Decomposing Treatment Effect Variation

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1080/01621459.2017.1407322</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:34785398">http://nrs.harvard.edu/urn-3:HUL.InstRepos:34785398</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP</a></td>
</tr>
</tbody>
</table>
Decomposing Treatment Effect Variation *

Peng Ding  Avi Feller  Luke Miratrix
UC Berkeley  UC Berkeley  Harvard GSE

May 24, 2016

Abstract

Understanding and characterizing treatment effect variation in randomized experiments has become essential for going beyond the “black box” of the average treatment effect. Nonetheless, traditional statistical approaches often ignore or assume away such variation. In the context of a randomized experiment, this paper proposes a framework for decomposing overall treatment effect variation into a systematic component that is explained by observed covariates, and a remaining idiosyncratic component. Our framework is fully randomization-based, with estimates of treatment effect variation that are fully justified by the randomization itself. Our framework can also account for noncompliance, which is an important practical complication. We make several key contributions. First, we show that randomization-based estimates of systematic variation are very similar in form to estimates from fully-interacted linear regression and two stage least squares. Second, we use these estimators to develop an omnibus test for systematic treatment effect variation, both with and without noncompliance. Third, we propose an $R^2$-like measure of treatment effect variation explained by covariates and, when applicable, noncompliance. Finally, we assess these methods via simulation studies and apply them to the Head Start Impact Study, a large-scale randomized experiment.

Key Words: Noncompliance; Heterogeneous treatment effect; Idiosyncratic treatment effect variation; Randomization inference; Systematic treatment effect variation.

*Peng Ding (Email: pengdingpku@berkeley.edu) is Assistant Professor, Department of Statistics, University of California, Berkeley. Avi Feller (Email: afeller@berkeley.edu) is Assistant Professor, Goldman School of Public Policy, University of California, Berkeley. Luke Miratrix (Email: lmiratrix@g.harvard.edu) is Assistant Professor, Harvard Graduate School of Education. We thank Alberto Abadie, Donald Rubin, participants at the Applied Statistics Seminar at the Harvard Institute of Quantitative Social Science, and colleagues at University of California, Berkeley and Harvard University for helpful comments. We gratefully acknowledge financial support from the Spencer Foundation through a grant entitled “Using Emerging Methods with Existing Data from Multi-site Trials to Learn About and From Variation in Educational Program Effects,” and from the Institute for Education Science (IES Grant #R305D150040).
1 Introduction

The analysis of randomized experiments has traditionally focused on the average treatment effect, often ignoring or assuming away treatment effect variation (e.g., Neyman 1923, Fisher 1935, Kempthorne 1952, Rosenbaum 2002). Today, understanding and characterizing treatment effect variation in randomized experiments has become essential for going beyond the “black box” of the average treatment effect. This is clear from the increasing number of papers on the topic in statistics and machine learning (Hill 2011, Athey and Imbens 2015, Wager and Athey 2015), biostatistics (Huang et al. 2012, Matsouka et al. 2014), education (Raudenbush and Bloom 2015), economics (Heckman et al. 1997, Crump et al. 2008, Djebbari and Smith 2008), political science (Green and Kern 2012, Imai and Ratkovic 2013), and other areas.

This paper proposes a framework for decomposing overall treatment effect variation in a randomized experiment into a systematic component that is explained by observed covariates, and an idiosyncratic component that is not explained by covariates (Heckman et al. 1997, Djebbari and Smith 2008). In doing so, we make several key contributions. First, we take a fully randomization-based perspective (cf. Rosenbaum 2002, Imbens and Rubin 2015), and propose estimators that are fully justified by the randomization itself. This is in contrast to much of the randomization-based method literature, where treatment effect variation is a nuisance (e.g. Rosenbaum 1999, 2007). As in Lin (2013), we show that the resulting estimator is very similar in form to linear regression with interactions between the treatment indicator and covariates. Unlike with linear regression, however, the proposed estimator does not require any modeling assumptions on the marginal outcomes.

Second, we extend these methods for intention-to-treat (ITT) analysis to incorporate noncompliance, proposing a randomized-based estimator for systematic treatment effect variation for the Local Average Treatment Effect (LATE) in the case of noncompliance (Angrist et al. 1996). We show that this estimator is nearly identical to the two-stage least squares estimator with interactions between the treatment and covariates. We believe that this is a particularly novel contribution to the recent literature seeking to reconcile the randomization-based tradition in statistics and the linear model-based perspective more common in econometrics (Abadie 2003, Imbens 2014, Imbens and Rubin 2015).

Armed with these estimators, we turn to two main practical tools of decomposing treatment effect variation. The first is an omnibus test for the presence of systematic treatment effect vari-
ation. While versions of this test have been proposed previously, largely in the context of linear models (Cox, 1984; Crump et al., 2008), our proposed test is fully randomization-based and can also account for noncompliance. The second is to develop and bound an $R^2$-like measure of the fraction of treatment effect variation explained by covariates. This builds on previous versions proposed in the econometrics literature (Heckman et al., 1997; Djebbari and Smith, 2008), again extending results to account for noncompliance. Finally, we apply these methods to the Head Start Impact Study, a large-scale randomized trial of Head Start, a Federally funded preschool program (Puma et al., 2010).

The paper proceeds as follows. Section 2 sets up a framework for decomposing treatment effect variation and randomization-based inference. Section 3 discusses estimation and testing for systematic treatment effect variation explained by the observed covariates. Section 4 gives sharp bounds and a sensitivity analysis technique for assessing idiosyncratic treatment effect variation, which are used to measure the fraction of treatment effect variation explained. Section 5 extends to the noncompliance setting, decomposing treatment effect variation into three components: a component explained by (partially observed) compliance status, a component explained by (fully observed) covariates, and a residual component. Section 6 illustrates the theory with numerical examples. Section 7 applies these methods to the Head Start Impact Study. Section 8 concludes. We relegate the technical details and some further extensions to the online Supplementary Material.

2 Framework for Treatment Effect Variation

2.1 Setup and notation

Assume that we have $n$ units in an experiment. For unit $i$, let $X_i \in \mathbb{R}^K$ denote the vector of pretreatment covariates, with the constant 1 as its first component; let $T_i$ denote the treatment indicator with 1 for treatment and 0 for control. We use the potential outcomes framework (Neyman, 1923; Rubin, 1974) to define causal effects. Under the Stable Unit Treatment Value Assumption (Rubin, 1980) that there is only one version of the treatment and no interference among units, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes of unit $i$ under treatment and control, respectively. The observed outcome, $Y_{i, \text{obs}} = T_iY_i(1) + (1 - T_i)Y_i(0)$, is a deterministic function of the treatment assignment and potential outcomes. On the difference scale, the individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. 

3
Importantly, we are conditioning on the \( n \) units at hand—the potential outcomes are fixed and pre-treatment. In other words, this is a finite population inference, rather than super population inference assuming that some variables or residuals are independent and identically distributed (IID) draws from some distribution. See, for example, Rosenbaum (2002) and Imbens and Rubin (2015). Under the potential outcomes framework, \( \{Y_i(1), Y_i(0)\}_{i=1}^n \) are all fixed numbers; the randomness of any estimator comes from the assignment mechanism only. We represent this mechanism as the vector of treatment assignments \( T = (T_1, \ldots, T_n)'. \)

### 2.2 Randomization inference for vector outcomes

To set up our overall framework, we first generalize Neyman (1923)’s classic results to vector outcomes. We consider a completely randomized experiment, with \( n_1 \) units receiving treatment and \( n_0 \) units receiving control. We are interested in estimating the finite population average treatment effect on a vector outcome \( V \in \mathbb{R}^K \):

\[
\tau_V = \frac{1}{n} \sum_{i=1}^{n} \{V_i(1) - V_i(0)\},
\]

where \( V_i(1) \) and \( V_i(0) \) are the potential outcomes of \( V \) for unit \( i \). The Neyman-type unbiased estimator for \( \tau_V \) is the difference between the sample mean vectors of the observed outcomes under treatment and control:

\[
\hat{\tau}_V = \bar{V}^{obs}_1 - \bar{V}^{obs}_0 = \frac{1}{n_1} \sum_{i=1}^{n} T_i V^{obs}_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) V^{obs}_i = \frac{1}{n_1} \sum_{i=1}^{n} T_i V_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) V_i(0).
\]

Finally, define the covariance operator:

\[
S(V) = \frac{1}{n-1} \sum_{i=1}^{n} (V_i - \bar{V})(V_i - \bar{V})^T.
\]

The following theorem, generalizing the results for scalar outcomes from Neyman (1923), demonstrates that \( \hat{\tau}_V \) is unbiased and gives its covariance matrix.

**Theorem 1.** Over all possible randomizations of a completely randomized experiment, \( \hat{\tau}_V \) is unbiased for \( \tau_V \), with \( K \times K \) covariance matrix:

\[
\text{cov}(\hat{\tau}_V) = \frac{S\{V(1)\}}{n_1} + \frac{S\{V(0)\}}{n_0} - \frac{S\{V(1) - V(0)\}}{n}.
\]
The diagonal elements of this matrix are the variances of the estimators of each component of $\tau_V$. The covariance matrix of $\hat{\tau}_V$ depends on the covariances of the potential outcomes under treatment and control, and the covariance of individual treatment effect. The last term depends on the correlation between the potential outcomes $V(1)$ and $V(0)$, and therefore cannot be identified from the observed data. When the individual treatment effects are constant for all components of $V$, the last term in the above covariance matrix vanishes, because $S\{V(1) - V(0)\} = 0_{p \times p}$. Under this assumption, we can unbiasedly estimate the sampling covariance matrix $\text{cov}(\hat{\tau}_V)$ by replacing the covariances of the potential outcomes by the sample analogues: 
\[ \hat{\text{cov}}(\hat{\tau}_V) = \frac{\hat{S}_t(V_{\text{obs}})}{n_1} + \frac{\hat{S}_0(V_{\text{obs}})}{n_0}, \]
where \[ \hat{S}_t(V_{\text{obs}}) = \frac{1}{n_t - 1} \sum_{i=1}^{n} I(T_i = t)(V_i - \bar{V}_t)(V_i - \bar{V}_t)^T \] are the sample covariance matrices of $V_{\text{obs}}$ in the treatment and control groups. Without the constant treatment effect assumption, the covariance estimator $\hat{\text{cov}}(\hat{\tau}_V)$ is conservative in the sense that the difference between the expectation of the variance estimator and the true variance is a non-negative definite matrix. In particular, the diagonal terms will all be larger.

Using the mathematical framework introduced in the Appendix, we can easily generalize Theorem 1 to more complicated experimental designs, e.g., cluster-randomized trials (Middleton and Aronow, 2015) and unbalanced $2^2$ split-plot designs (Zhao et al., 2016).

### 2.3 Decomposing Treatment Effect Variation

We now apply this general framework to treatment effect variation. We decompose the individual treatment effect, $\tau_i$, via
\[ \tau_i = Y_i(1) - Y_i(0) = X_i^T \beta + \epsilon_i \quad (i = 1, \ldots, n), \] (2)
with $\beta$ being the finite population linear regression coefficient of $\tau_i$ on $X_i$, defined by
\[ \beta = \arg \min_b \sum_{i=1}^{n} (\tau_i - X_i^T b)^2. \] (3)
Following Heckman et al. (1997) and Djebbari and Smith (2008), we call $\delta_i = X_i^T \beta$ the systematic treatment effect variation explained by the observed covariates, $X_i$, and call $\epsilon_i$ the idiosyncratic treatment effect variation unexplained by $X_i$. 

5
More generally, we can view this decomposition in a regression-style framework. Define
\[
S_{xx} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T, \quad S_{x\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} X_i \varepsilon_i, \quad S_{x\tau} = \frac{1}{n} \sum_{i=1}^{n} X_i \tau_i,
\]
where \(S_{xx} \in \mathbb{R}^{K \times K}\) and is non-degenerate, and where \(S_{x\varepsilon}, S_{x\tau} \in \mathbb{R}^{K}\). Also,
\[
S_{xt} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i(t), \quad (t = 0, 1)
\]
where \(S_{xt} \in \mathbb{R}^{K}\). These are all finite population quantities, as in they are fixed pre-randomization values. The definition of \(\beta\) gives \(S_{x\varepsilon} = 0\), i.e., \(\varepsilon_i\) and \(X_i\) have covariance zero. Therefore, in the spirit of the agnostic regression framework (e.g., Lin (2013)), the systematic component, \(\delta_i = X_i^T \beta\), is a projection of \(\tau_i\) onto the linear space spanned by \(X_i\), and the idiosyncratic treatment effect, \(\varepsilon_i\), is the corresponding residual.

Because of our finite population focus, if we observed all the potential outcomes, we could immediately calculate all individual treatment effects and apply standard linear regression theory to (2) and obtain \(\beta\). In particular, the solution of (3), i.e. the ordinary least squares (OLS) solution from regressing \(\tau\) on \(X\), is
\[
\beta = S_{xx}^{-1} S_{x\tau} = S_{xx}^{-1} S_{x1} - S_{xx}^{-1} S_{x0} \equiv \gamma_1 - \gamma_0,
\]
where \(\gamma_1 = S_{xx}^{-1} S_{x1}\) and \(\gamma_0 = S_{xx}^{-1} S_{x0}\) are the corresponding finite population regression coefficients of the potential outcomes on the covariates. Furthermore, \(e_i(1) = Y_i(1) - X_i^T \gamma_1\) and \(e_i(0) = Y_i(0) - X_i^T \gamma_0\) are the residual potential outcomes from the regression of \(Y_i(t)\) onto \(X\). Our idiosyncratic treatment variation is then the difference of residuals: \(\varepsilon_i = e_i(1) - e_i(0)\). In practice, we do not fully observe these components, but we can obtain unbiased or consistent estimates as we discuss below.

3 Systematic treatment effect variation for ITT

3.1 Randomization-based estimator

We now turn to estimating \(\beta\), as shown in (4). This has three components. The first term, \(S_{xx}\), is fully observed as all the covariates are observed. Our estimation then depends on the sample analogues of \(S_{x1}\) and \(S_{x0}\):
\[
\tilde{S}_{x1} = \frac{1}{n_1} \sum_{i=1}^{n} T_i X_i Y_i^{\text{obs}}, \quad \tilde{S}_{x0} = \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) X_i Y_i^{\text{obs}}.
\]
They are both $p \times 1$ matrices, capturing how the observed potential outcomes correlate with each of the covariates. Plug these into (4) to obtain an overall estimate of $\beta$. The physical randomization of $T$ justifies the following theorem.

**Theorem 2.** Under model (2), $S_{xx}^{-1} \hat{S}_{x1}$ and $S_{xx}^{-1} \hat{S}_{x0}$ are unbiased estimates of $\gamma_1$ and $\gamma_0$ respectively. Therefore

$$\hat{\beta}_{RI} = S_{xx}^{-1} \hat{S}_{x1} - S_{xx}^{-1} \hat{S}_{x0},$$

is an unbiased estimator for $\beta$ with covariance matrix

$$\text{cov}(\hat{\beta}_{RI}) = S_{xx}^{-1} \left[ \frac{S\{XY(1)\}}{n_1} + \frac{S\{XY(0)\}}{n_0} - \frac{S(X\tau)}{n} \right] S_{xx}^{-1}. \quad (5)$$

Here, for example, $S\{XY(0)\}$ denotes the covariance operator on new unit-level variables $X_i Y_i(0) \in \mathbb{R}^K$, made by scaling the $X_i$ vector of each unit by $Y_i(0)$. Similarly for $S\{XY(1)\}$ and $S\{X\tau\}$. This slight abuse of notation gives formulae less cluttered by subscripts and excessive annotation. As with the vector version of Neyman’s formula, the diagonals are the variances.

Furthermore, under a finite population asymptotic framework, $\hat{\beta}_{RI} \overset{a}{\sim} N(\beta, \text{cov}(\hat{\beta}_{RI}))$. We use this, and similar asymptotic relationships, to obtain confidence intervals and to conduct hypothesis testing as we describe further below. The finite asymptotic scheme embeds the finite population $\{(X_i, Y_i(1), Y_i(0), X_i, T_i)\}_{i=1}^n$ with size $n$ into a hypothetical sequence of finite populations with sizes approaching infinity, and assumes finite limiting values of the population means, variances and covariances of $X$, $Y(1)$ and $Y(0)$. For more technical discussion, see Ding (2014), Aronow et al. (2014), and Middleton and Aronow (2015); for regularity conditions of the finite population central limit theorems, see Hájek (1960) and Lehmann (1998).

The covariance formula (5) generalizes the result of Neyman (1923) for the average treatment effect, reducing to Neyman’s formula if $X_i = 1$ for all units. In Theorem 2, $S_{xx}$ is known for the finite population, rather than estimated. We can obtain a “conservative” estimate of $\text{cov}(\hat{\beta}_{RI})$ by

$$\hat{\text{cov}}(\hat{\beta}_{RI}) = S_{xx}^{-1} \left[ \frac{\hat{S}_1(XY_{obs})}{n_1} + \frac{\hat{S}_0(XY_{obs})}{n_0} \right] S_{xx}^{-1}.$$

Similar to Neyman (1923), this implicitly assumes $S(X\tau) = 0$. Under the assumption that $\varepsilon_i = 0$ (i.e., no idiosyncratic variation whatsoever) for all units, we can instead use $S(X\hat{\tau})$ with $\hat{\tau} = X_i^T \hat{\beta}_{RI}$ as a plug-in estimate for $S(X\tau)$, which will shrink the component standard errors made from the diagonal elements of the covariance matrix.
3.2 Regression with treatment-covariate interactions

We now use the results from the randomization inference to better understand the familiar case of linear regression with all treatment-covariate interactions. The classic approach assumes the model

\[ Y_{i}^{\text{obs}} = X_i^T \gamma + T_i X_i^T \beta + u_i \quad (i = 1, \ldots, n), \]  

(6)

where \( \{u_i\}_{i=1}^n \) are errors implicitly assumed to induce the randomness, and where \( \beta \) models systematic treatment effect variation, as in (2). Departing from much of the previous literature (e.g., Cox, 1984; Berrington de González and Cox, 2007; Crump et al., 2008), we study the properties of the least squares estimator under complete randomization, without assuming that model (6) is correctly specified. In particular, we do not assume any i.i.d. sampling; the assignment mechanism drives the distribution of the OLS estimator.

**Theorem 3.** The OLS estimator for \( \beta \) from fitting model (6) can be rewritten as

\[
\hat{\beta}_{\text{OLS}} = \hat{S}_{xx,1}^{-1} \hat{S}_{x1}^{-1} \hat{S}_{xx,0}^{-1} \hat{S}_{x0},
\]

where

\[
\hat{S}_{xx,t} = \frac{1}{n_t} \sum_{i=1}^{n} I(T_i = t) X_i X_i^T, \quad (t = 0, 1).
\]

Over all possible randomizations, \( \hat{S}_{xx,1}^{-1} \hat{S}_{x1}^{-1} \) and \( \hat{S}_{xx,0}^{-1} \hat{S}_{x0}^{-1} \) are consistent estimates of \( \gamma_1 \) and \( \gamma_0 \) respectively; \( \hat{\beta}_{\text{OLS}} \) therefore follows an asymptotic normal distribution with mean \( \beta \) and covariance matrix:

\[
\text{cov}(\hat{\beta}_{\text{OLS}}) = S_{xx}^{-1} \left[ \frac{S\{Xe(1)\}}{n_1} + \frac{S\{Xe(0)\}}{n_0} - \frac{S(Xe)}{n} \right] S_{xx}^{-1}.
\]

(7)

with \( e_i(1), e_i(0), \) and \( e_i \) as defined after (4).

This estimate is simply the difference between \( \hat{\gamma}_{1,\text{OLS}} = \hat{S}_{xx,1}^{-1} \hat{S}_{x1}^{-1} \) and \( \hat{\gamma}_{0,\text{OLS}} = \hat{S}_{xx,0}^{-1} \hat{S}_{x0}^{-1} \), two OLS regressions run separately on each treatment arm. For treated units, define residual \( \hat{e}_i = Y_i^{\text{obs}} - X_i^T \hat{\gamma}_{1,\text{OLS}} \), and for control units, define residual \( \hat{e}_i = Y_i^{\text{obs}} - X_i^T \hat{\gamma}_{0,\text{OLS}} \). We can drop the unidentifiable term \( S(Xe) \), estimate \( S\{Xe(1)\} \) and \( S\{Xe(0)\} \) by their sample analogues, and conservatively estimate the asymptotic covariance matrix (7) by

\[
\hat{\text{cov}}(\hat{\beta}_{\text{OLS}}) = \hat{S}_{xx,1} \left[ \frac{\hat{S}_{1}(Xe)}{n_1} \right] \hat{S}_{xx,1}^{-1} + \hat{S}_{xx,0} \left[ \frac{\hat{S}_{0}(Xe)}{n_0} \right] \hat{S}_{xx,0}^{-1}.
\]

(8)
This form of sandwich variance estimator is nearly identical to the Huber–White variance estimator for linear model \( \text{Lin, 2013} \) (Angrist and Pischke, 2008).

Importantly, \( \hat{\beta}_{RI} \) and \( \hat{\beta}_{OLS} \) are quite similar in form. In particular, \( \hat{\beta}_{RI} \) uses the true \( S_{xx} \) while \( \hat{\beta}_{OLS} \) separately estimates the covariance matrix for each treatment arm, \( \hat{S}_{xx,0} \) and \( \hat{S}_{xx,1} \). It is effectively a ratio estimator. Although this introduces some small bias (on the order of \( 1/n \)), using the estimated \( \hat{S}_{xx,t} \) rather than true \( S_{xx} \) can often lead to gains in precision, especially when covariates are strongly correlated with the potential outcomes. For related discussion, see Cochran (1977) on ratio estimators in surveys.

### 3.3 Omnibus test for systematic variation

Finally, we can use these results to develop an omnibus test for the presence of any systematic treatment effect variation. The null hypothesis of no treatment effect variation explained by the observed covariates can be characterized by

\[
H_0(X) : \beta_1 = 0,
\]

where \( \beta_1 \) contains all the components of \( \beta \) except the first component corresponding to the intercept. Under \( H_0(X) \), the individual treatment effect does not linearly depend on \( X \).

We then construct a Wald-type test for \( H_0(X) \) using an estimator \( \hat{\beta} \) and its covariance estimator \( \hat{\text{cov}}(\hat{\beta}) \); it could be \( \hat{\beta}_{RI} \) or \( \hat{\beta}_{OLS} \). Let \( \hat{\beta}_1 \) and \( \hat{\text{cov}}(\hat{\beta}_1) \) denote the sub-vector of \( \hat{\beta}_{RI} \) and sub-matrix of \( \hat{\text{cov}}(\hat{\beta}_{RI}) \), corresponding to the non-intercept coordinates of \( X \). We reject when

\[
\hat{\beta}_1^T \hat{\text{cov}}^{-1}(\hat{\beta}_1) \hat{\beta}_1 > q_{\chi^2, K-1}(1 - \alpha), \tag{8}
\]

where \( q_{\chi^2, K-1}(1 - \alpha) \) is the \( 1 - \alpha \) quantile of the \( \chi^2 \) random variable with degrees of freedom \( K - 1 \).

The test in (8) is nearly identical to the test proposed by Crump et al. (2008). They relax the parametric assumption by taking a “sieve estimator” approach, namely by using a quadratic form of the regression function, which allows for more flexible marginal distributions. Our approach differs in that we avoid modeling the marginal distributions entirely. As in Crump et al. (2008), we can add polynomials of \( X \) (or other basis functions) into the model for \( \delta \) to allow for more flexible systematic treatment effect variation.
3.4 Additional considerations

In the Supplementary Material, we describe two additional points about systematic treatment effect variation that we briefly address here.

First, we can use model-assisted estimation to improve both the randomization-based and OLS estimators. In particular, \( \widehat{S}_{xt} \) plays an important role in both \( \widehat{\beta}_{RI} \) and \( \widehat{\beta}_{OLS} \). We show that we can replace the standard sample estimator, \( \widehat{S}_{xt} \), by a more efficient, model-assisted estimator, as in survey sampling (Cochran 1977). As we show in simulations, this can lead to meaningful precision gains in practice. More importantly, this setup allows researchers to assess systematic variation across one set of covariates while adjusting for another set.

Second, under the assumption of no idiosyncratic variation (i.e., \( \varepsilon_i = 0 \) for all \( i \)), we can obtain exact inference for \( \beta \) by inverting a sequence of randomization-based tests. This complements previous work on randomization-based tests for the presence of idiosyncratic treatment effect variation (Ding et al., 2016).

4 Idiosyncratic treatment effect variation for ITT

After characterizing the systematic component of treatment effect variation, the next question is characterizing the idiosyncratic component. Since this quantity is inherently unidentifiable, we propose sharp bounds on this component and a framework for sensitivity analysis. We then leverage these results to bound an \( R^2 \)-like measure of treatment effect variation explained by covariates. In the Supplementary Material we also show that we can use these results to obtain sharp bounds on the variance of Neyman (1923)’s estimate of the Average Treatment Effect, extending previous work by Heckman et al. (1997) and Aronow et al. (2014).

4.1 Bounds

We first define the main quantities of interest:

\[
S_{\tau\tau} = \frac{1}{n} \sum_{i=1}^{n} (\tau_i - \tau)^2, \quad S_{\delta\delta} = \frac{1}{n} \sum_{i=1}^{n} (\delta_i - \tau)^2, \quad S_{\varepsilon\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2,
\]

with \( \delta_i \) and \( \varepsilon_i \) defined as in (2). Then \( S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon} \). We can immediately estimate \( S_{\delta\delta} \) via the sample variance of \( \{ \widehat{\delta}_i = X_i' \times \widehat{\beta} \}_{i=1}^{n} \), where \( \widehat{\beta} \) is a consistent estimator, e.g., \( \widehat{\beta}_{RI} \) or \( \widehat{\beta}_{OLS} \). However, the
idiosyncratic variance, $S_{\varepsilon\varepsilon}$, is inherently unidentifiable because it depends on the joint distribution of potential outcomes.

We can, however, derive sharp bounds for $S_{\varepsilon\varepsilon}$. Let $F_1(y)$ and $F_0(y)$ be the empirical cumulative distribution functions of $\{e_i(1)\}_{i=1}^n$ and $\{e_i(0)\}_{i=1}^n$. Below we denote $e(t)$ as a random variable taking equal probabilities on $n$ values of $\{e_i(t)\}_{i=1}^n$.

Based on Fréchet–Hoeffding bounds (Hoeffding, 1941; Fréchet, 1951; Nelsen, 2007), we can bound $S_{\varepsilon\varepsilon}$ as follows.

**Theorem 4.** $S_{\varepsilon\varepsilon}$ has sharp bounds $\underline{S}_{\varepsilon\varepsilon} \leq S_{\varepsilon\varepsilon} \leq \overline{S}_{\varepsilon\varepsilon}$, where

$$\underline{S}_{\varepsilon\varepsilon} = \int_0^1 \{F_1^{-1}(u) - F_0^{-1}(u)\}^2 du, \quad \overline{S}_{\varepsilon\varepsilon} = \int_0^1 \{F_1^{-1}(u) - F_0^{-1}(1-u)\}^2 du$$

are attainable when $e(1)$ and $e(0)$ have the same ranks and opposite ranks, respectively.

The lower bound of $S_{\varepsilon\varepsilon}$ corresponds to a rank-preserving relationship between $e(1)$ and $e(0)$, and the upper bound of $S_{\varepsilon\varepsilon}$ corresponds to an anti-rank-preserving relationship between $e(1)$ and $e(0)$. Equivalently, they correspond to the cases where the Spearman rank correlation coefficients between $e(1)$ and $e(0)$ are $+1$ and $-1$.

In practice, we can often sharpen these bounds because we are unlikely to have negatively associated potential outcomes after adjusting for covariates. If we assume a nonnegative correlation between $e(1)$ and $e(0)$, we have the following corollary:

**Corollary 1.** If the correlation between $e(1)$ and $e(0)$ is nonnegative, then the bounds for $S_{\varepsilon\varepsilon}$ become $\underline{S}_{\varepsilon\varepsilon} \leq S_{\varepsilon\varepsilon} \leq V_1 + V_0$, where $V_t$ is the variance of $e(t)$ for $t = 0, 1$.

We can consistently estimate each quantity: $S_{\delta\delta}$ by the sample variance of $X_i^T\hat{\beta}$, and $F_{e1}(y)$ and $F_{e0}(y)$ by $\hat{F}_1(y)$ and $\hat{F}_0(y)$, the empirical cumulative distribution functions of the residuals $\tilde{e}_i$ under treatment and control.

A variance ratio test. Finally, while the relationship between $e(0)$ and $e(1)$ is inherently unidentifiable, there is some information in the data about the relationship between $\varepsilon_i$, the individual-level idiosyncratic treatment effect, and $Y_i(0)$, the control potential outcome. In particular, Raudenbush and Bloom (2015) noted that if the variance of the treatment potential outcomes is smaller than the control potential outcomes, we can use the variance ratio test to infer the relationship between $\varepsilon_i$ and $Y_i(0)$.
the variance of the control potential outcomes, then the treatment effect must be negatively associated with the control potential outcomes. In the Supplementary Material, we extend this result to incorporate covariates and propose a formal test.

4.2 Sensitivity analysis

Going beyond worst-case bounds, we can assess the sensitivity of our estimate of $S_{\varepsilon\varepsilon}$ to different assumptions of the dependence between potential outcomes. Using the probability integral transformation, represent the residual potential outcomes as

$$e(1) = F_1^{-1}(U_1), \quad e(0) = F_0^{-1}(U_0), \quad U_1, U_0 \sim \text{Uniform}(0, 1),$$

where $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ is the quantile function. Therefore, the dependence of the potential outcomes is determined by the dependence of the uniform random variables $U_1$ and $U_0$, which are the standardized ranks of the potential outcomes. When $U_1 = U_0$, $S_{\varepsilon\varepsilon}$ attains the lower bound $S_{\varepsilon\varepsilon}$; when $U_1 = 1 - U_0$, $S_{\varepsilon\varepsilon}$ attains the upper bound $S_{\varepsilon\varepsilon}$; when $U_1 \perp\!\!\!\perp U_0$, $S_{\varepsilon\varepsilon}$ attains the improved upper bound $V_1 + V_0$.

Rather than simply examine extreme scenarios of $S_{\varepsilon\varepsilon}$, we can instead represent $U_1$ as a mixture of $U_0$ and another independent uniform random variable $V_0$:

$$U_1 \sim \rho U_0 + (1 - \rho) V_0, \quad U_0, V_0 \overset{\text{iid}}{\sim} \text{Uniform}(0, 1),$$

which the sensitivity parameter $\rho$ captures the association between $U_1$ and $U_0$. An immediate interpretation of $\rho$ is the proportion of rank preserved units, with the other $1 - \rho$ as the proportion of units with independent treatment and control residual outcomes. When $\rho = 0$, $U_1 \perp\!\!\!\perp U_0$, and the residual potential outcomes are independent; when $\rho = 1$, $U_1 = U_0$, and the residual potential outcomes have the same ranks. The values between $(0, 1)$ corresponds to positive rank correlation but not full rank preservation. Note that the representation of the joint distribution is not unique, because we can choose any copula as a joint distribution of $(U_1, U_0)$ (Nelsen 2007). We choose the above representation and notation $\rho$ for the following theorem.

**Theorem 5.** $\rho$ is Spearman’s rank correlation coefficient between $e(1)$ and $e(0)$. Furthermore, $S_{\varepsilon\varepsilon}$ is a linear function of $\rho$:

$$S_{\varepsilon\varepsilon}(\rho) = \rho S_{\varepsilon\varepsilon} + (1 - \rho)(V_1 + V_0).$$
In practice, we cannot extract any information about $\rho$ from the data, and therefore we treat $\rho$ as a sensitivity parameter. We can choose a plausible range of $\rho$, and obtain corresponding values $S_{\varepsilon\varepsilon}$.

### 4.3 Fraction of treatment effect variation explained

A natural question is the relative magnitudes of $S_{\delta\delta}$ and $S_{\varepsilon\varepsilon}$ [Djebbari and Smith, 2008]. Continuing the regression analogy, this is an $R^2$-like measure for the proportion of total treatment effect variation explained by the systematic component:

$$R^2_\tau = \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}},$$

which is the ratio between the finite population variances of $\delta$ and $\tau$. As above, we can directly estimate $S_{\delta\delta}$ but must bound $S_{\varepsilon\varepsilon}$. Applying Theorem 4, we obtain the following bounds on $R^2_\tau$.

**Corollary 2.** The sharp bounds on $R^2_\tau$ are

$$\frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}} \leq R^2_\tau \leq \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}}.$$

If we further assume that the correlation between $e(1)$ and $e(0)$ is nonnegative, the sharp bounds on $R^2_\tau$ are

$$\frac{S_{\delta\delta}}{S_{\delta\delta} + V_1 + V_0} \leq R^2_\tau \leq \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}}.$$

We estimate these bounds via plug-in estimates. Note that Djebbari and Smith (2008) explore a similar quantity by using a permutation approach to approximate the Fréchet–Hoeffding upper and lower bounds.

Finally, we can use the sensitivity results for $S_{\varepsilon\varepsilon}$, with values of $\rho \in [0, 1]$:

$$R^2_\tau(\rho) = \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}(\rho)}.$$

### 5 Noncompliance

#### 5.1 Setup

We now extend our results to allow for noncompliance. Let $T$ be the indicator of treatment assigned, $D$ be the indicator of treatment received, $Y$ be outcome of interest, and $X$ be pretreatment covariates. Under the Stable Unit Treatment Value Assumption, we define $D_i(t)$ and $Y_i(t)$ as
the potential outcomes for unit \( i \) under treatment assignment \( t \). Following Angrist et al. (1996) and Frangakis and Rubin (2002), we can classify units into four compliance types based on the joint values of \( D_i(1) \) and \( D_i(0) \):

\[
U_i = \begin{cases} 
\text{Always Taker} & \text{if } D_i(1) = 1, D_i(0) = 1, \\
\text{Never Taker} & \text{if } D_i(1) = 0, D_i(0) = 0, \\
\text{Complier} & \text{if } D_i(1) = 1, D_i(0) = 0, \\
\text{Defier} & \text{if } D_i(1) = 0, D_i(0) = 1.
\end{cases}
\]

We then make the monotonicity assumption, which requires \( D_i(1) \geq D_i(0) \) and rules out the existence of Defiers. We also assume exclusion restrictions for Always Takers and Never Takers, which states that \( Y_i(1) = Y_i(0) \) for all units with \( D_i(1) = D_i(0) \). In other words, we assume that treatment assignment has no effect on the outcome for Always Takers and Never Takers. As a result, treatment effect variation is trivially zero for Always Takers and Never Takers.

We are therefore interested in treatment effect variation among Compliers, which motivates the following decomposition:

\[
\tau_i = Y_i(1) - Y_i(0) = \begin{cases} 
0, & \text{if } U_i = a \text{ or } n, \\
X_i^T \beta_c + \varepsilon_i, & \text{if } U_i = c,
\end{cases}
\]

where \( \beta_c \) is the regression coefficient of \( \tau_i \) on \( X_i \) among Compliers, analogous to (2). Finally, we estimate the proportion of compliance types \( \pi_u \) of stratum \( U = u \) using the observed counts of units classified by \( T \) and \( D \). Let \( n_{td} = \# \{ i : T_i = t, D_i = d \} \) for \( t, d = 0, 1 \) and \( n_t = \# \{ i : T_i = t \} \) for \( t = 0, 1 \). Then \( \hat{\pi}_n = n_{10}/n_1, \hat{\pi}_a = n_{01}/n_0, \) and \( \hat{\pi}_c = n_{11}/n_1 - n_{01}/n_0 \).

### 5.2 Systematic treatment effect variation among Compliers

#### 5.2.1 Randomization inference

We now extend the results of Section 3 to estimate systematic treatment effect variation among Compliers. Define

\[
S_{xx,u} = \frac{1}{n_u} \sum_{i=1}^{n} I(U_i = u) X_i X_i^T, \quad S_{xt,u} = \frac{1}{n_u} \sum_{i=1}^{n} I(U_i = u) X_i Y_i(t) \quad (t = 0, 1)
\]

for \( u = a, c, n \). Then, analogous to Equation (4),

\[
\beta_c = S_{xx,c}^{-1} S_{x1,c} - S_{x0,c} = S_{xx,c}^{-1} S_{x1,c} - S_{xx,c}^{-1} S_{x0,c} \equiv \gamma_{1c} - \gamma_{0c},
\]

where

\[
\gamma_{1c} = S_{xx,c}^{-1} S_{x1,c}, \quad \gamma_{0c} = S_{xx,c}^{-1} S_{x0,c}
\]
are the linear regression coefficients of \( Y(1) \) and \( Y(0) \) on covariates \( X \) among Compliers.

Unlike in the ITT case, we cannot estimate these quantities directly. Instead, following standard results from noncompliance (e.g., [Angrist et al. 1996, Abadie 2003, Angrist and Pischke 2008]), we use estimates from observed subgroups to estimate the desired quantities of interest. Define sample moments:

\[
\tilde{S}_{xx,td} = \frac{1}{n_t} \sum_{i=1}^{n} I(T_i=t)I(D_i=d)X_iX_i', \quad \tilde{S}_{xt,td} = \frac{1}{n_t} \sum_{i=1}^{n} I(T_i=t)I(D_i=d)X_iY_i^{\text{obs}} \quad (t, d = 0, 1).
\]

The following theorem connects these quantities with the finite population quantities in (10).

**Theorem 6.** Over all possible randomizations of a completely randomized experiment, both \( \tilde{S}_{xx}(1) = \tilde{S}_{xx,11} - \tilde{S}_{xx,01} \) and \( \tilde{S}_{xx}(0) = \tilde{S}_{xx,00} - \tilde{S}_{xx,10} \) are unbiased for \( \pi_cS_{xx,c} \), and

\[
E(\tilde{S}_{x1,11} - \tilde{S}_{x0,01}) = \pi_cS_{x1,c}, \quad E(\tilde{S}_{x0,00} - \tilde{S}_{x1,10}) = \pi_cS_{x0,c}.
\]

This theorem shows that we can obtain unbiased estimates for all terms in Equation (10). The following corollary shows that we can then obtain consistent estimates for \( \gamma_{1c}, \gamma_{0c}, \) and \( \beta_c \).

**Corollary 3.** \( \hat{\gamma}_{1c,R1} = \tilde{S}_{xx}^{-1}(1)(\tilde{S}_{x1,11} - \tilde{S}_{x0,01}) \) and \( \hat{\gamma}_{0c,R1} = \tilde{S}_{xx}^{-1}(0)(\tilde{S}_{x0,00} - \tilde{S}_{x1,10}) \) are consistent for \( \gamma_{1c} \) and \( \gamma_{0c} \). Furthermore, \( \hat{\beta}_{c,R1} = \hat{\gamma}_{1c,R1} - \hat{\gamma}_{0c,R1} \) is consistent for \( \beta_c \) and follows an asymptotic normal distribution with covariance matrix

\[
\text{cov}(\hat{\beta}_{c,R1}) = (\pi_cS_{xx,c})^{-1}\left[ \frac{S\{Xe'(1)\}}{n_1} + \frac{S\{Xe'(0)\}}{n_0} - \frac{S(Xe)}{n} \right] (\pi_cS_{xx,c})^{-1},
\]

where we define the residual potential outcomes to be:

\[
\begin{align*}
e_i'(1) = \begin{cases} Y_i(1) - X_i'\gamma_{1c}, & U_i = a, \\
Y_i(1) - X_i'\gamma_{0c}, & U_i = c.
\end{cases}
\end{align*}
\]

\[
\begin{align*}
e_i'(0) = \begin{cases} Y_i(0) - X_i'\gamma_{1c}, & U_i = a, \\
Y_i(0) - X_i'\gamma_{0c}, & U_i = c.
\end{cases}
\end{align*}
\]

The idiosyncratic variation is \( \varepsilon_i = \varepsilon_i'(1) - \varepsilon_i'(0) \) for unit \( i \), with \( \varepsilon_i = 0 \) for Never Takers and Always Takers, and with \( \varepsilon_i \) for Compliers as in [9]. The two sets of residuals are not formed from a regression on all units, but instead the population regression on just Compliers. As in the ITT case, we can estimate \( S\{Xe'(1)\} \) and \( S\{Xe'(0)\} \) using their sample analogues; \( S(Xe) \), however, is unidentifiable. For units with \( D_i = 1 \), we define the residual \( \tilde{e}_i' = Y_i^{\text{obs}} - X_i'\gamma_{1c,R1} \), and for units with \( D_i = 0 \), we define the residual \( \tilde{e}_i' = Y_i^{\text{obs}} - X_i'\gamma_{0c,R1} \). Therefore, we can obtain a conservative estimate for the asymptotic covariance (13) by the following sandwich form:

\[
\hat{\text{cov}}(\hat{\beta}_{c,R1}) = \tilde{S}_{xx}^{-1}(1) \left[ \frac{\tilde{S}_1(X\tilde{e}')}{n_1} \right] \tilde{S}_{xx}^{-1}(1) + \tilde{S}_{xx}^{-1}(0) \left[ \frac{\tilde{S}_0(X\tilde{e}')}{n_0} \right] \tilde{S}_{xx}^{-1}(0).
\]
As with the ITT analog, so long as we have monotonicity and exclusion restrictions, randomization itself fully justifies the theorem and estimators without relying on a model of the observed outcomes.

5.2.2 Two-Stage Least Squares

We now turn to the standard two stage least squares (TSLS) setting in econometrics (e.g., Angrist and Pischke 2008). First, we impose a linear regression model with treatment-covariate interactions:

\[ Y_{i}^{\text{obs}} = X_i^\top \gamma + D_i X_i^\top \beta + u_i \quad (i = 1, \ldots, n). \]

Here, the randomness of the observed outcome comes from the randomness of \( D_i \) and \( u_i \). In the language of econometrics, the treatment received is “endogenous,” i.e., \( D_i \) and the error term \( u_i \) are assumed to be correlated; we therefore use \( T_i \) as an instrument for \( D_i \). The TSLS estimates \((\hat{\gamma}_{\text{TSLS}}, \hat{\beta}_{\text{TSLS}})\) is the solution of the following estimating equations:

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i}{T_i X_i} \right) \left( Y_{i}^{\text{obs}} - X_i^\top \hat{\gamma}_{\text{TSLS}} - D_i X_i^\top \hat{\beta}_{\text{TSLS}} \right) = 0.
\]

This approach is based on \( M \)-estimation, though there are many other ways to formalize the TSLS estimator (e.g., Imbens 2014). The following theorem shows that the fully-interacted TSLS estimator \( \hat{\beta}_{\text{TSLS}} \) is consistent for \( \beta_c \) across randomizations.

**Theorem 7.** Over all randomizations, the TSLS estimator \( \hat{\beta}_{\text{TSLS}} \) follows an asymptotic normal distribution with mean \( \beta_c \) and covariance matrix

\[
(\pi_c S_{xx,c})^{-1} \left[ \frac{S\{Xe''(1)\}}{n_1} + \frac{S\{Xe''(0)\}}{n_0} - \frac{S(X\varepsilon)}{n} \right] (\pi_c S_{xx,c})^{-1},
\]

where the residual potential outcomes are defined as

\[
e_i''(1) = \begin{cases} Y_i(1) - X_i^\top (\gamma_\infty + \beta_c), \\ Y_i(1) - X_i^\top \gamma_\infty, \\ Y_i(1) - X_i^\top (\gamma_\infty + \beta_c), \end{cases} \quad e_i''(0) = \begin{cases} Y_i(0) - X_i^\top (\gamma_\infty + \beta_c), \\ Y_i(0) - X_i^\top \gamma_\infty, \\ Y_i(0) - X_i^\top \gamma_\infty, \end{cases}
\]

where \( \gamma_\infty \) is the probability limit of the TSLS regression coefficient, \( \hat{\gamma}_{\text{TSLS}} \), and the idiosyncratic treatment effect is \( \varepsilon_i \equiv e_i''(1) - e_i''(0) \).

For variance estimation, define the residual as \( \hat{e}_i'' = Y_i^{\text{obs}} - X_i^\top (\hat{\gamma}_{\text{TSLS}} + \hat{\beta}_{\text{TSLS}}) \) for units with \( D_i = 1 \) and \( \hat{e}_i'' = Y_i^{\text{obs}} - X_i^\top \hat{\gamma}_{\text{TSLS}} \) for units with \( D_i = 0 \). We can then use the following sandwich
variance estimator

$$\widehat{\text{cov}}(\widehat{\beta}_{TSLS}) = \widehat{S}_{xx}^{-1}(1) \left[ \frac{\widehat{S}_{1}(X'e''n)}{n_1} \right] \widehat{S}_{xx}^{-1}(1) + \widehat{S}_{xx}^{-1}(0) \left[ \frac{\widehat{S}_{0}(X'e'0)}{n_0} \right] \widehat{S}_{xx}^{-1}(0),$$

which has the same probability limit as the Huber–White covariance estimator for \(\widehat{\beta}_{TSLS}\).

Therefore, the randomization itself effectively justifies the use of TSLS for estimating systematic treatment effect variation among Compliers, extending our ITT results. Because \(\widehat{\beta}_{c,RI}\) and \(\widehat{\beta}_{TSLS}\) are inherently ratio estimators, there is no guarantee of unbiasedness across randomizations. Finally, while \(\widehat{\beta}_{TSLS}\) is a consistent estimator for \(\beta_c\), \(\widehat{\gamma}_{TSLS}\) is not, in general, a consistent estimator for \(\gamma_{c0}\); that is, \(\gamma_{\infty} \neq \gamma_{c0}\). Instead, \(\widehat{\gamma}_{TSLS}\) converges to \(\gamma_{\infty} = \gamma_0 - \pi_a S_{xx}^{-1} S_{xx,a} \beta_c\), with \(\gamma_0\) being the population OLS regression coefficient, among all Compliers and Never Takers, of \(Y(0)\) on \(X\). In the special case of one-sided noncompliance (i.e., \(\pi_a = 0\)), \(\gamma_{\infty} = \gamma_0\).

5.2.3 Omnibus test for systematic treatment effect variation among Compliers

With a point estimator \(\widehat{\beta}\) and a covariance estimator \(\widehat{\text{cov}}(\beta)\) for \(\beta_c\), we can use the same Wald-type \(\chi^2\) test as in Equation (8) for the presence or absence of systematic treatment effect variation among Compliers. Here, the estimator can be either randomization-based \(\widehat{\beta}_{c,RI}\) or TSLS estimator \(\widehat{\beta}_{TSLS}\). The degrees of freedom are the same, \(K - 1\). Unlike in the ITT case, we are not aware of existing tests for systematic treatment effect variation among Compliers.

5.3 Idiosyncratic treatment effect variation with noncompliance

5.3.1 Bounding idiosyncratic variation

We now turn to decomposing the overall treatment effect in the presence of noncompliance. In this setting, we have three sources of treatment effect variation: (1) systematic treatment effect variation among Compliers, (2) idiosyncratic treatment effect variation among Compliers, and (3) treatment effect variation due to noncompliance.

First, recall that total treatment effect variation is \(S_{\tau \tau} = \sum_{i=1}^{n} (\tau_i - \tau)^2 / n\). We can define a similar quantity among Compliers:

$$S_{\tau \tau,c} = \frac{1}{n_c} \sum_{i=1}^{n} I(U_i = c) (\tau_i - \tau_c)^2.$$

As in Section 4, we can decompose this variation into systematic and idiosyncratic treatment effect variation.
variation for Compliers, respectively:

\[ S_{\delta c} = \frac{1}{n_c} \sum_{i=1}^{n_c} I(U_i = c) (\delta_i - \tau_c)^2, \quad S_{\varepsilon c} = \frac{1}{n_c} \sum_{i=1}^{n_c} I(U_i = c) \varepsilon_i^2. \]

Because treatment effects for Never Takers and Always Takers are zero, there is no treatment effect variation for these units. The component of treatment effect variation due to compliance status is

\[ S_{\tau c,U} = \sum_{u=c,a,n} \pi_u (\tau_u - \tau)^2. \]

Using \( \tau_a = \tau_n = 0 \) and \( \tau = \pi_c \tau_c \) due to the exclusion restrictions, we have the following theorem summarizing the relationships among the above components.

**Theorem 8.** \( S_{\tau c} = \pi_c S_{\tau c}, S_{\tau c,U} = S_{\delta c} + S_{\varepsilon c}, \) and \( S_{\tau c,U} = \pi_c (1 - \pi_c) \tau_c^2. \)

In words, total treatment effect variation has three parts: (1) systematic treatment effect variation among Compliers, \( \pi_c S_{\delta c}; \) (2) idiosyncratic treatment effect variation among Compliers, \( \pi_c S_{\varepsilon c}; \) (3) treatment effect variation due to noncompliance, \( S_{\tau c,U}. \)

As in the ITT case, even though \( S_{\varepsilon c} \) is not identifiable, we can derive bounds in terms of the marginal distributions of the residuals, \( \{e_i'(1) = Y_i(1) - X_i' \gamma_1 : U_i = c, i = 1, \ldots, n\} \) and \( \{e_i'(0) = Y_i(0) - X_i' \gamma_0 : U_i = c, i = 1, \ldots, n\} \), denoted by \( F_{1c}(y) \) and \( F_{0c}(y) \), and with marginal variances, \( V_{1c} \) and \( V_{0c} \). We show this with the following theorem.

**Corollary 4.** Sharp bounds on \( S_{\varepsilon c} \) are \( \underline{S}_{\varepsilon c} \leq S_{\varepsilon c} \leq \overline{S}_{\varepsilon c}, \) where

\[ \underline{S}_{\varepsilon c} = \int_0^1 \{F_{1c}^{-1}(u) - F_{0c}^{-1}(u)\}^2 \, du \quad \overline{S}_{\varepsilon c} = \int_0^1 \{F_{1c}^{-1}(u) - F_{0c}^{-1}(1 - u)\}^2 \, du \]

are attainable when \( \{e_i'(1) : U_i = c, i = 1, \ldots, n\} \) and \( \{e_i'(0) : U_i = c, i = 1, \ldots, n\} \) have the same ranks and opposite ranks, respectively. If we further assume that \( e_i'(1) \) and \( e_i'(0) \) for those units with \( U_i = c \) have nonnegative correlation, the upper bound can be further sharpened to \( V_{1c} + V_{0c}. \)

We defer discussion of estimating \( F_{tc}(y) \) and other technical details to the Supplementary Material.

**5.3.2 Treatment effect decomposition**

Finally, we return to estimating the fraction of treatment effect variation explained. Since there are two sources of variation—covariates and noncompliance—there are three possible \( R^2 \)-type measures.
First, we can measure the treatment effect variation explained by noncompliance alone (i.e., only \( U \)):

\[
R^2_{\tau,U} = \frac{S_{\tau\tau,U}}{S_{\tau\tau}} = \frac{S_{\tau\tau,U}}{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c} + \pi_c S_{\varepsilon\varepsilon,c}}.
\]

Second, we can measure the proportion of treatment effect variation among Compliers explained by covariates (i.e., only \( X \)):

\[
R^2_{\tau,c} = \frac{S_{\delta\delta,c}}{S_{\tau\tau,c}} = \frac{S_{\delta\delta,c}}{S_{\delta\delta,c} + S_{\varepsilon\varepsilon,c}}.
\]

Third, we can measure the treatment effect variation explained by covariates and noncompliance (i.e., both \( X \) and \( U \)):

\[
R^2_{\tau,U,X} = \frac{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c}}{S_{\tau\tau}} = \frac{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c}}{S_{\tau\tau,U} + \pi_c S_{\delta\delta,c} + \pi_c S_{\varepsilon\varepsilon,c}}.
\]

For each measure, we can use Corollary 4 to construct bounds, or conduct sensitivity analysis as Section 4.2, with the sensitivity parameter expressed as the Spearman correlation between the treatment and control potential outcomes among Compliers.

6 Simulation study

6.1 ITT estimators

We simulate completely randomized experiments to evaluate the finite sample performance of the tests for systematic treatment effect variation based on \( \hat{\beta}_{OLS} \), \( \hat{\beta}_{RI} \), and the model-assisted version, \( \hat{\beta}^w_{RI} \), discussed in the Supplementary Material.

We present simulation results for a data generation process inspired by the Head Start Impact Study (HSIS) study analyzed in the next section. For a given sample size, we first generate three covariates (one standard normal and two binary). The control potential outcomes are then generated from

\[
Y_i(0) = 0.3 + 0.5X_{1i} + 0.3X_{2i} - 0.6X_{3i} + u_i, \quad u_i \sim N(0, 0.8^2).
\]

The marginal variance of the control potential outcomes is approximately 1, with \((0.5^2)(1) + (0.3^2)(0.25) + (0.6^2)(0.25) + 0.8^2 \approx 1\); thus we can interpret impacts in “effect size” units. The treatment effects are \( \tau_i = \delta_i + \varepsilon_i \), with (1) either \( \delta_i = 0.3 \) for all \( i \), or \( \delta_i = 0.2 + 0.2X_{3i} \); and (2)
either $\varepsilon_i = 0$ for all $i$, or $\varepsilon_i \sim N(0, 1)$. All combinations of these two options give the four cases of (a) no treatment effect variation, (b) only systematic variation, (c) idiosyncratic variation with no systematic variation, and (d) both systematic and idiosyncratic variation. For an $\alpha$-level test of systematic variation, scenarios (a) and (c) should only reject at rate $\alpha$, while we would like to see high rejection rates for scenarios (b) and (d). To generate a synthetic dataset we generated all potential outcomes, randomized units into treatment with $p = 0.6$, and then calculated the corresponding observed outcomes. We finally calculated our four estimators using these data, and conducted a test for systematic variation using each one.

Figure 1 shows the power of the tests, with $\alpha = 0.05$, for different sample sizes. First, all estimators appear asymptotically valid, consistent with the theoretical results. Second, the OLS estimator appears to have the greatest power in this setting, which is unsurprising since the true data generating process is a linear model. Finally, covariate adjustment slightly improves the power of the RI estimator. Overall, sample sizes are considerable in order to achieve decent levels of power in this scenario, although there seems to be reasonable power for the sample size in the data application, $N = 3,586$.

6.2 LATE estimators

We next simulate completely randomized experiments with noncompliance to evaluate the finite sample performance of the tests for systematic treatment effect variation among Compliers based on $\hat{\beta}_{c,RI}$ and $\hat{\beta}_{TSLS}$. We generated data as in the ITT case above, and then generated strata membership for all units with probabilities proportional to the covariates. For both Always Takers and Never Takers, we then set $Y_i(0) = Y_i(1)$. The overall ITT is now reduced to 0.17 (due to the 0 effects of Never Takers and Always Takers). The resulting proportion of Compliers is approximately 68%.

We then manipulated (for Compliers) the systematic and idiosyncratic effects as above and tested for presence of systematic variation of Compliers under exclusion restrictions. Figure 2 shows the power of these tests for our RI and TSLS estimators. First, in this scenario, the 2SLS and the RI estimators are virtually equivalent; the additional adjustment provided by TSLS does not add significantly to the precision. We see the tests are valid (they even appear conservative) for cases (a) and (c). Power is reduced compared to ITT; this is reasonable as power is effectively a function of the number of Compliers, with additional uncertainty due to only partial observation of who the Compliers are.
7 Application to the Head Start Impact Study

The HSIS (Puma et al., 2010) is a large-scale randomized evaluation of Head Start, a Federal preschool program serving around 900,000 children each year at a cost of roughly $8 billion. Overall, Puma et al. (2010) find modest average effects of the program on a range of children’s cognitive and social-emotional outcomes. However, both the original study and several recent papers argue that these topline results mask important treatment effect variation (e.g., Bloom and Weiland, 2014; Bitler et al., 2014; Ding et al., 2016; Feller et al., 2016). We now use the methods outlined above to assess treatment effect variation in HSIS. Following earlier analyses (Ding et al., 2016) and to simplify exposition, we restrict our attention to a complete-case subset of HSIS, with \( n_1 = 2,238 \) in the treatment group and \( n_0 = 1,348 \) in the control group. For illustration, our outcome of interest is the Peabody Picture Vocabulary Test (PPVT), a widely used measure of cognitive ability in early education.
childhood. To assess treatment effect variation, we consider the rich set of child- and family-level covariates used in the original HSIS analysis of Puma et al. (2010), including pre-test score, child’s age, child’s race, mother’s education level, and mother’s marital status. After creating dummy variables for factors (i.e., re-coding race), the covariate matrix has 17 columns. See Figure 3b.

7.1 Decomposing variation in the ITT effect

We first turn to treatment effect variation for the ITT estimate, beginning with estimating the systematic treatment effect variation. We examine three estimators, the randomization-based and OLS estimators discussed in Section 3, $\hat{\beta}_{RI}$ and $\hat{\beta}_{OLS}$, and the corresponding model-assisted version of the RI estimator discussed in the Supplementary Material, $\hat{\beta}_{RI}^w$. For this latter estimator, we use all available covariates to adjust the standard estimators, that is, $W$ is the entire vector of covariates.
Omnibus test for systematic treatment effect variation. We begin by using these estimators for an omnibus test of whether any treatment effect variation is explained by the full set of covariates. The p-values for the unadjusted $\hat{\beta}_{RI}$ estimator and model-assisted $\hat{\beta}_{wRI}$ are $p = 0.39$ and $p = 0.25$, respectively, which do not show any evidence of treatment effect variation. The OLS and model-assisted OLS estimators, however, show much stronger evidence. The p-value for $\hat{\beta}_{OLS}$ is $p = 0.005$. This pattern is unsurprising given the simulation results showing greater statistical power for the OLS estimators over the RI estimators.

Treatment effect $R^2$. Next, we examine how much of the variation could be explained by our covariates. We turn to the overall treatment effect $R^2_T$ for ITT in HSIS. Figure 3a shows values of the treatment effect $R^2_T$ using $\hat{\beta}_{wRI}$ to estimate the systematic variation. Results are nearly identical using the other estimators. In the worst case of perfect negative dependence between potential outcomes (not shown), the treatment effect $R^2_T$ could be as low as 0.01. Assuming that this dependence is nonnegative, the treatment effect $R^2_T$ ranges from 0.03 to 0.76. While the estimate is clearly sensitive to the unidentifiable sensitivity parameter, the covariates explain a substantial proportion of treatment effect variation for values of $\rho$ near 1.

We can also use this framework to assess the relative importance of each covariate in terms of
explaining overall treatment effect variation. To do this, we use the model-assisted RI estimator, $\hat{\beta}^w_{RI}$, adjusting for all covariates (i.e., $\dim(W) = 17$) but restricting systematic treatment effect variation to one covariate at a time. Note that we consider factors (e.g., race) as a group. Figure 3b shows the resulting estimates for the upper bound of $R^2_\tau$, with lower bound estimates all below 0.01. Having a mother who is a recent immigrant and dual language learner status (which are highly correlated in practice) could each explain a substantial proportion of treatment effect variation, consistent with previous results from Bloom and Weiland (2014) and Bitler et al. (2014). This is not true for other covariates, like mother’s education level.

Negative correlation between treatment effect and control potential outcomes. Finally, we test whether the individual-level idiosyncratic treatment effects, $\{\varepsilon_i\}_{i=1}^n$, are negatively correlated with the control potential outcomes, $\{Y_i(0)\}_{i=1}^n$, extending results from Raudenbush and Bloom (2015). As outlined in the Supplementary Material, we do so by testing whether the variance of $\{Y^\text{obs}_i - X_i'\hat{\beta}^w_{RI} : T_i = 1\}$ is smaller than the variance of $\{Y^\text{obs}_i : T_i = 0\}$. This yields a $p$-value of $p = 0.02$, which suggests that the unexplained treatment effect is indeed larger for smaller values of the control potential outcomes. This result is consistent with findings from Bitler et al. (2014) who use a quantile treatment effect approach.

### 7.2 Incorporating noncompliance

As with many social experiments, there is substantial noncompliance with random assignment in HSIS. In the analysis sample we consider here, the estimated proportion of compliance types is $\hat{\pi}_c = 0.69$ for Compliers, $\hat{\pi}_a = 0.13$ for Always Takers, and $\hat{\pi}_n = 0.18$ for Never Takers. Given the exclusion restrictions for Always Takers and Never Takers, the treatment effect is therefore zero (by assumption) for over 30 percent of the sample, suggesting that noncompliance will be an important component of treatment effect variation.

In the setting with noncompliance, we focus on two estimators for systematic treatment effect variation among Compliers: the randomization-based estimator, $\hat{\beta}_{c, RI}$, and the Two-Stage Least Squares estimator, $\hat{\beta}_{TSLS}$. We first use these estimators to construct omnibus tests for systematic treatment effect variation among Compliers. Tests using both estimators show strong evidence for such variation, with $p = 0.02$ using $\hat{\beta}_{c, RI}$ and $p = 0.01$ using $\hat{\beta}_{TSLS}$.

Finally, we turn to decomposing the overall treatment effect. As in the ITT case, we assume that
the potential outcomes have a nonnegative correlation. Figure 3a shows the treatment effect $R^2$ among Compliers, which ranges from $R^2_{\tau,c} = 0.05$ to $R^2_{\tau,c} = 0.68$. Next, we can calculate treatment effect variation due to noncompliance, $R^2_{\tau,U}$. In the case of HSIS, this is relatively small—between 0.01 and 0.16—in part because the overall treatment effect is fairly small. Therefore, the overall treatment effect decomposition due to both covariates and noncompliance, $R^2_{\tau,U|X}$, is quite close to $R^2_{\tau,c}$, as shown in Figure 3a. Taken together, these estimates suggest that there is indeed important treatment effect variation that is neither captured by pre-treatment covariates nor by noncompliance, consistent with previous results in Ding et al. (2016).

8 Conclusion

In this paper, we propose a broad, flexible framework for assessing and decomposing treatment effect variation in randomized experiments with and without noncompliance. In general, we believe this is a natural setup for researchers to formulate and investigate a broad range of questions about impact heterogeneity (e.g., Heckman et al. 1997). Applications include assessing underlying causal mechanisms and targeting treatments based on individual-level characteristics. Understanding such variation is also important for the design of experiments. Djebbari and Smith (2008), for example, argue that characterizing the size of the idiosyncratic treatment effect is useful for determining the value of additional data collection.

We briefly note several directions for future work. First, we emphasized the role of randomization throughout in justifying the proposed framework. As a result, we focused on relatively simple estimators, essentially versions of linear regression and TSLS. While these estimators are simple, we do not fully explore their practical and finite-sample properties here. For example, determining when the model-assistance helps, and assessing the increased power of the OLS approach to the unbiased RI approach, are areas of future work. Similarly, there is still much potential improvement in determining ways of characterizing the degree of heterogeneity, such as with an effect size for the systematic variation.

Second, a natural extension is to use more complex methods to estimate systematic treatment effects, such as via hierarchical models (Feller and Gelman 2015) or via machine learning methods (Wager and Athey 2015), extending the results for the omnibus test and treatment effect $R^2_\tau$ accordingly. While the guarantees from randomization are clearly weaker in such settings,
researchers can assess these tradeoffs themselves. For example, hierarchical modeling would be especially useful in the Head Start Impact Study due to the multi-site design (Bloom and Weiland, 2014).

Third, a question of increasing practical importance is the generalizability of experimental results to a given target population (Stuart et al., 2011). We believe that the treatment effect \( R^2_\tau \) is a critical measure for assessing the credibility of these generalizations. In short, if there is substantial idiosyncratic treatment effect variation, i.e., \( R^2_\tau \) is small, then researchers should be wary of using observed covariates to extrapolate treatment effects.

Finally, a question is how to extend this treatment effect variation framework to non-randomized settings. While the results would necessarily rest on much stronger assumptions, many settings already use an as-if-randomized framework, such as in observational studies (Rosenbaum, 2002; Imbens and Rubin, 2015). Under this approach, extensions should be natural.
References


Supplementary Material

Appendix A gives all the proofs, and Appendix B provides the additional commentary mentioned in the main text.

Appendix A  Lemmas and Proofs

Derivations of the properties of estimators under the randomization framework tend to involve a lot of bookkeeping of sums of indicators variables and potential outcomes. At best, this bookkeeping is tedious, and at worst it can be prohibitively complex, making it virtually impossible to derive the properties of these estimators for moderately complex assignment mechanisms or data structures. We greatly simplify this math by representing relevant operations as matrix projections, which are of independent interest for other applications of randomization inference. The core idea is that, in order to characterize estimators based on \( T \), we first need to characterize the behavior of \( T \)'s mean and covariance in as simple a form as possible.

To begin, let \( 1_n = (1, \ldots, 1)^T \) and \( 0_n = (0, \ldots, 0)^T \) be column vectors of length \( n \), and \( I_n \) be the \( n \times n \) identity matrix. Then \( S_n = I_n - n^{-1}1_n1_n^T \) is the projection matrix orthogonal to \( 1_n \) with \( S_n1_n = 0_n \). Under this formulation, the covariance matrix of the treatment assignment vector is a scaled projection matrix orthogonal to \( 1_n \), as shown in the following lemma.

**Lemma A.1.** The treatment assignment vector \( T \) of a completely randomized experiment has

\[
E(T) = \frac{n_1}{n} 1_n, \quad \text{cov}(T) = \frac{n_1n_0}{n(n-1)} S_n.
\]

**Proof of Lemma A.1.** The conclusions follow from

\[
E(T_i) = \frac{n_1}{n}, \quad \text{var}(T_i) = \frac{n_1n_0}{n^2}, \quad \text{cov}(T_i, T_j) = -\frac{n_1n_0}{n^2(n-1)}, \quad (i \neq j).
\]

The projection matrix \( S_n \) enjoys nice mathematical properties, and it acts as a covariance operator as illustrated by the following lemma.

**Lemma A.2.** Let \( U_i, V_i \in \mathbb{R}^K \) be column vectors of length \( K \). Define \( U = [U_1, U_2, \ldots, U_n] \) and \( V = [V_1, V_2, \ldots, V_n] \in \mathbb{R}^{K \times n} \) as two matrices of dimension \( K \times n \). If \( \bar{U} = n^{-1} \sum_{i=1}^n U_i \) and \( \bar{V} = n^{-1} \sum_{i=1}^n V_i \) as two matrices of dimension \( K \times 1 \). If \( \bar{U} \) and \( \bar{V} \) are centered, then

\[
\text{cov}(U_i, V_j) = \frac{n_1n_0}{n^2} S_n, \quad (i \neq j).
\]
\[ \bar{V} = n^{-1} \sum_{i=1}^{n} V_i, \text{ then} \]
\[ U S_n V^T = \sum_{i=1}^{n} (U_i - \bar{U})(V_i - \bar{V})^T. \]

In particular, when \( U_i = V_i, \)
\[ V S_n V^T = \sum_{i=1}^{n} (V_i - \bar{V})(V_i - \bar{V})^T = (N - 1)S(V). \]

**Proof of Lemma A.2.** The left hand side is equal to
\[ U S_n V^T = UV^T - n^{-1} \left( \sum_{i=1}^{n} U_i V_i^T \right) = \sum_{i=1}^{n} U_i V_i^T - n^{-1} \left( \sum_{i=1}^{n} U_i V_i^T \right) = \sum_{i=1}^{n} U_i V_i^T - n\bar{U}\bar{V}^T, \]
which is the same as the right hand side. \( \square \)

**Proof of Theorem 1.** The Neymanian estimator has the following representation:
\[ \hat{\tau}_V = \frac{V_{1 \text{obs}}}{n_1} - \frac{V_{0 \text{obs}}}{n_0}, \]
\[ = \frac{1}{n_1} \sum_{i=1}^{n} T_i V_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) V_i(0) \]
\[ = \sum_{i=1}^{n} T_i \left\{ \frac{V_i(1)}{n_1} + \frac{V_i(0)}{n_0} \right\} - \frac{1}{n_0} \sum_{i=1}^{n} V_i(0). \]

The unbiasedness of \( \hat{\tau}_V \) follows from the linearity of the expectation and Lemma A.1. Define \( V_1 = [V_1(1), \ldots, V_n(1)] \) and \( V_0 = [V_1(0), \ldots, V_n(0)] \) as the matrices of the potential outcomes. The estimator \( \hat{\tau}_V \) can be represented as
\[ \hat{\tau}_V = \left( \frac{V_1}{n_1} + \frac{V_0}{n_0} \right) T - \frac{1}{n_0} \sum_{i=1}^{n} V_i(0). \]

Note the second term is constant, and so is not involved in the covariance. Applying Lemmas A.1 and A.2 we can obtain the covariance matrix of \( \hat{\tau}_V \):
\[ \text{cov}(\hat{\tau}_V) = \left( \frac{V_1}{n_1} + \frac{V_0}{n_0} \right) \text{cov}(T) \left( \frac{V_1}{n_1} + \frac{V_0}{n_0} \right)^\top \]
\[ = \frac{n_1 n_0}{n(n-1)} \left( \frac{V_1}{n_1} + \frac{V_0}{n_0} \right) S_n \left( \frac{V_1}{n_1} + \frac{V_0}{n_0} \right)^\top \]
\[ = \frac{n_1 n_0}{n(n-1)} \left( \frac{1}{n_1} V_1 S_n V_1^\top + \frac{1}{n_0} V_0 S_n V_0^\top + \frac{1}{n_1 n_0} V_0 S_n V_1^\top + \frac{1}{n_1 n_0} V_1 S_n V_0^\top \right) \]
\[ = \frac{n_0}{n n_1} S\{V(1)\} + \frac{n_1}{n n_0} S\{V(0)\} + \frac{1}{n(n-1)} (V_0 S_n V_1^\top + V_1 S_n V_0^\top). \]

2
Using the fact $ab^T + ba^T = aa^T + bb^T - (a - b)(a - b)^T$ for two column vectors $a$ and $b$, we have

$$
\{V_i(1) - \bar{V}(1)\} \{V_i(0) - \bar{V}(0)\}^T + \{V_i(0) - \bar{V}(0)\} \{V_i(1) - \bar{V}(1)\}^T
- \{V_i(1) - V_i(0) - \bar{V}(1) + \bar{V}(0)\} \{V_i(1) - V_i(0) - \bar{V}(1) + \bar{V}(0)\}^T.
$$

Summing over $i = 1, \ldots, n$ and applying Lemma A.2 we have

$$
\frac{V_0 S_n v_i^T}{n - 1} + \frac{V_1 S_n v_0^T}{n - 1} = S\{V(1)\} + S\{V(0)\} - S\{V(1) - V(0)\}.
$$

Therefore, the covariance of $\hat{\nu}$ can be simplified as:

$$
cov(\hat{\nu}) = \frac{n_0}{nn_1} S\{V(1)\} + \frac{n_1}{nn_0} S\{V(0)\} + \frac{1}{n} [S\{V(1)\} + S\{V(0)\} - S\{V(1) - V(0)\}]
- \frac{S\{V(1)\}}{n_1} + \frac{S\{V(0)\}}{n_0} - \frac{S\{V(1) - V(0)\}}{n}.
$$

The above proof is more concise than other proofs of Neyman’s theorem, such as in Imbens and Rubin (2015).

**Proof of Theorem 2.** Because $\hat{S}_{xt}$ is the sample mean for $\{X_i Y_i^{\text{obs}} : T_i = t, i = 1, \ldots, n\} = \{X_i Y_i(t) : T_i = t, i = 1, \ldots, n\}$, it is unbiased for the population mean $S_{xt}$. Thus, the estimator $\hat{\beta}_{RI}$ is also unbiased for $\beta$ as $S_{xx}^{-1}$ is fixed and the expectation is linear. Its sampling covariance over all possible randomizations is

$$
cov(\hat{\beta}_{RI}) = S_{xx}^{-1} \text{cov}(\hat{S}_{x1} - \hat{S}_{x0}) S_{xx}^{-1}.
$$

Therefore, we need only to obtain the covariance of

$$
\hat{S}_{x1} - \hat{S}_{x0} = \frac{1}{n_1} \sum_{i=1}^{n} T_i X_i Y_i^{\text{obs}} - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) X_i Y_i^{\text{obs}}.
$$

which is the difference between the sample means of $\{X_i Y_i(1) : i = 1, \ldots, n\}$ and $\{X_i Y_i(0) : i = 1, \ldots, N\}$ under treatment and control. Viewing $X_i Y_i^{\text{obs}}$ as a vector outcome in a completely randomized experiment, we can apply Theorem 4 to obtain

$$
cov(\hat{S}_{x1} - \hat{S}_{x0}) = \frac{S\{XY(1)\}}{n_1} + \frac{S\{XY(0)\}}{n_0} - \frac{S(X \tau)}{n},
$$

which completes the proof.
Proof of Theorem 3. Define sample proportions as $p_1 = n_1/n$ and $p_0 = n_0/n$, and sample covariances

$$
\hat{S}_{xx,1} = \frac{1}{n_1} \sum_{i=1}^{n} T_i X_i X_i^T, \quad \hat{S}_{xx,0} = \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) X_i X_i^T, \quad \hat{S}_{xy} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_{\text{obs}, i}.
$$

Therefore, $S_{xx} = p_1 \hat{S}_{xx,1} + p_0 \hat{S}_{xx,0}$ and $\hat{S}_{xy} = p_1 \hat{S}_{x1} + p_0 \hat{S}_{x0}$. The OLS estimators of the regression coefficients are

$$
\begin{align*}
\begin{pmatrix}
\hat{\gamma}_{\text{OLS}} \\
\hat{\beta}_{\text{OLS}}
\end{pmatrix}
&= \left\{ \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ T_i X_i \end{pmatrix} \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ T_i X_i \end{pmatrix} Y_{\text{obs}, i} \right\} \\
&= \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right) Y_{\text{obs}, i} \\
&= \left( p_1 \hat{S}_{xx,1} + p_0 \hat{S}_{xx,0} \right)^{-1} \left( p_1 \hat{S}_{x1} + p_0 \hat{S}_{x0} \right).
\end{align*}
$$

We will use the following formula for the inverse of a block matrix

$$
\begin{pmatrix} A & B \\ B & B \end{pmatrix}^{-1} = \begin{pmatrix} (A - B)^{-1} & -A^{-1} B (B - BA^{-1} B)^{-1} \\ -(A - B)^{-1} & (B - BA^{-1} B)^{-1} \end{pmatrix}.
$$

Take $A = p_1 \hat{S}_{xx,1} + p_0 \hat{S}_{xx,0}$ and $B = p_1 \hat{S}_{xx,1}$, and we can simplify each component above as $(A - B)^{-1} = p_0 \hat{S}_{xx,0}$, $(B - BA^{-1} B)^{-1} = (p_0 \hat{S}_{xx,0})^{-1} + (p_1 \hat{S}_{xx,1})^{-1}$, and $-A^{-1} B (B - BA^{-1} B)^{-1} = -(p_0 \hat{S}_{xx,0})^{-1}$. Therefore,

$$
\begin{align*}
\begin{pmatrix}
\hat{\gamma}_{\text{OLS}} \\
\hat{\beta}_{\text{OLS}}
\end{pmatrix}
&= \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right) Y_{\text{obs}, i} \\
&= \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right) Y_{\text{obs}, i} \\
&= \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right) Y_{\text{obs}, i}.
\end{align*}
$$

The above formula shows that $\hat{\gamma}_{\text{OLS}}$ can be obtained by running regression of $Y_{\text{obs}}$ onto $X$ using the control group data, and $\hat{\beta}_{\text{OLS}}$ can be obtained by running regression of $Y_{\text{obs}}$ onto $X$ using the treatment group data.

To prove the asymptotic covariance of $\hat{\beta}_{\text{OLS}}$, first we write

$$
\hat{\beta}_{\text{OLS}} - \beta = \hat{S}_{xx,1}^{-1} (\hat{S}_{x1} - \hat{S}_{xx,1} \gamma_1) - \hat{S}_{xx,0}^{-1} (\hat{S}_{x0} - \hat{S}_{xx,0} \gamma_0),
$$

second we introduce

$$
\begin{align*}
\bar{\beta}_{\text{OLS}} - \beta = S_{xx,1}^{-1} (\bar{S}_{x1} - \hat{S}_{xx,1} \gamma_1) - S_{xx,0}^{-1} (\bar{S}_{x0} - \hat{S}_{xx,0} \gamma_0),
\end{align*}
$$

(A.1)
and third we observe that the difference between \( \hat{\beta}_{\text{OLS}} \) and \( \tilde{\beta}_{\text{OLS}} \) is of higher order, because

\[
\tilde{\beta}_{\text{OLS}} - \hat{\beta}_{\text{OLS}} = (\tilde{S}_{xx,1}^{-1} - S_{xx,1}^{-1})(\tilde{S}_{x1} - \tilde{S}_{x,1} \gamma_1) - (\tilde{S}_{xx,0}^{-1} - S_{xx,0}^{-1})(\tilde{S}_{x0} - \tilde{S}_{x,0} \gamma_0) = O_P(n^{-1/2})O_P(n^{-1/2}) - O_P(n^{-1/2})O_P(n^{-1/2}) = O_P(n^{-1}). \tag{A.2}
\]

Therefore, \( \tilde{\beta}_{\text{OLS}} \) and \( \hat{\beta}_{\text{OLS}} \) have the same asymptotic covariance matrix, and we need only to find the covariance matrix of \( \tilde{\beta}_{\text{OLS}} \). In fact, we can further simplify (A.1) as

\[
\tilde{\beta}_{\text{OLS}} - \beta = S_{xx}^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} T_i X_i e_i(1) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) X_i e_i(0) \right], \tag{A.3}
\]

where \( e_i(1) = Y_i(1) - X_i \gamma_1 \) and \( e_i(0) = Y_i(0) - X_i \gamma_0 \) are the residual potential outcomes, satisfying \( e_i(1) - e_i(0) = \varepsilon_i \) and thus \( \sum_{i=1}^{n} X_i \{e_i(1) - e_i(0)\} / n = S_{xx} e = 0 \). Applying Theorem 1 to the vector outcome \( X e \), we can obtain the covariance matrix of \( \tilde{\beta}_{\text{OLS}} \). The asymptotic normality follows from the representation [A.3] and the finite population central limit theorem (Hájek, 1960). \( \square \)

To prove Theorem 4 we need to invoke the following Fréchet–Hoeffding inequality (Hoeffding, 1941; Fréchet, 1951; Heckman et al., 1997; Aronow et al., 2014).

**Lemma A.3.** If we know only the marginal distributions of two random variables \( X \sim F_X(x) \) and \( Y \sim F_Y(y) \), then \( E(XY) \) can be sharply bounded by

\[
\int_0^1 F_X^{-1}(u) F_Y^{-1}(1-u) du \leq E(XY) \leq \int_0^1 F_X^{-1}(u) F_Y^{-1}(u) du.
\]

**Lemma A.4.** If we know only the marginal distributions \( X \sim F_X(x) \), \( Y \sim F_Y(y) \), and \( E(X - Y) = 0 \), then \( \text{var}(X - Y) \) can be sharply bounded by

\[
\int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(1-u)\}^2 du \leq \text{var}(X - Y) \leq \int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(1-u)\}^2 du
\]

**Proof of Lemma A.4.** The variance \( \text{var}(X - Y) \) can be decomposed as \( \text{var}(X - Y) = E(X - Y)^2 = E(X^2) + E(Y^2) - 2E(XY) \), depending on the following three terms:

\[
E(X^2) = \int x^2 dF_X(x) = \int_0^1 \{F_X^{-1}(u)\}^2 du,
\]

\[
E(Y^2) = \int_0^1 \{F_Y^{-1}(u)\}^2 du = \int_0^1 \{F_Y^{-1}(1-u)\}^2 du,
\]

\[
\int_0^1 F_X^{-1}(u) F_Y^{-1}(1-u) du \leq E(XY) \leq \int_0^1 F_X^{-1}(u) F_Y^{-1}(u) du.
\]

Plugging the above expressions into the variance of \( X - Y \), we can obtain the desired bounds. \( \square \)
Applying Lemma A.4, we can easily prove Theorem 4.

Proof of Theorem 4. Because \( S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon} \), we need only to bound \( S_{\varepsilon\varepsilon} \), which is the finite population variance of \( \varepsilon_i = \{ Y_i(1) - X_i^T \gamma_1 \} - \{ Y_i(0) - X_i^T \gamma_0 \} = e_i(1) - e_i(0) \). We can identify the marginal distributions of \( \{ e_i(1) : i = 1, \ldots, n \} \) and \( \{ e_i(0) : i = 1, \ldots, n \} \), and know \( n^{-1} \sum_{i=1}^{n} \varepsilon_i = 0 \). Therefore, the bounds in Lemma A.4 imply the bounds in Theorem 4.

Proof of Theorem 5. The joint distribution of \((U_1, U_0)\) is

\[
C(u_1, u_0) = P(U_1 \leq u_1, U_0 \leq u_0) = \rho P(U_0 \leq u_1, U_0 \leq u_0) + (1 - \rho) P(V_0 \leq u_1, U_0 \leq u_0) \]

\[
= \rho \min(u_1, u_0) + (1 - \rho) u_1 u_0.
\]

Therefore, the distribution function \( C(u_1, u_0) \) is a weighted average of \( \min(u_1, u_0) = C_R(u_1, u_0) \) and \( u_1 u_0 = C_I(u_1, u_0) \), i.e., the joint distributions when \( U_1 = U_0 \) and \( U_1 \perp \! \! \! \perp U_0 \), respectively.

According to Nelsen (2007, Theorem 5.1.6), Spearman’s rank correlation coefficient between \( e(1) \) and \( e(0) \) is

\[
12 \int_0^1 \int_0^1 \{ C(u_1, u_0) - u_1 u_0 \} du_1 du_0 = 12 \rho \int_0^1 \int_0^1 \{ \min(u_1, u_0) - u_1 u_0 \} du_1 du_0 \]

\[
= 12 \rho \left( 2 \int_0^1 du_1 \int_0^{u_1} u_0 du_0 - \frac{1}{4} \right) \]

\[
= 12 \rho (1/3 - 1/4) = \rho.
\]

To complete the proof of the theorem, we need only to show that the covariance between \( e(1) \) and \( e(0) \) is linear in \( \rho \), which follows from

\[
\int_0^1 \int_0^1 F_{1}^{-1}(u_1) F_{0}^{-1}(u_0) dC(u_1, u_0) \]

\[
= \rho \int_0^1 \int_0^1 F_{1}^{-1}(u_1) F_{0}^{-1}(u_0) dC_R(u_1, u_0) + \rho \int_0^1 \int_0^1 F_{1}^{-1}(u_1) F_{0}^{-1}(u_0) dC_I(u_1, u_0) \]

\[
= \rho \int_0^1 F_{1}^{-1}(u) F_{0}^{-1}(u) du + (1 - \rho) \int_0^1 F_{1}^{-1}(u) du \int_0^1 F_{0}^{-1}(u) du.
\]

To prove Theorem 6, we need to introduce more notation. Because of the exclusion restrictions for Never Takers and Always Takers, we define the population covariance between \( X \) and \( Y(1) = \)
$Y(0)$ within stratum $U = a$ and $U = n$ as

$$S_{x,u} = \frac{1}{n_u} \sum_{i=1}^{n} I_{(U_i=u)} X_i Y(1) = \frac{1}{n_u} \sum_{i=1}^{n} I_{(U_i=u)} X_i Y(0) \quad (u = a, n).$$

**Proof of Theorem 6**  From the observed data with $(T_i, D_i) = (1, 1)$,

$$E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i D_i X_i X_i^\top \right\} = E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i I_{(U_i=a)} X_i X_i^\top + \frac{1}{n_1} \sum_{i=1}^{n} T_i I_{(U_i=c)} X_i X_i^\top \right\} = \pi_a S_{xx,a} + \pi_c S_{xx,c}. \tag{A.4}$$

Similar to (A.4),

$$E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i (1 - D_i) X_i X_i^\top \right\} = \pi_n S_{xx,n}, \tag{A.5}$$

$$E \left\{ \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) D_i X_i X_i^\top \right\} = \pi_a S_{xx,a}, \tag{A.6}$$

$$E \left\{ \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i)(1 - D_i) X_i X_i^\top \right\} = \pi_n S_{xx,n} + \pi_c S_{xx,c}. \tag{A.7}$$

Subtracting (A.6) from (A.4), and subtracting (A.5) from (A.7), we obtain unbiased estimators for $\pi_c S_{xx,c}$.

From the observed data with $(T_i, D_i) = (1, 1)$,

$$E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i D_i X_i Y_{i,\text{obs}} \right\} = E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i I_{(U_i=a)} X_i Y_i(1) + \frac{1}{n_1} \sum_{i=1}^{n} T_i I_{(U_i=c)} X_i Y_i(1) \right\} = \pi_a S_{x,a} + \pi_c S_{x,c}. \tag{A.8}$$

Similar to (A.8),

$$E \left\{ \frac{1}{n_1} \sum_{i=1}^{n} T_i (1 - D_i) X_i Y_{i,\text{obs}} \right\} = \pi_n S_{x,n}, \tag{A.9}$$

$$E \left\{ \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) D_i X_i Y_{i,\text{obs}} \right\} = \pi_a S_{x,a}, \tag{A.10}$$

$$E \left\{ \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i)(1 - D_i) X_i Y_{i,\text{obs}} \right\} = \pi_n S_{x,n} + \pi_c S_{x,c}. \tag{A.11}$$

Subtracting (A.10) from (A.8), and subtracting (A.9) from (A.11), we obtain the results in (12).
Proof of Corollary 3. First we write

$$\hat{\beta}_{c,RI} - \beta_c = (\hat{S}_{xx,11} - \hat{S}_{xx,01})^{-1}\{\hat{S}_{x1,11} - \hat{S}_{x0,01} - (\hat{S}_{xx,11} - \hat{S}_{xx,01})\gamma_{1c}\}$$
$$- (\hat{S}_{xx,00} - \hat{S}_{xx,10})^{-1}\{\hat{S}_{x0,00} - \hat{S}_{x1,10} - (\hat{S}_{xx,00} - \hat{S}_{xx,10})\gamma_{0c}\},$$

second we introduce

$$\tilde{\beta}_{c,RI} - \beta_c = (\pi_c S_{xx,c})^{-1}\{\hat{S}_{x1,11} - \hat{S}_{x0,01} - (\hat{S}_{xx,11} - \hat{S}_{xx,01})\gamma_{1c}\}$$
$$- (\pi_c S_{xx,c})^{-1}\{\hat{S}_{x0,00} - \hat{S}_{x1,10} - (\hat{S}_{xx,00} - \hat{S}_{xx,10})\gamma_{0c}\},$$

third we observed that the difference between $\tilde{\beta}_{c,RI}$ and $\beta_{c,RI}$ has higher order following the same argument as \([A.2]\). Therefore, we need only to find the covariance of $\tilde{\beta}_{c,RI}$. Simple algebra gives

$$\tilde{\beta}_{c,RI} - \beta_c$$
$$= (\pi_c S_{xx,c})^{-1}\left[\frac{1}{n_1}\sum_{i=1}^{n} T_i D_i X_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) D_i X_i Y_i(0)\right]$$
$$- \frac{1}{n_1}\sum_{i=1}^{n} T_i D_i X_i X_i^T \gamma_{1c} + \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) D_i X_i X_i^T \gamma_{1c}$$
$$- \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i)(1-D_i) X_i Y_i(0) + \frac{1}{n_1}\sum_{i=1}^{n} T_i (1-D_i) X_i Y_i(1)$$
$$+ \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i)(1-D_i) X_i X_i^T \gamma_{0c} - \frac{1}{n_1}\sum_{i=1}^{n} T_i (1-D_i) X_i X_i^T \gamma_{0c}\right]$$

$$= (\pi_c S_{xx,c})^{-1}\left[\frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)} X_i Y_i(1) + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)} X_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=a)} X_i Y_i(0)\right]$$
$$- \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=a)} X_i X_i^T \gamma_{1c} - \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=c)} X_i X_i^T \gamma_{1c} + \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=a)} X_i X_i^T \gamma_{1c}$$
$$- \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=n)} X_i Y_i(0) - \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=c)} X_i Y_i(0) + \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)} X_i Y_i(1)$$
$$+ \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=n)} X_i X_i^T \gamma_{0c} + \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) I_{(U_i=c)} X_i X_i^T \gamma_{0c} - \frac{1}{n_1}\sum_{i=1}^{n} T_i I_{(U_i=n)} X_i X_i^T \gamma_{0c}\right]$$

$$= (\pi_c S_{xx,c})^{-1}\left\{\frac{1}{n_1}\sum_{i=1}^{n} T_i X_i [I_{(U_i=a)}(Y_i(1) - X_i^T \gamma_{1c}) + I_{(U_i=n)}(Y_i(1) - X_i^T \gamma_{0c}) + I_{(U_i=c)}(Y_i(1) - X_i^T \gamma_{1c})]$$
$$- \frac{1}{n_0}\sum_{i=1}^{n} (1-T_i) X_i [I_{(U_i=a)}(Y_i(0) - X_i^T \gamma_{1c}) + I_{(U_i=n)}(Y_i(0) - X_i^T \gamma_{0c}) + I_{(U_i=c)}(Y_i(0) - X_i^T \gamma_{0c})] \right\}.$$
The above term \((\beta_{c,R1} = (\pi_c S_{xx,c})^{-1} \left( \frac{1}{n_1} \sum_{i=1}^{n_1} T_i X_i e_i'(1) - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) X_i e_i'(0) \right) \).

The representation in (A.12) implies the asymptotic covariance matrix according to Theorem 1 and the asymptotic normality of \(\beta_{c,R1}\) according to Hájek (1960).

**Proof of Theorem 7.** Recall the definitions \(p_1 = n_1/n\) and \(p_0 = n_0/n\). First, we find the probability limits of the TSLS estimators:

\[
\begin{align*}
(\hat{\gamma}_{\text{TSLS}}) & = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( X_i (X_i' D_i X_i)^{-1} X_i' D_i X_i' \right) \right\} ^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( (X_i D_i X_i)^{-1} Y_{i,}\text{obs} \right) \right\} \\
& = \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=a)} X_i X_i' \right\} ^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=c)} X_i X_i' \right\} ^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) I_{(U_i=a)} X_i X_i' \right\} \\
& \xrightarrow{p} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} \begin{pmatrix} G \\ H \end{pmatrix} .
\end{align*}
\]

The above term \(A\) is \(A = S_{xx}\), and terms \((B, C, D, G, H)\) are the population limits of the sample quantities. We will find each of them. Term \(B\) is

\[
B = E \left\{ \frac{1}{n} \sum_{i=1}^{n} D_i X_i X_i' \right\} = E \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i D_i X_i X_i' + \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) D_i X_i X_i' \right\} \\
= E \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=a)} X_i X_i' + \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=c)} X_i X_i' + \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) I_{(U_i=a)} X_i X_i' \right\} \\
= p_1 \pi_a S_{xx,a} + p_1 \pi_c S_{xx,c} + p_0 \pi_a S_{xx,a} = \pi_a S_{xx,a} + p_1 \pi_c S_{xx,c} .
\]

Term \(C\) is \(C = E \left\{ n^{-1} \sum_{i=1}^{n} T_i X_i X_i' \right\} = p_1 S_{xx}\). Term \(D\) is

\[
D = E \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i D_i X_i X_i' \right\} = E \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=a)} X_i X_i' + \frac{1}{n} \sum_{i=1}^{n} T_i I_{(U_i=c)} X_i X_i' \right\} \\
= p_1 \pi_a S_{xx,a} + p_1 \pi_c S_{xx,c} .
\]

Term \(G\) is

\[
G = E \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i Y_{i,}\text{obs} \right\} = E \left\{ \frac{1}{n} \sum_{i=1}^{n} T_i X_i Y_{i,}\text{obs} + \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) X_i Y_{i,}\text{obs} \right\} \\
= p_1 S_{x1} + p_0 S_{x0} .
\]

Term \(H\) is \(H = E \left\{ n^{-1} \sum_{i=1}^{n} T_i X_i Y_{i,}\text{obs} \right\} = p_1 S_{x1}\). We apply the following formula for the inverse of a block matrix:

\[
\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} S_D^{-1} - A^{-1} B S_A^{-1} \\ - D^{-1} C S_D^{-1} \end{pmatrix} ,
\]

9
where $S_D = A - BD^{-1}C$ and $S_A = D - CA^{-1}B$ are the Schur complements of blocks $D$ and $A$. Omitting some tedious matrix algebra, we obtain

$$S_D = p_0 \pi_c S_{xx,c}(\pi_a S_{xx,a} + \pi_c S_{xx,c})^{-1} S_{xx}, \quad S_A = p_1 p_0 \pi_c S_{xx,c},$$

and the inverse of the block matrix is

$$
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}^{-1} =
\begin{pmatrix}
(p_0^{-1} \pi_c^{-1} S_{xx}(\pi_a S_{xx,a} + \pi_c S_{xx,c}) S_{xx,c})^{-1} & -p_1^{-1} \pi_c^{-1} S_{xx}(\pi_a S_{xx,a} + p_1 \pi_c S_{xx,c}) S_{xx,c}^{-1} \\
-p_0^{-1} \pi_c^{-1} S_{xx,c}^{-1} & p_1^{-1} \pi_c^{-1} S_{xx,c}^{-1} 
\end{pmatrix}.
$$

Therefore, the probability limit of $\tilde{\gamma}_{TSLS}$ is

$$p_0^{-1} \pi_c^{-1} S_{xx,c}^{-1}(p_1 S_{x1} + p_0 S_{x0}) - p_1^{-1} \pi_c^{-1} S_{xx,c}^{-1}(p_1 S_{x1}) = \gamma_0 - p_1 D \beta_c \equiv \gamma_\infty,$$

and the probability limit of $\tilde{\beta}_{TSLS}$ is

$$-p_1^{-1} \pi_c^{-1} S_{xx,c}^{-1}(p_1 S_{x1} + p_0 S_{x0}) + p_1^{-1} p_0^{-1} \pi_c^{-1} S_{xx,c}^{-1}(p_1 S_{x1}) = \pi_c^{-1} S_{xx,c}^{-1}(S_{x1} - S_{x0}) = \beta_c,$$

where we use $S_{x1} - S_{x0} = \pi_c (S_{x1,c} - S_{x0,c})$ guaranteed by exclusion restrictions.

Below we will find the asymptotic distribution of $\tilde{\beta}_{TSLS}$. First we write

$$
\begin{pmatrix}
\tilde{\gamma}_{TSLS} \\
\tilde{\beta}_{TSLS}
\end{pmatrix} = 
\begin{pmatrix}
\gamma_\infty \\
\beta_c
\end{pmatrix} =
\begin{pmatrix}
\frac{1}{n} \sum_{i=1}^{n} \left( X_i \right) \left( X_i^T, D_i X_i^T \right)^{-1} \\
\frac{1}{n} \sum_{i=1}^{n} \left( X_i (Y_{i}^{obs} - X_i^T \gamma_\infty - D_i X_i^T \beta_c) \right)
\end{pmatrix}.
$$

Second we introduce

$$
\begin{pmatrix}
\tilde{\gamma}_{TSLS} \\
\tilde{\beta}_{TSLS}
\end{pmatrix} = 
\begin{pmatrix}
\gamma_\infty \\
\beta_c
\end{pmatrix} =
\begin{pmatrix}
\frac{1}{n} \sum_{i=1}^{n} \left( X_i (Y_{i}^{obs} - X_i^T \gamma_\infty - D_i X_i^T \beta_c) \right)
\end{pmatrix},
$$

third we recognize that the difference between the above two formulas has high order. Therefore, we need only to find the asymptotic distribution of $(\tilde{\gamma}_{TSLS}, \tilde{\beta}_{TSLS})$. According to the definitions of the residual potential outcomes,

$$
cov \left( \frac{n^{-1} \sum_{i=1}^{n} X_i (Y_{i}^{obs} - X_i^T \gamma_\infty - D_i X_i^T \beta_c)}{n^{-1} \sum_{i=1}^{n} T_i X_i (Y_{i}^{obs} - X_i^T \gamma_\infty - D_i X_i^T \beta_c)} \right)
$$

\begin{align*}
&= \frac{n^{-1} \sum_{i=1}^{n} T_i X_i e_i''(1) + n^{-1} \sum_{i=1}^{n} (1 - T_i) X_i e_i''(0)}{n^{-1} \sum_{i=1}^{n} T_i X_i e_i''(1)} \\
&= \frac{n^{-1} \sum_{i=1}^{n} T_i X_i \{e_i''(1) - e_i''(0)\}}{n^{-1} \sum_{i=1}^{n} T_i X_i e_i''(1)} \\
&= \frac{1}{n^2} \frac{n_1 n_0}{n} \left( \frac{S(\epsilon X)}{S(\epsilon X)} - \frac{1}{2} \left[ S\{X e''(1)\} - S\{X e''(0)\} + S(\epsilon X) \right] \right),
\end{align*}
where the off-diagonal term comes from the finite population covariance between $X\{e''(1) - e''(0)\}$ and $Xe''(1)$. Therefore, the asymptotic covariance of $\hat{\beta}_{TSL}$, or equivalently the covariance of $\hat{\beta}_{TSL}$, is

$$\frac{1}{n^2} \frac{n_1 n_0}{n} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} \begin{pmatrix} S(X\epsilon) \\ \frac{1}{2}[S\{Xe''(1)\} - S\{Xe''(0)\} + S(X\epsilon)] \end{pmatrix} \begin{pmatrix} \frac{1}{2}[S\{Xe''(1)\} - S\{Xe''(0)\} + S(X\epsilon)] \\ S(Xe''(1)) \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-\top},$$

which is

$$\frac{1}{n^2} \frac{n_1 n_0}{n} \left\{ (p_0^{-1} \pi^{-1}_c - 1)S(X\epsilon)(p_0^{-1} \pi^{-1}_c - 1)S(X\epsilon)^\top + (p_1^{-1} p_0^{-1} \pi^{-1}_c - 1)S\{Xe''(1)\}(p_1^{-1} p_0^{-1} \pi^{-1}_c - 1)S\{Xe''(1)\} \right\}$$

$$= (\pi_c S_{xx,c})^{-1} \left[ \frac{S\{Xe''(1)\}}{n_1} + \frac{S\{Xe''(0)\}}{n_0} - \frac{S(X\epsilon)}{n} \right] (\pi_c S_{xx,c})^{-1}.$$

Proof of Corollary 4. The following proof uses the facts: $\tau_a = \tau_n = 0$, and $\tau = \pi_c \tau_c$. We write the total treatment effect variation as

$$S_{\tau \tau} = \frac{1}{n} \sum_{i=1}^n (\tau_i - \tau)^2 = \frac{1}{n} \sum_{i=1}^n \tau_i^2 - \tau^2$$

$$= \frac{1}{n} \sum_{i=1}^n I(U_i = c) \tau_i^2 - \pi_c^2 \tau_c^2 = \pi_c \left( \frac{1}{n_c} \sum_{i=1}^n I(U_i = c) \tau_i^2 - \tau_c^2 \right) + \pi_c (1 - \pi_c) \tau_c^2,$$

the treatment effect variation explained by compliance status as

$$S_{\tau \tau, U} = \sum_{u=c,a,n} \pi_u (\tau_u - \tau)^2 = \pi_c (\tau_c - \pi_c \tau_c)^2 + \pi_a (0 - \pi_c \tau_c)^2 + \pi_n (0 - \pi_c \tau_c)^2$$

$$= \pi_c \tau_c^2 \left\{ (1 - \pi_c)^2 + \pi_c (\pi_a + \pi_n) \right\} = \pi_c (1 - \pi_c) \tau_c^2,$$

and the subtotal treatment effect variation for compliers as

$$S_{\tau \tau, c} = \frac{1}{n_c} \sum_{i=1}^n I(U_i = c) (\tau_i - \tau_c)^2 = \frac{1}{n_c} \sum_{i=1}^n I(U_i = c) \tau_i^2 - \tau_c^2.$$

Therefore, the above three terms has the relationship $S_{\tau \tau} = \pi_c S_{\tau \tau, c} + S_{\tau \tau, U}$.

The decomposition $S_{\tau \tau, c} = S_{\delta \delta, c} + S_{\epsilon \epsilon, c}$ follows immediately from the definition of $\beta_c$.

Proof of Corollary 4. The proof follows from the same logic as the proofs of Theorem 4 and Corollary 1.

\[ \square \]
Appendix B  More detailed comments

Appendix B.1  Covariate adjustment to improve efficiency

In the main text, the role of covariates has been to model the treatment effect alone. In general, we also want to use covariates to reduce sampling variability of $\hat{\beta}_{RI}$, just as we can use covariates to get more precise estimates of the average treatment effect. In particular, the goal is to more precisely estimate $\hat{S}_{xt}$ as these are important random components in $\hat{\beta}_{RI}$. Let $W_i \in \mathbb{R}^J$ denote a vector of pretreatment covariates. Because $X_i$ and $W_i$ have different roles in estimation, they may also contain different sets of covariates, though, in practice, $X$ is likely to be a subset of $W$.

Following the covariate adjustment approach in survey sampling (Cochran [1977]), we can therefore obtain a model-assisted estimator for $\beta$ that uses $W$ to reduce sampling variability. To see this, we need several definitions. Define $\bar{W} = \sum_{i=1}^n W_i/n$ and $S_{ww} = \sum_{i=1}^n W_i W_i^T/n$, with det($S_{ww}$) > 0; define $\bar{W}_t$ and $\hat{S}_{ww,t}$ as the sample mean and covariance of $W$ under treatment arm $t$; define $\hat{B}_t \in \mathbb{R}^J \times K$ as the regression coefficient of $Y_{obs}X$ on $W$ for treatment arm $t$: 

$$\hat{B}_t = \hat{S}_{-1}^{ww,t} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} I(T_i=t)W_i(Y_{obs i}X_i)^T \right\}.$$ 

The model-assisted estimator for $S_{xt}$ is then

$$\hat{S}_{xt}^w = \hat{S}_{xt} - \hat{B}_t^T(\bar{W}_t - \bar{W}) \quad (t = 0, 1).$$

As a result, we can improve the randomization-based estimator by

$$\hat{\beta}_{RI}^w = S_{xx}^{-1}(\hat{S}_{xt}^w - \hat{S}_{x0}^w).$$

Theorem A.1. The model-assisted estimator $\hat{\beta}_{RI}^w$ is consistent for $\beta$, and has asymptotic covariance matrix

$$S_{xx}^{-1} S\{E(1)\}_n / n_1 + S\{E(0)\}_n / n_0 - S(\Delta)_n / n \right] S_{xx}^{-1},$$

where $E_i(t) = Y_i(t)X_i - B_i^T(W_i - \bar{W})$ is the residual of the population regression of $Y_i(t)X_i$ on $W_i - \bar{W}$ for $t = 1$ and 0, and $\Delta_i = E_i(1) - E_i(0)$.

The resulting estimator, $\hat{\beta}_{RI}^w$, therefore uses covariates both to estimate treatment effect variation and to reduce sampling variability. Asymptotically, as long as $W$ is predictive of the marginal potential outcomes, the model-assisted estimator will improve precision over the unassisted estimators. Below we give the proof of Theorem A.1.
Proof of Theorem A.1. The population-level OLS regression matrix of $Y(t)$ $X$ onto $W$ is

$$B_t = S_{ww}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} W_i(Y_i(t)X_i)^T \right\} \in \mathbb{R}^{J \times K}.$$ 

Define $S_{xt} = \hat{S}_{xt} + B_t^T(\bar{W} - \bar{W}_t)$ and $\tilde{\beta}_{RI} = S_{xx}^{-1}(\hat{S}_{x1} - \hat{S}_{x0})$. We first observe that

$$\hat{\beta}_{RI} - \tilde{\beta}_{RI} = S_{xx}^{-1} \left\{ (\hat{B}_1 - B_1)^T(\bar{W} - \bar{W}_1) + (\hat{B}_0 - B_0)^T(\bar{W} - \bar{W}_0) \right\} = S_{xx}^{-1} \left\{ O_P(n^{-1/2})O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(n^{-1/2}) \right\} = O_P(n^{-1}), \quad (A.15)$$

based on the same rationale of regression estimator in surveys (Cochran, 1977). Therefore, $\hat{\beta}_{RI}$ and $\tilde{\beta}_{RI}$ have the same asymptotic covariance, and in the following we need only to discuss the covariance of $\tilde{\beta}_{RI}$. Because

$$\tilde{S}_{x1} - \tilde{S}_{x0} = \frac{1}{n_1} \sum_{i=1}^{n_1} T_i \{ Y_i(1)X_i + B_1^T(\bar{W} - \bar{W}_i) \} - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) \{ Y_i(0)X_i + B_0^T(\bar{W} - \bar{W}_i) \}$$

$$= \frac{1}{n_1} \sum_{i=1}^{n_1} T_i E_i(1) - \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) E_i(0)$$

can be represented as the difference between the sample means of $E_i(1)$ and $E_i(0)$, applying Theorem 2 we can obtain its variance:

$$\text{cov} \left( \tilde{S}_{x1} - \tilde{S}_{x0} \right) = S \left\{ \frac{E(1)}{n_1} \right\} + S \left\{ \frac{E(0)}{n_0} \right\} - S \left\{ \frac{\Delta}{n} \right\},$$

which completes the proof. 

\[ \Box \]

Appendix B.2 Fisherian exact inference

When $\varepsilon_i = 0$ for all $i$, we can obtain exact inference for $\beta$ based on the Fisher randomization test (Rubin, 1980; Rosenbaum, 2002; Ding et al., 2016). With a known $\beta$, the null hypothesis

$$H_0(\beta) : Y_i(1) - Y_i(0) = X_i^T \beta$$

(A.16)

is sharp in the sense of allowing for full imputation of all missing potential outcomes based on the observed data. We can perform randomization test using any sensible test statistic measuring the deviation from the null hypothesis $H_0(\beta)$, for example, the test statistic $t(T, Y^{obs}; \beta)$ can be the difference-in-means, difference-in-medians or the Kolmogorov–Smirnov statistics comparing two
samples \( \{ Y_{i}^{\text{obs}} - X_i^T \beta : T_i = 1, i = 1, \ldots, n \} \) and \( \{ Y_{i}^{\text{obs}} : T_i = 0, i = 1, \ldots, n \} \). Then we can obtain a \((1 - \alpha)\) level confidence region for \( \beta \) by inverting a sequence of randomization tests:

\[
\text{CR}_\alpha = \{ \beta : \text{Randomization test fails to reject } H_0(\beta) \text{ at significance level } \alpha \}.
\]

The confidence region \(\text{CR}_\alpha\) is exact regardless of the sample size, and it is valid for general designs of experiments if we use the corresponding treatment mechanism to simulate the null distribution of the test statistic. Due to the duality between testing and interval estimation, we reject \( H_0(X) \) if \(\text{CR}_\alpha \cap \{ \beta : \beta_1 = 0 \} \) is an empty set, which controls the type one error rate by \( \alpha \).

### Appendix B.3 Variance of the average treatment effect estimate

To estimate the average treatment effect \( \tau = \sum_{i=1}^{n} \tau_i / n \), Neyman (1923) proposed the difference-in-means statistic, \( \hat{\tau} = Y_{1}^{\text{obs}} - \bar{Y}_0^{\text{obs}} \), which is an unbiased estimator. Its sampling variance,

\[
\text{var}(\hat{\tau}) = \frac{S_{11}}{n_1} + \frac{S_{00}}{n_0} - \frac{S_{\tau\tau}}{n},
\]

depends on \( S_{11}, S_{00}, \) and \( S_{\tau\tau} \), the finite population variances of \( Y_i(1), Y_i(0) \) and \( \tau_i \), respectively. While \( S_{11} \) and \( S_{00} \) are estimable quantities, \( S_{\tau\tau} \) depends on the correlation of potential outcomes and is unidentified.

There are a range of variance estimators that circumvent this unidentifiability. Neyman (1923) initially proposed a lower bound for \( \text{var}(\hat{\tau}) \) under the assumption of a constant treatment effect, \( S_{\tau\tau} = 0 \). More recently, Aronow et al. (2014) proposed to bound \( S_{\tau\tau} \) via Fréchet–Hoeffding bounds rather than to assume \( S_{\tau\tau} = 0 \) (cf. Heckman et al., 1997).

We propose a modest extension here. Using the results in Section 4, we have

\[
\text{var}(\hat{\tau}) = \frac{S_{11}}{n_1} + \frac{S_{00}}{n_0} - \left( \frac{S_{\delta\delta}}{n} + \frac{S_{\varepsilon\varepsilon}}{n} \right).
\]

We can then estimate \( S_{\delta\delta} \) directly and apply Fréchet–Hoeffding bounds of \( S_{\varepsilon\varepsilon} \) in Theorem 4 rather than to \( S_{\tau\tau} \). So long as \( S_{\delta\delta} > 0 \), this yields strictly tighter bounds on \( \text{var}(\hat{\tau}) \) than the corresponding bounds that do not incorporate covariate information. This gives a tighter estimate of the standard error without changing the estimated value \( \hat{\tau} \); the estimator \( \hat{\tau} \) is still the simple difference estimator.

While this is of theoretical interest, we note that we do not see meaningful gains in practice, at least in our applied example. In that context, relative to the estimator that assumes \( S_{\tau\tau} = 0 \), the reduction in variance is roughly 1 percent using Fréchet–Hoeffding bounds either for \( S_{\tau\tau} \) or for \( S_{\varepsilon\varepsilon} \).
To extend the result to the noncompliance setting, we need to derive the Neyman-type variance estimator for the LATE, which is beyond the scope of the current paper and is the topic of ongoing research.

Appendix B.4 A Variance Ratio Test

Raudenbush and Bloom (2015) have noticed that if the variance of the treatment potential outcome is smaller than the control potential outcome, then the correlation between the individual treatment effect and the control potential outcome is negative. This statement does not involve any covariates, which, in fact, can be generalized to incorporate systematic and idiosyncratic treatment effect variation. Below we give a finite population version of their result.

Theorem A.2. If the finite population variance of \( \{Y_i(1) - X_i'\beta\}_{i=1}^n \) is smaller than \( \{Y_i(0)\}_{i=1}^n \), then the idiosyncratic treatment effect variation, \( \{\varepsilon_i(0)\}_{i=1}^n \), is negatively correlated with the control potential outcomes.

Because the condition in Theorem A.2 depends only on the marginal distributions of the potential outcomes, we propose a formal test of it using the observed data. While many tests are possible, we propose a variance ratio test, which is a generalization of a similar theorem in Ding et al. (2016):

Theorem A.3. The variance ratio test with rejection region

\[
\frac{\log s_1^2 - \log s_0^2}{\sqrt{(\hat{\kappa}_1 - 1)/n_1 + (\hat{\kappa}_0 - 1)/n_0}} < \Phi^{-1}(\alpha),
\]

has size at least as large as \( \alpha \), where \( s_1^2 \) and \( \hat{\kappa}_1 \) are the sample variance and kurtosis of \( \{Y_{i,\text{obs}} - X_i\hat{\beta}_{RI} : T_i = 1, i = 1, \ldots, n\} \), and \( s_0^2 \) and \( \hat{\kappa}_0 \) are the sample variance and kurtosis of \( \{Y_{i,\text{obs}} : T_i = 0, i = 1, \ldots, n\} \), and \( \Phi^{-1}(\alpha) \) is the \( \alpha \)-th quantile of the standard normal distribution.

For finite population inference, the above test in Theorem A.3 is generally conservative, but for superpopulation inference, it is asymptotically exact. The conservativeness is not a problem of our test, but rather a feature of finite population inference as suggested by Neyman (1923).

Note that Raudenbush and Bloom (2015) and Theorem A.2 are only about detecting negative association. Unfortunately, there is no testable condition for positive association between them. Below we give proofs for Theorems A.2 and A.3.
Proof of Theorem A.2. For simplicity, we abuse the variance and covariance notation for finite population. For example, \( \text{var}\{Y(0)\} = \sum_{i=1}^{n}(Y_i(0) - \bar{Y}(0))^2/(n-1) \). If \( \text{var}\{Y(1) - X^T \beta\} \leq \text{var}\{Y(0)\} \), then \( \text{var}\{Y(0) + \varepsilon\} \leq \text{var}\{Y(0)\} \). Expanding the left hand side,

\[
\text{var}\{Y(0)\} + \text{var}\{\varepsilon\} + 2\text{cov}\{Y(0), \varepsilon\} \leq \text{var}\{Y(0)\},
\]

which implies \( 2\text{cov}\{Y(0), \varepsilon\} \leq -\text{var}\{\varepsilon\} < 0. \)

Although it is straightforward to prove the conclusion for super population inference of Theorem A.3 by using Ding et al. (2016, Theorem 2, Supplementary Material) and Slutsky’s Theorem, it is less obvious to prove the conclusion for finite population inference. To simplify the proof, we first prove the following lemma. Let \((c_1, \ldots, c_n)^T\) and \((d_1, \ldots, d_n)^T\) be two vectors of nonnegative constants with the same mean \( m > 0 \) but different variances \( S^2_c \) and \( S^2_d \). The difference vector \((c_1 - d_1, \ldots, c_n - d_n)^T\) has mean zero and variance \( S^2_{c-d} \). Let

\[
\hat{\theta}_c = \frac{1}{n_1} \sum_{i=1}^{n} T_i c_i, \quad \hat{\theta}_d = \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) d_i
\]

be two sample means of the treatment and control group, respectively.

Lemma A.5. Under the regularity conditions for the finite population Central Limit Theorem (Hájek 1960; Lehmann 1998), \( \log \hat{\theta}_c - \log \hat{\theta}_d \) has asymptotic mean zero and variance

\[
\frac{1}{m^2} \left( \frac{S^2_c}{n_1} + \frac{S^2_d}{n_0} - \frac{S^2_{c-d}}{n} \right). \quad \text{(A.17)}
\]

Proof of Lemma A.5. According to the finite population central limit theorem (Hájek 1960; Lehmann 1998), we have the following joint asymptotic normality of \( \hat{\theta}_c \) and \( \hat{\theta}_d \):

\[
\begin{pmatrix} \hat{\theta}_c \\ \hat{\theta}_d \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} \sum_{i=1}^{n} T_i c_i \\ \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) d_i \end{pmatrix} \sim N \left[ \begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} V_c & V_{cd} \\ V_{cd} & V_d \end{pmatrix} \right],
\]

where

\[
V_c = \frac{n_0}{n_1 n} S^2_c, \quad V_d = \frac{n_1}{n_0 n} S^2_d, \quad V_{cd} = -\frac{1}{2n} (S^2_c + S^2_d - S^2_{c-d}).
\]

Applying Taylor expansion at \( m \), we have \( \log \hat{\theta}_c - \log \hat{\theta}_d = \{(\hat{\theta}_c - m) - (\hat{\theta}_d - m)\}/m + o_P(n^{-1/2}), \) which, coupled with Neyman (1923)'s variance formula, gives the asymptotic variance of \( \log \hat{\theta}_c - \log \hat{\theta}_d \) in (A.17). \( \square \)
Proof of Theorem A.3. First, as a direct consequence of Lemma A.5, the finite sample variance is always larger than the super population variance, unless \( S_{c-d}^2 = 0 \). Therefore, we need only to show that the test in Theorem A.3 is asymptotically exact for super population inference, and the asymptotic size of the test is no larger than \( \alpha \) for finite population inference.

Second, replacing \( \beta \) by its consistent estimator \( \hat{\beta}_{RI} \) does not affect the asymptotic distribution of the test statistic, due to the Slutsky’s Theorem. For simplicity, we treat \( \beta \) as known in our asymptotic analysis.

With the two ingredients above, Theorem A.3 follows directly from the variance ratio test in Ding et al. (2016, Theorem 2, Supplementary Material).

Appendix B.5 More on noncompliance: estimating the bounds of \( R^2 \)

The component \( S_{\tau \tau, U} \) and the probability \( \pi_c \) are directly identifiable according to previous discussion. Furthermore, \( S_{\delta \delta, c} \) is also identifiable according to the following result.

**Corollary A.1.** \( S_{\delta \delta, c} \) can be expressed as the expectation of the following quantity:

\[
\frac{1}{\pi_c} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\delta_i - \tau_c)^2 - \frac{1}{n_1} \sum_{i=1}^{n} T_i (1 - D_i) (\delta_i - \tau_c)^2 - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) D_i (\delta_i - \tau_c)^2 \right\}.
\]

Because \( \pi_c, \delta_i = X_i^\top \beta_c \) and \( \tau_c \) can be estimated by a plug-in approach, \( S_{\delta \delta, c} \) can also be estimated from the observed data.

In the ITT case, estimation of the residual distributions are straightforward. In the noncompliance case, however, we need more discussion about the estimation of \( F_{1c}(y) \) and \( F_{0c}(y) \), because \( U_i \) is a latent variable. To avoid notational clatter, we assume that \( \gamma_{c1} \) and \( \gamma_{c0} \) are known; in practice we can replace them by the randomization-based estimators \( \hat{\gamma}_{c1,Ri} \) and \( \hat{\gamma}_{c0,Ri} \), and the consistency of the final estimator will not be affected. Recall the potential residuals \( e'_i(1) \) and \( e'_i(0) \) defined in (14), and its observed value \( e'_i = T_i e'_i(1) + (1 - T_i) e'_i(0) \). We define the following quantities

\[
\begin{align*}
\hat{F}_{11}(y) &= \frac{1}{n_1} \sum_{i=1}^{n} T_i D_i I(e'_i \leq y), \\
\hat{F}_{01}(y) &= \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) D_i I(e'_i \leq y), \\
\hat{F}_{10}(y) &= \frac{1}{n_1} \sum_{i=1}^{n} T_i (1 - D_i) I(e'_i \leq y), \\
\hat{F}_{00}(y) &= \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) (1 - D_i) I(e'_i \leq y).
\end{align*}
\]

(A.18)

Similar to Corollary 3 we have the following results.

**Corollary A.2.** For any \( y \),

\[
E\{\hat{F}_{11}(y) - \hat{F}_{01}(y)\} = \pi_c F_{1c}(y), \quad E\{\hat{F}_{00}(y) - \hat{F}_{10}(y)\} = \pi_c F_{0c}(y).
\]
Therefore, we can estimate $F_{1c}(y)$ by $\{\hat{F}_{11}(y) - \hat{F}_{01}(y)\}/\hat{\pi}_c$, and estimate $F_{0c}(y)$ by $\{\hat{F}_{00}(y) - \hat{F}_{10}(y)\}/\hat{\pi}_c$. As we mentioned before, in practice, we use $\hat{e}_i'$ instead of $e_i'$ in the formulas in (A.18).

We end this subsection with the proofs of the above corollaries.

**Proof of Corollary A.1.** The conclusion follows from

$$E\left\{\frac{1}{n_1} \sum_{i=1}^{n_1} T_i (1 - D_i) (\delta_i - \tau_c)^2\right\} = E\left\{\frac{1}{n_1} \sum_{i=1}^{n_1} T_i I(U_i = n) (\delta_i - \tau_c)^2\right\} = \frac{1}{n} \sum_{i=1}^{n} I(U_i = n) (\delta_i - \tau_c)^2,$$

$$E\left\{\frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) D_i (\delta_i - \tau_c)^2\right\} = E\left\{\frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) I(U_i = a) (\delta_i - \tau_c)^2\right\} = \frac{1}{n} \sum_{i=1}^{n} I(U_i = a) (\delta_i - \tau_c)^2.$$

**Proof of Corollary A.2.** We rewrite

$$\hat{F}_{11}(y) = \frac{1}{n_1} \sum_{i=1}^{n_1} T_i I(U_i = c) I(e_i(1) \leq y) + \frac{1}{n_1} \sum_{i=1}^{n_1} T_i I(U_i = a) I(e_i(1) \leq y),$$

$$\hat{F}_{10}(y) = \frac{1}{n_1} \sum_{i=1}^{n_1} T_i I(U_i = n) I(e_i(1) \leq y),$$

$$\hat{F}_{01}(y) = \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) I(U_i = a) I(e_i(0) \leq y),$$

$$\hat{F}_{00}(y) = \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) I(U_i = c) I(e_i(0) \leq y) + \frac{1}{n_0} \sum_{i=1}^{n_0} (1 - T_i) I(U_i = n) I(e_i(0) \leq y).$$

In the above formulas, the random components are the $T_i$'s, and therefore, the corollary follows from Lemma A.1 and the linearity of expectations.