}_C) \quad (28)$$

Inversely weight the observations, giving weight $\frac{1}{n_T}$ to the control observations, and $\frac{1}{n_C}$ to the treatment, so that (28) will hold true when $\text{Cov}(\bar{\mathbf{X}}_T, \hat{\boldsymbol{\beta}}_T) = \text{Cov}(\bar{\mathbf{X}}_C, \hat{\boldsymbol{\beta}}_C)$. When $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$, then, since the $\bar{\mathbf{X}}_T$ and $\bar{\mathbf{X}}_C$ are identically distributed, the above equality will hold. $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$ when there is a constant offset.

Proof of 4.3

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

$$\begin{aligned} \hat{\tau}_{\text{diff}} - \tau &= \bar{T} - \bar{C} - (\beta_T^0 - \beta_C^0) \\ &= [\bar{T} - (\beta_T^0 + \bar{X}_T \boldsymbol{\beta}_T)] - [\bar{C} - (\beta_C^0 + \bar{X}_C \boldsymbol{\beta}_C)] \\ &+ \bar{X}_T \boldsymbol{\beta}_T - \bar{X}_C \boldsymbol{\beta}_C \end{aligned} \quad (29)$$

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

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

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

Proof of 4.4

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 $\text{Var}[X] = \Sigma$. Denote the ATE by τ . The ATE in the population, τ , equals $\mathbb{E}[T] - \mathbb{E}[C] = (\beta_T^0 - \beta_C^0) + \boldsymbol{\mu}(\beta_T - \beta_C)$. Then

$$\hat{\tau}_{\text{regression}} = \hat{\beta}_T^0 - \hat{\beta}_C^0 + \hat{\boldsymbol{\mu}}(\hat{\beta}_T - \hat{\beta}_C)$$

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

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

$$\begin{aligned}
\hat{\tau}_{\text{regression}} - \tau &= \bar{T} - \bar{C} - (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)(p_C\hat{\beta}_T + p_T\hat{\beta}_C) - \beta_T^0 + \beta_C^0 \\
&= [\bar{T} - (\beta_T^0 + \bar{\mathbf{X}}_T\beta_T)] - [\bar{C} - (\beta_C^0 + \bar{\mathbf{X}}_C\beta_C)] \\
&\quad - (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)[p_C(\hat{\beta}_T - \beta_T) + p_T(\hat{\beta}_C - \beta_C)] \\
&\quad + (p_T\bar{\mathbf{X}}_T + p_C\bar{\mathbf{X}}_C)(\beta_T - \beta_C) \\
&= R_1 + R_2 + R_3x \tag{30}
\end{aligned}$$

R_1, R_2 , and R_3 are independent: R_1 is a function of the errors, which are independent of the covariates, while R_2 and R_3 lie in the column space of the covariates. R_2 is uncorrelated with R_3 because [we have the correlation between sums and differences of i.i.d variables. Check the math again]. Moreover, each of the terms has expectation $\mathbf{0}_p$: the first, R_1 , is a difference of average errors, equal to $(\bar{\epsilon}_T + \bar{f}_T) - (\bar{\epsilon}_C + \bar{f}_C)$. 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 $\mathbb{E}[\bar{\mathbf{X}}] = \mathbf{0}$, so the only component of R_2 not equal for all n to $\mathbf{0}$ in expectation is $p_C \bar{\mathbf{X}}_T \hat{\boldsymbol{\beta}}_T - p_T \bar{\mathbf{X}}_C \hat{\boldsymbol{\beta}}_C$. We'll now show that $\mathbb{E}[\bar{\mathbf{X}}_T \hat{\boldsymbol{\beta}}_T] \rightarrow \mathbf{0}$:

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

The first terms is equal to $\mathbf{0}$ because $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ by assumption. The second term is equal to $\mathbf{0}$ because $f_T(\mathbf{X}_T)$ is uncorrelated with the covariates and itself has expectation zero.

$\mathbb{E}[R_3] = 0$ because $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. So

$$\begin{aligned}
\text{Var}(\hat{\tau}_{\text{regression}}) &= \mathbb{E}[R_1^2] + \mathbb{E}[R_2^2] + \mathbb{E}[R_3^2] \\
&= \{(\mathbb{E}[\bar{\epsilon}_T^2] + \mathbb{E}[\bar{f}_T^2]) + (\mathbb{E}[\bar{\epsilon}_C^2] + \mathbb{E}[\bar{f}_C^2])\} + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(p_T^2 \frac{\Sigma_{X_T}}{n_T} + p_C^2 \frac{\Sigma_{X_C}}{n_C} \right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= \left(\frac{\sigma_T^2}{n_T} + \frac{\text{Var}[f_T]}{n_T} \right) + \left(\frac{\sigma_C^2}{n_C} + \frac{\text{Var}[f_C]}{n_C} \right) + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(p_T \frac{\Sigma_{X_T}}{N} + p_C \frac{\Sigma_{X_C}}{N} \right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= \left[\frac{\sigma_T^2 + \text{Var}[f_T]}{n_T} + \frac{\sigma_C^2 + \text{Var}[f_C]}{n_C} \right] + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N} \right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \tag{31}
\end{aligned}$$

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

Proof of 4.5.1 Suppose now that the distribution of \mathbf{X} is known. Its mean can be assumed to be $\mathbf{0}$ WLOG. Then $\tau = \beta_T^0 - \beta_C^0$ and $\hat{\tau}_{\text{regression}} = \hat{\beta}_T^0 - \hat{\beta}_C^0$, so that, using a similar rearrangement as before,

$$\begin{aligned}
\hat{\tau}_{\text{regression}} - \tau &= (\bar{T} - \hat{\boldsymbol{\beta}}_T \bar{\mathbf{X}}_T) - (\bar{C} - \hat{\boldsymbol{\beta}}_C \bar{\mathbf{X}}_C) - (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= [\bar{T} - (\beta_T^0 + \bar{\mathbf{X}}_T \boldsymbol{\beta}_T)] - [\bar{C} - (\beta_C^0 + \bar{\mathbf{X}}_C \boldsymbol{\beta}_C)] \\
&\quad + \bar{\mathbf{X}}_T (\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T) - \bar{\mathbf{X}}_C (\boldsymbol{\beta}_C - \hat{\boldsymbol{\beta}}_C) \\
&= R_1 + R_2^* \tag{32}
\end{aligned}$$

Direct comparison of 32 with 30 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

$$\sqrt{\frac{\text{MSE}_T}{n_T} + \frac{\text{MSE}_C}{n_C}}$$

Proof of 4.5

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

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

to

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

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.

$$\begin{aligned} \left(\sqrt{\frac{n_C}{n_T}} \boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}} \boldsymbol{\beta}_C \right)' \Sigma_X \left(\sqrt{\frac{n_C}{n_T}} \boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}} \boldsymbol{\beta}_C \right) &\geq 0 & (33) \\ \frac{n_C}{n_T} (\boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_T) + 2 \boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_C + \frac{n_T}{n_C} (\boldsymbol{\beta}'_C \Sigma_X \boldsymbol{\beta}_C) &\geq 0 \\ \frac{N}{n_T} \boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_T + \frac{N}{n_C} \boldsymbol{\beta}'_C \Sigma_X \boldsymbol{\beta}_C &\geq \boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_T - 2 \boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_C + \boldsymbol{\beta}'_C \Sigma_X \boldsymbol{\beta}_C \\ \frac{1}{n_T} [\boldsymbol{\beta}'_T \Sigma_X \boldsymbol{\beta}_T] + \frac{1}{n_C} [\boldsymbol{\beta}'_C \Sigma_X \boldsymbol{\beta}_C] &\geq (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N} \right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \blacksquare \end{aligned}$$

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 (33).

Proof of remark on R^2 following equation (22):

$\text{Var}(\bar{T}) = \frac{SST}{n_T}$, 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

- [1] Buja, Andreas et al. 2013. "A Conspiracy of Random X and Model Violation against Classical Inference in Linear Regression." University of Pennsylvania working paper (<http://stat.wharton.upenn.edu/~buja/PAPERS/Notes-on-Wrong-Models-and-Random-X.pdf>).
- [2] Cochran, William G. (1977). Sampling Techniques, 3rd Edition, Wiley
- [3] Freedman, David A. 2008a. "On regression adjustments to experimental data." Advances in Applied Mathematics 40:180-93.

- [4] Freedman, David A. 2008b. "On regression adjustments in experiments with several treatments." *Annals of Applied Statistics* 2: 176-96
- [5] Freedman, David A. 1981. "Bootstrapping Regression Models." *The Annals of Statistics*, Vol. 9, No. 6, pp. 1218-1228
- [6] Freedman, David A. and Berk, Richard A. 2008 "On weighting regressions by propensity scores. *Evaluation Review* vol. 32 (2008) pp. 392-409
- [7] Dehejia, Rajeev H. and Wahba, Sadek 1999 "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, Vol. 94, No. 448, pp. 1053-1062
- [8] Freedman, David, Robert Pisani and Roger Purves. 1998 "Statistics (3rd edition)"
- [9] Heritier SR, Gebiski VJ, Keech AC 2003 "Inclusion of patients in clinical trial analysis: The intention-to-treat principle." *Med J Aust.* 179:43840
- [10] Imai, Kosuke et al. 2008 "Misunderstandings between experimentalists and observationalists about causal inference." *J. R. Statistic. Soc. A* 171, Part 2, 481-502
- [11] Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47(1): 5-86
- [12] Imbens, Guido W. 2011. "Experimental design for unit and cluster randomized trials." *International Initiative for Impact Evaluations*
- [13] Lalonde, Robert, 1986. "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, Vol. 76, pp. 604-620.
- [14] Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *Annals of Applied Statistics*, forthcoming.
- [15] Miratrix, Luke W., Sekhon, J. S. and Yu, B. 2013. Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 75 369396
- [16] Neyman, Jerzy, translated by Dabrowska, D. M., Speed, T.P. Published in 1923, translated in 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science*, Vol. 5, No. 4, 465-472
- [17] Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science*, Vol. 5, No. 4. 472-480
- [18] Yang, Li and Tsiatis, Anastasios A. 2001. "Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial." *The American Statistician*, Vol. 55, No. 4 (Nov., 2001), 314-321

- [19] Tsiatis, AA, Davidian, M, Zhang M, and Lu, X. "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach." *Statistics in Medicine* special issue on "Statistical methods in HIV/AIDS and its practical application." 2008; 27:4658-4677