Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

February 5, 2026

Publisher: Online
Version 0.1.2

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Causal Inference via Randomized Experiments

2

> "Let us divide them in halves, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation; but do you do as ye know [...]. We shall see how many funerals both of us shall have."
>
> – Jan Baptist van Helmont (17th Century)

In this chapter we begin discussion of causal inference by focusing on Randomized Control Trials (RCTs). In an RCT, units are randomly divided into those that receive a treatment and those that receive no treatment. Under randomization and other assumptions, the difference in average outcomes between the treated and untreated groups is an average treatment (causal) effect (ATE). By considering pre-treatment covariates, we can improve the precision of the ATE estimate, explore heterogeneity across subgroups, or both. We describe methods for doing so and apply them to several RCTs. We introduce causal diagrams as a means of visualizing RCTs and their underlying causal assumptions. We conclude by outlining some limitations of RCTs.

2.1 Potential Outcomes Framework and Average Treatment Effects

In this section, we discuss the potential outcomes (POs) framework for analyzing causality and treatment effects.[1] We begin by introducing the two *latent* (unobserved) variables

$$Y(1) \text{ and } Y(0).$$

They represent the potential or counterfactual random outcomes for an observational unit when the unit is subject to treatment (treatment state $d = 1$) or no treatment (control or untreated state $d = 0$).[2] In an economic context, the treatment might be a training program or a policy intervention, and the outcome might be an individual's wage or employment status. In what follows, it is also useful to introduce the potential response or structural function:

$$d \mapsto Y(d),$$

which maps the potential treatment state $d \in \{0, 1\}$ to the random potential outcome $Y(d)$.

The quantities $Y(1)$ and $Y(0)$ are "counterfactual" because they can't be simultaneously observed. That is, we generally do not have identical replicas of the observational units that are simultaneously subject to both treatment and control.

The individual treatment effect is

$$Y(1) - Y(0).$$

This effect will vary across individuals. As mentioned above, typically we observe only one of the two terms, and hence it is generally infeasible to uncover the individual treatment effect.[3] However, we can hope to estimate averages and the distribution of $Y(d)$ at the population level to compute quantities such as the average treatment effect (ATE):

$$\delta = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)].$$

Let D denote the actual *assigned treatment*, a random variable, which takes a value of 1 if the observational unit participated in the treatment and 0 otherwise.

Assumption 2.1.1 (Consistency) *We observe*

$$Y := Y(D).$$

1: POs were introduced by Donal Rubin [1], builing upon the earlier work of R.A.Fisher and J. Neymann in the 1920s.

2: For simplicity, here we focus on binary treatments; the ideas are similar formultivalued and continuous treatments.

3: As an example, we could uncover individual treatment effects if we had identical twins (for example, a pair of cars) that could be put in treatment and control groups.

In the binary case, this can also be written as $Y = DY(1) + (1 - D)Y(0)$. For example, suppose the treatment group ($D = 1$) consists of individuals who completed a job training program, while the control group ($D = 0$) comprises those who did not. According to Assumption 2.1.1, an individual's observed wage outcome equals $Y(1)$ if she completed the program ($D = 1$) and equals $Y(0)$ if she did not ($D = 0$). Although this assumption might seem almost tautological, it plays a crucial role by ruling out hidden variations in treatment. In other words, it requires that the treatment and control conditions are well defined and clearly correspond to the observed treatment status, D.

Assumption 2.1.2 (No Interference) *Potential outcomes for any observational unit depend only on the treatment status of that unit and not on the treatment status of any other unit.*

Assumption 2.1.2 is implicitly embedded in our definition of the potential outcomes, $Y(d)$, which specify each unit's outcome under treatment state d.[4] This formulation rules out scenarios in which the treatment administered to one unit affects the outcome of another unit. For example, in social networks, treating an individual might influence the outcomes of all of that person's friends. Some types of spillover effects can be accommodated by expanding the treatment definition and adjusting the potential outcomes accordingly,[5] but addressing these extensions is beyond the scope of this book.[6]

The following analytical example may help gain better understanding of the potential outcomes framework.

Example 2.1.1 [Analytical Example] Consider the following model:

$$\begin{aligned} Y(d) &:= \eta_0 + \eta_1 d, \\ D &:= 1(\nu > 0), \\ Y &:= \eta_0 + \eta_1 D, \end{aligned}$$

where (η_0, η_1, ν) are jointly normal stochastic variables. Here, ν represents factors that influence selection into the treatment state. In this example $\mathrm{E}[Y(1)] = \mathrm{E}[\eta_0 + \eta_1]$, $\mathrm{E}[Y(0)] = \mathrm{E}\eta_0$, and the ATE is $\delta = \mathrm{E}\eta_1$. Importantly, only D and Y are observed.

Under Assumption 2.1.1, population data directly provide the conditional averages

$$\mathrm{E}[Y \mid D = d] = \mathrm{E}[Y(d) \mid D = d], \text{ for } d \in \{0, 1\}.$$

4: Assumptions 2.1.1 and 2.1.2 together constitute what is commonly known as the Stable Unit-Treatment Value Assumption (SUTVA); see, e.g., Imbens and Rubin [2].

5: For example, if each individual has two friends, we could define potential outcomes with spillovers as $Y(d_0, d_1, d_2)$, where d_0 represents the individual's treatment status, d_1 denotes the treatment status of friend 1, and d_2 denotes that of friend 2.

6: For further reading, see, among many others, [3], [4], [5], [6], and [7].

The difference of the two averages gives us the average predictive effect (APE) of treatment status on the outcome:

$$\pi = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0].$$

It measures the association of the treatment status with the outcome.

While the APE is identified – meaning computable from the population data – it may seem surprising (or not at all) that the APE in general does not agree with the ATE δ:

$$\delta \neq \pi. \tag{2.1.1}$$

The difference between the APE and ATE is generally said to be due to *selection bias*. The meaning of selection bias is clarified through the following example, and clarified theoretically below.

Example 2.1.2 (Selection Bias in Observational Data) Suppose we want to study the impact of smoking marijuana on life longevity. Suppose that smoking marijuana has no causal effect on life longevity:

$$Y = Y(0) = Y(1),$$

so that

$$\delta = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = 0.$$

However, the observed smoking behavior, D, is not assigned in an experimental study. Suppose that the behavior determining D is associated with poor health choices such as drinking alcohol, which are known to cause shorter life expectancy, so that $\mathrm{E}[Y \mid D = 1] < \mathrm{E}[Y \mid D = 0]$. In this case, we have a negative predictive effect:

$$\pi = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0] < 0 = \delta,$$

which differs from the true causal effect $\delta = 0$.

To sum up, in the smoking example, the chosen "treatment" variable D is potentially negatively associated with the potential health outcome, inducing the selection bias – the difference between the predictive effect and the causal effect.

Example 2.1.3 (Analytical Version of the Smoking Example) To capture dependence between $Y(d)$ and v in the smoking context analytically, we can go back to Example 2.1.1. Here

we have the case that that the TE is $\eta_1 = 0$, and we choose variables η_0 and ν to be negatively associated:

$$\mathrm{E}[\eta_0 \nu] < 0.$$

The negative association between the η_0 and ν then results in the observed smoking status, D, being negatively associated with $Y = Y(d)$. Specifically, we have

$$\mathrm{E}[Y|D = 1] < \mathrm{E}[Y|D = 0].$$

We can verify this analytically or via simulation (a homework).

It is useful to emphasize the main reason for having selection bias is that

$$\mathrm{E}[Y(d)|D = 1] \neq \mathrm{E}[Y(d)]$$

whenever D is not independent of $Y(d)$. If D and $Y(d)$ were independent,

$$\mathrm{E}[Y(d)|D = 1] = \mathrm{E}[Y(d)]$$

would hold since in this case D is uninformative about the potential outcome and drops out from the conditional expectation.

To sum up, the problem with observational studies like our contrived example is that the "treatment" variable D is determined by individual behaviors which may be linked to potential outcomes. This linkage generates selection bias - the disagreement between APE and ATE. There are many ways of addressing selection bias, one of which is through an experiment, where we randomly assign the treatment to the units.

Random Assignment/Randomized Controlled Trials

A way to remove selection bias is through random assignment of treatment.

Assumption 2.1.3 (Random Assignment/Exogeneity) *Suppose that treatment status is randomly assigned. Namely, D is statistically independent of each potential outcome $Y(d)$ for $d \in \{0, 1\}$, which is denoted as*

$$D \perp\!\!\!\perp Y(d)$$

and $0 < \mathrm{P}(D = 1) < 1$.

This assumption states that the treatment assignment mechanism is purely random, and ensures that there are units in treatment and in control.

Example 2.1.4 (Analytical Example Continued) In the analytical example 2.1.1, Assumption 2.1.3 is satisfied if the stochastic shock ν determining D is independent of stochastic shocks η_0 and η_1 determining $Y(1)$ and $Y(0)$, i.e.

$$\nu \perp\!\!\!\perp (\eta_0, \eta_1).$$

A key result is that selection bias is removed under random assignment, which allows us to learn summaries of causal effects.

Theorem 2.1.1 (Randomization Removes Selection Bias) *Under Assumption 2.1.3, the average outcome in treatment group d recovers the average potential outcome under the treatment status d:*

$$\mathrm{E}[Y \mid D = d] = \mathrm{E}[Y(d) \mid D = d] = \mathrm{E}[Y(d)],$$

for each $d \in \{0, 1\}$. Hence the average predictive effect and average treatment effect coincide:

$$\begin{aligned} \pi &:= \mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0] \\ &= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] =: \delta. \end{aligned}$$

Assumption 2.1.3 is often not plausible for observational data. In a *randomized controlled trial* (RCT)[7] , the aim is to ensure the plausibility of Assumption 2.1.3 by direct random assignment of treatment D. That is, subjects are randomly assigned a treatment state D by the experimenter without regard to any of their characteristics. Because the random assignment of the treatment is unrelated to all subject characteristics by construction, well-executed RCTs guarantee that Assumption 2.1.3 is satisfied. Because of this property, many consider RCTs as the gold standard in causal inference, and RCTs are routinely employed in a variety of important settings.[8] Examples include evaluating the efficacy of medical treatment, vaccinations, training programs, marketing campaigns, and other kinds of interventions.

7: Synonyms are experiments and A/B tests.

8: Of course, RCTs must be correctly done to guarantee Assumption 2.1.3. For example, RCTs where experimental protocols are not followed may suffer from selection bias.

Example 2.1.5 (No Selection Bias in Experimental Data) Suppose that in the smoking example (Example 2.1.2), we worked with data where smoking or non-smoking was generated by perfectly enforced random assignment. In this case, we would have agreement between average predictive and treatment

effects: $\pi = \delta$. While it is difficult to imagine a long-run RCT where study participants could be forced to smoke or not smoke marijuana (we discuss such limitations as well as ethical considerations in Section 2.4), RCTs are routinely employed in a variety of other important settings.

Statistical Inference with Two Sample Means

Inference is based on the independent sample $\{(Y_i, D_i)\}_{i=1}^n$ obtained from an RCT, where index i denotes the observational unit. We assume that each (Y_i, D_i) has the same distribution as (Y, D). Estimation of the two means $\theta_d = \mathrm{E}[Y \mid D = d]$ for $d = 0$ and $d = 1$ can be done by considering two group means

$$\hat{\theta}_d = \frac{\mathbb{E}_n[Y1(D = d)]}{\mathbb{E}_n[1(D = d)]}.$$

The two means example can also be treated as a special case of linear regression,[9] but we find it instructive to work out the details directly for the two group means. We provide these details in Section 2.A.

9: Indeed, we can regress Y on D and $1 - D$; that is, estimate the model $Y = \theta_1 D + \theta_0(1-D) + U$. We can then apply the inferential machinery developed in the previous chapter.

Under mild regularity conditions, we have that

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_0 - \theta_0 \\ \hat{\theta}_1 - \theta_1 \end{pmatrix} \overset{\text{a}}{\sim} N(0, \mathsf{V}),$$

where

$$\mathsf{V} = \begin{pmatrix} \frac{\mathrm{Var}(Y|D=0)}{P(D=0)} & 0 \\ 0 & \frac{\mathrm{Var}(Y|D=1)}{P(D=1)} \end{pmatrix}$$

so that $\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_0$ obeys

$$\sqrt{n}(\hat{\delta} - \delta) \overset{\text{a}}{\sim} N(0, \mathsf{V}_{11} + \mathsf{V}_{22}).$$

To use this result in practice, variance components are usually estimated using the *plug-in principle*, which amounts to using the sample analogues of the expressions above.

Sometimes we are interested in relative effectiveness of treatment effects (for example, vaccine efficiency):

$$f(\theta) = (\theta_1 - \theta_0)/\theta_0 = \delta/\theta_0.$$

Relative effectiveness can be estimated by $\hat{\delta}/\hat{\theta}_0 = f(\hat{\theta})$, where $\hat{\theta} = \{\hat{\theta}_d\}_{d\in\{0,1\}}$ and $\theta = \{\theta_d\}_{d\in\{0,1\}}$, with approximate distribution obtained using the *delta method*:

$$\sqrt{n}(f(\hat{\theta}) - f(\theta)) \approx G'\sqrt{n}(\hat{\theta} - \theta) \overset{a}{\sim} N(0, G'\mathsf{V}G),$$

where $G = \nabla f(\theta)$, $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)'$, $\theta = (\theta_0, \theta_1)$.[10] .

10: The approximation follows from application of the first order Taylor expansion and continuity of the derivative ∇f at θ.

Pfizer/BioNTech Covid Vaccine RCT

Pfizer/BNTX was the first vaccine approved for emergency use in the EU and US to reduce the risk of Covid-19 disease. See the Food and Drug Administration (FDA) briefing for details about the RCT and the summary data. Volunteers were randomly assigned to receive either a treatment (2-dose vaccination) or a placebo, without knowing which they received, and the doctors making the diagnoses did not know whether a given volunteer received a vaccination or not. In other words, the trial was a double-blind randomized control trial. The results of the study are presented in the following table.

Figure 2.1: Tozinameran (Pfizer-BioNTech Covid-19 vaccine); Image Source: Wikipedia / Arne Müseler

The Notebooks 2.6.1 contain the analysis of the Pfizer-BioNTech Covid-19 Vaccine RCTs.

Efficacy Endpoint Subgroup	BNT162b2 N[a]=19965 Cases n1[b] Surveillance Time[c] (n2[d])	Placebo N[a]=20172 Cases n1[b] Surveillance Time[c] (n2[d])	Vaccine Efficacy % (95% CI)[e]
Overall	9 2.332 (18559)	169 2.345 (18708)	94.6 (89.6, 97.6)
Age group (years)			
16 to 17	0 0.003 (58)	1 0.003 (61)	100.0 (-3969.9, 100.0)
18 to 64	8 1.799 (14443)	149 1.811 (14566)	94.6 (89.1, 97.7)
65 to 74	1 0.424 (3239)	14 0.423 (3255)	92.9 (53.2, 99.8)
≥75	0 0.106 (805)	5 0.109 (812)	100.0 (-12.1, 100.0)

Figure 2.2: The aggregate data from the Pfizer RCT; source: FDA briefing.

We see that the rate of Covid-19 infection was relatively low at the time. Specifically, the treatment group saw 9 Covid-19 cases per 19,965, while the control group saw 169 cases per 20,172.

The estimated average treatment effect is about

$$-792.7 \text{ cases per } 100{,}000,$$

and the 95% confidence band is[11]

$$[-922, -664].$$

Under Assumptions 2.1.3 and 2.2.1 the confidence band suggests that the Covid-19 vaccine caused a reduction in the risk of contracting Covid-19.

11: In this example, we don't need the underlying individual data to evaluate the effectiveness of the vaccine because the potential outcomes are Bernoulli random variables with mean $\mathrm{E}[Y(d)]$ and variance $\mathrm{Var}(Y(d)) = \mathrm{E}[Y(d)](1 - \mathrm{E}[Y(d)])$.

We also compute the Vaccine Efficacy metric, which according to [8], refers to the following measure:

$$\mathsf{VE} = \frac{\text{Risk for Unvaccinated - Risk for Vaccinated}}{\text{Risk for Unvaccinated}}.$$

It describes the relative reduction in risk caused by vaccination. Estimating the VE is simple as we can plug-in the estimated group means. We can compute standard errors using the delta method or by simulation. We obtain that the overall vaccine efficacy is 94.6%, replicating the results shown in Figure 2.2. Our 95% confidence interval for VE, based on the normal approximation, is

$$[90.9\%, 98.2\%],$$

which differs only slightly from the FDA briefing table.[12]

12: The analysis in the FDA table is based on the inversion of exact binomial tests, the Cornfield procedure.

Remark 2.1.1 We notice that the confidence intervals for the VE for the two age groups of seniors are very wide, so to increase precision we pool them together and calculate the effectiveness of the vaccine for the two groups that are 65 or older. The resulting VE estimate is 95% and the two-sided confidence interval based on the normal approximation is

$$[82\%, 106\%]$$

A more refined approach is possible, based on the inversion of exact binomial ratio of Cornfield [9], which we report in Notebooks 2.6.1. This approach, using Vaccination RCT R Notebook 2.6.1, yields a confidence interval of

$$[69\%, 99\%].$$

The reason is that the accumulated counts of binomials are too few for the Gaussian approximations to provide a high-quality approximation, so the exact binomial ratio test inversion delivers a more accurate confidence interval.

2.2 Pre-treatment Covariates and Heterogeneity

Sometimes we also have additional *pre-treatment* or *pre-determined* covariates W. We might be interested in either using these covariates to estimate average effects more precisely or to describe heterogeneity of the treatment effects. For example, we might

be interested in the impact of a treatment across age or income groups.

For this purpose, we consider conditional average treatment effects (CATE):

$$\delta(W) = \mathrm{E}[Y(1) \mid W] - \mathrm{E}[Y(0) \mid W],$$

which compare the average potential outcomes conditional on a set of covariates W.

We can directly learn the conditional predictive effects (CAPE),

$$\pi(W) = \mathrm{E}[Y \mid D = 1, W] - \mathrm{E}[Y \mid D = 0, W],$$

from population data. However, these CAPE will generally not agree with the CATE. One assumption that will be sufficient for the CAPE and CATE to agree is having treatment assigned randomly and independently of covariates. As before, the use of RCTs help ensure the plausibility of this assumption.

Assumption 2.2.1 (Random Assignment Independent of Covariates) *Suppose that treatment status is randomly assigned. Namely, D is statistically independent of both the potential outcomes and a set of pre-determined covariates:*

$$D \perp\!\!\!\perp (Y(0), Y(1), W),$$

and $0 < \mathrm{P}(D = 1) < 1$.

This assumption spells out that, if we plan to use covariates in the analysis, randomization has to be made with respect to these covariates as well.

In practice, it is often tempting to use post-treatment covariates in regression analysis, but the use of such variables runs the danger of violating Assumption 2.2.1. In the extreme case, conditioning on the post-treatment observed outcome Y, we find that $\pi(Y) = 0$, even when there is a treatment effect. In a less extreme case, conditioning on post-treatment variables related to the outcome can "control-away" part of the effect, diminishing estimates.

A common scenario where accidentally using a post-treatment covariate may occur is when researchers encounter missing data from imperfect data collection in following-up with control and treated units to collect demographic information. When we drop observations with missing data, we implicitly condition

on a post-treatment variable (missingness) which can cause violations of Assumption 2.2.1.

The desire to assess randomization with respect to covariates motivates the following diagnostic procedure.

For random variables A and B, $A \sim B$ denotes that A and B have the same distribution.

Testing Covariance Balance. The random assignment assumption induces covariate balance. Namely, the distribution of covariates should be the same under both treatment and control:

$$W|D = 1 \sim W|D = 0,$$

and, equivalently,

$$D|W \sim D.$$

A useful implication is that D is not predictable by W:

$$\mathrm{E}[D \mid W] = \mathrm{E}[D].$$

This latter condition is testable using regression tools. It amounts to saying that the R^2 of a regression of D on W is 0.

Under Assumption 2.2.1, Theorem 2.1.1 continues to hold, but we now have a stronger result.

Theorem 2.2.1 (Randomization with Covariates) *Under Assumption 2.2.1, the expected value of Y conditional on treatment status $D = d$ and covariates W coincides with the expected value of potential outcome $Y(d)$ conditional on covariates W:*

$$\mathrm{E}[Y \mid D = d, W] = \mathrm{E}[Y(d) \mid D = d, W] = \mathrm{E}[Y(d)|W],$$

for each d. Hence the conditional predictive and average treatment effects agree:

$$\pi(W) = \delta(W).$$

Regression and Statistical Inference for ATEs

Empirical researchers often base statistical inference on the ATE using the classical additive linear regression model, where covariates enter additively in the model. This approach has some good practical properties and often empirically leads to improvements in precision over the simple two-means approach, though this precision improvement is not guaranteed. Another approach that we will emphasize is the interactive regression

approach, where de-meaned covariates are also interacted with the base treatment. Including interactions of de-meaned covariates with the treatment always improves precision, and it also allows us to discover treatment effect heterogeneity.

Classical Additive Approach

We begin explaining the classical additive approach. Here, to simplify the exposition, we make the strong assumption that the conditional expectation function is exactly linear:

$$\mathrm{E}[Y \mid D, W] = D\alpha + \beta' X, \tag{2.2.1}$$

where $X = (1, W)$ contains an intercept and pre-treatment covariates W. This setup is clearly restrictive, but the statistical inference result will be valid without this assumption.[13] Later in the book, we will consider fully nonlinear models.

13: See Section 2.B for details.

We assume that covariates are centered:[14]

14: Theoretically, this is implemented by redefining $W := W - \mathrm{E}[W]$. In estimation, this is implemented by redefining $W_i := W_i - \mathbb{E}_n[W]$. Recentering by empirical means is asymptotically equivalent to recentering by the true means. This is true here but is not true more generally. This can be verified using the concept of Neyman orthogonality that we develop later.

$$\mathrm{E}[W] = 0.$$

By Assumption 2.2.1, there is covariate balance:

$$\mathrm{E}[W \mid D = 1] = \mathrm{E}[W \mid D = 0].$$

Using centered covariates implies that

$$\mathrm{E}[Y(0)] = \mathrm{E}[\mathrm{E}[Y \mid D = 0, X]] = \beta_1$$

$$\mathrm{E}[Y(1)] = \mathrm{E}[\mathrm{E}[Y \mid D = 1, X]] = \beta_1 + \alpha.$$

That is, the average outcome in the untreated state is β_1, and the average treatment effect $\delta = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ equals α.

Equation (2.2.1) implies that

$$Y = D\alpha + \beta' X + \epsilon, \quad \epsilon \perp (D, X), \tag{2.2.2}$$

implying that α coincides with the coefficient in the BLP of Y on D and X.

In fact, even if we don't assume the linearity (2.2.1), we still have that $\alpha = \delta$. That is, the projection coefficient α recovers the ATE δ without the linearity assumption as we detail in Section 2.B. Furthermore the statistical inference result stated below will hold without requiring linear conditional expectation functions as it is simply a statement about inference on the BLP.

We are interested in statistical inference on the ATE and Relative ATE[15]

$$\alpha \quad \text{and} \quad \alpha/\beta_1.$$

15: Relative ATE is often called *lift* in business applications.

Under regularity conditions, application of the OLS theory from Chapter 1 gives us

$$\left(\sqrt{n}(\hat{\alpha} - \alpha),\ \sqrt{n}(\hat{\beta}_1 - \beta_1) \right)' \overset{a}{\sim} N(0, \mathsf{V}),$$

where covariance matrix V has components:

$$\mathsf{V}_{11} = \frac{\mathrm{E}[\epsilon^2 \tilde{D}^2]}{(\mathrm{E}[\tilde{D}^2])^2},\ \ \mathsf{V}_{22} = \frac{\mathrm{E}[\epsilon^2 \tilde{1}^2]}{(\mathrm{E}[\tilde{1}^2])^2},\ \ \mathsf{V}_{12} = \mathsf{V}_{21} = \frac{\mathrm{E}[\epsilon^2 \tilde{D}\tilde{1}]}{\mathrm{E}[\tilde{1}^2]\mathrm{E}[\tilde{D}^2]},$$

where $\tilde{D} = D - \mathrm{E}[D]$ is the residual after partialling out X from D linearly and $\tilde{1} := (1 - D)$ is the residual after partialling out D and W from 1.

We also obtain the approximate normality for the Relative ATE using the delta method:

$$\sqrt{n}(\hat{\alpha}/\hat{\beta}_1 - \alpha/\beta_1) \overset{a}{\sim} N(0, G'\mathsf{V}G),$$

where

$$G = [1/\beta_1, -\alpha/\beta_1^2]'.$$

Improvement in Precision under Linearity

Now we explain the role of covariates in potentially delivering improvements in precision of estimating the ATE. The underlying idea is that of "denoising." This improvement, however, hinges on the linear model (2.2.1). In the next section, we will obtain improvement without linearity assumptions.

We consider what happens when we do not include covariates in the regression. In this case, the OLS estimator $\bar{\alpha}$ estimates the projection coefficient α in the BLP using $(1, D)$ alone:[16]

$$Y = \alpha D + \beta_1 + U, \quad \mathrm{E}[U] = \mathrm{E}[UD] = 0,$$

16: Here $U = Y - \alpha D - \beta_1$ obeys

$$\begin{aligned}\mathrm{E}[U \mid D = d] &= \mathrm{E}[Y(d) - \alpha d - \beta_1 \mid D = d] \\ &= \mathrm{E}[Y(d) - \alpha d - \beta_1] = 0,\end{aligned}$$

invoking random assignment and the definition of α and β_1.

where the noise

$$U = \beta'(X - \mathrm{E}[X]) + \epsilon$$

contains the part of Y that is linearly predicted by X, $\beta'(X - \mathrm{E}[X]) = \beta' X - \beta_1$. We then have that $\bar{\alpha}$ obeys

$$\sqrt{n}(\bar{\alpha} - \alpha) \overset{a}{\sim} N(0, \bar{\mathsf{V}}_{11}),\ \ \bar{\mathsf{V}}_{11} = \frac{\mathrm{E}[U^2 \tilde{D}^2]}{(\mathrm{E}[\tilde{D}^2])^2}.$$

Under the linear model (2.2.1), it follows that

$$\mathsf{V}_{11} \leq \bar{\mathsf{V}}_{11},$$

with the inequality being strict ("<") if $\mathrm{Var}(\beta' X) > 0$.[17] That is, under (2.2.1), using pre-determined covariates improves the precision of estimating the ATE α.

17: Verify this as a reading exercise.

However, this improvement theoretically hinges on the correctness of the additive linear model. Statistical inference on the ATE based on the the normal approximation provided above remains valid without this assumption as long as robust standard errors are used.[18] However, the precision can be either higher or lower than that of the classical two-sample approach without covariates. That is, without (2.2.1), V_{11} and $\bar{\mathsf{V}}_{11}$ are not generally comparable.

18: We always use robust variance formulas throughout the book. However, the default inferential algorithms in R and Python often report the classical Student's formulas as variances, which critically rely on the linearity assumption.

Remark 2.2.1 While the inferential result we derived is robust with respect to the linearity assumption on the CEF, the improvement in precision itself is **not** guaranteed in general and hinges on the validity of the linearity assumption.

The Interactive Approach: Always Improves Precision and Discovers Heterogeneity

We can also consider estimation of CATE through the lens of an interactive linear regression model, which interacts treatment indicator D with regressors X constructed from original raw regressors W. Including these interactions respects the logic of approximating the conditional expectation of Y given D and raw regressors using linear functional forms. To simplify exposition, we first assume that the interactive model is exactly correct for the CEF:

$$\mathrm{E}[Y \mid D, W] = \alpha' X D + \beta' X. \tag{2.2.3}$$

However, this approach works without this assumption.

As before, we assume

$$X = (1, W')', \quad \mathrm{E}[W] = 0,$$

which can be achieved in practice by recentering.[19] Here, we

19: In this interactive model, using re-centering by empirical means is not equivalent to using re-centering true means. This requires adding additional variance term to the conventional variance output of least squares, which is variance of CAPE $X_i'\alpha$.

recover CATE via

$$\begin{aligned}\delta(W) &= \mathrm{E}[Y(1) \mid W] - \mathrm{E}[Y(0) \mid W] \\ &= \mathrm{E}[Y \mid D = 1, W] - \mathrm{E}[Y \mid D = 0, W] = \alpha' X.\end{aligned}$$

Using that $\mathrm{E}W = 0$, the ATE is then

$$\delta = \mathrm{E}[\delta(W)] = \mathrm{E}[\alpha' X] = \alpha_1,$$

where α_1 is the first component of α. The function $\alpha_2' W$, where α_2 is the vector all elements of α excluding α_1, therefore describes the deviation of CATE away from the ATE.

We can verify that α is the coefficient of the linear projection equation:

$$Y = \alpha' DX + \beta' X + \epsilon, \quad \epsilon \perp (X, DX).$$

Therefore, we can treat

$$\bar{D} := DX$$

as a vector of technical treatments[20] and invoke the "partialling out" approach for inference on components of α.

20: A technical treatment refers to any variable obtained as a transformation of the original treatment variable.

Remark 2.2.2 (Improvement in Precision Guarantee) Unlike the previous approach, the "interactive" approach always delivers improvements in precision for estimating δ, even if the linearity in (2.2.3) does not hold; this was demonstrated by Lin [10] and Cytrynbaum [11].

Reemployment Bonus RCT

The Notebooks 2.6.2 explore the use of covariates to improve precision and learn about heterogeneity in a Reemployment Bonus RCT.

Here we re-analyze the Pennsylvania re-employment bonus experiment [12], which was conducted in the 1980s by the U.S. Department of Labor to test the incentive effects of alternative compensation schemes for unemployment insurance (UI). In these experiments, UI claimants were randomly assigned either to a control group or one of five treatment groups. We focus our discussion on treatment group 4. In the control group the current rules of the UI applied. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period of time (qualification period), provided that the job was retained for a specified duration; see the Penn Data Codebook for further details on the data.

We consider the

- classical 2-sample approach, no adjustment (CL)
- classical linear regression adjustment (CRA)
- interactive regression adjustment (IRA)
- interactive regression adjustment with double lasso (partialling out by lasso) (IRA-DL)

We use the last approach in the spirit of exploration and experimentation. We describe the last approach and establish its validity in Chapter 4.

Estimates of the ATE on (log) unemployment duration and corresponding estimated standard errors are given in Table 2.1.

	CL	CRA	IRA	IRA-DL
Estimate	-0.0855	-0.0797	-0.0755	-0.0789
Std. Error	0.0359	0.0356	0.0356	0.0356

Table 2.1: Estimates of the ATE of the reemployment bonus on log unemployment duration..

The different estimators deliver fairly similar point estimates suggesting that treatment group 4 experiences an average decrease in unemployment duration of around 8%. The three regression estimators deliver estimates that are slightly more precise (have lower standard errors) than the simple difference in means estimator.[21]

We also see that the regression estimators offer slightly lower estimates of the ATE than the difference in means estimator. These differences likely occur due to minor imbalances in the treatment allocation: People older than 54 tended to receive the treatment more than other groups of qualified UI claimants during the later period of the experiment. Loosely speaking, the regression estimators try to correct for this imbalance by "partialling out" the effect of this oversampling[22] and averaging over differences net of these "imbalancing" effects. We will explain how regression adjustment corrects for imbalances in Chapter 5.

21: The standard errors are computed using the standard Eicker-Huber-White formula (HC0). Surprisingly, IRA performs worse than CL when we use the HC1 formula, which multiplies the HC0 standard errors by a factor of $n/(n-p)$, where p is the total number of regressors in the model. This adjustment is heuristic and is motivated to capture the effect of overfitting, assuming a homoscedastic model.

22: See the Reemployment Bonus RCT Notebooks 2.6.2 for the results from the balance check.

2.3 Drawing RCTs via Causal Diagrams

RCTs can be represented using causal diagrams, which clearly display the assumptions underlying our treatment effect model. Causal diagrams were first introduced by Sewall Wright in the 1920s ([13], [14]) and later formalized by Judea Pearl and James M. Robins ([15], [16]).

In these diagrams, nodes represent random variables, and arrows indicate causal effects (and related statistical dependencies). In our RCT setup, the assigned treatment variable D causally affects the outcome variable Y, while pre-treatment variables W also affect Y but do not influence D. The missing arrow between D and W encodes their statistical independence.

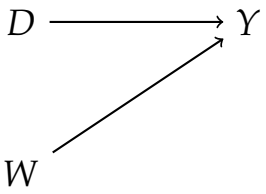

Figure 2.3: Causal Diagram for an RCT

Figure 2.4 extends this diagram by including potential outcomes as nodes.

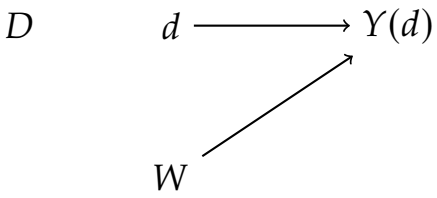

Figure 2.4: Causal Diagram for the RCT Research Design

In Figure 2.4, the potential outcomes $Y(d)$ are represented as a single node influenced by W (indicated by the arrow from W to $Y(d)$). The absence of an arrow from D to $Y(d)$ indicates that the treatment assignment is independent of the potential outcomes. The arrow from the deterministic node d to $Y(d)$ captures the causal dependency inherent in the potential outcome process. Together, the process $d \mapsto Y(d)$ and the treatment assignment D determine the realized outcome via $Y := Y(D)$.

We further develop these concepts and employ causal diagrams as a formal tool in Chapters Chapter 7 and Chapter 11.

2.4 The Limitations of RCTs

We outline key limitations of RCTs. First, we discuss identification threats by examining settings in which the Stable Unit Treatment Value Assumption (SUTVA) may fail, and the resulting implications for inference. We then address ethical, practical, and generalizability concerns.

Externalities, Stability, and Equilibrium Effects

Rubin's causal model relies on SUTVA (see Section 2.1), which requires that one unit's potential outcomes remain unaffected by others' treatment assignments [17]. However, there are settings where this assumption may not hold.

For instance, in vaccine trials SUTVA holds when treatment and control groups are "small" (infinitesimal) subpopulations of the general population, allowing us to measure average vaccine effects. If a large fraction of the population is vaccinated—thus achieving herd immunity—the outcomes for the control group may mirror those for the treated, and SUTVA fails.[23]

23: Because SUTVA does not hold in the vaccination context, it is customary to use relative measures of impact like "vaccine efficiency" because they may be a somewhat more stable measure when generalizing from "small" treated subpopulations to a "large" treated population.

In economics, such spillover effects are called externalities or, in some cases, general equilibrium effects. For example, if only a small subpopulation earns a college degree, general equilibrium wage effects are minimal. Conversely, if many obtain a degree, the equilibrium wage may adjust (reducing the college wage premium). Similarly, in large-scale training programs, an individual's outcomes may depend on the number of people trained for the same job.

Ethical, Practical, and Generalizability Concerns

Many RCTs are infeasible because implementing them can be unethical. The 1978 Belmont Report ([18]) outlines ethical principles—"Respect for persons," "Beneficence," and "Justice"—that govern research with human subjects and are enforced by institutional review boards. For example, a hypothetical RCT assigning individuals to a smoking treatment would violate "Beneficence" by causing harm, rendering such studies unethical.

RCTs may also encounter practical issues. They can be prohibitively expensive when treatment costs, data collection, or the necessary sample size for adequate power are high. Long-term RCTs are particularly challenging due to attrition, and randomizing access to a desirable treatment may be politically unfeasible.

Even when successfully implemented, RCT findings may be difficult to generalize. Local conditions or differences in implementation and intervention scale may limit the external validity of the estimated average treatment effect.

2.5 Notes

RCTs have had a profound influence on business, economics, and science. They are standard in testing drug efficacy and various programs in labor and development economics. The FDA adopted RCTs as the gold standard in the 1970s–80s, and in the tech industry and marketing, RCTs are known as "A/B Tests" and are widely used. Many major tech companies have dedicated experimental platforms.[24]

24: See, for example, ExP platform at Microsoft and the WebLab platform at Amazon.

The expansion of experimentation in economics is linked to the work of Richard Thaler (2017 Alfred Nobel Memorial Prize in Economics), Abhijit Banerjee, Esther Duflo, and Michael Kremer (2019 Alfred Nobel Memorial Prize in Economics), John List, among others.

For real examples of how RCTs are done and designed in practice, see, for example, the FDA registry of RCTs, the American Economic Association's registry of RCTs in economics, or the The Poverty Action Lab.

We have presented basic ideas here. For a detailed analysis of experimental design, please see Art Owen's lecture notes ([19]). For further reading on RCTs and causal analysis, refer to Imbens and Rubin [2] and Duflo et al. [20].

2.6 Notebooks

Notebook 2.6.1 (Vaccination RCT) Vaccination RCT R Notebook and Vaccination RCT Python Notebook contain the analysis of vaccination examples.

Notebook 2.6.2 (Reemployment Bonus) Reemployment Bonus RCT R Notebook and Reemployment Bonus RCT Python Notebook explore the use of covariates to improve precision and learn about heterogeneity in a Reemployment Bonus RCT.

2.7 Exercises

Exercise 2.7.1 (Selection Bias) Set-up a simulation experiment that illustrates the contrived smoking example, following the analytical example we've presented in the text. Illustrate the difference between estimates obtained via an RCT (smoking generated independently of potential outcomes) and an ob-

servational study (smoking choice is correlated with potential outcomes).

Exercise 2.7.2 (Vaccinations RCT) Study the notebook on vaccinations RCTs. Try to replicate the results in the FDA briefing table for age group 18-64 (exact replication is not required). Explain your calculations.

Exercise 2.7.3 (Reemployment example) Study the notebook on the reemployment example. Experiment with putting even more flexible controls (e.g. use extra interactions of some controls). Experiment with using HC0 vs. HC1 standard errors. Report and explain your findings.

Exercise 2.7.4 (RCT Design) Skim over the information on the Pfizer RCT design briefing. Write down one paragraph summarizing the study design.

Exercise 2.7.5 (AEA RCT Registry) Skim over one of the RCTs registered with AEA RCT Registry. Write down one paragraph summarizing the study design.

Exercise 2.7.6 (Stability) Think of some RCTs where stability is likely to hold and some RCTs where it likely does not.

2.A Approximate Distribution of the Two Sample Means

To demonstrate the result in the text, we note that

$$\hat{\theta}_d - \theta_d = \frac{\mathbb{E}_n[(Y(d) - \mathrm{E}Y(d))1(D = d)]}{\mathbb{E}_n[1(D = d)]}$$

for $d \in \{0, 1\}$ because we can re-write the population group average as

$$\theta_d = \mathrm{E}[Y(d)] = \mathrm{E}[Y(d)]\frac{\mathbb{E}_n[1(D = d)]}{\mathbb{E}_n[1(D = d)]}.$$

Hence, for each $d \in \{0, 1\}$,

$$\sqrt{n}(\hat{\theta}_d - \theta_d) = \sqrt{n}\frac{\mathbb{E}_n[(Y(d) - \mathrm{E}Y(d))1(D = d)]}{\mathbb{E}_n[1(D = d)]}.$$

By the law of large numbers, $\mathbb{E}_n[1(D = d)] \approx \mathrm{P}(D = d)$; so we have the approximation

$$\sqrt{n}\{\hat{\theta}_d - \theta_d\}_{d\in\{0,1\}} \approx \sqrt{n}\frac{\mathbb{E}_n[(Y(d) - \mathrm{E}Y(d))1(D = d)]}{\mathrm{P}(D = d)}.$$

Note that the terms being averaged are

$$\frac{(Y_i(d) - \mathrm{E}[Y(d)])1(D_i = d)}{\mathrm{P}(D = d)}.$$

These terms have zero mean[25] and variance

$$\frac{\mathrm{E}[(Y(d) - \mathrm{E}[Y(d)])^2 1(D = d)^2]}{\mathrm{P}(D = d)^2} = \frac{\mathrm{Var}(Y \mid 1(D = d) = 1)}{\mathrm{P}(D = d)}.$$

25: Why? Hint: Use the law of iterated expectations.

Also note the zero covariance:

$$\mathrm{E}\left[\frac{(Y(1) - \mathrm{E}[Y(1)])1(D = 1)}{\mathrm{P}(D = 1)}\frac{(Y(0) - \mathrm{E}[Y(0)])1(D = 0)}{\mathrm{P}(D = 0)}\right] = 0.$$

The application of the central limit theorem then yields the claimed result.

2.B Statistical Properties of the Classical Additive Approach*

Here we analyze statistical inference on ATE using OLS and adjusting for $X = (1, W)$, without making the linearity assumptions we made in Section 2.2.

We consider the linear projection equation in the population:

$$Y = D\alpha + X'\beta + \epsilon, \quad \epsilon \perp (D, X).$$

Here, we have that D and $X = (1, W)$ with $\mathrm{E}[W] = 0$, so that $\beta' X = \beta_1 + \beta_2' W$. Moreover, we have that $D \perp W$ in the RCT setting.

First, we'd like to verify that $\alpha = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ and $\beta_1 = \mathrm{E}[Y(0)]$. For $U := \beta_2' W + \epsilon$, we can write

$$Y = D\alpha + \beta_1 + U, \quad U \perp (1, D).$$

$U \perp (1, D)$ holds because $(1, D) \perp (W, \epsilon)$ using that $\mathrm{E}[W] = 0$ and that $D \perp (W, \epsilon)$. Therefore, $D\alpha + \beta_1$ coincides with the population projection of Y onto $(1, D)$. Hence, the projection coefficients are the same as those obtained by the 2-sample

approach in the population. Therefore, $\beta_1 = \mathrm{E}[Y(0)]$ and $\alpha = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$.

Second, we'd like to explain the details of the approximate normality for the estimators of sample OLS coefficients $\hat{\beta}_1$. The OLS theory of the first chapter implies that the OLS estimator $\hat{\alpha}$ obeys

$$\sqrt{n}(\hat{\alpha} - \alpha) \approx \sqrt{n}\frac{\mathbb{E}_n[\epsilon\tilde{D}]}{\mathbb{E}_n[\tilde{D}^2]} \overset{\text{a}}{\sim} N(0, \mathsf{V}_{11}),$$

where $\tilde{D} = D - \mathrm{E}[D]$ is the residual after partialling out X from D linearly,[26] and

$$\mathsf{V}_{11} = \frac{\mathrm{E}[\epsilon^2\tilde{D}^2]}{(\mathrm{E}[\tilde{D}^2])^2}.$$

Applying the same theory for β_1 (the intercept coefficient), yields[27]

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \approx \sqrt{n}\frac{\mathbb{E}_n[\epsilon\tilde{1}]}{\mathbb{E}_n[\tilde{1}^2]} \overset{\text{a}}{\sim} N(0, \mathsf{V}_{22}),$$

where $\tilde{1} := (1 - D)$ is the residual after partialling out D and X from 1 and

$$\mathsf{V}_{22} = \frac{\mathrm{E}[\epsilon^2\tilde{1}^2]}{(\mathrm{E}[\tilde{1}^2])^2}.$$

We can also establish that the estimators are jointly approximately normal with covariance

$$\mathsf{V}_{12} = \frac{\mathrm{E}[\epsilon^2\tilde{D}\tilde{1}]}{\mathrm{E}[\tilde{1}^2]\mathrm{E}[\tilde{D}^2]}.$$

26: Derive that $\tilde{D} = D - \mathrm{E}[D]$ from Assumption 2.2.1.

27: To explain the derivation, note that by partialling out D and W (recall that $X = (1, W)$) from 1 and Y, we obtain

$$\tilde{Y} = \beta_1\tilde{1} + \epsilon; \quad \tilde{1} := (1 - D).$$

The projection of 1 on D and W is given by D since D is binary and we've assumed $\mathrm{E}[W] = 0$.

Bibliography

[1] Donald B. Rubin. 'Estimating causal effects of treatments in randomized and nonrandomized studies.' In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701 (cited on page 42).

[2] Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015 (cited on pages 43, 59).

[3] Tyler J. VanderWeele, Guanglei Hong, Stephanie M. Jones, and Joshua L. Brown. 'Mediation and Spillover Effects in Group-Randomized Trials: A Case Study of the 4Rs Educational Intervention'. In: *Journal of the American Statistical Association* 108.502 (2013), pp. 469–482. (Visited on 02/17/2024) (cited on page 43).

[4] Peter M. Aronow and Cyrus Samii. 'Estimating average causal effects under general interference, with application to a social network experiment'. In: *The Annals of Applied Statistics* 11.4 (2017), pp. 1912 –1947. DOI: 10.1214/16-AOAS1005 (cited on page 43).

[5] Michael P. Leung. 'Treatment and Spillover Effects Under Network Interference'. In: *The Review of Economics and Statistics* 102.2 (2020), pp. 368–380 (cited on page 43).

[6] Francis J. DiTraglia, Camilo García-Jimeno, Rossa O'Keeffe-O'Donovan, and Alejandro Sánchez-Becerra. 'Identifying causal effects in experiments with spillovers and non-compliance'. In: *Journal of Econometrics* 235.2 (2023), pp. 1589–1624. DOI: https://doi.org/10.1016/j.jeconom.2023.01.008 (cited on page 43).

[7] Gonzalo Vazquez-Bare. 'Identification and estimation of spillover effects in randomized experiments'. In: *Journal of Econometrics* 237.1 (2023), p. 105237. DOI: https://doi.org/10.1016/j.jeconom.2021.10.014 (cited on page 43).

[8] Walter A Orenstein, Roger H Bernier, Timothy J Dondero, Alan R Hinman, James S Marks, Kenneth J Bart, and Barry Sirotkin. *Field evaluation of vaccine efficacy / Walter A. Orenstein ... [et al.]* 1984 (cited on page 49).

[9] Jerome Cornfield. 'A statistical problem arising from retrospective studies'. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 4. University of California Press Berkeley, CA. 1956, pp. 135–148 (cited on page 49).

[10] Winston Lin. 'Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique'. In: *Annals of Applied Statistics* 7.1 (2013), pp. 295–318 (cited on page 55).

[11] Max Cytrynbaum. 'Covariate adjustment in stratified experiments'. In: *Quantitative Economics* 15.4 (2024), pp. 971–998 (cited on page 55).

[12] Yannis Bilias. 'Sequential testing of duration data: The case of the Pennsylvania 'reemployment bonus' experiment'. In: *Journal of Applied Econometrics* 15.6 (2000), pp. 575–594 (cited on page 55).

[13] Philip G. Wright. *The Tariff on Animal and Vegetable Oils*. New York: The Macmillan company, 1928 (cited on page 56).

[14] Sewall Wright. 'Correlation and Causation'. In: *Journal of Agricultural Research* 20.7 (Jan. 1921), pp. 557–585 (cited on page 56).

[15] Judea Pearl. 'Causal diagrams for empirical research'. In: *Biometrika* 82.4 (1995), pp. 669–688 (cited on page 56).

[16] Sander Greenland, Judea Pearl, and James M. Robins. 'Causal diagrams for epidemiologic research'. In: *Epidemiology* 10.1 (1999), pp. 37–48 (cited on page 56).

[17] David R. Cox. *Planning of experiments*. Wiley, 1958 (cited on page 58).

[18] *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Tech. rep. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978 (cited on page 58).

[19] Art Owen. 'A First Course in Experimental Design: Notes from Stat 263/363'. Lecture notes. Accessed 1/17/2024. 2020 (cited on page 59).

[20] Esther Duflo, Rachel Glennerster, and Michael Kremer. 'Using randomization in development economics research: A toolkit'. In: *Handbook of Development Economics* 4 (2007), pp. 3895–3962 (cited on page 59).