RECENT PROGRESS ON REGRESSION

Michal Kolesár*

April 11, 2024

We have data $\mathcal{D}_n = \{Y_i, X_i\}_{i=1}^n$, where Y_i is a scalar outcome, and $X_i = (D_i, W_i')'$ is a k-vector of covariates, composed of a scalar D_i of interest, and a vector of $k-1$ controls W_i (including the intercept). We regress the outcome Y_i onto X_i, obtaining the ordinary least squares (OLS) estimator $\hat{\theta} = (X'X)^{-1}X'Y$ (we use the usual matrix notation that rows of a matrix A are given by A_i'). By the Frisch–Waugh–Lovell (FWL) theorem, the coefficient on the scalar D_i is given by

$$\hat{\beta} = \hat{\theta}_1 = \frac{\sum_{i=1}^n \ddot{D}_i Y_i}{\sum_{i=1}^n \ddot{D}_i^2},$$

where $\ddot{D} = D - H_W D$ is the residual from projecting D onto W, and $H_A = A(A'A)^{-1}A'$ is the hat matrix (also called the projection matrix). We will consider two questions that turn out to be quite a bit more complicated that one may at first think:

1. What are $\hat{\beta}$ (and $\hat{\theta}$) estimating?
2. What standard errors for $\hat{\beta}$ should we report?

For the first question, the answer is simple if we assume that the regression function $\mu(D_i, W_i) := E[Y_i \mid D_i, W_i]$ is linear: $\hat{\beta}$ estimates the marginal effect of a unit increase in D_i. If we assume that D_i is as good as randomly assigned conditional on the controls W_i, then this marginal effect also has a causal interpretation. But the linearity assumption may be suspect: it rules out that the marginal effect varies with W_i, for instance. Our task is to think about what $\hat{\beta}$ may be estimating when the regression is *misspecified* in the sense that $\mu(D_i, W_i)$ is not linear. As we'll see, once we allow for the possibility of misspecification, we will have three choices for the estimand.

For the second question, there are two leading options. First, we could use the Eicker-Huber-White (EHW) variance estimator,

$$\hat{V}_{\text{EHW}} = (X'X)^{-1} \sum_{i=1}^n \hat{\epsilon}_i^2 X_i X_i' (X'X)^{-1}, \qquad \hat{\epsilon}_i = Y_i - X_i'\hat{\theta},$$

*Email: mkolesar@princeton.edu.

with the standard error given by the square root of its $(1,1)$ element,

$$\text{se}_{\text{EHW}}(\hat{\beta}) = \hat{V}_{\text{EHW},11}^{1/2} = \frac{\sqrt{\sum_{i=1}^{n} \hat{\epsilon}_i^2 \ddot{D}_i^2}}{\sum_{i=1}^{n} \ddot{D}_i^2}. \tag{1}$$

Alternatively, if observation i belongs to cluster $s(i) \in \{1, \ldots, S\}$, we may use the Liang-Zeger (LZ) variance estimator (Liang and Zeger 1986)[1]

$$\hat{V}_{\text{LZ}} = (X'X)^{-1} \sum_{s} \sum_{i,j:\ s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j X_i X_j' (X'X)^{-1},$$

with the standard error given by the square root of its $(1,1)$ element,

$$\text{se}_{\text{LZ}} = \hat{V}_{\text{LZ},11}^{1/2} = \frac{\sqrt{\sum_{s=1}^{S} \sum_{i,j:\ s(i)=s(j)=s} \hat{\epsilon}_i \hat{\epsilon}_j \ddot{D}_i^2}}{\sum_{i=1}^{n} \ddot{D}_i^2}.$$

Stata's `cluster` option uses this standard error multiplied by a finite-sample adjustment, $(n-1)/(n-k) \times S/(S-1)$.

While settling these two issues, we will also consider the regularity conditions needed for se_{EHW} and se_{LZ} to deliver asymptotically valid (exact or conservative) inference. Next time, we will consider the diagnostics implied by these regularity conditions, and what to do if the diagnostics are questionable.

1. ESTIMANDS AND THE POPULATION OF INTEREST

Our first agenda item is to make precise what $\hat{\beta}$ and $\hat{\theta}$ are estimating. Saying that they estimate β or θ in the "linear regression"

$$Y_i = D_i \beta + W_i' \gamma + \epsilon_i = X_i' \theta + \epsilon_i, \tag{2}$$

is somewhat vacuous, since eq. (2) *does not define* these coefficients. This is because, as we'll see shortly, there are several ways in which we can define β and θ. The definition will determine how we think about repeated sampling when we consider the statistical properties of the OLS estimator, what assumptions we need for inference, as well as whether the usual EHW or LZ standard errors yield exact coverage in large samples, are conservative, or perhaps misleading.

To define the estimand, recall that broadly speaking, econometric analysis may have one of three goals:

PREDICTION Here the interest lies in estimating the conditional mean $E[Y_i \mid X_i]$. If Y_i is binary, we may also be interested in predicting the actual value of Y_i. For instance, we may want to predict whether a defendant would commit pretrial misconduct,

1. The second most cited paper published by Biometrika after Rosenbaum and Rubin (1983).

such as failing to appear in court, if released on bail. Within this context, we may then be interested in if our "machine predictions" beat the predictions of bail judges, as in Kleinberg et al. (2018).

DESCRIPTION Assuming D_i is binary, we would like to identify some weighted average of covariate-specific contrasts $\beta(W_i) := E[Y_i \mid D_i = 1, W_i] - E[Y_i \mid D_i = 0, W_i]$.

For example, we may be interested in identifying the racial disparity between whites and blacks, adjusted for covariates, or in studying gender disparities. A classic example is Fryer and Levitt (2013), who are interested in whether there are racial differences in cognitive ability of children at 6 months and at 2 years.

CAUSAL INFERENCE Assume again that D_i is binary. There are potential outcomes $Y_i(1)$ and $Y_i(0)$, and we're interested in some average of conditional average treatment effects (ATEs) $\tau(W_i) := E[Y_i(1) \mid W_i] - E[Y_i(0) \mid W_i]$, or perhaps of individual treatment effects $\tau_i = Y_i(1) - Y_i(0)$.

We will leave the prediction problem to the end of the course, and focus on the other two possibilities here. As we will see, unless the population from which the units are drawn is finite (see Remark 13), the distinction between descriptive and causal inference will only matter for the *interpretation* of the estimand, but not for how we conduct inference.

Remark 1. Recall that by the Holland and Rubin motto, "no casuation without manipulation" (Holland 1986), questions concerning racial or gender disparities all fall into the second category: it doesn't make sense to ask whether the more women would be offered lucrative job if they were male. *Immutable attributes* by definition cannot be manipulated! However, we can ask whether the *perception* of someone's gender or race has a causal impact, as in audit studies (e.g. Bertrand and Mullainathan 2004), or blind interviews (Goldin and Rouse 2000). If D_i is a manipulable variable rather than an immutable attribute, such as union membership or an industry indicator, we may be interested in both descriptive and causal questions.

Remark 2 (Fundamental problem of causal inference). For the most part, we'll be interested in averages of conditional ATEs. Why? Recall the *fundamental problem of causal inference*: we can only ever observe $Y_i(1)$ or $Y_i(0)$, but not both. Thus, individual treatment effects τ_i are not identified unless we make homogeneity restrictions—what Holland (1986) calls a *scientific solution*, and what economists may call a *structural model*, which may allow us to infer counterfactual outcomes for a given unit by extrapolating from similar units, or from other time periods. More generally, since the data is at best informative only about the marginal distributions of $Y_i(1)$ and $Y_i(0)$,

1.1. *Descriptive estimands: Classic approach*

The standard approach you saw in 517 is to think of the sample $\mathcal{D}_n$ as drawn from some large superpopulation (with an infinite number of units i), and to define θ as the best linear predictor for this superpopulation,

$$\theta_{\text{u}} = \underset{\theta}{\text{argmin}}\, E[(Y_i - X_i'\theta)^2] = \underset{\theta}{\text{argmin}}\, E[(\mu(X_i) - X_i'\theta)^2] = E[X_i X_i']^{-1} E[X_i Y_i],$$

where "u" stands for unconditional (superpopulation) inference. By the FWL theorem,

$$\beta_{\text{u}} = \theta_{\text{u},1} = \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{E[\tilde{D}_i \mu(X_i)]}{E[\tilde{D}_i^2]},$$

where $\tilde{D}_i$ is the population residual from projecting D_i onto W_i—the population analog of $\ddot{D}_i$:

$$\tilde{D}_i = D_i - W_i'\delta, \qquad \delta = E[W_i W_i']^{-1} E[W_i D_i].$$

When does β_{u} correspond to some weighted average of covariate-specific contrasts $\beta(W_i)$? Let us first consider the additively separable case where $\mu(X_i) = D_i\beta + g(W_i)$, which is called the *partially linear model*. Then $\beta(W_i)$ does in fact not depend on W_i, and questions about how to weight the covariate-specific contrasts are moot. Then, since $E[\tilde{D}_i D_i] = E[\tilde{D}_i^2]$, the previous display simplifies to

$$\beta_{\text{u}} = \beta + \frac{E[\tilde{D}_i g(W_i)]}{E[\tilde{D}_i^2]}.$$

There are two ways to kill the second term.

Assumption 1 (Linearity). At least one of the following conditions holds:

(i) the propensity score $p(W_i) := E[D_i \mid W_i]$ is linear in W_i; or

(ii) $\mu(0, W_i) = E[Y_i \mid D_i = 0, W_i]$ is linear in W_i.

Under Assumption 1(i), the second term suffers death by iterated expectations since $E[\tilde{D}_i \mid W_i] = 0$. Under version (ii), we use orthogonality of $\tilde{D}_i$ and W_i to kill it. Regression is, in this sense, *doubly robust*: we need either the regression function or the propensity score to be linear, but not both. This property was observed, for instance, by Robins, Mark, and Newey (1992).

What if $\mu(X_i)$ is not additively separable? Let us suppose that

$$\mu(X_i) = D_i\beta(W_i) + \mu(0, W_i), \tag{3}$$

so that conditional on W_i, μ is linear in D_i. Under eq. (3), the average of covariate-specific contrasts, $\beta(W_i) = E[Y_i \mid D_i = k, W_i] - E[Y_i \mid D_i = k-1, W_i]$, doesn't depend on

k. Observe eq. (3) holds automatically if the treatment is binary. Then

$$\beta_{\text{u}} = \frac{E[\tilde{D}_i D_i \beta(W_i)]}{E[\tilde{D}_i^2]} + \frac{E[\tilde{D}_i \mu(0, W_i)]}{E[\tilde{D}_i^2]} = \frac{E[\lambda_{\text{u}}(W_i)\beta(W_i)]}{E[\lambda_{\text{u}}(W_i)]},$$
$$\lambda_{\text{u}}(W_i) = E[\tilde{D}_i D_i \mid W_i] = \text{var}(D_i \mid W_i) + p(W_i)(p(W_i) - W_i'\delta), \quad (4)$$

where the second equality in the expression for β_{u} uses Assumption 1 and iterated expectations. So under Assumption 1, regression identifies a weighted average of covariate-specific contrasts $\beta(W_i)$, with weights $\lambda_{\text{u}}(W_i)/E[\lambda_{\text{u}}(W_i)]$ that have expectation one. Furthermore, under Assumption 1(i), we have $\lambda_{\text{u}}(W_i) = \text{var}(D_i \mid W_i) > 0$, so the weights are convex. But they are not necessarily convex under Assumption 1(ii), so to the extent that we care about convex weights (see Remark 11 below) the double robustness property doesn't fully generalize. In the case with a binary treatment, in particular, we can simplify the weights to $\lambda_{\text{u}}(W_i) = p(W_i)(1 - W_i'\delta)$, and get a simple necessary and sufficient condition for convexity of the weights: the projection of D_i onto W_i fits probabilities that lie below one. That is, we need that for all W_i such that $p(W_i) > 0$,

$$W_i'\delta \leq 1. \quad (5)$$

The sample analog of this condition is easy to check. The condition is also intuitive—by FWL, we implicitly estimate the propensity score by projecting D_i onto W_i, and we just require that our propensity score model doesn't yield nonsensical predictions.

Remark 3 (Double robustness). We can actually make the double robustness result a little more general, which will have important implications for "double machine learning" we'll discuss later in the course. Let $r_p(W_i) = p(W_i) - W_i'\delta$ be the non-linearity in the propensity score, and let $r_\mu(W_i) = \mu(0, W_i) - W_i' E[W_i W_i'] E[W_i Y_i(0)]$ be the non-linearity in the conditional mean function. Then we can write the functional-form bias term in eq. (4) as

$$\frac{E[\tilde{D}_i \mu(0, W_i)]}{E[\tilde{D}_i^2]} = \frac{E[r_p(W_i) r_\mu(W_i)]}{E[\tilde{D}_i^2]} = O(E[r_p W_{(i)}^2]^{1/2} E[r_\mu(W_i)^2]^{1/2}),$$

For this bias to be negligible relative to the variance of the OLS estimator, which will typically be of the order $O(n^{-1/2})$, we need the *product* of the non-linearity residuals to be of smaller order than $n^{-1/2}$: $E[r_p W_{(i)}^2]^{1/2} E[r_\mu(W_i)^2]^{1/2} = o(n^{-1/2})$. For instance, if both the propensity score and the conditional mean of the outcome are non-linear, but the non-linearities are of the order $o(n^{-1/4})$, we're good. So regression is "doubly robust" in the sense that we can allow for misspecification in the implicit model for the propensity score, as well as in the conditional mean function. Unless *both* are substantial, the non-linearity bias will not really matter. In practice, this means that small functional form mistakes, like omitting to put in interactions between elements of W_i should not really matter for the estimand β_{u}. In my view, this double robustness property is one of the main advantages of regression.

Question 1. What if eq. (3) doesn't hold? Hints are provided in Yitzhaki (1996, Proposition 2), Angrist and Pischke (2009, p. 78), or Angrist and Krueger (1999, p. 1311–2).

SUMMARY To *interpret* the best linear predictor β_u as a weighted average of covariate-specific contrasts, we need Assumption 1, which we can ensure by controlling for W_i "flexibly" (allowing for interactions etc), or at least sufficiently flexibly so that the remaining non-linearities are small. Only the first version of the assumption guarantees convexity of the weights. In practice, checking the weights $\lambda_u(W_i)$ seems like a good idea.

Separate from the interpretation of the estimand is the question of inference. For this we need a sampling assumption to ensure that our sample is representative of the population, and we don't have selection bias. The simplest one is:

Assumption 2 (Sampling). The units i are drawn i.i.d. from a large superpopulation.

Given our definition of the regression slope θ_u, we define the residual as $\epsilon_{u,i} = Y_i - X_i'\theta_u = Y_i - X_i'E[X_iX_i']^{-1}E[X_iY_i]$. From this definition, it follows that $X_i\epsilon_{u,i}$ is mean zero and i.i.d. across i. Therefore, by the central limit theorem (CLT), and law of large numbers,

$$\mathcal{V}_u^{-1/2}(\hat{\theta} - \theta_u) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_u), \tag{6}$$

where

$$\mathcal{V}_u = (X'X)^{-1}\sum_{i=1}^{n}\epsilon_{u,i}^2 X_iX_i'(X'X)^{-1}.$$

The regularity conditions we need are quite mild:

Assumption 3. $E[X_iX_i']$ is non-singular, and Y_i, X_i have finite fourth moments.

This ensures that θ_u is well-defined in the first place, that the law of large numbers applies to $X'X/n$, and that the variance $\text{var}(\epsilon_{u,i}X_i) = E[(Y_i - X_i'\theta_u)^2X_iX_i']$ is finite. Under further regularity conditions, the EHW variance estimator will be consistent for $\mathcal{V}_u$, using the standard error (1) will yield asymptotically exact inference.

1.2. *Descriptive estimand: conditional approach*

Another possibility, explored in Abadie, Imbens, and Zheng (2014) is to define the estimand of interest to be the conditional best linear predictor

$$\theta_{cx} = \underset{\theta}{\text{argmin}}\,\frac{1}{n}\sum_i(\mu(X_i) - X_i'\theta)^2 = (X'X)^{-1}X'\mu(X),$$

so that the residual is now defined as $\epsilon_{cx,i} = Y_i - X_i(X'X)^{-1}X'\mu(X)$. Here "cx" stands for inference that is conditional on X.

The distinction between the unconditional and conditional best linear predictor does not matter for estimation (we use OLS in each case), but, as we'll see, it will matter for inference. One way of thinking about the distinction between θ_{cx} and θ_{u} is to think about the sampling perspective we wish to take: do we redraw the regressors in repeated sampling, or do we treat them as fixed? Do we want to do inference conditionally on X, or unconditionally? In textbooks, the starting point is often to treat the regressors as fixed: it underlies, for instance, the Gauss-Markov theorem. So it seems natural to keep them fixed when we generalize to allow for misspecification in the conditional mean. However, just because it seems "natural", it doesn't mean it's always the right thing to do.

Example 1. As an example of where θ_{cx} may be of interest, Abadie, Imbens, and Zheng (2014) consider the setting of Karlan and List (2007), who, using direct mail solicitations to over 50,000 prior donors of a nonprofit organization, randomly offer matching incentives at different match ratios (1:1, 2:1, 3:1, or none) to prior donors. They report probit estimates where the object of interest is the regression coefficient on the indicator for being offered a matching incentive for charitable giving, and they control for the characteristics of the matching incentives in these estimates. They argue that since the distribution of these incentives is fixed by the researchers there appears to be no reason to take this uncertainty into account. ⊠

Question 2. Is this a compelling example?

Remark 4. See a discussion of this issue in Wooldridge (2010, Chapter 1.4). Do you agree with the points? Do you disagree?

Under this definition of θ, by arguments analogous to eq. (4), the coefficient on D_i now becomes

$$\beta_{\text{cx}} = \frac{\sum_i \ddot{D}_i \mu(X_i)}{\sum_i \ddot{D}_i^2} = \frac{\sum_i \lambda_{\text{cx}}(X_i)\beta(W_i)}{\sum_i \lambda_{\text{cx}}(X_i)} + \frac{\sum_i \ddot{D}_i \mu(0, W_i)}{\sum_i \ddot{D}_i^2}, \qquad \lambda_{\text{cx}}(X_i) = \ddot{D}_i D_i.$$

Now, killing the second term requires version (ii) of Assumption 1 to hold, version (i) doesn't help us. Intuitively, once we condition on X_i, we can't really make use of the propensity score. The condition in eq. (5) on convexity of the weights for the first term is replaced by the condition that $W_i'\hat{\delta} \leq 1$ whenever $D_i = 1$, where $\hat{\delta} = (W'W)^{-1}W'D$ is the slope from the regression of D_i onto D_i, which is easy enough to check. The weights λ_{cx} can be thought of as a finte-sample analog of the weights λ_{u} we defined in eq. (4). One important difference is that, when D_i is binary, λ_{cx} only put positive weights on units with $D_i = 1$: so β_{cx} defines a weighted average of conditional contrasts only among treated units.

Again, the question of interpretation of the estimand is separate from the question of inference—for inference, we don't need Assumption 1, but we instead need some sampling assumptions. Abadie, Imbens, and Zheng (2014) analyze the properties of $\hat{\theta}$ as an estimator of θ_{cx} under Assumption 2. However, given the definition of the estimand,

it may be more natural here to do inference that is conditional on X, which is a stronger requirement. For this we can weaken Assumption 2 and instead impose:

Assumption 4 (Conditional sampling). Conditional on X, the outcomes Y_i are independent.

This, in particular, allows for dependence between the regressors. To apply a CLT to $\hat{\theta}$, on the other hand, we need a stronger condition on the regressors, which will become useful later when we think of finite-sample issues with the EHW estimator. Recall the diagonal elements $H_{X,ii}$ of the projection matrix H_X are called the leverage of the ith observation. In analogy, diagonal elements of $H_{\ddot{D}}$ are called *partial leverage* (e.g. Velleman and Welsch 1981; Chatterjee and Hadi 1986).

Assumption 5. $E[|Y_i - \mu(X_i)|^{2+\eta} \mid X]$ is bounded above for some $\eta > 0$, and $\sigma^2(X_i) := \operatorname{var}(Y_i \mid X_i)$ is bounded away from zero. In addition, the maximum leverage $\max_i H_{X,ii}$ converges to zero (for inference on θ_{cx}); or the maximum partial leverage $\max_i H_{\ddot{D},ii}$ converges to zero (for inference on β_{cx}).

Then, under Assumptions 4 and 5,

$$\mathcal{V}_{\text{cx}}^{-1/2}(\hat{\theta} - \theta_{\text{cx}}) \xrightarrow{d} \mathcal{N}(0, I)$$
$$\mathcal{V}_{\text{cx},11}^{-1/2}(\hat{\beta} - \beta_{\text{cx}}) \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\mathcal{V}_{\text{cx}} = (X'X)^{-1}\sum_i (Y_i - \mu(X_i))^2 X_i X_i' (X'X)^{-1}$, with its $(1,1)$ element given by $\mathcal{V}_{\text{cx},11} = \frac{\sum_i (Y_i - \mu(X_i))^2 \ddot{D}_i^2}{(\sum_i \ddot{D}_i^2)^2}$.

Proof. First consider

$$\frac{\sum_i (\epsilon_i^2 - \sigma^2(X_i))\ddot{D}_i^2}{\sum_i \sigma^2(X_i)\ddot{D}_i^2}.$$

Now by inequality of von Bahr and Esseen (1965)[2],

$$E\left[\left(\sum_i (\epsilon_i^2 - \sigma^2(X_i))\ddot{D}_i^2\right)^{1+\eta/2} \mid X\right] \leq 2\sum_i E[\epsilon_i^{2+\eta} \mid X]\ddot{D}_i^{2+\eta} \leq 2\max_i \ddot{D}_i^{\eta} \sum_i \ddot{D}_i^2.$$

Therefore, by Markov's inequality, the expression is of the order

$$\frac{2\max_i \ddot{D}_i^{\eta} \sum_i \ddot{D}_i^2}{(\sum_i \sigma^2(X_i)\ddot{D}_i^2)^{1+\eta/2}} \preceq \max_i H_{\ddot{D},ii}^{\eta/2} \to 0.$$

Similar argument applies to the full variance expression. Therefore, it suffices to prove the theorem with $\mathcal{V}_{\text{cx}} = (X'X)^{-1}\sum_i \operatorname{var}(Y_i \mid X_i) X_i X_i' (X'X)^{-1}$. Write $\mathcal{V}_{\text{cx}}^{-1/2}(\hat{\theta} - \theta_{\text{cx}}) = \sum_i w_i (Y_i - \mu(X_i))$, where $w_i = (\sum_i \operatorname{var}(Y_i \mid X_i) X_i X_i')^{-1/2} X_i$. We need to apply a CLT to $\sum_i w_i (Y_i - \mu(X_i))$. Since $w_i(Y_i - \mu(X_i))$ is not i.i.d., we need the Lindeberg-Feller CLT (e.g. Davidson (1994), Theorem 23.6), which says that if A_i is mean zero, independent, with variance Ω_i, then $\sum_i A_i \Rightarrow \mathcal{N}(0, \Omega)$ so long as the Lindeberg condition $\sum_i E[\|A_i\|^2 \mathbb{1}\{\|A_i\| > \epsilon\}] \to 0$ holds, and $\sum_i \Omega_i \to \Omega$. The Lindeberg condition is implied by the Lyapunov condition $\sum_i E\|A_i^{2+\eta}\| \to 0$. In our case, since

2. Let $1 \leq p \leq 2$, and let X_i be independent random variables with zero mean and finite pth moment. Then $E|\sum_i X_i|^p \leq c_{p,n} \sum_i E|X_i|^p$, where $c_{p,n} \leq 2 - n^{-1}$.

$E[|Y_i - \mu(X_i)|^{2+\eta} \mid X]$ is bounded, $\sum_i E[\|w_i(Y_i - \mu(X_i))\|^{2+\eta} \mid X]$ is bounded by a constant times $\sum_i \|w_i\|^{2+\eta} \leq \max_i \text{var}(Y_i \mid X_i)^{-(1+\eta/2)} \cdot \sum_i H_{X,ii}^{1+\eta/2}$. Now, $\sum_i H_{X,ii}^{1+\eta/2}$ converges to zero if and only if $\max_i H_{X,ii} \to 0$: In one direction, observe $\sum_i H_{X,ii}^{1+\eta/2} \leq \max_j H_{X,jj}^{\delta} \sum_i H_{X,ii} = \max_j H_{X,jj}^{\delta}$. In the other direction, $\sum_i H_{X,ii}^{1+\eta/2} \geq \max_i H_{X,ii}^{1+\eta/2}$.

For $\hat{\beta}$, we write $\mathcal{V}_{\text{cx},11}^{-1/2}(\hat{\beta} - \beta_{\text{cx}}) = \sum_i w_i(Y_i - \mu(X_i))$, with $w_i = (\sum_i \text{var}(Y_i \mid X_i)\ddot{D}_i^2)^{-1/2}\ddot{D}_i$, and use the same arguments.

Note that for consistency of $\hat{\theta}$, or $\hat{\beta}$, we can allow the leverages not to converge to zero. By Chebyshev inequality, the estimator is consistent (indeed, $\sqrt{n}$-consistent), if $X'X/n$ converges to a positive definite matrix. □

There are three takeaways from this result.

FORM OF VARIANCE The EHW variance estimator targets $\mathcal{V}_{\text{u}}$ defined in eq. (6), which replaces $Y_i - \mu(X_i)$ with $Y_i - X_i'\theta_{\text{u}}$ in the middle of the sandwich. But since $Y_i - \mu(X_i)$ is conditionally mean zero,

$$(Y_i - X_i'\theta_{\text{u}})^2 = (Y_i - \mu(X_i))^2 + (\mu(X_i) - X_i'\theta_{\text{u}})^2$$

plus a term that's asymptotically negligible since it's mean zero conditional on X_i. Thus, the EHW variance estimator will be conservative whenever the conditional mean $\mu(X_i)$ is not linear.

Abadie, Imbens, and Zheng (2014) propose an alternative variance estimator based on nearest neighbor matching that is consistent, and that therefore leads to smaller standard errors in large samples than those in eq. (1). While the possibility that you can reduce your standard errors is always tantalizing, remember that the gains in precision come from changing the estimand. Furthermore, if the Abadie, Imbens, and Zheng (2014) standard errors are meaningfully smaller than those based on $\hat{V}_{\text{EHW}}$, this indicates that the regression function is misspecified. Either way, one would need a clear argument in defense of the particular estimand β_{cx}.

HIGH-DIMENSIONAL CONTROLS For inference on β_{cx}, there is no condition that restricts the dimensionality of the controls W_i. So long as the partial leverage goes to zero, the asymptotic normality result goes through even when $\dim(W_i)$ is proportional to the sample size n.

LEVERAGE CONDITION The assumption on the leverages is needed to ensure that the contribution of any individual observation to the estimate $\hat{\beta}$ is asymptotically negligible—otherwise, if any one observation makes a non-negligible contribution, the distribution of the estimator will depend in a non-negligible way on the distribution of the outcome for that particular observation.

To think about more about the leverage condition, which does not appear in Assumption 3, consider regressing log wages on education and gender. We know that in the population, 50% of individuals are women. But in our sample of size 100, we only

have 2 women. Is this an issue for unconditional inference? Is this an issue for conditional inference on the coefficient on gender? Is this an issue for conditional inference on the coefficient on education?

Question 3. How does this relate to the measurement problem considered by Cox (1958) (see also Cox and Hinkley (1974), page 96).[3]

Remark 5. How do the regularity conditions in Assumption 5 fit in with Young (2019)? What is the practical takeaway?

Remark 6. As we'll discuss in the lecture on regression discontinuity (RD) designs, inference on the RD parameter using local polynomial regression methods amounts to just running OLS for observations near the cutoff (or perhaps weighted least squares if kernel weights are used). In that design, conditioning on X_i is well-motivated. As essentially a special case of the results above, we'll see that the EHW variance estimator is conservative, and that nearest-neighbor methods are preferred. But we'll have the additional complication that we'll need to deal with bias, since our parameter of interest will be the RD parameter, and not β_{cx}.

SUMMARY The conditional estimand β_{cx} seems harder to motivate than the estimand β_{u}. But the conditional approach gives us a simple diagnostic on the partial leverage matrix, and it seems like a good idea to always check the maximal partial leverage $\max_i H_{\ddot{D},ii}$ in practice. If it's high, say larger than 0.1, one needs to be careful with standard inference. We'll discuss what to do in this case in the next set of notes. The other nice thing about this framework is that we get to allow for high-dimensional covariates at basically no cost, one just needs to be careful in estimating the variance—again, we'll come back to this next time.

1.3. *Causal estimands*

Consider now a setting in which we're interested in the causal effect of D_i on Y_i. The observed outcome is then a function of the potential outcomes $Y_i = Y_i(D_i)$.

Remark 7. Indexing potential outcomes by treatment only imposes what's known a Stable unit treatment value assumption (SUTVA), which involves two restrictions.[4] The first is a particular type of an *exclusion restriction*: treatment values of other units do not

3. There are two measurement instruments, both mean zero with normal error, the first one has variance σ^2 and the second one has variance $100\sigma^2$. We flip a coin to decide which one to use. A_i is an indicator that we're using the first one. The natural confidence interval (CI) is $Y \pm (1.96A_i + 196(1 - A_i))\sigma$. But the CI $Y \pm (5A_i + 164(1 - A_i))\sigma$ has shorter average length. Should we prefer it? For our purposes, if the regression function is linear, the marginal distribution of X is ancillary, so Cox's argument implies we should condition on it.

4. The name of the assumption comes from a scathing comment by Don Rubin (Rubin 1980) that begins: "Basu's article on Fisher's randomization test for experimental data is certainly entertaining" and gets better from there.

affect the potential outcomes of unit i. This rules out peer, network, and general equilibrium effects. There is a growing literature studying casual inference on networks that relaxes this restriction. Second, it posits that only the treatment dose as measured by D_i matters, it doesn't matter how it is administered. Suppose that D_i is years of education. We're implicitly saying that school quality doesn't matter, only school quantity: a strong restriction that is refuted by the vast literature on school quality such as Card and Krueger (1992), who argue that 20% of the narrowing of the black-white wage gap can be attributed to improvements in school quality. To carefully define what we're estimating, returns to schooling studies should therefore arguably index potential outcomes also by school quality, or perhaps by which school the individual attended to recover SUTVA.

There are three estimands that can be of interest. First, we may be interested in treatment effects for the superpopulation from which $\mathcal{D}_n$ is drawn. Second, $\mathcal{D}_n$ may itself be the population of interest, such as when our sample comprises all US States, or all countries in the world, etc. It is also relevant if we don't in fact have a random sample from a population, but a convenience sample (since then it's not clear how to extrapolate to the population of interest). Then the only statistical uncertainty we have is that we don't observe all potential outcomes for each unit. This is what Abadie et al. (2020) call *design-based uncertainty*. We observe $Y_i(D_i)$, but we could have observed some other potential outcome if the treatment assignment was different. However, there is no *sampling uncertainty*: we observe the entire population. In such a case we can do inference that is internally valid, but lacks external validity in that is unclear how the results translate to other populations. The third possibility is that we think of eq. (2) as a "structural model" (computer scientists may say "generative model"), with ϵ_i being "structural shocks" that are unobserved by the econometrician, and ask what happens if this structural model is misspecified.

CAUSAL INFERENCE ON SUPERPOPULATION If we're interested in treatment effects for a superpopulation of units, then the only remaining question is what assumptions on the potential outcomes and the treatment assignment process we need in order to interpret β_u as a causal object (inference on β_u is as in Section 1.1). For this we just need to make sure that $\beta(W_i)$ has a causal meaning. There are two paths to it, a "model-based" path, and "design-based" path. To simplify things, let us assume a causal version of eq. (3),

$$Y_i(d) = Y_i(0) + d\tau_i, \tag{7}$$

so that the treatment effect is linear in the "dose" d, but the effect per dose is allowed to vary by individual. Again, this is not restrictive if D_i is binary, and this assumption could be dropped following the hints in question 1.

Assumption 6 (Unconfoundedness). At least one of the following conditions holds:

(i) $E[D_i \mid W_i, Y_i(\cdot)] = E[D_i \mid W_i]$ and version (i) of Assumption 1 holds.

(ii) $E[Y_i(d) \mid D_i, W_i] = E[Y_i(d) \mid W_i]$ and version (ii) of of Assumption 1 holds.

While we could have assumed that the treatment is as good as randomly assigned conditional on covariates in the sense that $\{Y(d) \colon d \in \mathbb{R}\} \perp\!\!\!\perp D_i \mid W_i$, I like stating the random assignment assumption (conditional on covariates) as in Assumption 6 because it makes the distinction between a *model-based* and *design-based* approaches to identification clearer.

In a *design-based* approach, we only make restrictions on how the treatment is assigned, but we do not restrict the potential outcomes. Correspondingly, version (i) of Assumption 6 says that conditional on covariates, the assignment doesn't depend on potential outcomes, and that, to ensure we don't have functional-form bias, the covariates affect the propensity score linearly. If we have quasi-experimental or experimental data, then it's easy to verify this assumption.

In contrast, in a *model-based* approach, we make assumptions about the distribution of potential outcomes given (D_i, W_i), and we do not restrict how the treatment is assigned: the propensity scores are unrestricted. Correspondingly, version (ii) of Assumption 6 says that conditional on covariates, the treatment doesn't help predict the potential outcomes, and we make a functional form restriction on the form of $\mu(0, W_i)$ to ensure that we don't have functional form bias. As we'll discuss later, differences-in-differences designs are a special case of this approach, where the functional form restriction takes the form of a parallel trends assumption, and the assumption that treatment doesn't help predict potential outcomes rules out things like the Ashenfelter dip.

Remark 8. The "design-based" vs "model-based" nomenclature is a bit confusing: both approaches require a model which we restrict. It's just that in one case, it's a model for treatment assignment, while in the other, it's a model for the potential outcomes.

Assumption 6 formalizes the idea that assignment is as good as random conditional on the covariates: the only selection is on W_i; there is no selection on unobservables. Correspondingly, iterated expectations arguments imply that in the decomposition in eq. (4), $\mu(0, W_i) = E[Y_i(0) \mid W_i]$, and $\beta(w) = \tau(w)$, where $\tau(w) := E[\tau_i \mid W_i = w]$ is the average effect of increasing the treatment by one unit among observations with $W_i = w$. Thus, under Assumption 6, we estimate a weighted average of covariate-specific treatment effects with weights $\lambda_u(W_i)$. This is perhaps the most important result in these lecture notes, so let's record it as a theorem:

Theorem 9. Suppose Assumption 6 and eq. (7) holds, and let $\tau(w) := E[\tau_i \mid W_i = w]$. Then regression estimates a weighted average of $\tau(w)$ with weights $\lambda_u(w)$ given in eq. (4).

Let's parse through the implications:

Remark 10 (Unconfoundedness and bad controls). *Good controls* that aid the unconfoundedness assumptions are things that are fixed at the time of our hypothetical experiment that defines the potential outcomes. In contrast, W_i should not include variables that are themselves outcomes of our hypothetical experiment and therefore potentially mediate the causal effect of D_i onto the outcome of interest: Angrist and Pischke (2009,

Chapter 3.2.2) call those *bad controls*, and including them will violate Assumption 6 (to see this, consider the extreme case where we condition on a proxy for the observed outcome). This is why it's important to define the potential outcomes carefully, they help you figure out what should to into W_i.

Note that the coefficients on the controls W_i do not have any causal meaning, $\gamma_u := E[W_i W_i']^{-1} E[W_i(Y_i - D_i\beta_u)]$ is just a projection.

The result that under Assumption 6(i), we estimate a weighted average of conditional ATEs dates back to at least Angrist (1998). The generalization of this result in Theorem 9 is important since it allows for a causal interpretation of linear regression estimates, even in the presence of treatment effect heterogeneity. We don't estimate the population average treatment effect, but at least we estimate *some* weighted average of treatment effects. We may motivate the linear regression approach in a number of ways:

1. We think that $\tau(w)$ is (approximately) constant, but at the same time wish to remain somewhat robust to failure of that assumption.

2. We don't particularly care about which weighted average of $\tau(w)$ we estimate, we just want to get small standard errors—then it turns out that the $\lambda_u(w)$ weighting indeed delivers the smallest possible standard errors under heteroskedasticity, see Goldsmith-Pinkham, Hull, and Kolesár (2024).

   Research Question. Does this result extend to non-binary D_i? ⊠

3. We are considering "incremental propensity score interventions" (Kennedy 2019) where the policy increases the log odds of treatment by a constant amount (imagine the treatment is determined by a logit model, and we increase the intercept). Then, β_u identifies the marginal effect of this policy.

4. In spite of the potential policy-relevance of β_u, it may seem unappealing to weight by $\lambda_u(w)$ in many contexts. However, the feature of the weighting that observations with extreme propensity scores are downweighted is in some sense necessary if we want to avoid issues with weak overlap (we can discuss this in class).

Example 2 (No covariates). In the special case in which there are no controls W_i, $\beta_u = E[\tau(W_i)]$ (why?). Furthermore, if the treatment is binary, then

$$\mathcal{V}_u = \frac{S_0^2}{n_0/n} + \frac{S_1^2}{n_1/n} + o_p(1), \tag{8}$$

where $S_d = n^{-1}\sum_i (Y_i(d) - \overline{Y}(d))^2$, and n_d is the number of units in treatment group d.⊠

Example 3. Consider a binary treatment D_i, with two covariates: a male indicator M_i, and indicators for completing high school and college, H_i and C_i. To make things simple, suppose a third of the population completed college, and a third dropped out of high school, and that these fractions are the same between men and women. Suppose men comprise half the population. Suppose also that men who completed high

school, and women who completed college are treated; others are not treated. As an exercise, you can show that $\tilde{D}_i = (-1/6, 1/3, -1/6)$ for men who are high-school dropouts, completed high school, and completed college, respectively. For women, $\tilde{D}_i = (1/6, -1/3, 1/6)$. Hence, $E[\tilde{D}_i D_i] = 1/18$, and the weights are equal to $(0, 6, -3)$ for men, and $(0, 0, 3)$ for women. Suppose that $Y_i(0) = M_i + H_i + C_i$, while $Y_i(1) = M_i + H_i + C_i + M_i C_i$, so that only college males benefit from the treatment. Then β_{u} is negative. ⊠

Remark 11 (Convex weights). There are at least two ways to motivate the interest in convex weights. First, convexity ensures the estimand captures average effects for *some* well-defined (and characterizable) subpopulation. Second, it prevents what Small et al. (2017) call a sign-reversal: that if $\tau(w)$ has the same sign for all w ($+, 0$ or $-$), then the estimand will also have this sign. Blandhol et al. (2022) call such estimands "weakly causal". On the other hand, convexity is neither necessary nor sufficient for policy relevance. In particular, if some policy induces flows both in and out of treatment, then we'd want to put negative weight on the policy "defiers" to estimate the effect of the policy.

Remark 12. Of course, the $\lambda_{\text{u}}(w)$ weighting is not the only possible choice—depending on the ultimate research goal, we may be interested in the unweighted ATE, the ATE for the treated, or a particular policy counterfactual. These would all require different approaches. The simplest one is to regress Y_i onto W_i separately in the samples with $D_i = 1$ and $D_i = 0$; the fitted values will estimate $\mu(d, W_i)$, and differences between the fitted values will yield estimates of covariate-specific treatment effects $\tau(W_i)$. We then average these estimates over the people in our sample to obtain an estimate of the unweighted ATE, $E[\tau_i]$.

We can actually do this all in one step by running a regression with interactions, where we demean the covariates before interacting (we don't interact D_i with the intercept)

$$Y_i = W_i'\gamma + D_i'\tau + \sum_i D_i(W_i - \overline{W})\kappa + u_i.$$

As an exercise, you can show that if $E[\tau_i \mid W_i]$ is linear in W_i, then $\tau + (W_i - E[W_i])'\kappa = E[\tau_i \mid W_i]$. Thus $\tau = E[\tau_i]$–the coefficient on D_i in the above regression identifies the ATE, while $(W_i - \overline{W})'\hat{\kappa}$ gives the estimate of the difference between the covariate-specific treatment effect and the ATE.

In practice, people don't run this interacted regression as often as they should. Perhaps one reason is that under weak overlap, it gives large standard errors, and that it's not even feasible if overlap fails. In my view, while targeting the estimand β_{u} is a reasonable first step of any analysis, it is often useful to explore the role of treatment effect heterogeneity in a second step. In practice, this doesn't have to take the form of a regression with full interactions if such regression is infeasible or too noisy, but one may want to interact the treatment with a few covariates one suspects may be key drivers of heterogeneity.

Research Question. If we put in some interactions, but not all, what does τ in the above regression estimate? ⊠

Question 4. What are some alternatives to a regression with interactions if we want to estimate the (unweighted) average treatment effect $E[\tau(W_i)]$?

CAUSAL INFERENCE CONDITIONAL ON X We take the estimand of interest to be β_{cx}. This means we can't make use of any propensity score restrictions. Instead, we need to take a model-based approach, and assume version (ii) of Assumption 6. Then, in direct analogy to the previous case, this estimand β_{cx} can be expressed as a weighted average of conditional ATEs $\tau(w)$, but with weights $\lambda_{\text{cx}}(X_i) = \ddot{D}_i D_i$.

The previous discussion applies. In particular, even if we run a randomized experiment, we are not guaranteed that the weights will be positive or that version (ii) of Assumption 1 will hold.

CAUSAL INFERENCE ON FINITE POPULATION If $\mathcal{D}_n$ is the population of interest, it is natural to treat W_i and the potential outcomes as fixed—we're interested in what changes if we change the treatment status of the units, but keep everything else fixed. In other words, the only uncertainty comes from the fact that we observe the units under a particular realization of the treatment vector D, but we *could have observed* a different realization—I'll call this design-based uncertainty. In other words, the descriptive estimand of interest is

$$\beta_{\text{ce}} = \frac{\sum_i E[\ddot{D}_i Y_i \mid W_i, Y_i(\cdot)]}{\sum_i E[\ddot{D}_i^2 \mid W_i, Y_i(\cdot)]}, \qquad \theta_{\text{ce}} = \operatorname*{argmin}_\theta \sum_{i=1}^n E[(Y_i - X_i'\theta)^2 \mid W_i, Y_i(\cdot)],$$

with $\gamma_{\text{ce}} = (W'W)^{-1}W'E[(Y - D\beta_{\text{ce}}) \mid W, Y(\cdot)]$, and $Y(\cdot) = \{Y_i(d) \colon i = 1, \ldots, n,\ d \in \mathbb{R}\}$.

Suppose version (i) of Assumption 6 holds. Since we're conditioning on the potential outcomes, version (ii) will not do us any good. Then an iterated expectations argument shows that

$$\beta_{\text{ce}} = \frac{\sum_i \lambda_{\text{u}}(W_i)\tau_i}{\sum_i \lambda_{\text{u}}(W_i)},$$

a finite-sample analog of the expression for β_{u}. In particular, we're estimating a weighed average of treatment effects *for the people in the sample*, that is of $\tau_1, \ldots, \tau_n$.

Define $\epsilon_{\text{ce},i} = Y_i - D_i\beta_{\text{ce}} - W_i'\gamma_{\text{ce}}$.

- If the treatment effects τ_i are constant, so that $\beta = \tau$, then under this approach, we're doing inference conditional on $\epsilon_{\text{ce},i} = Y_i(0) - W_i'\gamma_{\text{ce}}$, so "ce" accordingly stands for "conditional on epsilon". In this case, $\gamma_{\text{ce}} = (W'W)^{-1}W'Y(0)$. This is effectively the opposite of doing inference conditional on X.

Like in the two previous cases, the unconfoundedness and linearity conditions that Assumption 6 entails serve only to ascribe casual meaning to β_{ce}—inference remains the same whether the condition holds. We'll impose regularity conditions similar to those in Abadie et al. (2020),

Assumption 7. $n^{-1}\sum_i \|W_i\|^{4+\eta}$, $n^{-1}\sum_i E[Y_i \mid Y(d), W]^{4+\eta}$, and $n^{-1}\sum_i E[D_i^{4+\eta} \mid W_i]$ are bounded. $E[X'X \mid W]/n$ converges to a positive definite limit. Conditional on $Y(\cdot)$ and W, the treatment is assigned independently.

Then under Assumption 7,

$$\sqrt{n}(\hat{\beta} - \beta_{\text{ce}}) = \mathcal{N}(0, \mathcal{V}_{\text{ce},n}) + o_p(1),$$

where

$$\mathcal{V}_{\text{ce},n} = n\frac{\sum_i \text{var}(\epsilon_{\text{ce},i}\ddot{D}_i \mid W_i)}{(\sum_i E[\ddot{D}_i^2 \mid W_i])^2}.$$

Proof. All expectations are conditional on $Y(d)$ and W. We first establish a couple of preliminary results

1. $n^{-1/2}W'\epsilon = O_p(1)$. To show this, note $n^{-1/2}W'\epsilon$ has mean zero, with variance given by $n^{-1}\sum_i E[\epsilon_i^2]W_{ij}^2$, which is bounded by Assumption 7, since $E[\epsilon_i^2] \leq E[Y_i^2] + E[D_i]^2\beta_{\text{ce}} + (W_i'\gamma_{\text{ce}})^2$ by the C_r inequality. Therefore, the claim follows by Chebyshev's inequality.
2. Let $\hat{\delta} = (W'W)^{-1}W'D$. Then $\hat{\delta} - \delta = o_p(1)$. This follows because $\hat{\delta} - \delta = (W'W/n)^{-1} \cdot n^{-1}W'(D - W\delta)$. The first term is $O(1)$ by Assumption 7. The second term is mean zero with variance of the jth term given by

   $$n^{-2}\sum_i W_{ij}^2 \,\text{var}(D_i \mid W) \leq n^{-1}\sqrt{n^{-1}\sum_i W_{ij}^4 \cdot n^{-1}\sum_i \text{var}(D_i \mid W)^2},$$

   which converges to zero since the term inside the square root is bounded by Assumption 7.
3. $n^{-(1+\eta/4)}\sum_i E|\ddot{D}_i\epsilon_i|^{2+\eta/2} \to 0$. This again follows since by Assumption 7, the quantity $n^{-1}\sum_i E|\ddot{D}_i\epsilon_i|^{2+\eta/2}$ is bounded.
4. $(n^{-1}\sum_i \tilde{D}_i^2)^{-1} = (n^{-1}\sum_i \sigma_D^2(W_i)) + o_p(1)$. This follows since $\sum_i \tilde{D}_i^2 = \sum_i(\ddot{D}_i - W_i(\hat{\gamma} - \gamma))^2$.

Then, using the first two results yields,

$$\begin{aligned} n^{-1/2}\tilde{D}'(Y - \tilde{D}\beta_{\text{ce}}) &= n^{-1/2}D'(I - H_W)\epsilon = n^{-1/2}\ddot{D}'\epsilon - n^{-1/2}(\hat{\delta} - \delta)'W'\epsilon \\ &= n^{-1/2}\ddot{D}'\epsilon + o_p(1). \end{aligned}$$

Using the third result to verify the Lyapunov condition then yields that

$$n^{-1/2}\ddot{D}'\epsilon = n^{-1/2}\sum_i(\ddot{D}_i\epsilon_i - E[\ddot{D}_i\epsilon_i]) = \mathcal{N}(0, n^{-1}\sum_i \text{var}(\epsilon_i\ddot{D}_i)) + o_p(1),$$

since $\sum_i E[\ddot{D}_i\epsilon_i] = 0$. Combining this with the last result then concludes the proof for asymptotic normality of $\hat{\beta}$. □

Again, EHW standard errors will be conservative in this case. Abadie et al. (2020) propose an alternative standard error estimator that projects $\ddot{D}_i\hat{\epsilon}_i$ onto the covariates W_i to take out part of its mean. There is a related literature on getting tighter variance estimates in the context of randomized experiments (see, for example Aronow, Green, and Lee 2014).

Example 2 (continued). In the special case in which there are no controls and the treatment is binary, $\beta_{\text{ce}} = \frac{1}{n}\sum_i \tau_i$, and the variance simplifies to

$$\mathcal{V}_{\text{ce},n} = \frac{S_1^2}{n_1/n} + \frac{S_0^2}{n_0/n} - S_\tau^2, \tag{9}$$

where $S_\tau^2 = n^{-1}\sum_i(\tau_i - \beta_{\text{ce}})^2$ is the variance of the treatment effect. This is the additional factor in the variance relative to eq. (8). This variance is known as the Neyman (1923) variance. ⊠

- If the treatment effects are constant, then $\mathcal{V}_{\text{ce},n} = \mathcal{V}_{\text{u}} + o_p(1)$. So again, like it was the case for descriptive estimands, EHW standard errors will be conservative only if the regression is "misspecified" (in the sense that it implicitly assumes constant treatment effects, which is incorrect).

Remark 13 (Finite superpopulation). For some descriptive exercises, we may observe a non-negligible fraction of the superpopulation of interest, or perhaps the whole superpopulation. Examples include tracking the evolution of the gender wage gap over time (perhaps adjusted for education and other observables) using Census data, descriptive analysis of student outcomes when we have data on all students in a district, or observing a 20% random sample of Medicate part D beneficiaries, as in Einav, Finkelstein, and Schrimpf (2015).

Clearly, in absence of measurement error, for descriptive inference, if we observe the whole population, there is no sampling uncertainty left: the standard errors are zero, in contrast to the "causal" standard errors $\mathcal{V}_{\text{ce}}$ above. More generally, suppose the superpopulation is finite, and the sample comprises a fraction ρ of the population—it'll make a difference whether we're doing descriptive or causal inference. In particular, Abadie et al. (2020) show that the asymptotic variance for descriptive inference is given by $(1-\rho)\mathcal{V}_{\text{u},11}$, while the asymptotic variance for causal inference is given by $(1-\rho)\mathcal{V}_{\text{u},11} + \rho\mathcal{V}_{\text{ce},11}$.

1.4. *Causal estimands with multiple treatments*

See Goldsmith-Pinkham, Hull, and Kolesár (2024).

2. CLUSTERED STANDARD ERRORS AND WHEN TO USE THEM

In many cases of practical interest, one may worry that the standard errors based on $\hat{V}_{\text{EHW}}$ do not adequately reflect the uncertainty in the estimate $\hat{\beta}$.

2.1. *Clustered sampling*

The simplest instance when $\hat{V}_{\text{EHW}}$ leads to misleading inference arises when we're interested in inference on β_{u}, and Assumption 2 is violated. Instead, the sampling is clustered: a subset S of clusters were sampled randomly from an infinite superpopulation of clusters, and in the second stage, n_s units were sampled randomly from the sampled clusters (potentially all units are sampled in the second stage, and n_s may depend on what type of cluster we sampled).

Question 5. What if the set S of the sampled clusters comprises the whole population?

The total sample size is $n = \sum_{s=1}^{S} n_s$. In this case the components $X_i \epsilon_{\text{u},i} = X_i(Y_i - X_i'\theta_{\text{u}})$ of the sum in eq. (6) are not independent across i in repeated samples—they are only independent across clusters $s(i)$ that the observations belong to. Instead, when we are applying the CLT, we need to treat the sums $\sum_{i:\, s_i = s} X_i \epsilon_{\text{u},i}$ as independent. Then, under regularity conditions (we'll discuss them next time),

$$\mathcal{V}_{\text{ce}}^{-1/2}(\hat{\theta} - \theta_{\text{u}}) \xrightarrow{d} \mathcal{N}(0, I),$$

where

$$\mathcal{V}_{\text{uc}} = (X'X)^{-1} \sum_{s} \sum_{i,j:\, s(i)=s(j)=s} \epsilon_{\text{u},i}\epsilon_{\text{u},j} X_i X_j' (X'X)^{-1}.$$

The asymptotic variance can be estimated using the LZ estimator.

- Notice that here the uncertainty primarily comes from the fact that we sampled the S clusters at hand, but we could have sampled a different set of S clusters.
- For inference on β_{ce}, there is no need to cluster so long as the treatment D_i is assigned independently across units.

2.2. *Clustered assignment*

The second clear reason for clustering occurs when the treatment assignment D_i is clustered, and we are interested in β_{ce} or β_{u}.

2.3. *General considerations*

There are many other cases apart from clustered sampling when we may be worried that standard errors based on $\hat{V}_{\text{EHW}}$ may be misleading. *The key to thinking through whether one should cluster is to consider correlations of $X_i \epsilon_i = X_i(Y_i - X_i'\theta)$ across units within a cluster.* This, in turn, depends on how we think about repeated sampling, and how we defined θ and hence ϵ_i.

1. For inference on β_{u}, we need to worry about the sampling process. How is the sample (Y_i, D_i, W_i) drawn from the population? Is the treatment assigned in a way that's different from how the unit i is drawn (say, by the experimenter)? If there is clustering in either the sampling of the units, or assignment of the treamtent, we need to cluster.

   For example, suppose we want to predict test scores using some background characteristics X_i of students. If we sample classrooms of students, so that students within the same classroom are sampled together, and there are classrooms missing in our data, then we need to cluster the standard errors.

   Similarly, we need to cluster the standard errors if we draw the units i.i.d., but offer treatments to students that only vary across classrooms. Then W_i are i.i.d., but not (Y_i, D_i).

   Note that with clustered sampling, clustering is necessary even if we include cluster fixed effects (for cases where the assignment is correlated, but not perfectly correlated withing clusters).[5]

2. For inference on β_{ce} we don't need to worry about how (Y_i, W_i) is correlated across units. We only worry about assignment of the treatment D_i: if it's clustered, then we generally need to cluster the standard errors. In this case, the clustered standard errors will generally be conservative (unless the treatment effect is constant), just like EHW standard errors under independent treatment assignment. Interestingly, Young (2019) reports that in 12 papers in his sample, the authors didn't cluster even though the treatment is applied to groups: so the standard errors in those papers are not correct.

   ~~**Research question**~~: Is it straightforward to adapt some alternative standard error formulas, such as those proposed in Abadie et al. (2020) to the case with clustered assignment? What about descriptive inference under misspecification? **Answer:** see the job market paper by Ruonan Xu.

   We do not need to worry about the correlation structure of ϵ_i here. In particular, if X_i is randomly assigned and independent across i, then we do not need to cluster even if ϵ_i is correlated across i. Similarly, as pointed out in Barrios et al. (2012), if X_i is independent across known clusters (e.g. states), then we don't need to worry about whether the correlations in ϵ_i spill over state boundaries: clustering on state will be sufficient.

3. For inference on β_{cx} that's valid conditional on X, things are more complicated. What matters here is whether, conditional on X, the errors ϵ_i are correlated. But since ϵ_i is a residual, it's a bit harder to think through.

   Either explicitly or implicitly, this is the case that people most often have in mind when they discuss clustering. For example Cameron and Miller (2015, p. 320) write:

5. Unless there is no treatment effect heterogeneity; then $\beta_{\text{u}} = \beta_{\text{ce}}$, so have now effectively eliminated sampling/extrapolation uncertainty. Now, if the units are assigned to treatment with a cluster-specific probability, and we include fixed effects, $\ddot{D}_i\epsilon_i$ will not be correlated within clusters. See Abadie et al. (2023).

"The key assumption is that the errors are uncorrelated across clusters while errors for individuals belonging to the same cluster may be correlated" Often, researchers also take this approach of doing inference that's conditional on X even for causal inference—what they have in mind is that eq. (2) is a structural model, with ϵ_i being "unobserved shocks". We want to do inference under a sampling process in which the individuals we observe receive different unobserved shocks—this is the logic underlying "model-based" approaches to inference. We want to do inference conditional on X (treat it as fixed, and treat the sample at hand as the population of interest), and think of "repeated sampling" as drawing different realizations of ϵ_i. Then, if there are within-cluster correlations of these shocks, we need to cluster.[6]

However, because the correlations 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. How do we know what level to cluster at? Should it be counties, or states? How do we know that we shouldn't instead use Conley (1999) standard errors, or some other spatial approach? In addition, as we discussed in Section 1.3, there are issues with interpretation of the estimand under treatment effect heterogeneity.

If we have a clear idea about the population of interest, and about the sampling or assignment mechanism, the issues discussed in the last point can be avoided if we do not insist that inference be conditional on X. This is not always so clear.

Research Question. How do we think about these issues in economic history papers? ☒

Example 4. As an example, consider Becker and Woessmann (2009), who are interested in seeing whether the economic prosperity of Protestant regions is higher because instruction in reading the Bible led to generation of human capital.

They use county-level data from late-nineteenth-century Prussia, using distance to Wittenberg as an instrument for Protestantism to see the effect of Protestantism on economic prosperity, and also on literacy. If the instrument is as good as randomly assigned, how do we think about the assignment process? ☒

Remark 14. We have seen that to decide whether to cluster, we need to think about whether $X_i\epsilon_i$ is correlated within clusters. **The data are only partially informative about this: one cannot use the data alone to decide whether to cluster**, because the data only tell us about correlations in $X_i\hat{\epsilon}_i$. Fox instance, the data don't tell us whether there are clusters in the population of interest that we have not sampled. In other words, the data are informative about whether clustering matters, but not whether one should cluster.

Example 5. To illustrate this point, consider the following example due to Abadie et al. (2023): suppose we assign treatment with probability $1/2$ to everyone, independently

6. Indeed, this is the approach taken in early paper on clustering, such as Kloek (1981) or Moulton (1990), who propose to fix the standard errors by modeling the correlation structure of the errors.

of everything else. There is a large population of clusters, and we sample 100 of them. We then sample 1000 units from each cluster. The treatment effect in half the clusters is 1, and it's -1 in the other half. $Y_i(0)$ has mean zero in each cluster.

The clustering in this example matters: it changes the standard errors by an order of magnitude. Furthermore, both clustered and robust standard errors are correct, but for different estimands. What are these estimands?

This example can be thought of as a stylized version of project STAR, which contains data on 80 schools, but assignment to a small or regular classroom was done at individual level. Here it's possible to cluster the standard errors by classroom (done by Krueger 1999), school, or not cluster (done in a reanalysis by Goldsmith-Pinkham, Hull, and Kolesár 2024). ⊠

Remark 15 (Testing correlations). Whether clustering adjustment will *matter* in a given sample depends on the within-cluster correlation of the product $X_i\hat{\epsilon}_i$. If ϵ_i and X_i are independent (as is the case, say, under random assignment of X_i when treatment effects are constant), then one can check the correlations in the residuals and correlations in X_i separately—if both are present, then clustering will matter. However, in general, as pointed out in Abadie et al. (2023), the presence of these correlations is neither sufficient nor necessary for clustering adjustments to matter. In Example 5, D_i is not correlated within clusters, and neither is $\hat{\epsilon}_i \approx D_i\hat{\tau}$. But the product $X_i\hat{\epsilon}_i$ is correlated.

REFERENCES

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* 138, no. 1 (February): 1–35. https://doi.org/10.1093/qje/qjac038.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Marc Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88, no. 1 (January): 265–296. https://doi.org/10.3982/ECTA12675.

Abadie, Alberto, Guido W. Imbens, and Fanyin Zheng. 2014. "Inference for Misspecified Models With Fixed Regressors." *Journal of the American Statistical Association* 109 (508): 1601–1614. https://doi.org/10.1080/01621459.2014.928218.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66, no. 2 (March): 249–288. https://doi.org/10.2307/2998558.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." Chap. 23 in *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, vol. 3A, 1277–1366. Amsterdam: Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72.

Aronow, Peter M., Donald P. Green, and Donald K. K. Lee. 2014. "Sharp Bounds on the Variance in Randomized Experiments." *The Annals of Statistics* 42, no. 3 (June): 850–871. https://doi.org/10.1214/13-AOS1200.

Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár. 2012. "Clustering, Spatial Correlations, and Randomization Inference." *Journal of the American Statistical Association* 107, no. 498 (June): 578–591. https://doi.org/10.1080/01621459.2012.682524.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124, no. 2 (May): 531–596. https://doi.org/10.1162/qjec.2009.124.2.531.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94, no. 4 (September): 991–1013. https://doi.org/10.1257/0002828042002561.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. *When Is TSLS Actually LATE?* Working Paper 29709. Cambridge, MA: National Bureau of Economic Research, August. https://doi.org/10.3386/w29709.

Cameron, Colin A., and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372. https://doi.org/10.3368/jhr.50.2.317.

Card, David, and Alan B. Krueger. 1992. "School Quality and Black-White Relative Earnings: A Direct Assessment." *The Quarterly Journal of Economics* 107, no. 1 (February): 151–200. https://doi.org/10.2307/2118326.

Chatterjee, Samprit, and Ali S. Hadi. 1986. "Influential Observations, High Leverage Points, and Outliers in Linear Regression." *Statistical Science* 1, no. 3 (August): 379–416. https://doi.org/10.1214/ss/1177013622.

Conley, Timothy G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92, no. 1 (September): 1–45. https://doi.org/10.1016/S0304-4076(98)00084-0.

Cox, David Roxbee. 1958. "Some Problems Connected with Statistical Inference." *The Annals of Mathematical Statistics* 29, no. 2 (June): 357–372. https://doi.org/10.1214/aoms/1177706618.

Cox, David Roxbee, and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman & Hall.

Davidson, James. 1994. *Stochastic Limit Theory.* Oxford, UK: Oxford University Press. https://doi.org/10.1093/0198774036.001.0001.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf. 2015. "The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D." *The Quarterly Journal of Economics* 130, no. 2 (May): 841–899. https://doi.org/10.1093/qje/qjv005.

Fryer, Roland G, and Steven D Levitt. 2013. "Testing for Racial Differences in the Mental Ability of Young Children." *American Economic Review* 103, no. 2 (April): 981–1005. https://doi.org/10.1257/aer.103.2.981.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *American Economic Review* 90, no. 4 (September): 715–741. https://doi.org/10.1257/aer.90.4.715.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2024. "On Estimating Multiple Treatment Effects with Regression" (February). arXiv: 2106.05024.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396 (December): 945–960. https://doi.org/10.1080/01621459.1986.10478354.

Karlan, Dean, and John A. List. 2007. "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment." *The American Economic Review* 97, no. 5 (December): 1774–1793. https://doi.org/10.1257/aer.97.5.1774.

Kennedy, Edward H. 2019. "Nonparametric Causal Effects Based on Incremental Propensity Score Interventions." *Journal of the American Statistical Association* 114, no. 526 (April): 645–656. https://doi.org/10.1080/01621459.2017.1422737.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133, no. 1 (February): 237–293. https://doi.org/10.1093/qje/qjx032.

Kloek, T. 1981. "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated." *Econometrica* 49, no. 1 (January): 205–207. https://doi.org/10.2307/1911134.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114, no. 2 (May): 497–532. https://doi.org/10.1162/003355399556052.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis for Generalized Linear Models." *Biometrika* 73, no. 1 (April): 13–22. https://doi.org/10.1093/biomet/73.1.13.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *The Review of Economics and Statistics* 72, no. 2 (May): 334–338. https://doi.org/10.2307/2109724.

Neyman, Jerzy. 1923. "Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych." *Roczniki Nauk Rolniczych* 9 (1): 1–51. Dorota M. Dabrowska and Terence P. Speed, trans. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5, no. 4 (1990): 465–472. https://doi.org/10.1214/ss/1177012031.

Robins, James M., Steven D. Mark, and Whitney K. Newey. 1992. "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders." *Biometrics* 48, no. 2 (June): 479–495. https://doi.org/10.2307/2532304.

Rosenbaum, Paul R., and Donald Bruce Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70, no. 1 (April): 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment." *Journal of the American Statistical Association* 75, no. 371 (September): 591–593. https://doi.org/10.2307/2287653.

Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science* 32, no. 4 (November): 561–579. https://doi.org/10.1214/17-STS623.

Velleman, Paul F., and Roy E. Welsch. 1981. "Efficient Computing of Regression Diagnostics." *The American Statistician* 35, no. 4 (November): 234–242. https://doi.org/10.1080/00031305.1981.10479362.

von Bahr, Bengt, and Carl-Gustav Esseen. 1965. "Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$." *The Annals of Mathematical Statistics* 36, no. 1 (February): 299–303. https://doi.org/10.1214/aoms/1177700291.

Wooldridge, Jeffrey Marc. 2010. *Econometric Analysis of Cross Section and Panel Data.* 2nd ed. Cambridge, MA: MIT Press. ISBN: 0-262-23258-8.

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics* 14, no. 4 (October): 10. https://doi.org/10.1080/07350015.1996.10524677.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134, no. 2 (May): 557–598. https://doi.org/10.1093/qje/qjy029.