NBER WORKING PAPER SERIES

WHEN SHOULD YOU ADJUST STANDARD ERRORS FOR CLUSTERING?

Alberto Abadie
Susan Athey
Guido W. Imbens
Jeffrey Wooldridge

Working Paper 24003
http://www.nber.org/papers/w24003

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2017

The questions addressed in this paper partly originated in discussions with Gary Chamberlain. We are grateful for questions raised by Chris Blattman. We are grateful to seminar audiences at the 2016 NBER Labor Studies meeting, CEMMAP, Chicago, Brown University, the Harvard-MIT Econometrics seminar, Ca' Foscari University of Venice, the California Econometrics Conference, the Erasmus University Rotterdam, and Stanford University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w24003.ack

When Should You Adjust Standard Errors for Clustering?
Alberto Abadie, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge
NBER Working Paper No. 24003
November 2017
JEL No. C21

ABSTRACT

In empirical work in economics it is common to report standard errors that account for clustering of units. Typically, the motivation given for the clustering adjustments is that unobserved components in outcomes for units within clusters are correlated. However, because correlation may occur across more than one dimension, this motivation makes it difficult to justify why researchers use clustering in some dimensions, such as geographic, but not others, such as age cohorts or gender. This motivation also makes it difficult to explain why one should not cluster with data from a randomized experiment. In this paper, we argue that clustering is in essence a design problem, either a sampling design or an experimental design issue. It is a sampling design issue if sampling follows a two stage process where in the first stage, a subset of clusters were sampled randomly from a population of clusters, and in the second stage, units were sampled randomly from the sampled clusters. In this case the clustering adjustment is justified by the fact that there are clusters in the population that we do not see in the sample. Clustering is an experimental design issue if the assignment is correlated within the clusters. We take the view that this second perspective best fits the typical setting in economics where clustering adjustments are used. This perspective allows us to shed new light on three questions: (i) when should one adjust the standard errors for clustering, (ii) when is the conventional adjustment for clustering appropriate, and (iii) when does the conventional adjustment of the standard errors matter.

Alberto Abadie
Department of Economics, E52-546
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
abadie@mit.edu

Guido W. Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
Imbens@stanford.edu

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Jeffrey Wooldridge
Department of Economics
Michigan State University
wooldri1@msu.edu

1 Introduction

In empirical work in economics, it is common to report standard errors that account for clustering of units. The first issue we address in this manuscript is the motivation for this adjustment. Typically the stated motivation is that unobserved components of outcomes for units within clusters are correlated (Moulton [1986, 1990], Moulton and Randolph [1989], Kloek [1981], Hansen [2007], Cameron and Miller [2015]). For example, Hansen [2007] writes: "The clustering problem is caused by the presence of a common unobserved random shock at the group level that will lead to correlation between all observations within each group" (Hansen [2007], p. 671). Similarly Cameron and Miller [2015] write: "The key assumption is that the errors are uncorrelated across clusters while errors for individuals belonging to the same cluster may be correlated" (Cameron and Miller [2015], p. 320). This motivation for clustering adjustments in terms of within-group correlations of the errors makes it difficult to justify clustering by some partitioning of the population, but not by others. For example, in a regression of wages on years of education, this argument could be used to justify clustering by age cohorts just as easily as clustering by state. Similarly, this motivation makes it difficult to explain why, in a randomized experiment, researchers typically do not cluster by groups. It also makes it difficult to motivate clustering if the regression function already includes fixed effects. The second issue we address concerns the appropriate level of clustering. The typical answer is to go for the most aggregate level feasible. For example, in a recent survey Cameron and Miller [2015] write: "The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters." (Cameron and Miller [2015], p. 333). We argue in this paper that there is in fact harm in clustering at too aggregate a level, We also make the case that the confusion regarding both issues arises from the dominant model-based perspective on clustering.

We take the view that clustering is in essence a *design* problem, either a *sampling design* or an *experimental design* issue. It is a sampling design issue when the sampling follows a two stage process, where in the first stage, a subset of clusters is sampled randomly from a population of clusters, and in the second stage, units are sampled randomly from the sampled clusters. Although this clustered sampling approach is the perspective taken most often when a formal justification is given for clustering adjustments to standard errors, it actually rarely fits applications in economics. Angrist and Pischke [2008] write: "Most of the samples that we work with are close enough to random that we typically worry more about the dependence due to a group structure than clustering due to stratification." (Angrist and Pischke [2008], footnote 10, p. 309). Instead of a sampling issue, clustering can also be an *experimental design* issue, when clusters of units, rather than units, are assigned to a treatment. In the view developed in this manuscript, this perspective fits best the typical application in economics, but surprisingly it is rarely explicitly presented as the motivation for cluster adjustments to the standard errors.

We argue that the design perspective on clustering, related to randomization inference (e.g., Rosenbaum [2002], Athey and Imbens [2017]), clarifies the role of clustering adjustments to standard errors and aids in the decision whether to, and at what level to, cluster, both in standard clustering settings and in more general spatial correlation settings (Bester et al.

[2009], Conley [1999], Barrios et al. [2012], Cressie [2015]). For example, we show that, contrary to common wisdom, correlations between residuals within clusters are neither necessary, nor sufficient, for cluster adjustments to matter. Similarly, correlations between regressors within clusters are neither necessary, not sufficient, for cluster adjustments to matter or to justify clustering. In fact, we show that cluster adjustments can matter, and substantially so, even when both residuals and regressors are uncorrelated within clusters. Moreover, we show that the question whether, and at what level, to adjust standard errors for clustering is a substantive question that cannot be informed solely by the data. In other words, although the data are informative about *whether* clustering matters for the standard errors, but they are only partially informative about whether one *should* adjust the standard errors for clustering. A consequence is that in general clustering at too aggregate a level is not innocuous, and can lead to standard errors that are unnecessarily conservative, even in large samples.

One important theme of the paper, building on Abadie et al. [2017], is that it is critical to define estimands carefully, and to articulate precisely the relation between the sample and the population. In this setting that means one should define the estimand in terms of a finite population, with a finite number of clusters and a finite number of units per clusters. This is important even if asymptotic approximations to finite sample distributions involve sequences of experiments with an increasing number of clusters and/or an increasing number of units per cluster. In addition, researchers need to be explicit about the way the sample is generated from this population, addressing two issues: (i) how units in the sample were selected and, most importantly whether there are clusters in the population of interest that are not represented in the sample, and (ii) how units were assigned to the various treatments, and whether this assignment was clustered. If either the sampling or assignment varies systematically with groups in the sample, clustering will in general be justified. We show that the conventional adjustments, often implicitly, assume that the clusters in the sample are only a small fraction of the clusters in the population of interest. To make the conceptual points as clear as possible, we focus in the current manuscript on the cross-section setting. In the panel case (e.g., Bertrand et al. [2004]), the same issues arise, but there are additional complications because of the time series correlation of the treatment assignment. Analyzing the uncertainty from the experimental design perspective would require modeling the time series pattern of the assignments, and we leave that to future work.

The practical implications from the results in this paper are as follows. First, the researcher should assess whether the sampling process is clustered or not, and whether the assignment mechanism is clustered. If the answer to both is no, one should *not* adjust the standard errors for clustering, irrespective of whether such an adjustment would change the standard errors. Second, in general, the standard iang-Zeger clustering adjustment is conservative unless one of three conditions holds: (i) there is no heterogeneity in treatment effects; (ii) we observe only a few clusters from a large population of clusters; or (iii) a vanishing fraction of units in each cluster is sampled, e.g. at most one unit is sampled per cluster. Third, the (positive) bias from standard clustering adjustments can be corrected if all clusters are included in the sample and further, there is variation in treatment assignment within each cluster. For this case we propose a new variance estimator. Fourth, if one estimates a fixed effects regression (with fixed effects

at the level of the relevant clusters), the analysis changes. Then, heterogeneity in the treatment effects is a requirement for a clustering adjustment to be necessary.

2 Simple Example and Two Misconceptions

In this section we discuss two misconceptions about clustering that appear common in the literature. The first misconception is about when clustering matters, and the second about whether one ought to cluster. Both misconceptions are related to the common model-based perspective of clustering which we outline briefly below. We argue that this perspective obscures the justification for clustering that is relevant for most empirical work.

2.1 The Model-based pproach to Clustering

First let us briefly review the textbook, model-based approach to clustering (e.g., Cameron and Miller [2015], Wooldridge [2003, 2010], Angrist and Pischke [2008]). ater, we contrast this with the design-based approach starting from clustered randomized experiments (Donner and Klar [2000], Murray [1998], Fisher [1937]). Consider a setting where we wish to model a scalar outcome Y_i in terms of a binary covariate $W_i \in \{0,1\}$, with the units belonging to clusters, with the cluster for unit i denoted by $C_i \in \{1,\ldots,C\}$. We estimate the linear model

$$Y_i = \alpha + \tau W_i + \varepsilon_i = \beta^\top X_i + \varepsilon_i,$$

where $\beta^\top = (\alpha,\tau)$ and $X_i^\top = (1, W_i)$, using least squares, leading to

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^N \left(Y_i - \beta^\top X_i\right)^2 = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\mathbf{Y}\right).$$

In the model-based perspective, the N-vector with ith element equal to ε_i, is viewed as the stochastic component. The $N \times 2$ matrix $\mathbf{X}$ with ith row equal to $(1, W_i)$ and the N-vector $\mathbf{C}$ with ith element equal to C_i are viewed as non-stochastic. Thus the repeated sampling thought experiment is redrawing the vectors , keeping fixed $\mathbf{C}$ and $\mathbf{W}$.

Often the following structure is imposed on the first two moments of ε,

$$\mathbb{E}[\ \mathbf{X},\mathbf{C}] = 0, \qquad \mathbb{E}\left[\ \ ^\top \middle| \mathbf{X},\mathbf{C}\right] = \Omega,$$

leading to the following expression for the variance of the ordinary least squares (O S) estimator:

$$\mathbb{V}(\hat{\beta}) = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^\top\Omega\mathbf{X}\right)\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.$$

In the setting without clustering, the key assumption is that Ω is diagonal. If one is also willing to assume homoskedasticity the variance reduces to the standard O S variance:

$$\mathbb{V}_{\text{O S}} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1},$$

where $\sigma = \Omega_{ii} = \mathbb{V}(\varepsilon_i)$ for all i. Often researchers allow for general heteroskedasticity and use the robust Eicker-Huber-White (EHW) variance (White [2014], Eicker [1967], Huber [1967])

$$\mathbb{V}_{\text{EHW}}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\sum_{i=1}^{N} \Omega_{ii} X_i X_i^\top\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}.$$

In settings with clusters of units, the assumption that Ω is diagonal is often viewed as not credible. Instead, Kloek [1981], Moulton and Randolph [1989], Moulton [1990] use the (homoskedastic) structure

$$\Omega_{ij} = \begin{cases} 0 & \text{if } C_i \neq C_j, \\ \rho\sigma^2 & \text{if } C_i = C_j,\ i \neq j, \\ \sigma^2 & \text{if } i = j. \end{cases}$$

Assuming the clusters are equal size this leads to the following variance for the slope coefficient $\hat{\tau}$:

$$\mathbb{V}_{\text{K OEK}}(\hat{\tau}) = \mathbb{V}_{\text{O S}} \times \left(1 + \rho_\varepsilon \rho_W \frac{N}{C}\right), \tag{2.1}$$

where ρ_ε and ρ_W are the within-cluster correlation of the errors and covariates respectively. Often researchers (*e.g.,* iang and Zeger [1986], Diggle et al. [2013], Bertrand et al. [2004], Stock and Watson [2008], William [1998]) further relax this model by allowing the Ω_{ij} for pairs (i, j) with $C_i = C_j$ to be unrestricted. et the units be ordered by cluster, and let the $N_c \times N_c$ submatrix of Ω corresponding to the units from cluster c be denoted by Ω_c, and the submatrix of $\mathbf{X}$ corresponding to cluster c by $\mathbf{X}_c$. Then:

$$\mathbb{V}_{\text{Z}}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\sum_{c=1}^{C} \mathbf{X}_c^\top \Omega_c \mathbf{X}_c\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}.$$

This can be viewed as the extension to robust variance estimator from the least squares variance, applied in the case with clustering.

The estimated version of the EHW variance is

$$\hat{\mathbb{V}}_{\text{EHW}}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\sum_{i=1}^{N} (Y_i - \hat{\beta}^\top X_i)^2 X_i X_i^\top\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}. \tag{2.2}$$

The estimated version of the Z variance is

$$\hat{\mathbb{V}}_{\text{Z}}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\sum_{c=1}^{C} \left(\sum_{i:C_i=c} (Y_i - \hat{\beta}^\top X_i) X_i\right) \left(\sum_{i:C_i=c} (Y_i - \hat{\beta}^\top X_i) X_i\right)^\top\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}. \tag{2.3}$$

2.2 Clustering Matters Only if the Residuals and the Regressors are both Correlated Within Clusters

There appears to be a view, captured by the expression in equation (2.1), that whether the cluster correction to the standard errors matters depends on two objects. First, it depends on the within-cluster correlation of the residuals, ρ_ε, and second, it depends on the within-cluster correlation of the regressors of interest, ρ_W. It has been argued that clustering does not matter if either of the two within-cluster correlations are zero. If this were true, an implication would be that in large samples the cluster adjustment makes no difference in a randomized experiment with completely randomly assigned treatments. This would follow because in that case the within-cluster correlation of the regressor of interest is zero by virtue of the random assignment. A second implication would be that, in a cross-sectional data context, if one includes fixed effects in the regression function to account for the clusters, there is no reason to cluster standard errors, because the fixed effects completely eliminate the within-cluster correlation of the residuals. Although the latter implication is known to be false (*e.g.*, Arellano [1987]), the perception has lingered.

To illustrate the fallacy of this view, we simulated a single data set with $N = 100,323$ units, partitioned into $C = 100$ clusters or strata with an average approximately 1,000 units per cluster, where the actual number of units per cluster ranges from 950 to 1063. Both the number of clusters and the number of units per cluster are substantial to avoid small sample problems of the type analyzed in Donald and ang [2007] and Ibragimov and Müller [2010, 2016]. Below we discuss exactly how the sample was generated, here we wish to make the basic point that whether the clustering adjustment matters in a given sample is not simply a matter of inspecting the within-cluster correlation of the errors and covariates. For each unit in our sample we observe an outcome Y_i, a single binary regressor $W_i \in \{0, 1\}$, and the cluster label $C_i \in \{1, \ldots, C\}$. We estimate a linear regression function,

$$Y_i = \alpha + \tau W_i + \varepsilon_i,$$

by O S.

For our sample set we first calculate the within-cluster correlation of the residuals and the within-cluster correlation of the regressors. We estimate these by first calculating the sample variance of the residuals (regressors) with and without demeaning by cluster, and then taking the ratio of the difference of these two to the overall variance of the residuals (regressors), leading to:

$$\hat{\rho}_{\hat{\varepsilon}} = 0.001, \qquad \hat{\rho}_W = 0.001.$$

Both within-cluster correlations are close to zero, and because there is only modest variation in cluster sizes, the standard Moulton-Kloek (Kloek [1981], Moulton [1986], Moulton and Randolph [1989], Moulton [1990]) adjustment given in (2.1) would essentially be zero. However, when we calculate the least squares estimator for τ and both the EHW and Z standard errors, we find that the clustering does matter substantially:

$$\hat{\tau}^{\text{ls}} = -0.120 \qquad (\text{se}_{\text{EHW}} = 0.004) \qquad [\text{se}_{\text{Z}} = 0.100].$$

This demonstrates that inspecting the within-cluster correlation of the residuals and the within-cluster correlation of the regressors is not necessarily informative about the question whether clustering the standard errors using the iang-Zeger variance estimator matters.

Instead, what is relevant for whether the iang-Zeger variance adjustment matters is the within-cluster correlation of the product of the residuals and the regressors. Calculating that correlation, we find

$$\rho_{\hat{\varepsilon}W} = 0.500.$$

This correlation is substantial, and it explains why the clustering matter. Note that this does not mean one *should* adjust the standard errors, merely that doing so will matter.

If we use a fixed effects regression instead of O S, the same conclusion arises, not surprisingly given the Arellano [1987] results. We estimate the fixed effects regression

$$Y_i = \tau W_i + \sum_{c=1}^{C} \alpha_c C_{ic} + \varepsilon_i,$$

where $C_{ic} = \mathbf{1}_{C_i=c}$ is a binary indicator equal to one if unit i belongs to cluster c, and zero otherwise. We run this regression and estimate both the regular and the clustered standard errors (without degrees of freedom corrections, which do not matter here given the design), leading to:

$$\hat{\tau}^{\text{fe}} = -0.120 \qquad (\text{se}_{\text{EHW}} = 0.003) \qquad [\text{se}_{\text{C}} = 0.243].$$

Again, the clustering of the standard errors makes a substantial difference, despite the fact that the within-cluster correlation of the residuals is now exactly equal to zero.

2.3 If Clustering Matters, One Should Cluster

There is also a common view that there is no harm, at least in large samples, to adjusting the standard errors for clustering. Therefore, one should cluster at the highest level of aggregation possible, subject to finite sample issues: if clustering matters, it should be done, and if it does not matter, clustering the standard errors does no harm, at least in large samples. Based on this perception, many discussions of clustering adjustments to standard errors recommend researchers to calculate diagnostics on the sample to inform the decision whether or not one should cluster. These diagnostics often amount to simply comparing standard errors with and without clustering adjustments. We argue that such attempts are futile, and that a researcher should decide whether to cluster the standard errors based on substantive information, not solely based on whether it makes a difference.

To discuss whether one ought to cluster, we step back from the previously analyzed sample and consider both the population this sample was drawn from, and the manner in which it was drawn. We had generated a population with 10,000,000 units, partitioned $C = 100$ clusters, each cluster with exactly 100,000 units. Units were assigned a value $W_i \in \{0, 1\}$, with probability 1/2 for each value, independent of everything else. The outcome for unit i was generated as

$$Y_i = \tau_{C_i} W_i + \nu_i.$$

where ν_i was drawn from a normal distribution with mean zero and unit variance independent across all units. The slope coefficients τ_c are cluster-specific coefficients, equal to $\tau_c = -1$ for exactly half the clusters and equal to $\tau_c = 1$ for the other half, so that the average treatment in the population is exactly zero. We sample units from this population completely randomly, with the probability for each unit of being sampled equal to 0.01.

In this example, the EHW standard errors are the appropriate ones, even though the Z standard errors are substantially larger. We first demonstrate this informally, and present some formal results that cover this case in the next section. For the informal argument, let us 10,000 times draw our sample, and calculate the least squares estimator and both the EHW and Z standard errors. Table 1 gives the coverage rates for the associated 95% confidence intervals for the true average effect of zero. The Z standard errors are systematically substantially larger than the EHW standard errors, and the Z-based confidence intervals have substantial over-coverage, whereas the EHW confidence intervals are accurate. This holds for the simple regressions and for the fixed effect regressions.

Table 1: Standard Errors and Coverage Rates Random Sampling, Random ssignment (10,000 replications)

No Fixed Effects				Fixed Effects			
EHW variance		Z variance		EHW variance		Z variance	
$\sqrt{\mathbb{V}_{\text{EHW}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{Z}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{EHW}}}$	cov rate	$\sqrt{\mathbb{V}_{\text{Z}}}$	cov rate
0.007	0.950	0.051	1.000	0.007	0.950	0.131	0.986

The reason for the difference between the EHW and Z standard errors is simple, but reflects the fundamental source of confusion in this literature. Given the random assignment both standard errors are correct, but for different estimands. The Z standard errors are based on the presumption that there are clusters in the population of interest beyond the 100 clusters that are seen in the sample. The EHW standard errors assume the sample is drawn randomly from the population of interest. It is this presumption underlying the Z standard errors of existence of clusters that are not observed in the sample, but that are part of the population of interest, that is critical, and often implicit, in the model-based motivation for clustering the standard errors. It is of course explicit in the sampling design literature (*e.g.,* Kish [1965]). If we changed the set up to one where the population of 10,000,000 consisted of say 1,000 clusters, with 100 clusters drawn at random, and then sampling units randomly from those sampled clusters, the Z standard errors would be correct, and the EHW standard errors would be incorrect. Obviously one cannot tell from the sample itself whether there exist such clusters that are part of the population of interest that are not in the sample, and therefore one needs to choose between the two standard errors on the basis of substantive knowledge of the study design.

3 Formal Result

In this section we consider a special case with a single binary covariate to formalize the ideas from the previous subsection. We derive the exact variance to an approximation of the least squares estimator, taking into account both sampling variation and variation induced by the experimental design, that is, by the assignment mechanism. This will allow us to demonstrate exactly when the EHW and Z variances are appropriate, and why they fail when they do so. To make the arguments rigorous, we do need large sample approximations. To do so, we build a sequence of finite populations where the sample size and the number of clusters goes to infinity. However, the estimands are defined for finite populations.

We start with the existence of a pair of potential outcomes for each unit. This implicitly makes the stable-unit-treatment-value assumption (sutva, Rubin [1980]) that rules out peer effects and versions of the treatment. There is a part of the clustering literature that is concerned with clusters of units receiving different treatments (e.g., clusters of individuals receiving services from the same health care provider, where the exact set of services provided may vary by provider), see for example ee and Thompson [2005], Roberts and Roberts [2005], Weiss et al. [2016]. Our analysis can be thought of applying to that case keeping fixed the health care provider associated with each individual, rather than focusing on the average effect over all possible health care providers that an individual might receive care from.

3.1 The Sequence of Populations

We have a sequence of populations indexed by n. The n-th population has M units, indexed by $i = 1, \ldots, M$, with M strictly increasing in n. The population is partitioned into C strata or clusters, with C weakly increasing in n. $C_i \in \{1, \ldots, C\}$ denotes the stratum that unit i belongs to. $C_{ic} = \mathbf{1}_{C_i = c}$ is a binary indicator, equal to 1 if unit i in population n belongs to cluster c and zero otherwise. The number of units in cluster c in population n is $M_c = \sum_{i=1} C_{ic}$, with $M = n = \sum_{c=1}^{C} M_c$. For each unit there are two potential outcomes, $Y_i(0)$ and $Y_i(1)$ for unit i, corresponding to a control and treated outcome (*e.g.*, Imbens and Rubin [2015]). We are interested in the population average effect of the treatment in population n,

$$\tau = \frac{1}{M} \sum_{i=1}^{M} \Big(Y_i(1) - Y_i(0) \Big) = \overline{Y}(1) - \overline{Y}(0),$$

where, for $w = 0, 1$

$$\overline{Y}(w) = \frac{1}{M} \sum_{i=1}^{M} Y_i(w).$$

It is also useful to define the population average treatment effect by cluster,

$$\tau_c = \frac{1}{M_c} \sum_{i: C_i = c} \Big(Y_i(1) - Y_i(0) \Big), \qquad \text{so that } \tau = \sum_{c=1}^{C} \frac{M_c}{M} \tau_c .$$

Define also the treatment-specific residuals and their cluster averages, for $w = 0, 1$,

$$\varepsilon_i\ (w) = Y_i\ (w) - \frac{1}{M}\sum_{j=1}^{M} Y_j\ (w), \qquad \overline{\varepsilon}_c\ (w) = \frac{1}{M_c}\sum_{i=1} C_i\ _c\varepsilon_i\ (w).$$

All these objects, $Y_i\ (w)$, $\varepsilon_i\ (w)$, $\overline{\varepsilon}_c\ (w)$, and functions thereof are non-stochastic.

There are some restrictions on the sequence of populations. These are mild regularity conditions, and most can be weakened. As n increases, the number of clusters increases without limit, the relative cluster sizes are bounded, and the potential outcomes do not become too large in absolute value.

Assumption 1. *The sequence of populations satisfies* (i)

$$\lim_{\to\infty} C^{-1} = 0,$$

(ii) *for some finite* K*,*

$$\frac{\max_c M_c}{\min_c M_c} \leq K,$$

and (iii) *for some* ,

$$\max_{i,w} Y_i\ (w)\ \leq\ \ , \quad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M} Y_i\ (w)^{\ k} \longrightarrow \mu_w^k,$$

with μ_w^k *finite for* $k \leq 2$.

3.2 The Sampling Process and the ssignment Mechanism

We do not observe $Y_i\ (0)$ and $Y_i\ (1)$ for all units in the population, and so we cannot directly infer the value of τ . In this section we describe precisely the two components of the stochastic nature of the sample. There is a stochastic binary treatment for each unit in each population, $W_i\ \in \{0, 1\}$. The realized outcome for unit i in population n is $Y_i\ = Y_i\ (W_i\ )$, with $\varepsilon_i\ = \varepsilon_i\ (W_i\ )$ the realized residual. We observe for a subset of the M units in the n-th population the triple $(Y_i\ , W_i\ , C_i\ )$, with stochastic sampling indicator $R_i\ \in \{0, 1\}$ describing whether $(Y_i\ , W_i\ , C_i\ )$ is observed $(R_i\ = 1)$, or not $(R_i\ = 0)$. The number of sampled units is $N\ = \sum_{i=1}^{M} R_i$.

Table 2 illustrates the set up. We observe $Y_i\ (0)$ or $Y_i\ (1)$ for a subset of units in the population: we observe $Y_i\ (0)$ if $R_i\ = 1$ and $W_i\ = 0$, we observe $Y_i\ (1)$ if $R_i\ = 1$ and $W_i\ = 1$, and we observe neither $Y_i\ (0)$ nor $Y_i\ (1)$ if $R_i\ = 0$, irrespective of the value of W_i . Uncertainty reflects the fact that our sample could have been different. In the table under the columns "Alternative Sample I" a different sample is given. This sample differs from the actual sample in two ways: different units are sampled, and different units are assigned to the treatment. Given an estimand, *e.g.,* the average effect of the treatment τ , standard errors are intended to capture both sources of variation.

Table 2: Random Sampling, Random ssignment (✓ is observed, ? is missing)

Unit	Actual Sample					Alternative Sample I					...
	R_i	Y_i (0)	Y_i (1)	W_i	C_i	R_i	Y_i (0)	Y_i (1)	W_i	C_i	...
1	1	✓	?	0	1	0	?	?	?	1	...
2	0	?	?	?	1	1	✓	?	0	1	...
3	0	?	?	?	1	0	?	?	?	1	...
⋮	⋮	⋮	⋮	⋮	1	⋮	⋮	⋮	⋮	1	...
M_1	1	?	✓	1	1	0	?	?	?	1	...
M_1 +1	1	✓	?	0	2	0	?	?	?	2	...
M_1 +2	0	?	?	?	2	0	?	?	?	2	...
M_1 +3	0	?	?	?	2	0	?	?	?	2	...
⋮	⋮	⋮	⋮	⋮	2	⋮	⋮	⋮	⋮	2	...
$M_1 + M_2$	1	?	✓	1	2	0	?	?	?	2	...
$M_1 + M_2$ +0	0	?	?	?	3	0	?	?	?	3	...
$M_1 + M_2$ +2	0	?	?	?	3	1	✓	?	0	3	...
$M_1 + M_2$ +3	0	?	?	?	3	1	?	✓	0	3	...
⋮	0	⋮	⋮	⋮	3	⋮	⋮	⋮	⋮	3	...
$M_1 + M_2 + M_3$	0	?	?	?	3	0	?	?	?	3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
$\sum_{c=1}^{C-1} M_c$ +1	1	?	✓	1	C	1	✓	?	0	C	...
$\sum_{c=1}^{C-1}$ +2	0	?	?	?	C	1	✓	?	0	C	...
$\sum_{c=1}^{C-1}$ +3	0	?	?	?	C	1	?	✓	1	C	...
⋮	⋮	⋮	⋮	⋮	C	⋮	⋮	⋮	⋮	C	...
M	1	?	✓	1	C	0	?	?	?	C	...

The sampling process that determines the values of R_i is independent of the potential outcomes and the assignment. It consists of two stages. First clusters are sampled with cluster sampling probability P_C . Second, we randomly sample units from the subpopulation consisting of all the sampled clusters, with unit sampling probability P_U . Both P_C and P_U may be equal to 1, or close to zero. If $P_C = 1$, we have completeley random sampling. If $P_U = 1$, we sample all units from the sampled clusters. If both $P_C = P_U = 1$, all units in the population are sampled. P_C close to zero is the case that is covered by the Z standard errors: we only observe units from a few clusters randomly drawn from a population consisting of a large

number of clusters.

The assignment process that determines the values of W_i for all i and n, also consists of two stages. In the first stage, for cluster c in population n, an assignment probability $q_c \in [0,1]$ is drawn randomly from a distribution $f(\cdot)$ with mean μ and variance σ^2. To simplify the algebra, we focus here on the case with $\mu = 1/2$. The variance $\sigma^2 \geq 0$ is key. If it is zero, we have random assignment. For positive values of σ^2 we have correlated assignment within the clusters, and if $\sigma^2 = 1/4$ then $q_c \in \{0,1\}$, all units with a cluster have the same assignments. In the second stage, each unit in cluster c is assigned to the treatment independently, with cluster-specific probability q_c .

The parameters σ^2, P_C and P_U are indexed by the population n to stress that they can be population specific. The sequences are assumed to converge to some limits, which may include zero and one for p_C and p_U to capture random sampling from a large population.

Formally we can summarize the conditions on the sampling and assignment processes as follows.

Assumption 2. *The vector of assigments* $\mathbf{W}$ *is independent of the vector of sampling indicators* $\mathbf{R}$.

Assumption 3. (Sampling)

$$\text{pr}(R_i = 1) = P_C P_U ,$$

$$\text{pr}(R_i = 1 \, R_j = 1, C_i \neq C_j) = P_C P_U ,$$

$$\text{pr}(R_i = 1 \, R_j = 1, C_i = C_j) = P_U .$$

Assumption 4. (ssignment)

$$\text{pr}(W_i = 1) = \mu = 1/2.$$

$$\text{pr}(W_i = 1 \, W_j = 1, C_i \neq C_j) = \mu = 1/2.$$

$$\text{pr}(W_i = 1 \, W_j = 1, C_i = C_j) = \mu + \sigma^2/\mu = 1/2 + 2\sigma^2.$$

Assumption 5. (Population Sequences) *The sequences* σ^2, P_C *and* P_U *satisfy*

$$\sigma^2 \in [0, 1/4], \text{ and } \sigma^2 \to \sigma^2 \in [0, 1/4],$$

$$P_C > 0, \text{ and } P_C \to P_C \in [0,1],$$

$$P_U > 0, \text{ and } P_U \to P_U \in [0,1],$$

$$(nP_C P_U)^{-1} \to 0.$$

Table 3: First Two Moments and Within-Cluster Covariances for Selected Random Variables

Variable	Expected Value	Variance	Within Cluster Covariance
R_i	$P_C\ P_U$	$P_C\ P_U\ (1 - P_C\ P_U\ )$	$P_C\ (1 - P_C\ )P_U^2$
W_i	$1/2$	$1/4$	σ^2
$R_i\ W_i$	$P_C\ P_U\ /2$	$P_C\ P_U\ (2 - P_C\ P_U\ )/4$	$P_C\ P_U^2\ (1 - P_C\ )/4 + \sigma^2 P_C\ P_U^2$

3.3 First and Second Moments of the ssignment and Sampling Indicators

We are interested in the distribution of the least squares estimator for τ and in particular in its approximate mean and variance. The estimator is stochastic through its dependence on two stochastic components, the sampling indicators $\mathbf{R}$ and the assignment indicators $\mathbf{W}$. The approximate mean and variance depend on the first and second (cross) moments of R_i and W_i . The first two moments, and the within-cluster covariance of R_i , W_i , and the product $R_i\ W_i$ are presented for reference in Table 3. Note that the covariance between any of these variables not in the same cluster is zero.

The within-cluster covariance of R_i is zero if $P_C\ = 0$ or $P_C\ = 1$, that is, if either all clusters are sampled or a vanishing number is sampled. The within-cluster covariance of W_i is zero if the assignment probability is constant across clusters ($\sigma^2 = 0$).

3.4 The Estimator

We are interested in the least squares estimator for τ in the regression

$$Y_i\ = \alpha + \tau W_i\ + \varepsilon_i\ .$$

Define the averages

$$\overline{R}\ = \frac{1}{M}\sum_{i=1}^{M} R_i\ , \qquad \overline{W}\ = \frac{1}{N}\sum_{i=1}^{M} R_i\ W_i\ ,$$

$$\overline{Y}\ = \frac{1}{N}\sum_{i=1}^{M} R_i\ Y_i\ .$$

Note that except for $\overline{R}$ these averages are defined over the units in the sample, not the units in the population. Now we can write the least squares estimator $\hat{\tau}$ as

$$\hat{\tau} = \frac{\sum_{i=1} R_i\ (W_i\ - \overline{W}\ )Y_i}{\sum_{i=1} R_i\ (W_i\ - \overline{W}\ )^2} = \overline{Y}\ _1 - \overline{Y}\ _0,$$

where

$$\overline{Y}\ _1 = \frac{1}{N\ _1}\sum_{i=1}^{M} R_i\ W_i\ Y_i\ , \qquad N\ _1 = \sum_{i=1}^{M} R_i\ W_i\ ,$$

$$\overline{Y}_0 = \frac{1}{N_0}\sum_{i=1}^{M} R_i\ (1-W_i\ )Y_i\ , \qquad N_0 = \sum_{i=1}^{M} R_i\ (1-W_i\ ).$$

We are interested in the variance of $\hat{\tau}$, and how it compares to the two standard variance estimators, the Eicker-Huber-White (EHW) variance estimator given in (2.2) and the iang-Zeger (Z) variance estimator given in (2.3).

The first step is to approximate the estimator by a sample average. This is where the large sample approximation is important.

emma 1. *Suppose ssumptions 1-5 hold. Then:*

$$\sqrt{N}\,(\hat{\tau}\ -\tau\ ) - \frac{2}{\sqrt{M\ P_C\ P_U}}\sum_{i=1}^{M} R_i\ (2W_i\ -1)\varepsilon_i\ = o_p(1).$$

emma 1 implies we can focus on properties of the η , the linear approximation to $\sqrt{N}\,(\hat{\tau}\ -\tau\ )$, defined as:

$$\eta\ = \frac{2}{\sqrt{nP_C\ P_U}}\sum_{i=1}^{M}\eta_i\ , \qquad \text{where}\ \ \eta_i\ = R_i\ (2W_i\ -1)\varepsilon_i\ .$$

We can calculate the exact (finite sample) variance of η for various values of the parameters and the corresponding normalized EHW and Z variance estimators, in order to analyze the implications of the two types of clustering and the importance (or not) of adjusting the standard errors for clustering.

Proposition 1. *Suppose ssumptions 1-5 hold. Then* (i)*, the exact variance of* η *is*

$$\mathbb{V}(\eta\ ) = \frac{1}{M}\sum_{i=1}^{M}\Big\{2\left(\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2\right) - P_U\ (\varepsilon_i\ (1)-\varepsilon_i\ (0))^2 + 4P_U\ \sigma^2(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2\Big\}$$

$$+\frac{P_U}{M}\sum_{c=1}^{C} M_c^2\ \Big\{(1-P_C\ )(\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0))^2 + 4\sigma^2(\overline{\varepsilon}_c\ (1)+\overline{\varepsilon}_c\ (0))^2\Big\},$$

(ii) *the difference between the limit of the normalized* Z *variance estimator and the correct variance is*

$$\mathbb{V}_{\ \mathrm{Z}} - \mathbb{V}(\eta\ ) = \frac{P_C\ P_U}{M}\sum_{c=1}^{C} M_c^2\ (\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0))^2 \geq 0, \tag{3.1}$$

and (iii)*, the difference between the limit of the normalized* Z *and* EHW *variance estimators is*

$$\mathbb{V}_{\ \mathrm{Z}} - \mathbb{V}_{\mathrm{EHW}} = -\frac{2P_U}{M}\sum_{i=1}^{M}\left\{(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2 + 4\sigma^2(\varepsilon_i\ (1)+\varepsilon_i\ (0))^2\right\}$$

$$+\frac{P_U}{M}\sum_{c=1}^{C} M_c^2\ \Big\{(\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0))^2 + 4\sigma^2\left(\overline{\varepsilon}_c\ (1)+\overline{\varepsilon}_c\ (0)\right)^2\Big\}.$$

This result follows from emma 1 and Appendix emmas A.1-A.3. Part (i) gives the exact variance for the linear approximation of $\sqrt{N}(\hat{\tau} - \tau)$, which is the correct variance of interest. The first sum in $\mathbb{V}(\eta)$ is approximately the EHW variance. If the sample is small relative to the population, so that P_U is close to zero, this first term simplifies to $\mathbb{V}_{\text{EHW}} = \sum_{i=1}^{N}(\varepsilon_i(1)^2 + \varepsilon_i(0)^2)/M$. The second sum in $\mathbb{V}(\eta)$ captures the effects of clustered sampling and assignment on the variance. There are two components to that sum. The first set of terms has a factor $1 - P_C$. The presence of this $1 - P_C$ factor captures the fact that these terms disappear if we have a random (non-clustered) sample (in which case $P_C = 1$). The second set of terms has a factor σ^2, which implies they vanish if there is no clustering in the assignment.

Part (ii) of the proposition compares the Z variance to the correct variance. It highlights the fact that the Z variance estimator captures correctly the component of the clustering due to clustered assignment (the component that depends on σ^2). However, the Z variance does not capture component due to clustered sampling correctly unless P_C is close to zero: implicitly the Z variance estimator relies on the assumption that the sampled clusters are a small proportion of the population of clusters of interest. This leads to the difference between the Z variance and the true variance being proportional to P_C.

Part (iii) of the proposition compares the Z variance to the EHW variance, highlighting the conditions under which using the Z variance makes a difference relative to using the EHW variance. Note that this is different from the question whether one *should* cluster, which is captured by part (*ii*) of the proposition. The first sum in the difference $\mathbb{V}_{\text{Z}} - \mathbb{V}_{\text{EHW}}$ is small relative to the second term when there is a substantial number of units per cluster relative to the number of clusters. For example, if the number of units per cluster $M_c = M/C$ is constant across clusters and large relative to the number of clusters, then the second sum is proportional to M/C^2, and large relative to the first sum. In that case, the clustering matters if there is heterogeneity in the treatment effects ($\overline{\varepsilon}_c(1) - \overline{\varepsilon}_c(0)$ differs from zero) or there is clustering in the assignment. Note that the difference does not depend on whether the sampling is clustered: this follows directly from the fact that one cannot tell from the data whether or not the sampling was clustered.

The following corollary describes two special cases under which clustering is not necessary.

Corollary 1. *Standard errors need to account for clustering unless one of the following two pairs of conditions hold: (i) there is no clustering in the sampling ($P_C = 1$ for all n) and there is no clustering in the assignment ($\sigma^2 = 0$); or (ii) there is no heterogeneity in the treatment effects ($Y_i(1) - Y_i(0) = \tau$ for all units) and there is no clustering in the assignment ($\sigma^2 = 0$).*

Our next result highlights three special cases where the Z clustering is correct.

Corollary 2. *The Z variance is approximately correct if one of three conditions hold: (i) there is no heterogeneity in the treatment effects, $Y_i(1) - Y_i(0) = \tau$ for all units; (ii) P_C is close to zero for all n, so that we observe only few clusters in the population of clusters; (iii) P_U is close to zero so that there is at most one sampled unit per cluster (in which case clustering adjustments do not matter).*

The first of these three conditions (no heterogeneity in the treatment effects) is unlikely to hold in practice. The third condition is easily verifiable by assessing the distribution of the number of sampled units per cluster, or by comparing the standard errors with and without clustering adjustments. The second condition cannot be assessed using the actual data. To assess this condition one needs to consider the facts about the sampling process and investigate whether there are clusters in the population of interest that are not included in the sample.

If one were to conclude that all the clusters in the population are included in the sample, the Z variance is in general conservative. Then, there are two possibilities. If the assignment is perfectly correlated within the clusters, there is no general improvement over the Z variance available. However, if there is variation in the treatment within the clusters, one can estimate $\mathbb{V}_{\text{Z}} - \mathbb{V}(\eta)$ and subtract that from $\hat{\mathbb{V}}_{\text{Z}}$. Define

$$\hat{\tau}_c = \overline{Y}_{c1} - \overline{Y}_{c0}$$

to be the difference in average outcomes by treatment status in cluster c, an estimator for the average treatment effect within the cluster. Then our proposed cluster-adjusted variance estimator is

$$\hat{\mathbb{V}}_{\text{CA}}(\hat{\tau}) = \hat{\mathbb{V}}_{\text{Z}}(\hat{\tau}) - \frac{1}{N^2}\sum_{c=1}^{C} N_c^2(\hat{\tau}_c - \hat{\tau})^2.$$

4 The Fixed Effects Case

The importance of clustering adjustments to standard errors in settings where the regression includes fixed effects is also a source of confusion. Arellano [1987] shows clearly that even with fixed effects included in the regression, the clustering adjustment may matter. Here we extend the results from the previous section to the case with fixed effects. In the fixed effect case the assignment within clusters cannot be perfectly correlated, so we focus on the case with $\sigma^2 < 1/4$. We consider the regression of the outcome on the cluster dummies and the treatment indicator:

$$Y_i = \alpha_{C_i} + \tau W_i + \dot{\varepsilon}_i \ .$$

First we strengthen the assumptions on the sequence of populations. The main difference is that we require the number of units per cluster to go to infinity so that we can estimate the fixed effects consistently.

Assumption 6. *The sequence of populations satisfies* (i)

$$\lim_{\to\infty} C^{-1} = 0,$$

(ii) *for some finite* K,

$$\frac{\max_c M_c}{\min_c M_c} \leq K,$$

and (iii)

$$\max_c M_c^{-1} \to 0.$$

Assumption 7. *The sequences* σ^2, P_C *and* P_U *satisfy*

$$\sigma^2 \in [0, 1/4), \text{ and } \sigma^2 \to \sigma^2 \in [0, 1/4),$$
$$P_C > 0, \text{ and } P_C \to P_C \in [0, 1],$$
$$P_U > 0, \text{ and } P_U \to P_U \in [0, 1],$$
$$\min_c (M_c \; P_C \; P_U \;)^{-1} \to 0.$$

Define the cluster specific treatment rate:

$$q_c = \mathbb{E}[W_i \mid C_i = c].$$

Also define

$$\kappa = \mathbb{V}(q_{C_i} \; (1 - q_{C_i} \;)), \qquad \text{and} \quad \kappa_{j,k} = \mathbb{E}[q_{C_i}^j \; (1 - q_{C_i} \;)^k],$$

Note that

$$\mathbb{E}[q_{C_i} \; (1 - q_{C_i} \;)] = \frac{1 - 4\sigma^2}{4},$$

and note that κ can only be positive if $\sigma^2 > 0$.

Define the adjusted residual as

$$\dot{\varepsilon}_i = \varepsilon_i - q_c \; \overline{\varepsilon}_{C_i} \; (1) - (1 - q_c \;)\overline{\varepsilon}_{C_i} \; (0).$$

emma 2. *Suppose ssumptions 1-7 hold. Then:*

$$\sqrt{N} \; (\hat{\tau}^{\text{fe}} - \tau \;) - \frac{4}{(1 - 4\sigma^2)\sqrt{M \; P_C \; P_U}} \sum_{i=1}^{M} R_i \; (W_i - q_{C_i} \;)\dot{\varepsilon}_i = o_p(1).$$

Analogous to our analysis of the case without fixed effects, we can now focus on the properties of the linear approximation to $\sqrt{N} \; (\hat{\tau}^{\text{fe}} - \tau \;)$, where

$$\eta^{\text{fe}} = \frac{4}{(1 - 4\sigma^2)\sqrt{M \; P_C \; P_U}} \sum_{i=1}^{M} R_i \; (W_i - q_{C_i} \;)\dot{\varepsilon}_i \; .$$

Proposition 2. *Suppose ssumptions 1-7 hold. Then* (i)*, the exact variance of* η^{fe} *is*

$$\mathbb{V}\left(\eta^{\text{fe}}\right) = \frac{1}{M} \sum_{i=1}^{M} \Bigg\{ (1 - P_U \;) \left(1 + \kappa \frac{16}{(1 - 4\sigma^2)^2}\right) (\varepsilon_i \; (1) - \varepsilon_i \; (0))^2$$
$$+ \frac{16\kappa_{3,1}}{(1 - 4\sigma^2)^2} (\varepsilon_i \; (1) - \overline{\varepsilon}_{C_i} \; (1))^2 + \frac{16\kappa_{1,3}}{(1 - 4\sigma^2)^2} (\varepsilon_i \; (0) - \overline{\varepsilon}_{C_i} \; (0))^2 \Bigg\}$$
$$+ \frac{P_U}{M} \sum_{c=1}^{C} M_c^2 \; \left\{ (1 - P_C \;) + \frac{16\kappa}{(1 - 4\sigma^2)^2} \right\} (\overline{\varepsilon}_c \; (1) - \overline{\varepsilon}_c \; (0))^2,$$

and (ii) *the difference between the limit of the normalized* Z *variance estimator and the correct variance is*

$$\mathbb{V} \; \text{Z} - \mathbb{V}(\eta^{\text{fe}}) = \frac{P_C \; P_U}{M} \sum_{c=1}^{C} M_c^2 \; (\overline{\varepsilon}_c \; (1) - \overline{\varepsilon}_c \; (0))^2. \tag{4.1}$$

Compared to the case without fixed effects given in (3.1), there is no difference in the relation between the Z variance estimator and the true variance, given in (4.1).

Compared to the case without fixed effects, however, there is a difference in when one should adjust the standard errors for clustering. Without fixed effects, one should cluster if either (i) both $P_C < 1$ (clustering in the sampling) and there is heterogeneity in the treatment effects, or (ii) $\sigma^2 > 0$ (clustering in the assignment). With fixed effects, one should cluster if either (i) both $P_C < 1$ (clustering in the sampling) and there is heterogeneity in the treatment effects, or (ii) $\sigma^2 > 0$ (clustering in the assignment) and there is heterogeneity in the treatment effects. In other words, heterogeneity in the treatment effects is now a requirement for clustering adjustments to be necessary, and beyond that, either clustering in sampling or assignment makes the adjustments important.

5 Conclusion

We develop a new perspective on clustering adjustments to standard errors. We argue that there are two potential motivations for such adjustments, one based on clustered sampling, and one based on clustered assignment. Although when researchers look for formal justification for clustering, they typically rely on clustered sampling justifications, we argue that clustered assignment is more commonly the setting of interest. This leads to new conclusions about when to adjust standard errors for clustering, and at what level to do the adjustment.

The practical implications from the results in this paper are as follows. The researcher should assess whether the sampling process is clustered or not, and whether the assignment mechanism is clustered. If the answer to both is no, one should *not* adjust the standard errors for clustering, irrespective of whether such an adjustment would change the standard errors. We show that the standard iang-Zeger cluster adjustment is conservative, and further, we derive an estimator for the correct variance that can be used if there is variation in treatment assignment within clusters and the fraction of clusters that is observed is known. This analysis extends to the case where fixed effects are included in the regression at the level of a cluster, with the provision that if there is no heterogeneity in the treatment effects, one need not adjust standard errors for clustering once fixed effects are included.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based vs. design-based uncertainty in regression analysis. arXiv preprint arXiv:1706.01778, 2017.

Joshua Angrist and Steve Pischke. Mostly Harmless Econometrics: An Empiricists' Companion. Princeton University Press, 2008.

Manuel Arellano. Computing robust standard errors for within group estimators. Oxford bulletin of Economics and Statistics, 49(4):431–434, 1987.

Susan Athey and Guido W Imbens. The econometrics of randomized experiments. Handbook of Economic Field Experiments, 1:73–140, 2017.

Thomas Barrios, Rebecca Diamond, Guido W Imbens, and Michal Kolesár. Clustering, spatial correlations, and randomization inference. Journal of the American Statistical Association, 107(498):578–591, 2012.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1):249–275, 2004.

C Bester, T Conley, and C Hansen. Inference with dependent data using clustering covariance matrix estimators. Unpublished Manuscript, University of Chicago Business School, 2009.

A Colin Cameron and Douglas Miller. A practitioners guide to cluster-robust inference. Journal of Human Resources, 50(2):317–372, 2015.

Timothy G Conley. Gmm estimation with cross sectional dependence. Journal of econometrics, 92(1):1–45, 1999.

Noel Cressie. Statistics for spatial data. John Wiley & Sons, 2015.

P. Diggle, P. Heagerty, K.Y. iang, and S. Zeger. Analysis of ongitudinal Data. Oxford Statistical Science Series. OUP Oxford, 2013. ISBN 9780191664335. UR `https://books.google.com/books?id=z iK-gWUqDUC`.

Stephen G Donald and Kevin ang. Inference with difference-in-differences and other panel data. The review of Economics and Statistics, 89(2):221–233, 2007.

Allan Donner and Neil Klar. Design and analysis of cluster randomization trials in health research. 2000.

Friedhelm Eicker. imit theorems for regressions with unequal and dependent errors. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 59–82, 1967.

Ronald Aylmer Fisher. The design of experiments. Oliver And Boyd; Edinburgh; ondon, 1937.

Christian B Hansen. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. Journal of Econometrics, 140(2):670–694, 2007.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 221–233, 1967.

Rustam Ibragimov and Ulrich K Müller. t-statistic based correlation and heterogeneity robust inference. Journal of Business & Economic Statistics, 28(4):453–468, 2010.

Rustam Ibragimov and Ulrich K Müller. Inference with few heterogeneous clusters. Review of Economics and Statistics, 98(1):83–96, 2016.

Guido W Imbens and Donald B Rubin. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, 2015.

eslie Kish. Survey sampling. 1965.

Teunis Kloek. Ols estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. Econometrica: Journal of the Econometric Society, pages 205–207, 1981.

Katherine J ee and Simon G Thompson. The use of random effects models to allow for clustering in individually randomized trials. Clinical Trials, 2(2):163–173, 2005.

Kung-Yee iang and Scott Zeger. ongitudinal data analysis using generalized linear models. Biometrika, 73(1):13–22, 1986.

Brent R Moulton. Random group effects and the precision of regression estimates. Journal of econometrics, 32(3):385–397, 1986.

Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. The review of Economics and Statistics, pages 334–338, 1990.

Brent R Moulton and William C Randolph. Alternative tests of the error components model. Econometrica: Journal of the Econometric Society, pages 685–693, 1989.

David M Murray. Design and analysis of group-randomized trials, volume 29. Monographs in Epidemiology & B, 1998.

Chris Roberts and Stephen A Roberts. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials, 2(2):152–162, 2005.

Paul R Rosenbaum. Observational studies. In Observational Studies. Springer, 2002.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association, 75(371):591–593, 1980.

James H Stock and Mark W Watson. Heteroskedasticity-robust standard errors for fixed effects panel data regression. Econometrica, 76(1):155–174, 2008.

Michael J Weiss, JR ockwood, and Daniel F McCaffrey. Estimating the standard error of the impact estimator in individually randomized trials with clustering. Journal of Research on Educational Effectiveness, 9(3):421–444, 2016.

Halbert White. Asymptotic theory for econometricians. Academic press, 2014.

S. William. Comparison of standard errors for robust, cluster, and standard estimators. 1998. UR `http://www.stata.com/support/faqs/stat/cluster.html`.

Jeffrey M Wooldridge. Cluster-sample methods in applied econometrics. The American Economic Review, 93(2):133–138, 2003.

Jeffrey M Wooldridge. Econometric analysis of cross section and panel data. MIT press, 2010.

PPENDIX

It is useful to work with a transformation of W_i :

$$T_i \ = 2W_i \ - 1 \qquad \text{so that } W_i \ = \frac{T_i \ + 1}{2}, \ \ 1 - W_i \ = \frac{1 - T_i}{2}.$$

Note that in terms of T_i we can write

$$Y_i \ = W_i \ Y_i(1) + (1 - W_i \ )Y_i(0) = T_i \ \frac{Y_i \ (1) - Y_i \ (0)}{2} + \frac{Y_i \ (1) + Y_i \ (0)}{2},$$

$$\varepsilon_i \ = T_i \ \frac{\varepsilon_i \ (1) - \varepsilon_i \ (0)}{2} + \frac{\varepsilon_i \ (1) + \varepsilon_i \ (0)}{2},$$

and

$$T_i \ \varepsilon_i \ = \frac{\varepsilon_i \ (1) - \varepsilon_i \ (0)}{2} + T_i \ \frac{\varepsilon_i \ (1) + \varepsilon_i \ (0)}{2}.$$

Proof of emma 1: First,

$$\begin{aligned}
&\sqrt{N} \ (\hat{\tau} \ - \tau \ ) - \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ (2W_i \ - 1)\varepsilon_i \\
&\qquad = \sqrt{N} \ (\hat{\tau} \ - \tau \ ) - \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \\
&\qquad = \sqrt{N} \ \left(2 \frac{\frac{1}{n} \sum_{i=1} R_i \ Y_i \ (T_i \ - \overline{T} \ )}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2} - \frac{1}{n} \sum_{i=1} (Y_i \ (1) - Y_i \ (0)) \right) - \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \ .
\end{aligned} \tag{A.1}$$

Substituting

$$\begin{aligned}
Y_i \ &= T_i \ \frac{Y_i \ (1) - Y_i \ (0)}{2} + \frac{Y_i \ (1) + Y_i \ (0)}{2} \\
&= T_i \ \frac{\varepsilon_i \ (1) - \varepsilon_i \ (0)}{2} + T_i \ \frac{\overline{Y} \ (1) - \overline{Y} \ (0)}{2} + \frac{\varepsilon_i \ (1) + \varepsilon_i \ (0)}{2} + \frac{\overline{Y} \ (1) + \overline{Y} \ (0)}{2} \\
&= \varepsilon_i \ + T_i \ \frac{\tau}{2} + \frac{\overline{Y} \ (1) + \overline{Y} \ (0)}{2},
\end{aligned}$$

into (A.1) leads to

$$\begin{aligned}
&\sqrt{N} \ \left(2 \frac{\frac{1}{n} \sum_{i=1} R_i \ \varepsilon_i \ (T_i \ - \overline{T} \ )}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2} + \tau \ \frac{\frac{1}{n} \sum_{i=1} R_i \ T_i \ (T_i \ - \overline{T} \ )}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2} - \tau \right. \\
&\qquad \left. + (\overline{Y} \ (1) + \overline{Y} \ (0)) \frac{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2} \right) - \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \\
&= \sqrt{N} \ \frac{2 \frac{1}{n} \sum_{i=1} R_i \ T_i \ \varepsilon_i}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2} - \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \\
&\qquad - \sqrt{N} \ 2\overline{T} \ \frac{\frac{1}{n} \sum_{i=1} R_i \ \varepsilon_i}{\frac{1}{n} \sum_{i=1} R_i \ (T_i \ - \overline{T} \ )^2}
\end{aligned}$$

$$+\sqrt{N}\left(\tau\ \frac{\frac{1}{}\sum_{i=1}R_i}{\frac{1}{}\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}-\tau\ \right)$$
$$-\sqrt{N}\,\overline{T}\ \frac{\frac{1}{}\sum_{i=1}R_i\ T_i}{\frac{1}{}\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}.$$

To prove that this is $o_p(1)$, it is sufficient to prove the following four claims,
(i)

$$\sqrt{N}\,\frac{\sum_{i=1}R_i\ T_i\ \varepsilon_i}{\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}-\frac{2}{\sqrt{nP_C\ P_U}}\sum_{i=1}R_i\ T_i\ \varepsilon_i\ =o_p(1), \tag{A.2}$$

(ii)

$$\sqrt{N}\,\overline{T}\ \frac{\frac{1}{}\sum_{i=1}R_i\ \varepsilon_i}{\frac{1}{}\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}=o_p(1), \tag{A.3}$$

(iii)

$$\sqrt{N}\left(\tau\ \frac{\sum_{i=1}^{N}R_i}{\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}-\tau\ \right)=o_p(1), \tag{A.4}$$

(iv)

$$\sqrt{N}\,\overline{T}\ \frac{\frac{1}{}\sum_{i=1}R_i\ T_i}{\frac{1}{}\sum_{i=1}R_i\ (T_i\ -\overline{T}\ )^2}=o_p(1). \tag{A.5}$$

First a couple of preliminary observations. By the assumptions it follows that

$$\frac{1}{n}\sum_{i=1}(R_i\ -P_C\ P_U\ )\xrightarrow{p}0 \tag{A.6}$$

and so that

$$\frac{N}{nP_C\ P_U}\xrightarrow{p}1. \tag{A.7}$$

In addition,

$$\sqrt{N}\,\overline{T}\ =O_p(1), \tag{A.8}$$

and

$$\sqrt{N}\,\frac{1}{n}\sum_{i=1}R_i\ \varepsilon_i\ =O_p(1). \tag{A.9}$$

emma A.1. *Suppose ssumptions 1 and 5 hold. Then* (i)

$$N\ \hat{\mathbb{V}}_{\text{EHW}}\to\ \ \mathbb{V}_{\text{EHW}}=\frac{4}{nP_C\ P_U}\sum_{i=1}=\frac{2}{n}\sum_{i=1}\left\{\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2\right\},$$

and

$$N\ \hat{\mathbb{V}}_{\ \text{Z}}\to\ \ \mathbb{V}_{\ \text{Z}}$$

$$= \frac{2}{n}\sum_{i=1}\left\{\varepsilon_i\ (1)^2\left(1-P_U\ (1+4\sigma^2)\right)+\varepsilon_i\ (0)^2\left(1-P_U\ (1+4\sigma^2)\right)+\varepsilon_i\ (0)\varepsilon_i\ (1)P_U\ (2-8\sigma^2)\right\}$$

$$+\frac{P_U}{n}\sum_{c=1}^{C} n_c^2\ \left\{\left(\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0)\right)^2+4\sigma^2\left(\overline{\varepsilon}_c\ (1)+\overline{\varepsilon}_c\ (0)\right)^2\right\}.$$

Proof of emma A.1: First (i):

$$\mathbb{V}_{\text{EHW}} = \frac{4}{nP_C\ P_U}\sum_{i=1}\mathbb{E}[\eta_i^2\ ].$$

Because

$$\mathbb{E}[\eta_i^2\ ] = \mathbb{E}[R_i^2\ T_i^2\ \varepsilon_i^2\ ] = \mathbb{E}[R_i\ \varepsilon_i^2\ ] = P_C\ P_U\ \mathbb{E}[\varepsilon_i^2\ ]$$

$$= P_C\ P_U\ \left\{T_i\ \frac{\varepsilon_i\ (1)-\varepsilon_i\ (0)}{2}+\frac{\varepsilon_i\ (1)+\varepsilon_i\ (0)}{2}\right\}^2$$

$$= \frac{1}{4}P_C\ P_U\ \left\{\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2-2\varepsilon_i\ (1)\varepsilon_i\ (0)+\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2+2\varepsilon_i\ (1)\varepsilon_i\ (0)\right\}$$

$$= \frac{1}{2}P_C\ P_U\ \left\{\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2\right\}$$

it follows that

$$\mathbb{V}_{\text{EHW}} = \frac{2}{n}\sum_{i=1}\left\{\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2\right\},$$

finishing the proof for part (i).
Next, consider (ii). The normalized Z variance estimator is

$$\mathbb{V}_{\text{Z}} = \frac{4}{nP_C\ P_U}\sum_{i=1}\sum_{j=1}\mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \right]$$

$$= \frac{4}{nP_C\ P_U}\sum_{c=1}^{C}\sum_{i=1}\sum_{j=1}C_i\ C_j\ \mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ =C_j\ \right].$$

Consider the expectations:

$$\mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ =C_j\ \right]$$

$$= \mathbb{E}\left[R_i\ T_i\ \left\{T_i\ \frac{\varepsilon_i\ (1)-\varepsilon_i\ (0)}{2}+\frac{\varepsilon_i\ (1)+\varepsilon_i\ (0)}{2}\right\}R_j\ T_j\ \left\{T_j\ \frac{\varepsilon_j\ (1)-\varepsilon_j\ (0)}{2}+\frac{\varepsilon_j\ (1)+\varepsilon_j\ (0)}{2}\right\}\middle| C_i\ =C_j\ \right]$$

$$= \mathbb{E}\left[R_i\ \left\{\frac{\varepsilon_i\ (1)-\varepsilon_i\ (0)}{2}+T_i\ \frac{\varepsilon_i\ (1)+\varepsilon_i\ (0)}{2}\right\}R_j\ \left\{\frac{\varepsilon_j\ (1)-\varepsilon_j\ (0)}{2}+T_j\ \frac{\varepsilon_j\ (1)+\varepsilon_j\ (0)}{2}\right\}\middle| C_i\ =C_j\ \right].$$

If $i=j$, the expectation is, per the earlier calculation for $\mathbb{V}_{\text{EHW}}$, equal to

$$\mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ =C_j\ ,i=j\right] = \frac{1}{2}P_C\ P_U\ \left\{\varepsilon_i\ (1)^2+\varepsilon_i\ (0)^2\right\}.$$

If the $i\neq j$, the expectation is

$$\mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ =C_j\ ,i\neq j\right]$$

$$
\begin{aligned}
&= \frac{1}{4}\left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) - \varepsilon_j\ (0)\right)\mathbb{E}[R_i\ R_j\ \ C_i\ = C_j\ , i \neq j]\right. \\
&\qquad + \left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) + \varepsilon_j\ (0)\right)\mathbb{E}[R_i\ R_j\ T_j\ \ C_i\ = C_j\ , i \neq j] \\
&\qquad + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) - \varepsilon_j\ (0)\right)\mathbb{E}[R_i\ R_j\ T_i\ \ C_i\ = C_j\ , i \neq j] \\
&\qquad \left. + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) + \varepsilon_j\ (0)\right)\mathbb{E}[R_i\ R_j\ T_i\ T_j\ \ C_i\ = C_j\ , i \neq j]\right\} \\
&= \frac{1}{4}\left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) - \varepsilon_j\ (0)\right)P_C\ P_U^2\right. \\
&\qquad \left. + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) + \varepsilon_j\ (0)\right)4P_C\ P_U^2\ \sigma^2\right\}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\sum_{i=1}\sum_{j=1} C_i\ {}_cC_j\ {}_c\mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ = C_j\ , i \neq j\right] \\
&= \frac{P_C\ P_U^2}{4}\sum_{i=1}\sum_{j=1} C_i\ {}_cC_j\ {}_c\left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) - \varepsilon_j\ (0)\right) + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)\left(\varepsilon_j\ (1) + \varepsilon_j\ (0)\right)4\sigma^2\right\} \\
&\qquad - \frac{P_C\ P_U^2}{4}\sum_{i=1} C_i\ \left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)^2 + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)^2 4\sigma^2\right\} \\
&= \frac{P_C\ P_U^2}{4}\sum_{i=1}\sum_{j=1} n_c^2\ \left\{\left(\overline{\varepsilon}_c\ (1) - \overline{\varepsilon}_c\ (0)\right)^2 + \left(\overline{\varepsilon}_c\ (1) + \overline{\varepsilon}_c\ (0)\right)^2 4\sigma^2\right\} \\
&\qquad - \frac{P_C\ P_U^2}{4}\sum_{i=1} C_i\ \left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)^2 + \left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)^2 4\sigma^2\right\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\mathbb{V}\ \mathrm{z} &= \frac{4}{nP_C\ P_U}\sum_{c=1}^{C}\sum_{i=1}\sum_{j=1} C_i\ C_j\ \mathbb{E}\left[R_i\ T_i\ \varepsilon_i\ R_j\ T_j\ \varepsilon_j\ \ C_i\ = C_j\ \right] \\
&= \frac{2}{n}\sum_{i=1}\left\{\varepsilon_i\ (1)^2 + \varepsilon_i\ (0)^2\right\} \\
&\qquad - \frac{P_U}{n}\sum_{i=1}\left\{\left(\varepsilon_i\ (1) - \varepsilon_i\ (0)\right)^2 + 4\sigma^2\left(\varepsilon_i\ (1) + \varepsilon_i\ (0)\right)^2\right\} \\
&\qquad + \frac{P_U}{n}\sum_{c=1}^{C} n_c^2\ \left\{\left(\overline{\varepsilon}_c\ (1) - \overline{\varepsilon}_c\ (0)\right)^2 + 4\sigma^2\left(\overline{\varepsilon}_c\ (1) + \overline{\varepsilon}_c\ (0)\right)^2\right\} \\
&= \frac{1}{n}\sum_{i=1}\left\{\varepsilon_i\ (1)^2\left(2 - P_U\ (1 + 4\sigma^2)\right) + \varepsilon_i\ (0)^2\left(2 - P_U\ (1 + 4\sigma^2)\right) + \varepsilon_i\ (0)\varepsilon_i\ (1)P_U\ (2 - 8\sigma^2)\right\} \\
&\qquad + \frac{P_U}{n}\sum_{c=1}^{C} n_c^2\ \left\{\left(\overline{\varepsilon}_c\ (1) - \overline{\varepsilon}_c\ (0)\right)^2 + 4\sigma^2\left(\overline{\varepsilon}_c\ (1) + \overline{\varepsilon}_c\ (0)\right)^2\right\}.
\end{aligned}
$$

Next we split η into two uncorrelated sums.

emma A.2.

$$\eta \ = \frac{2}{\sqrt{np_C p_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \ = S \ + D \ ,$$

where

$$S \ = \frac{1}{\sqrt{nP_C \ P_U}} \sum_{i=1} (R_i \ - P_C \ P_U \ )(\varepsilon_i \ (1) - \varepsilon_i \ (0)),$$

and

$$D \ = \frac{1}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ (\varepsilon_i \ (1) + \varepsilon_i \ (0)).$$

Proof of emma A.2: Substituting $\varepsilon_i \ = T_i \ (\varepsilon_i \ (1) - \varepsilon_i \ (0))/2 + (\varepsilon_i \ (1) + \varepsilon_i \ (0))/2$, we have

$$\frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \ = \frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} \left\{ R_i \ \frac{\varepsilon_i \ (1) - \varepsilon_i \ (0)}{2} + R_i T_i \ \frac{\varepsilon_i \ (1) + \varepsilon_i \ (0)}{2} \right\}.$$

Because $\sum_{i=1} \varepsilon_i \ (0) = \sum_{i=1} \varepsilon_i \ (1) = 0$, this is equal to

$$\frac{1}{\sqrt{nP_C \ P_U}} \sum_{i=1} (R_i \ - p_C p_U)(\varepsilon_i \ (1) - \varepsilon_i \ (0)) + \frac{1}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ (\varepsilon_i \ (1) + \varepsilon_i \ (0)) = S \ + D \ .$$

COMMENT: The S here refers to sampling, because S captures the sampling part of the clustering, and D refers to design, as D captures the design part of the clustering. For S only the clustering in the sampling (in R_i) matters, and the clustering in the assignment (in T_i) does not matter. For D it is the other way around. Even if R_i is clustered, if T_i is not, the covariance terms in the variance of D vanish.

emma A.3. *The first two moments of S and D are*

$$\mathbb{E}[S \ ] = 0, \qquad \mathbb{E}[D \ ] = 0,$$

$$\mathbb{E}[S^2] = \frac{1 - P_U}{n} \sum_{i=1} (\varepsilon_i \ (1) - \varepsilon_i \ (0))^2 + \frac{P_U \ (1 - P_C \ )}{n} \sum_{c=1}^{C} n_c^2 (\overline{\varepsilon}_c \ (1) - \overline{\varepsilon}_c \ (0))^2,$$

$$\mathbb{E}[D^2] = \frac{1 - 4\sigma^2 P_U}{n} \sum_{i=1} (\varepsilon_i \ (1) + \varepsilon_i \ (0))^2 + \frac{4\sigma^2 P_U}{n} \sum_{c=1}^{C} n_s^2 (\overline{\varepsilon}_c \ (1) + \overline{\varepsilon}_c \ (0))^2,$$

and

$$\mathbb{E}[S \ D \ ] = 0$$

so that

$$\mathbb{E}\left[\left(\frac{2}{\sqrt{nP_C \ P_U}} \sum_{i=1} R_i \ T_i \ \varepsilon_i \ \right)^2 \right]$$

$$= \frac{1}{n} \sum_{i=1} \left\{ (2 - P_U \ (1 + 4\sigma^2)) \varepsilon_i \ (1)^2 + (2 - P_U \ (1 + 4\sigma^2)) \varepsilon_i \ (0)^2 + P_U \ (2 - 8\sigma^2) \varepsilon_i \ (1) \varepsilon_i \ (0) \right\}$$

$$+ \frac{P_U}{n} \sum_{c=1}^{C} n_c^2 \left\{ (1 - P_C \ ) (\overline{\varepsilon}_c \ (1) - \overline{\varepsilon}_c \ (0))^2 + 4\sigma^2 (\overline{\varepsilon}_c \ (1) + \overline{\varepsilon}_c \ (0))^2 \right\}$$

Proof of emma A.3: Because $\mathbb{E}[R_i\ ] = P_C\ P_U\ $, it follows immediately that $\mathbb{E}[S\ ] = 0$. Because $\mathbb{E}[R_i\ T_i\ ] = 0$, it follows that $\mathbb{E}[D\ ] = 0$. Because $\mathbb{E}[(R_i\ - P_C\ P_U\ )R_i\ T_i\ ] = \mathbb{E}[(R_i\ - P_C\ P_U\ )R_i\ ]\mathbb{E}[T_i\ ] = 0$, it follows that $\mathbb{E}[S\ D\ ] = 0$. Next, consider $\mathbb{E}[S^2]$:

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{1}{nP_C\ P_U} \sum_{i=1}\sum_{j=1} \mathbb{E}\left[(R_i\ - P_C\ P_U\ )(\varepsilon_i\ (1) - \varepsilon_i\ (0))(R_j\ - P_C\ P_U\ )(\varepsilon_j\ (1) - \varepsilon_j\ (0))\right] \\
&= \frac{1}{nP_C\ P_U} \sum_{i=1} \left(P_C\ P_U\ (1 - P_C\ P_U\ ) - P_U^2\ P_C\ (1 - P_C\ )\right)(\varepsilon_i\ (1) - \varepsilon_i\ (0))^2 \\
&\quad + \frac{1}{nP_C\ P_U} \sum_{c=1}^{C}\sum_{i=1}\sum_{j=1} C_i\ C_j\ \left(P_U^2\ P_C\ (1 - P_C\ )\right)(\varepsilon_i\ (1) - \varepsilon_i\ (0))(\varepsilon_j\ (1) - \varepsilon_j\ (0)) \\
&= \frac{1 - P_U}{n} \sum_{i=1} (\varepsilon_i\ (1) - \varepsilon_i\ (0))^2 \\
&\quad + \frac{P_U\ (1 - P_C\ )}{n} \sum_{c=1}^{C}\sum_{i=1}\sum_{j=1} C_i\ C_j\ (\varepsilon_i\ (1) - \varepsilon_i\ (0))(\varepsilon_j\ (1) - \varepsilon_j\ (0)) \\
&= \frac{1 - P_U}{n} \sum_{i=1} (\varepsilon_i\ (1) - \varepsilon_i\ (0))^2 + \frac{P_U\ (1 - P_C\ )}{n} \sum_{c=1}^{C} n_c^2\ (\overline{\varepsilon}_c\ (1) - \overline{\varepsilon}_c\ (0))^2.
\end{aligned}$$

Next, consider $\mathbb{E}[D^2]$.

$$\begin{aligned}
\mathbb{E}[D^2] &= \frac{1}{nP_C\ P_U} \sum_{i=1}\sum_{j=1} \mathbb{E}\left[R_i\ T_i\ (\varepsilon_i\ (1) + \varepsilon_i\ (0))R_j\ T_j\ (\varepsilon_j\ (1) + \varepsilon_j\ (0))\right] \\
&= \frac{1}{nP_C\ P_U} \sum_{i=1} \left(P_C\ P_U\ - 4\sigma^2 P_C\ P_U^2\ \right)(\varepsilon_i\ (1) + \varepsilon_i\ (0))^2 \\
&\quad + \frac{1}{nP_C\ P_U} \sum_{c=1}^{C}\sum_{i=1}\sum_{j=1} C_i\ C_j\ 4\sigma^2 P_C\ P_U^2\ (\varepsilon_i\ (1) + \varepsilon_i\ (0))(\varepsilon_j\ (1) + \varepsilon_j\ (0)) \\
&= \frac{1 - 4\sigma^2 P_U}{n} \sum_{i=1} (\varepsilon_i\ (1) + \varepsilon_i\ (0))^2 + \frac{4\sigma^2 P_U}{n} \sum_{c=1}^{C} n_c^2\ (\overline{\varepsilon}_c\ (1) + \overline{\varepsilon}_c\ (0))^2.
\end{aligned}$$

emma A.4.

$$\eta^{\text{fe}} = \frac{2}{\sqrt{np_C p_U}} \sum_{i=1} R_i\ (T_i\ - q_{C_i}\ )\dot{\varepsilon}_i\ = S^{\text{fe}} + D^{\text{fe}},$$

where

$$S^{\text{fe}} = \frac{1}{\sqrt{nP_C\ P_U}} \sum_{i=1} (R_i\ - P_C\ P_U\ )(1 - q_{C_i}^2\ )(\dot{\varepsilon}_i\ (1) - \dot{\varepsilon}_i\ (0)),$$

and

$$D^{\text{fe}} = \frac{1}{\sqrt{nP_C\ P_U}} \sum_{i=1} R_i\ (T_i\ - q_{C_i}\ \left\{(\dot{\varepsilon}_i\ (1) + \dot{\varepsilon}_i\ (0)) - q_{C_i}\ (\dot{\varepsilon}_i\ (1) - \dot{\varepsilon}_i\ (0))\right\}.$$

The proof follows the same argument as the proof for emma A.2 and is omitted.
Proof of Proposition 2: By definition

$$\begin{aligned}\dot{\varepsilon}_i \;&= \varepsilon_i \;- q_c\;\overline{\varepsilon}_{C_i}\;(1) - (1-q_c\;)\overline{\varepsilon}_{C_i}\;(0)\\ &= W_i\;\varepsilon_i\;(1) + (1-W_i\;)\varepsilon_i\;(0) - q_c\;\overline{\varepsilon}_{C_i}\;(1) - (1-q_c\;)\overline{\varepsilon}_{C_i}\;(0)\\ &= (W_i\;- q_c\;)(\varepsilon_i\;(1) - \varepsilon_i\;(0)) + q_c\;(\varepsilon_i\;(1) - \overline{\varepsilon}_{C_i}\;(1)) + (1-q_c\;)(\varepsilon_i\;(0) - \overline{\varepsilon}_{C_i}\;(0)).\end{aligned}$$

Hence

$$\frac{1}{\sqrt{M}}\sum_{i=1}^{M} R_i\;(W_i\;- q_c\;)\dot{\varepsilon}_i\;= \frac{1}{\sqrt{M}}\sum_{i=1}^{M}\eta_{i1}\;+\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\eta_{i2}\;+\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\eta_{i3}$$

where

$$\begin{aligned}\eta_{i1}\;&= R_i\;(W_i\;- q_c\;)^2(\varepsilon_i\;(1) - \varepsilon_i\;(0))\\ \eta_{i2}\;&= R_i\;(W_i\;- q_c\;)q_c\;(\varepsilon_i\;(1) - \overline{\varepsilon}_{C_i}\;(1))\\ \eta_{i3}\;&= R_i\;(W_i\;- q_c\;)(1-q_c\;)(\varepsilon_i\;(0) - \overline{\varepsilon}_{C_i}\;(0)).\end{aligned}$$

Note that for the covariance terms the we can look at the three terms separately because

$$\mathbb{E}\left[\eta_{i1}\;\eta_{j2}\;\;i\neq j, C_i\;= C_j\;\right] = 0,$$

$$\mathbb{E}\left[\eta_{i1}\;\eta_{j3}\;\;i\neq j, C_i\;= C_j\;\right] = 0,$$

and

$$\mathbb{E}\left[\eta_{i2}\;\eta_{j3}\;\;i\neq j, C_i\;= C_j\;\right] = 0.$$

In addition,

$$\mathbb{E}\left[\eta_{i2}\;\eta_{j2}\;\;i\neq j, C_i\;= C_j\;\right] = 0,$$

and

$$\mathbb{E}\left[\eta_{i3}\;\eta_{j3}\;\;i\neq j, C_i\;= C_j\;\right] = 0,$$

so that we only need to consider the covariance terms from the first term. For this first term note that because

$$\sum_{i=1}^{M}(\varepsilon_i\;(1) - \varepsilon_i\;(0)) = 0,$$

it follows that

$$\begin{aligned}\frac{1}{\sqrt{M}}\sum_{i=1}^{M}\eta_{i1}\;&= \frac{1}{\sqrt{M}}\sum_{i=1}^{M} R_i\;(W_i\;- q_c\;)^2(\varepsilon_i\;(1) - \varepsilon_i\;(0))\\ &= \frac{1}{\sqrt{M}}\sum_{i=1}^{M}\left(R_i\;(W_i\;- q_c\;)^2 - P_C\;P_U\;\frac{1-4\sigma^2}{4}\right)(\varepsilon_i\;(1) - \varepsilon_i\;(0)).\end{aligned}$$

et us first look at the covariance terms:

$$\mathbb{E}\left[\eta_{i1}\;\eta_{j1}\;\;C_i\;= C_j\;, i\neq j\right]$$

$$= \mathbb{E}\left[\left(R_i\ (W_i\ -q_{C_i}\ )^2 - P_C\ P_U\ \frac{1-4\sigma^2}{4}\right)(\varepsilon_i\ (1)-\varepsilon_i\ (0)) \right.$$
$$\left.\left(R_j\ (W_i\ -q_{C_j}\ )^2 - P_C\ P_U\ \frac{1-4\sigma^2}{4}\right)(\varepsilon_j\ (1)-\varepsilon_j\ (0))\middle| C_i\ = C_j\ , i\neq j\right]$$
$$= \left\{P_C\ P_U^2\ \left(\kappa+\left(\frac{1-4\sigma^2}{4}\right)^2\right) - P_C^2\ P_U^2\ \left(\frac{1-4\sigma^2}{4}\right)^2\right\}(\varepsilon_i\ (1)-\varepsilon_i\ (0))(\varepsilon_j\ (1)-\varepsilon_j\ (0))$$
$$= \left\{P_C\ P_U^2\ (1-P_C\ )\left(\frac{1-4\sigma^2}{4}\right)^2 + \kappa P_C\ P_U^2\ \right\}(\varepsilon_i\ (1)-\varepsilon_i\ (0))(\varepsilon_j\ (1)-\varepsilon_j\ (0)).$$

Hence

$$\mathbb{E}\left[\sum_{i=1}^{M}\sum_{j=1,j=i}^{M}\eta_{i1}\ \eta_{j1}\ \right]$$
$$= \sum_{c=1}^{C}\left\{P_C\ P_U^2\ (1-P_C\ )\left(\frac{1-4\sigma^2}{4}\right)^2 + \kappa P_C\ P_U^2\ \right\}M_c^2\ (\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0))^2$$
$$-\sum_{i=1}^{M}\left\{P_C\ P_U^2\ (1-P_C\ )\left(\frac{1-4\sigma^2}{4}\right)^2 + \kappa P_C\ P_U^2\ \right\}(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2.$$

In addition we need to collect the variance terms:

$$\mathbb{E}[\eta_{i1}^2\ ] = \mathbb{E}\left[\left(R_i\ (W_i\ -q_c\ )^2 - P_C\ P_U\ \frac{1-4\sigma^2}{4}\right)^2(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2\right]$$
$$= \left\{P_C\ P_U\ \left(\kappa+\left(\frac{1-4\sigma^2}{4}\right)^2\right) - P_C^2\ P_U^2\ \left(\frac{1-4\sigma^2}{4}\right)^2\right\}(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2$$
$$= \left\{P_C\ P_U\ (1-P_C\ P_U\ )\left(\frac{1-4\sigma^2}{4}\right)^2 + P_C\ P_U\ \kappa\right\}(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2$$
$$\mathbb{E}[\eta_{i2}^2\ ] = \mathbb{E}\left[\{R_i\ (W_i\ -q_c\ )q_c\ (\varepsilon_i\ (1)-\overline{\varepsilon}_{C_i}\ \ (1))\}^2\right] = P_C\ P_U\ \kappa_{3,1}(\varepsilon_i\ (1)-\overline{\varepsilon}_{C_i}\ \ (1))^2,$$

and

$$\mathbb{E}[\eta_{i3}^2\ ] = \mathbb{E}\left[\{R_i\ (W_i\ -q_c\ )(1-q_c\ )(\varepsilon_i\ (0)-\overline{\varepsilon}_{C_i}\ \ (0))\}^2\right] = P_C\ P_U\ \kappa_{1,3}(\varepsilon_i\ (0)-\overline{\varepsilon}_{C_i}\ \ (0))^2.$$

Thus

$$\mathbb{E}[(\eta^{\mathrm{fe}})^2] = \mathbb{E}\left[\left(\frac{4}{(1-4\sigma^2)\sqrt{M\ P_C\ P_U}}\sum_{i=1}^{M}R_i\ (W_i\ -q_{C_i}\ )\dot{\varepsilon}_i\ \right)^2\right]$$
$$= \frac{16}{(1-4\sigma^2)^2 M\ P_C\ P_U}\left\{\sum_{i=1}^{M}\mathbb{E}[\eta_{i1}^2\ ] + \sum_{i=1}^{M}\sum_{j=1,j=i}^{M}\mathbb{E}[\eta_{i1}\ \eta_{j1}\ ] + \sum_{i=1}^{M}\mathbb{E}[\eta_{i2}^2\ ] + \sum_{i=1}^{M}\mathbb{E}[\eta_{i3}^2\ ]\right\}$$
$$= \frac{1}{M}\sum_{i=1}^{M}\left\{(1-P_U\ )\left(1+\kappa\frac{16}{(1-4\sigma^2)^2}\right)(\varepsilon_i\ (1)-\varepsilon_i\ (0))^2\right.$$
$$\left.+\frac{16\kappa_{3,1}}{(1-4\sigma^2)^2}(\varepsilon_i\ (1)-\overline{\varepsilon}_{C_i}\ \ (1))^2 + \frac{16\kappa_{1,3}}{(1-4\sigma^2)^2}(\varepsilon_i\ (0)-\overline{\varepsilon}_{C_i}\ \ (0))^2\right\}$$
$$+\frac{1}{M}\sum_{c=1}^{C}\left\{P_U\ (1-P_C\ ) + \kappa P_U\ \frac{16}{(1-4\sigma^2)^2}\right\}M_c^2\ (\overline{\varepsilon}_c\ (1)-\overline{\varepsilon}_c\ (0))^2.$$