THE BOOTSTRAP AND SUBSAMPLING

Michal Kolesár*

March 18, 2024

In this lecture, we discuss the bootstrap: a resampling method for approximating the distribution (or particular features of the distribution) of a given statistic. First we discuss the mechanics of how it works and how to implement it in practice. Second, we discuss why one may want to use it. Finally, we discuss situations in which it doesn't work at all, and what one can do in such situations.

1. THE BOOTSTRAP

1.1. *The usual approach to inference*

Let us first review the usual approach to inference, using slightly different notation to set us up for the bootstrap. I follow the setup in Lehmann and Romano (2005, Chapter 15.4). We observe $X_i \sim F \in \mathcal{F}$, $i = 1, \dots, n$. The parameter space $\mathcal{F}$ can be parametric or non-parametric. Parametric just means here that there is a finite-dimensional parameter $\gamma \in \Gamma \subseteq \mathbb{R}^k$ such that $\mathcal{F} = \{F_\gamma \colon \gamma \in \Gamma\}$; the parameter space is non-parametric otherwise. We're interested in some parameter $\theta(F) \in \Theta = \{\theta(F) \colon F \in \mathcal{F}\} \subseteq \mathbb{R}^d$. Denote the sample by $\mathbf{X}_n = (X_1, \dots, X_n)$.

We'd like to construct a confidence interval (CI) for θ using the real-valued function $R_n(\mathbf{X}_n, \theta(F))$. In the bootstrap literature, this function is called a *root* (the name is due to Beran 1984). In principle, R_n can be any real-valued function, but in practice, it typically takes one of the four following forms:

Example 1. If we have available a $\sqrt{n}$-consistent estimator $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ of θ, then there are two options for the root:

1. $R_n(\mathbf{X}_n, \theta) = \sqrt{n}(\hat{\theta}_n - \theta)$; or
2. $R_n(\mathbf{X}_n, \theta) = |\sqrt{n}(\hat{\theta}_n - \theta)|$.

If we also have an estimator of its asymptotic variance, $\hat{\sigma}_n^2 = \hat{\sigma}^2(\mathbf{X}_n)$, then there are two further options:

*Email: mkolesar@princeton.edu

3. the t-statistic, $R_n(\mathbf{X}_n, \theta) = \sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_n$,

4. the absolute value of the t-statistic, $R_n(\mathbf{X}_n, \theta) = \sqrt{n}|\hat{\theta}_n - \theta|/\hat{\sigma}_n$. ⊠

Since both $\mathbf{X}_n$ and θ depend on F, the distribution of R_n will also in general depend on F. Let $J_n(\cdot, F)$ denote this distribution, so that $J_n(r, F) = P_F(R_n \leq r)$. Define the quantiles of J_n,

$$J_n^{-1}(\alpha, F) = \inf\{r \colon J_n(r, F) \geq \alpha\}.$$

If we knew these quantiles, we could construct a (finite-sample valid) CI for θ as

$$\mathrm{CI}^* = \{\theta \in \Theta \colon R_n(\mathbf{X}_n, \theta) \leq J_n^{-1}(1-\alpha, F)\},$$

or, depending on the root, we could make use of both tails of the distribution, yielding

$$\widetilde{\mathrm{CI}}^* = \{\theta \in \Theta \colon J_n(\alpha/2, F) \leq R_n(\mathbf{X}_n, \theta) \leq J_n^{-1}(1-\alpha/2, F)\}.$$

Example 1 (continued). For two-sided CIs, using CI^* for cases 2 and 4, and $\widetilde{\mathrm{CI}}^*$ for cases 1 and 3, we get

$$\begin{aligned}
\widetilde{\mathrm{CI}}_1^* &= [\hat{\theta}_n - J_n^{-1}(1-\alpha/2, F)/\sqrt{n}, \hat{\theta}_n - J_n(\alpha/2, F)/\sqrt{n}], \\
\mathrm{CI}_2^* &= \{\hat{\theta}_n \pm J_n^{-1}(1-\alpha, F)/\sqrt{n}\}, \\
\widetilde{\mathrm{CI}}_3^* &= [\hat{\theta}_n - J_n^{-1}(1-\alpha/2, F)\hat{\sigma}_n/\sqrt{n}, \hat{\theta}_n - J_n(\alpha/2, F)\hat{\sigma}_n/\sqrt{n}], \\
\mathrm{CI}_4^* &= \{\hat{\theta}_n \pm J_n^{-1}(1-\alpha, F)\hat{\sigma}_n/\sqrt{n}\}.
\end{aligned}$$

Note the reversal of the quantiles! ⊠

This observation leads to the idea of constructing CIs using the pivot method, which you saw in 517. The idea of the pivot method is to be clever about choosing the root appropriately so that its distribution J_n doesn't depend on F.

Example 2. Suppose $X_i \sim \mathcal{N}(\theta, \sigma^2)$, $i = 1, \ldots, n$. If we set $R_n = \sqrt{n}(\overline{X}_n - \theta)/S_n$, where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the sample mean, and $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ is the sample variance, then the t-statistic R_n is distributed t with $n-1$ degrees of freedom, independently of θ or σ^2. So R_n is a pivot. Similarly, $R_n = |\sqrt{n}(\overline{X}_n - \theta)/S_n|$ is a pivot. Plugging in quantiles of the t-distribution to $\widetilde{\mathrm{CI}}_3^*$ or CI_4^* then leads to the usual two-sided CI, $\overline{X}_n \pm t_{1-\alpha/2}(n-1) \cdot S_n/\sqrt{n}$. ⊠

Exact pivots are rare. The standard way of doing inference (e.g. inference based on extremum estimators) is to find R_n that is asymptotically pivotal, that is find R_n such that $J_n(r, F) \to J_\infty(r)$ at all continuity points of J_∞ (this is just the definition of convergence in distribution). Then, provided J_∞ is continuous, we can use quantiles of J_∞ for inference.[1]

1. By Lemma 21.2 in van der Vaart (1998), $J_n(r, F) \to J_\infty(r)$ at all continuity points of J_∞ if and only if $J_n^{-1}(r, F) \to J_\infty^{-1}(r)$ at all continuity points.

Example 1 (continued). In "regular" models, we are typically able to show that $\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2)$, and that $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$, so that for all $F \in \mathcal{F}$, $R_n = \sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_n \Rightarrow \mathcal{N}(0,1)$, and J_∞ is thus the standard normal distribution. This leads to the "standard" asymptotic CIs for θ given by

$$\widetilde{\mathrm{CI}}_3^A = \mathrm{CI}_4^A = \{\hat{\theta}_n \pm z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n}\}. \qquad \boxtimes$$

1.2. *The bootstrap idea*

The bootstrap idea is due to Efron (1979), who gave it the name.[2] The idea of the bootstrap is to approximate the distribution $J_n(\cdot, F)$ by $\hat{J}_n = J_n(\cdot, \hat{F}_n)$, where $\hat{F}_n$ is some estimate of F. There are many versions of the bootstrap, each with different estimator of F. In the *nonparametric (or empirical) bootstrap* considered in Efron (1979), $\hat{F}_n$ is the empirical cumulative distribution function (CDF). The *parametric bootstrap* uses the estimator $F_{\hat{\gamma}_n}$, where $\hat{\gamma}_n$ is some reasonable estimator of γ.

Question 1. Such as?

This yields the bootstrap CIs

$$\mathrm{CI}^B = \{\theta \in \Theta \colon R_n(\mathbf{X}_n, \theta) \leq \hat{J}_n^{-1}(1-\alpha)\},$$

or, using both tails of the distribution,

$$\widetilde{\mathrm{CI}}^B = \{\theta \in \Theta \colon \hat{J}_n(\alpha/2) \leq R_n(\mathbf{X}_n, \theta) \leq \hat{J}_n^{-1}(1-\alpha/2)\}.$$

In practice, it is hard to compute the quantiles $\hat{J}_n^{-1}(\cdot)$ directly (a brute force approach requires evaluation of n^n samples), but one can always use simulation. In particular, we can:

1. Draw a sample $\mathbf{X}_n^*$ of size n from $\hat{F}_n$. For the nonparametric bootstrap, this just means sampling from the observed data $\mathbf{X}_n$ with replacement. For the parametric bootstrap, if we have a parametric model $F_\gamma(\cdot \mid z)$ for an outcome Y given a vector of regressors Z, this amounts to drawing $\mathbf{Z}_n^*$ with replacement from the empirical distribution, and drawing the outcomes Y_i^* from the distribution $F_{\hat{\gamma}}(\cdot \mid Z_i^*)$. Alternatively, one can put $\mathbf{Z}_n^* = \mathbf{Z}_n$

   Question 2. What's the difference?

2. Compute $R_n^* = R_n(\mathbf{X}_n^*, \theta(\hat{F}_n))$.

2. Tukey suggested "shotgun" because, as explained in Efron (1979), the bootstrap can "blow the head off any problem if the statistician can stand the resulting mess".

3. Repeat N times to obtain $(R^*_{n,1}, \ldots, R^*_{n,N})$. Use quantiles of this empirical distribution to approximate $\hat{J}_n^{-1}(t)$. We can make the approximation arbitrarily close by making N large.

Example 2 (continued). Suppose we didn't realize that $\sqrt{n}(\overline{X}_n - \theta)/S_n$ was a pivot, and instead we bootstrapped $R_n = \sqrt{n}(\overline{X}_n - \theta)$. Consider first the nonparametric bootstrap. Under the empirical CDF $\hat{F}_n$, the mean is $\theta(\hat{F}_n) = \int x \, \mathrm{d}\hat{F}_n(x) = \overline{X}_n$, so that $R^*_n = \sqrt{n}(\overline{X}^*_n - \overline{X}_n)$, where $\overline{X}^*_n = n^{-1}\sum_i X^*_i$ and X^*_i are drawn from $\mathbf{X}_n$ with replacement. This yields (note the reversal of the "natural" quantiles)

$$\widetilde{\mathrm{CI}}^B = [\overline{X}_n - \hat{J}_n^{-1}(0.975)/\sqrt{n}, \overline{X}_n - \hat{J}_n^{-1}(0.025)/\sqrt{n}].$$

Now consider the parametric bootstrap. Suppose we use the estimators $\hat{\gamma}_n = (\overline{X}_n, S_n^2)$ of θ and σ^2. Again, $\theta(F_{\hat{\gamma}_n}) = \overline{X}_n$. Thus, $R^*_n = \sqrt{n}(\overline{X}^*_n - \overline{X}_n)$, where X^*_i is drawn from $\mathcal{N}(\overline{X}_n, S_n^2)$. Thus, $R^*_n \sim \mathcal{N}(0, S_n^2)$. Therefore, the bootstrap CI is given by $\overline{X}_n \pm z_{1-\alpha/2}S_n/\sqrt{n}$, which is just the usual CI (except we use quantiles of the normal, rather than those of the t-distribution). ⊠

Example 1 (continued). Unless $\hat{\theta}_n$ is a plug-in estimator, $\hat{\theta}_n = \theta(\hat{F}_n)$, it will not generally be the case that $\theta(\hat{F}_n) = \hat{\theta}_n$. Nonetheless, to construct CIs for θ_n, people often approximate the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ by the bootstrap distribution of $R^*_n = \sqrt{n}(\hat{\theta}_n(\mathbf{X}^*_n) - \hat{\theta}_n)$. This leads to the *percentile bootstrap CI* (note again the reversal of the quantiles)

$$\widetilde{\mathrm{CI}}^B_1 = [\hat{\theta}_n - \hat{J}_n^{-1}(1-\alpha/2)/\sqrt{n}, \hat{\theta}_n - \hat{J}_n^{-1}(\alpha/2)/\sqrt{n}],$$

a bootstrap analog of $\widetilde{\mathrm{CI}}^*_1$. Using the t-statistic instead leads to the *percentile-t bootstrap CI*. In this case, we take bootstrap draws $R^*_n = \sqrt{n}(\hat{\theta}_n(\mathbf{X}^*_n) - \hat{\theta}_n)/\hat{\sigma}_n(\mathbf{X}^*)$, leading to the CI

$$\widetilde{\mathrm{CI}}^B_3 = [\hat{\theta}_n - \hat{J}_n^{-1}(1-\alpha/2)\hat{\sigma}_n/\sqrt{n}, \hat{\theta}_n - \hat{J}_n^{-1}(\alpha/2)\hat{\sigma}_n/\sqrt{n}],$$

which is similar to the usual CI ($\widetilde{\mathrm{CI}}^A_3$), except we replaced the normal quantiles $z_{1-\alpha/2}$ and $z_{\alpha/2}$ (i.e. 1.96 and -1.96 for $\alpha = 0.05$) with quantiles from the bootstrap distribution. In practice, people often instead use the CI $[\hat{\theta}_n + \hat{J}_n^{-1}(\alpha/2)/\sqrt{n}, \hat{\theta}_n + \hat{J}_n^{-1}(1-\alpha/2)/\sqrt{n}]$, which is known as *Efron's percentile method*. The logical justification for this method is less strong, but it often works fine in practice, and the difference is asymptotically negligible if the asymptotic distribution of $\hat{\theta}_n - \theta$ is symmetric around zero. One could also of course bootstrap the absolute values, leading to the CIs:

$$\mathrm{CI}^B_2 = \{\hat{\theta}_n \pm \hat{J}_n^{-1}(1-\alpha)/\sqrt{n}\},$$
$$\mathrm{CI}^B_4 = \{\hat{\theta}_n \pm \hat{J}_n^{-1}(1-\alpha)\hat{\sigma}_n/\sqrt{n}\},$$

which ensures that the CI is symmetric around $\hat{\theta}_n$. ⊠

In practice, the most common method of using the bootstrap is to use it to estimate the standard error of an estimator, so that the reader can form a CI at whatever level

they want. For this to work, of course, we need that $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal. Then:

1. Draw a sample $\mathbf{X}_n^*$ of size n from $\hat{F}_n$, and compute $\hat{\theta}_n(\mathbf{X}_n^*)$.
2. Do this N times to obtain $(\hat{\theta}_n(\mathbf{X}_{n,1}^*), \ldots, \hat{\theta}_n(\mathbf{X}_{n,N}^*))$. Use the sample standard deviation $\hat{\sigma}_{n,*} = (N^{-1}\sum_{j=1}^N (\hat{\theta}_n(\mathbf{X}_{n,j}^*) - \overline{\hat{\theta}_n}^*)^2)^{1/2}$, $\overline{\hat{\theta}_n}^* = N^{-1}\sum_{j=1}^N \hat{\theta}_n(\mathbf{X}_{n,j}^*)$ of this bootstrap sample, an estimate of the standard error of $\hat{\theta}_n$, which leads to the CI

$$\mathrm{CI}_5^B = \{\hat{\theta}_n \pm z_{1-\alpha/2}\hat{\sigma}_{n,*}\}.$$

A variant of this approach is to approximate the second moment of R_n, replacing $\hat{\sigma}_{n,*}^2$ with $\hat{s}_n^2 = \frac{1}{N}\sum_{j=1}^N (\hat{\theta}_n(\mathbf{X}_{n,j}^*) - \hat{\theta}_n)^2$

Question 3. What if $\hat{\theta}_n$ is not asymptotically normal? What is the relationship between $\hat{\sigma}_n^*$ and s_n^*?

Example 2 (continued). We want to bootstrap the standard error of $\overline{X}_n$. We know that for the non-parametric bootstrap, the bootstrap distribution has mean $\overline{X}_n$, and second moment $n^{-1}\sum_i X_i^2$ (why?). So $\hat{\sigma}_{n,*}^2 = \mathrm{var}(\overline{X}_n^*) = \mathrm{var}(X_i^*)/n = \sum_i (X_i - \overline{X}_n)^2/n^2$, which is the usual standard error for $\overline{X}_n$. What if we use the parametric bootstrap? ⊠

1.3. *Consistency of the bootstrap*

It's usually the case that the R_n converges in distribution to some limit, so that $J_n(t, F) \to J_\infty(t, F)$ at all continuity points t of J_∞. Also, $\hat{F}_n$ converges to F for most reasonable estimators $\hat{F}_n$. For instance, for the nonparametric bootstrap, we know that $\|\hat{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$ by the Glivenko-Cantelli theorem, where $\|F\|_\infty = \sup_x |F(x)|$ is the supremum norm. Thus, $\hat{J}_n = J_n(\cdot, \hat{F}_n)$ should be close to $J_\infty(\cdot, F)$, provided that $J_n(t, F)$ is suitably continuous in its second argument. So proving that the bootstrap distribution will be close to J_∞, the asymptotic distribution of R_n requires three ingredients:

1. $\hat{F}_n$ converges to F
2. $J_n(\cdot, F)$ converges to $J_\infty(\cdot, F)$ (i.e. there exists a well-defined asymptotic distribution).
3. J_n is "continuous" in the sense that if F_n converges to F, then $J_n(\cdot, F_n)$ converges to $J_\infty(\cdot, F)$.

If J_∞ is continuous, the consistency of the bootstrap, that is

$$J_n(x, \hat{F}_n) \xrightarrow{\text{a.s.}} J_\infty(x, F) \quad \text{for all } x \tag{1}$$

then follows by an extended version of a continuous mapping theorem.[3] It is then not much more work to show that this implies that this implies consistency of the bootstrap CIs $\widetilde{\text{CI}}_1^B, \text{CI}_2^B, \widetilde{\text{CI}}_3^B, \text{CI}_4^B$. See for example, Politis, Romano, and Wolf (1999, Theorem 1.2.1).

Remark 1 (Bootstrap standard error). For showing consistency of bootstrapped standard errors (i.e. that $n\hat{\sigma}^2_{n,*} \xrightarrow{p} \sigma^2$, if $\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2)$), we need slightly stronger conditions than showing that the percentile bootstrap leads to asymptotically valid CIs. This is because convergence in distribution doesn't imply convergence of moments (we need a uniform integrability condition in addition, which may be difficult to establish). In fact, by arguments you may have seen in 517 (Remark 6 of 517 lecture note 3 when I last taught it), the asymptotic variance is not necessarily the limit of finite-sample variances, instead $\liminf_{n\to\infty} \text{var}(\sqrt{n}(\hat{\theta}_n - \theta)) \geq \sigma^2$. Hahn and Liao (2021) use the same observation to show that CI_5^B is generally conservative—the bootstrap standard error is too large in general. Much like the Monte Carlo standard error for an instrumental variables (IV) estimator will be too large in general relative to its asymptotics variance, the bootstrap variance in IV will be too large in general. If you want to report a standard error, a better idea is to take one of the bootstrap CIs and divide its length by $2 \times z_{1-\alpha}$ (take, say $\alpha = 0.05$, or put $\alpha = 0.25$, which corresponds to using interquartile range to back out the asymptotic variance).

Remark 2 (Rule of thumb). If $R_n(\mathbf{X}_n, \theta) = \sqrt{n}(\hat{\theta}_n - \theta)$, and $\hat{\theta}_n$ is an extremum estimator that's asymptotically linear by the large-sample distribution theorems for extremum estimators that you may have discussed in 518 or 519, then the bootstrap is typically valid.[4] For example, the fact that the percentile bootstrap for quantile regression is valid was shown in Hahn (1995) (this was part of his PhD dissertation). Conversely, if the distribution is not asymptotically normal, the bootstrap will typically fail. For example, the bootstrap fails to consistently estimate the asymptotic distribution of the maximum score estimator, as first pointed out in Abrevaya and Huang (2005).[5]

Research question: Is it possible to formalize some part of this rule of thumb?

Remark 3 (Parametric vs. nonparametric bootstrap). The parametric bootstrap exploits the parametric model structure, and so will often perform better than the nonparametric bootstrap. On the other hand, if the model used in the parametric bootstrap is misspecified, then the parametric bootstrap will generally be inconsistent. The nonparametric bootstrap will typically remain to yield correct inference for the appropriate pseudo-parameter.

3. The theorem says (see van der Vaart 1998, Theorem 18.11) that if $g_n(x_n) \to g(x)$ whenever $x_n \to x$, then $X_n \xrightarrow{\text{a.s.}} X$ implies $g_n(X_n) \xrightarrow{\text{a.s.}} g(X)$. Here X_n are random elements.

4. This rule of thumb is based on Theorem 1 in Mammen (1992, p. 9): Consider bootstrapping a linear functional of F (i.e. for some h, $\theta = \int h(x)\,\mathrm{d}F(x)$, and $\hat{\theta}_n = n^{-1}\sum_i h(x_i)$). Then the non-parametric bootstrap is consistent if and only if $\hat{\theta}_n$ is asymptotically normal. This suggests the bootstrap should be consistent if the statistic of interest is linear asymptotically, $\sqrt{n}(\hat{\theta}_n - \theta) = n^{-1/2}\sum_i h(x_i; \theta) + o_p(1)$, as discussed in 519.

5. It appears that some of their results are incorrect; see Kosorok (2008) and Sen, Banerjee, and Woodroofe (2010) for discussion as well as cleaner arguments for bootstrap inconsistency in a related class of models.

2. WHY USE THE BOOTSTRAP

So there are three different bootstrap CIs: the percentile bootstrap CI $\widetilde{\text{CI}}_1^B$ (or the symmetric version CI_2^B), the percentile-t bootstrap $\widetilde{\text{CI}}_3^B$ (or the symmetric version CI_4^B), and CI_5^B, based on bootstrapping the standard errors. If the asymptotic distribution is normal, we can also use the usual CI, $\widetilde{\text{CI}}_3^A$. We now discuss the relative merits of the different CIs.

2.1. *Why use the percentile bootstrap*

A clear advantage of the $\widetilde{\text{CI}}_1^B$ is that it doesn't require the user to construct a consistent estimator of the asymptotic variance. This is especially useful when an analytic approach to constructing such an estimator is complicated or intractable, as the following two examples demonstrate.

Example 3 (Quantile regression). Consider the linear quantile regression model. That is, letting $Q_\tau(Y \mid X) = F^{-1}_{Y|X}(\tau \mid X)$ denote the conditional quantile function of Y given X at quantile τ, suppose $Q_\tau(Y \mid X) = X'\theta$. Recall from 517 that $Q_\tau(Y \mid X)$ solves the minimization problem $\min_{q(X)} E\rho_\tau(Y - q(X))$, where

$$q_\tau(u) = (\tau - \mathbb{1}\{u \leq 0\})u = \tau\,\mathbb{1}\{u > 0\}u + (1-\tau)\,\mathbb{1}\{u \leq 0\}(-u).$$

is the check function. The quantile regression estimator minimizes the sample analog:[6]

$$\hat{\theta}_n = \operatorname*{argmin}_\theta \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - X_i'\theta).$$

One can show (using the tools discussed in 519) that

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0, \tau(1-\tau)E[f_u(0 \mid X_i)X_iX_i']^{-1}E[X_iX_i']E[f_u(0 \mid X_i)X_iX_i']^{-1}),$$

where $f_u(0 \mid X)$ is the conditional density of the residual $u_i = Y_i - X_i'\theta$ at 0. Here the asymptotic variance is tricky to estimate; we'd like to avoid having to obtain nonparametric estimates of conditional densities. The bootstrap allows us to do exactly that (in Stata, `bsqreg` implements these bootstrapped standard errors). ⊠

Example 4 (Counterfactuals). Consider a discrete choice problem, where $U_{ij} = X_{ij}'\theta + \epsilon_{ij}$ is the utility of person i from choice $j = 0, \ldots, J$. Suppose that ϵ_{ij} has a type-I extreme value distribution $F(\epsilon) = e^{-e^{-\epsilon}}$. Person i makes choice j if and only if $U_{ij} \geq U_{ik}$ for $k = 0, \ldots, J$. This is called the conditional logit model. By evaluating some integrals,

6. In practice, we can compute $\hat{\theta}_n$ by casting the minimization problem as a linear programming problem:

$$\min_{\theta, u_+, u_-} \tau\iota'u_+ + (1-\tau)\iota'u_-, \qquad Y = X'\theta + u_+ + u_-, \qquad u_+, u_- \geq 0.$$

one can show that the *conditional choice probabilities* are given by

$$P(Y_i = j) = \frac{e^{X_{ij}'\theta}}{\sum_{k=0}^{J} e^{X_{ik}'\theta}}.$$

We can use this to build a likelihood (how?) and estimate θ. What happens if $J = 1$? The coefficients are hard to interpret, and typically not themselves of interest. Instead, we're interested in some counterfactuals, or average elasticities, $E[h(X_{i0}, \ldots, X_{iJ}, \theta)]$, or $n^{-1}\sum_i h(X_{i0}, \ldots, X_{iJ}, \theta)$, where h is some complicated function (Exercise: suppose we want to compute the average own price elasticity; how does h look like?). So even if we make use of the usual theory to get standard errors for $\hat{\theta}_n$, using the delta method may be tedious. This issue is even more severe in more complicated discrete choice models.⊠

Question 4. What is the general lesson here?

$\widetilde{\text{CI}}_1^B$ has two further advantages relative to the percentile-t bootstrap and the usual CI:

1. It is robust to model misspecification. If the model is misspecified, it still appropriately reflects the sampling uncertainty, if we view $\hat{\theta}_n$ as an estimator of the pseudo-parameter. In contrast, the percentile-t bootstrap and the usual CI don't have this interpretation unless we use misspecification-robust standard errors.

2. Because implementing the bootstrap resampling is typically much simpler than implementing asymptotic variance estimators, the approach is also more robust to coding errors.

2.2. *Why use the percentile-t bootstrap*

If we bootstrap t-statistics, then it will typically be the case that $\hat{J}_n$ is consistent and J_∞ is pivotal. So one may use either one as an approximation to J_n, and it's not clear which approximation is better, although it's clear that the bootstrap critical values are more work.

In some cases, however, one can show that the bootstrap approximation is better, in which case the bootstrap is said to provide an *asymptotic refinement*. The intuition is that while the finite-sample distribution of an estimator will be skewed, the normal asymptotic approximation is symmetric. On the other hand, the bootstrap may be able to capture the skewness. Typically, a necessary condition for showing that the bootstrap gives such a refinement is that J_∞ is pivotal. Takeaway: it's better to bootstrap t-statistics, if you can, rather than estimators or standard errors.

Formalizing the intuition requires Edgeworth expansions—refinements of a central limit theorem (CLT). The classic reference for this is Hall (1992). The main takeaways are:

1. Typically, if J_∞ is pivotal, then the bootstrap error in approximating the distribution of $J_n(\cdot, F)$ is of the order $O(n^{-1})$, while the CLT approximation (or the percentile bootstrap) has error $O(n^{-1/2})$.

   More technical arguments are needed to show that this better approximation to the asymptotic distribution translates into more accurate coverage probabilities for confidence intervals.

2. To be able to do these Edgeworth expansion in the first place, we need more moment assumptions. For example, in the case of bootstrapping the mean, Hall (1988) showed that finiteness of three absolute moments is in fact necessary and sufficient for higher-order accuracy of the bootstrap (and if the third moment doesn't exist, the bootstrap can in fact do worse than the CLT).

3. Furthermore, we also need $J_n(\cdot, F)$ to be smooth enough. For example, while the bootstrap is consistent in Example 3, the bootstrap error is of bigger order than the usual rate $O(n^{-1/2})$; (de Angelis, Hall, and Young 1993). One could smooth the objective function or smooth the bootstrap to improve the rate (see Section 4 in Horowitz (2019) for a discussion), but doing so requires a judicious choice of bandwidth.

Remark 4. See section 3.7 of Horowitz (2001) for discussion of how to get a refinement in generalized method of moments (GMM) models: the usual approach does not give it. The original reference is Hall and Horowitz (1996).

2.3. *Less common reasons to use the bootstrap*

The bootstrap is also used to provide critical values for testing procedures. For instance, Kitagawa (2015) uses the bootstrap to develop a test for instrument validity that allows for heterogeneous treatment effects.

The bootstrap can also sometimes estimate the bias of an estimator up to some asymptotic order. The bootstrap bias estimate can then be subtracted from the original parameter estimator to give rise to a less biased estimator (at the expense of increasing the variance). See Horowitz (2001, Chapter 3.1).

2.4. *Key things to remember*

- A key requirement for bootstrap validity is that the bootstrap distribution $\hat{F}_n$ is close to the true DGP F. This means that in models with dependence, the bootstrap needs to respect the dependence. For example, with clustered data, we need to bootstrap the clusters. In panel applications, we need to resample the individuals, preserving the time-series dependence of the data. In time-series applications,

we resample blocks of the data ("block bootstrap") to preserve the dependence structure in the data

- You only get an asymptotic refinement with the percentile-t method, which is asymptotically pivotal. Otherwise, if the percentile method, or the bootstrapped standard errors yield a CI that's different from the usual CI, it is not clear which one we should prefer. Remember the reversal of the quantiles (bootstrap the absolute value of the t-statistics if you want the CI to be symmetric).
- Don't bootstrap statistics that are non-smooth or not asymptotically normal without due diligence.
- Use the sandwich formula when bootstrapping t-statistics with the non-parametric bootstrap.

3. FAILURE OF THE BOOTSTRAP

Provided that we respect the structure of the data as discussed in Section 2.4, it follows from the discussion in Section 1.3 that bootstrap failure occurs if, heuristically, small differences between $\hat{F}_n$ and F can translate into large differences between $J_n(t, \hat{F}_n)$ and $J_n(t, F)$. We now go through some examples of when this happens. All of these examples are in line with the "rule of thumb" in Remark 2.

Example 5 (Parameter on the boundary). This example is due to Andrews (2000). Suppose $X_i \sim F$, we're interested in the mean $\theta(F) = \int x \, \mathrm{d}F(x)$, and we know that $\theta(F) \geq 0$. Suppose $\mathrm{var}_F(X_i) = 1$.

- A natural estimator is $\hat{\theta}_n = \max(0, \overline{X}_n)$. Let $R_n = n^{1/2}(\hat{\theta}_n - \theta)$.
- If $\theta > 0$, then $R_n \Rightarrow Z \sim \mathcal{N}(0,1)$, but if $\theta = 0$, then $R_n \Rightarrow \max\{Z, 0\}$.

Now, suppose that $\theta = 0$ and fix $c > 0$. Then, on the event $A_n = \{\sqrt{n}\overline{X}_n \geq c\}$,

$$\begin{aligned} R_n^* = n^{1/2}(\hat{\theta}_n^* - \theta(\hat{F}_n)) = n^{1/2}(\max\{\overline{X}_n^*, 0\} - \overline{X}_n) &= \max\{n^{1/2}(\overline{X}_n^* - \overline{X}_n), -n^{1/2}\overline{X}_n\} \\ &\leq \max\{n^{1/2}(\overline{X}_n^* - \overline{X}_n), -c\}. \end{aligned}$$

We know that conditional on $\mathbf{X}_n$, $n^{1/2}(\overline{X}_n^* - \overline{X}_n)$ is standard normal in large samples by CLT, so the RHS in the display above converges to $\max\{Z, -c\}$. Consequently, $\lim_{n\to\infty} P(R_n^* \leq x \mid A_n) \geq P(\max\{Z, -c\} \leq x) > P(\max\{Z, 0\} \leq x) = J_\infty(x)$. Thus, the bootstrap is incorrect in large samples whenever $n^{1/2}\overline{X}_n$ is positive (which happens with probability $\Phi(-c)$).

- By the same argument, we can't rescue the bootstrap by setting $\theta(\hat{F}_n) = \overline{X}_n$. The parametric bootstrap also fails (with $X_i^* \sim \mathcal{N}(\hat{\theta}_n, 1)$), and in this case, the asymptotic statements hold in finite samples.

The cause of the problem is that $J_\infty(t, F)$ is not continuous in F: consider a sequence $F_n = \mathcal{N}(\theta_n, 1)$ with $\theta_n \downarrow 0$. Then $\max_x |F_n(x) - \Phi(x)|_\infty \to 0$, but $J_\infty(\cdot, F_n)$ is standard normal, while $J_\infty(\cdot, \Phi)$ is the distribution of $\max\{0, Z\}$. ⊠

- The broad lesson is that whenever possible, one should try to choose R_n to be "as pivotal as possible"—not only to obtain asymptotic refinements, but for overall approximation quality.
- The narrow lesson is that the bootstrap will not work when applied naïvely in problems in which boundary issues are important, such as moment inequality problems.

Example 6 (Heavy tails). The nonparametric bootstrap fails for estimating the mean of distributions with fewer than two moments (as does the usual approach based on CLT). If the distribution is parametric, the parametric bootstrap can be made to work. ⊠

Example 7 (Sample maximum). Another classic counterexample due to Bickel and Freedman (1981) is estimation of a sample maximum. Let $X_i \sim F$ with support $[0, \theta]$, let $\hat{\theta}_n = X_{(n)}$ be the sample maximum, and $R_n = n(\hat{\theta}_n - \theta(F))$. The nonparametric bootstrap analog is $R_n^* = n(X_{(n)}^* - \hat{\theta}_n)$. Now, $P(R_n^* = 0 \mid \mathbf{X}_n) = 1 - (1 - 1/n)^n \to 1 - e^{-1}$ as $n \to \infty$. However, $J_\infty(t, F) = 1 - e^{tf(\theta)}$, where $f(x)$ is the density of F (if X_i are $U[0, 1]$, this limiting distribution is the exponential distribution). Hence, $P(R_n = 0) \to 0$ and the bootstrap is inconsistent. More generally, the bootstrap tends to fail if the support of $\theta(F)$ determines the support of X_i.

The parametric bootstrap is consistent. ⊠

Example 8. Instrumental variables model with weak instruments, and more generally, weakly identified models. ⊠

Example 9 (Matching estimator). This example is an exception to the rule of thumb, in that the bootstrap fails even though the estimator is asymptotically normal.

Suppose we observe $\{Y_i, D_i, X_i\}_{i=1}^n$, where Y_i is an outcome, D_i a treatment indicator, and X_i a vector of covariates. We're interested in estimating the average treatment effect for the treated under the assumption that the treatment is as good as randomly assigned conditional on covariates. A matching estimator (with a single match) estimates the untreated outcome $Y_j(0)$ for each treated individual with the outcome Y_i of the untreated individual i who is closest to j in terms of covariate distance $d(X_i, X_j)$. The estimator takes the form $N_1^{-1} \sum_i (D_i - (1 - D_i)K_i)Y_i$, where N_1 is the number of treated people, and K_i is the weighted number of times an untreated individual i is used as a match (if two units are closest, both are used as a match with weight $1/2$). Abadie and Imbens (2008) show that the bootstrap fails to reproduce the distribution of K_i. The intuition Abadie and Imbens (2008) give is that if the number of treated units is small, most control units are matched no more than once. But in the bootstrap world, a treated unit can appear multiple times, in which case the control is matched more than once. Otsu and Rai (2017) propose a weighted bootstrap modification that works. Adusumilli

(2018) proposes a bootstrap procedure for the propensity score matching estimator that's consistent. ⊠

4. SUBSAMPLING

An alternative way of approximating $J_n(\cdot, F)$ is to use subsamples of size $m < n$ from the original data. These methods often work even in cases in which the bootstrap fails. However, they tend to be less accurate than the bootstrap when it does work. A good summary of these alternative methods is in Horowitz (2001, Section 2.2). Politis, Romano, and Wolf (1999) is the standard technical reference. Much of the theoretical underpinning is due to Wu (1990) and Politis and Romano (1994) and Bickel, Götze, and van Zwet (1997).

The first alternative is called *replacement subsampling*, *m out of n bootstrap*, or *rescaled bootstrap*. It works just like the nonparametric bootstrap in that we draw i.i.d. from $\mathbf{X}_n$, but we only draw a sample of size $m < n$, so that we approximate $J_n(\cdot, F)$ by $J_m(\cdot, \hat{F}_n)$ rather than $J_n(\cdot, \hat{F}_n)$.

Intuitively, the reason this works is that if m is small relative to n, then the sampling error in $\hat{F}_n$ is negligible relative to sampling variability in the bootstrap subsample, making the m out of n bootstrap less sensitive to continuity of J_n in F.

Example 5 (continued). Consider the parameter at a boundary example. The m out of n bootstrap uses the distribution of $m^{1/2}(\hat{\theta}^*_m - \hat{\theta}_n)$ to approximate the distribution of $n^{1/2}(\hat{\theta}_n - \theta)$. We have

$$\begin{aligned}\sqrt{m}(\hat{\theta}_m(\mathbf{X}^*_m) - \hat{\theta}_n(\mathbf{X}_n)) &= \max\{m^{1/2}\overline{X}^*_m - m^{1/2}\theta, -m^{1/2}\theta\} - m^{1/2}(\hat{\theta}_n - \theta) \\ &= \max\{\sqrt{m}(\overline{X}^*_m - \overline{X}_n) + \sqrt{m}(\overline{X}_n - \theta), -m^{1/2}\theta\} - \sqrt{m}(\hat{\theta}_n - \theta).\end{aligned}$$

By the law of iterated logarithm, $\limsup_n \sqrt{n}(\overline{X}_n - \theta)/\sqrt{\log\log n} \xrightarrow{\text{a.s.}} \sqrt{2}$, so that if $\log\log n \cdot m/n \to 0$, then with probability one,

$$m^{1/2}(\hat{\theta}_m(\mathbf{X}^*_m) - \hat{\theta}_n(\mathbf{X}_n)) = \max\{m^{1/2}(\overline{X}^*_m - \overline{X}_n) + o(1), -m^{1/2}\theta\} - o(1).$$

Now, by the central limit theorem if $m \to \infty$, $m^{1/2}(\overline{X}^*_m - \overline{X}_n) \Rightarrow Z \sim \mathcal{N}(0,1)$, conditional on $\mathbf{X}_n$, so that conditionally on $\mathbf{X}_n$, $m^{1/2}(\hat{\theta}_m(\mathbf{X}^*_m) - \hat{\theta}_n(\mathbf{X}_n))$ converges to Z if $\theta > 0$ and to $\max\{Z, 0\}$ if $\theta = 0$ as required. In summary, if $m \to \infty$ and $m(\log\log n)/n \to 0$, then the m out of n bootstrap is consistent. ⊠

The second alternative is called *(non-replacement) subsampling*. Assume that the root has the form $R_n = \tau_n(\hat{\theta}_n(\mathbf{X}_n) - \theta(F))$, where τ_n is a normalizing constant (usually equal to $\sqrt{n}$). In subsampling, we draw a subsample of size m from $\mathbf{X}_n$ *without replacement*. Formally, enumerate the $N = \binom{n}{m}$ subsets of $\{1, \ldots, n\}$, and let $\hat{\theta}_{m,k}$ correspond the

statistic $\hat{\theta}(\mathbf{X}_{m,k})$ based on the kth subset. We approximate $J_n(t, F)$ by

$$\hat{J}_{n,m}(t) = N^{-1} \sum_{k=1}^{N} \mathbb{1}\{\tau_m(\hat{\theta}_{m,k} - \hat{\theta}_n) \leq t\}.$$

The intuition for this method is that $\mathbf{X}_{m,k}$ is a random sample of size m from F. Therefore, the *exact distribution* of $\tau_m(\hat{\theta}_{m,k} - \theta)$ is $J_m(\cdot, F)$, where

$$J_m(t, F) = E[\mathbb{1}\{\tau_m(\hat{\theta}_{m,k} - \theta) \leq t\}].$$

The subsampling distribution replaces θ by $\hat{\theta}_n$, and replaces the population expectation by subsample average. If m/n is small, then the error induced by the first substitution should be negligible relative to the randomness in $\hat{\theta}_{m,k}$ (same intuition as with the m out of n bootstrap). Similarly, if N is large, we should be able to appeal to some sort of law of large numbers so that the subsample average is close to the population expectation.

Theorem 5 (Politis and Romano, 1994, Theorem 2.1). Assume that the distribution $J_n(\cdot, F)$ converges weakly to some limiting distribution $J_\infty(\cdot, F)$, and that $\tau_m/\tau_n \to 0$ and $m \to \infty$ as $n \to \infty$. Then:

1. *Subsampling is consistent in the sense that $\hat{J}_{n,m}(t) \xrightarrow{p} J_\infty(t, F)$ at all continuity points t of J_∞ (and hence by an application of Pólya's Theorem, if J_∞ is continuous, then $\|\hat{J}_{n,m} - J_\infty\| \xrightarrow{p} 0$)*
2. *If J_∞ is continuous at its $1-\alpha$ quantile, then the one-sided subsampling CI is asymptotically valid,*

$$P_F(\tau_n(\hat{\theta}_n - \theta(F)) \leq \hat{J}_{m,n}^{-1}(1-\alpha)) \to 1-\alpha.$$

So with subsampling, we don't need to worry about continuity of J_n in F (at least as far as pointwise asymptotics are concerned).

- Because the distribution $J_\infty(\cdot, F)$ is well-defined for all F in Examples 5 to 7 and 9, subsampling will still work (as long as we know the rate of convergence). For instance, in Example 7, $P(R_m^* = 0 \mid \mathbf{X}_n) = 1 - (1 - 1/n)^m \approx 1 - e^{-m/n} \to 0$ if $m/n \to 0$, and the point-mass problem goes away. Consistency of subsampling in Example 6 was shown in Romano and Wolf (1999).
- By the same logic, subsampling still works in the maximum score example (Delgado, Rodríguez-Poo, and Wolf 2001). Although the limit distribution is not standard, all we need to know that it exists, and that the rate of convergence is $n^{-1/3}$ (and that the limit distribution has no mass points, which they fail to show).
- Subsampling (as well as the m out of n bootstrap) *fails* in Example 8—see Andrews and Guggenberger (2010). Here $\tau_n = 1$, so the condition $\tau_m/\tau_n \to 0$ fails.

On the other hand, choosing m appropriately can be hard in practice. Furthermore, it can be shown that in many cases in which the bootstrap works, it is also more accurate: the bootstrap error is at most $O_p(n^{-1/2})$ versus $O_p(m/n + m^{-1/2})$ for subsampling,

which is at least as big as $O_p(n^{-1/3})$ (see Section 2.4 in Politis and Romano 1994), with the optimal choice of m being $m = O_p(n^{2/3})$. This makes subsampling attractive mostly just in situations in which the bootstrap fails (and we don't have available a bootstrap modification that works).

REFERENCES

Abadie, Alberto, and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76, no. 6 (November): 1537–1557. https://doi.org/10.3982/ECTA6474.

Abrevaya, Jason, and Jian Huang. 2005. "On the Bootstrap of the Maximum Score Estimator." *Econometrica* 73, no. 4 (July): 1175–1204. https://doi.org/j.1468-0262.2005.00613.x.

Adusumilli, Karun. 2018. "Bootstrap Inference for Propensity Score Matching." Working paper, University of Pennsylvania, October. https://www.dropbox.com/s/e4n2wct32uopsyi/PSM-Bootstrap-10.pdf.

Andrews, Donald W. K. 2000. "Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space." *Econometrica* 68, no. 2 (March): 399–405. https://doi.org/10.1111/1468-0262.00114.

Andrews, Donald W. K., and Patrik Guggenberger. 2010. "Applications of Subsampling, Hybrid, and Size-Correction Methods." *Journal of Econometrics* 158, no. 2 (October): 285–305. https://doi.org/10.1016/j.jeconom.2010.01.002.

Beran, Rudolf. 1984. "Bootstrap Methods in Statistics." *Jahresbericht der Deutschen Mathematiker-Vereinigung* 86, no. 1 (January): 14–30.

Bickel, Peter J., and David A. Freedman. 1981. "Some Asymptotic Theory for the Bootstrap." *The Annals of Statistics* 9, no. 6 (November): 1196–1217. https://doi.org/10.1214/aos/1176345637.

Bickel, Peter J., Friedrich Götze, and Willem R. van Zwet. 1997. "Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses." *Statistica Sinica* 7, no. 1 (January): 1–31. https://www.jstor.org/stable/26432490.

de Angelis, D., Peter Hall, and G. A. Young. 1993. "Analytical and Bootstrap Approximations to Estimator Distributions in L^1 Regression." *Journal of the American Statistical Association* 88, no. 424 (December): 1310–1216. https://doi.org/10.1080/01621459.1993.10476412.

Delgado, Miguel A., Juan M. Rodríguez-Poo, and Michael Wolf. 2001. "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator." *Economics Letters* 73, no. 2 (November): 241–250. https://doi.org/10.1016/S0165-1765(01)00494-3.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7, no. 1 (January): 1–26. https://doi.org/10.1214/aos/1176344552.

Hahn, Jinyong. 1995. "Bootstrapping Quantile Regression Estimators." *Econometric Theory* 11, no. 1 (February): 105–121. https://doi.org/10.1017/S0266466600009051.

Hahn, Jinyong, and Zhipeng Liao. 2021. "Bootstrap Standard Error Estimates and Inference." *Econometrica* 89, no. 4 (July): 1963–1977. https://doi.org/10.3982/ECTA17912.

Hall, Peter. 1988. "Rate of Convergence in Bootstrap Approximations." *The Annals of Probability* 16, no. 4 (October): 1665–1684. https://doi.org/10.1214/aop/1176991590.

———. 1992. *The Bootstrap and Edgeworth Expansion.* New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4384-7.

Hall, Peter, and Joel L. Horowitz. 1996. "Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators." *Econometrica* 64, no. 4 (July): 891–916. https://doi.org/10.2307/2171849.

Horowitz, Joel L. 2001. "The Bootstrap." Chap. 52 in *Handbook of Econometrics,* edited by James J. Heckman and Edward E. Leamer, 5:3159–3228. Amsterdam: Elsevier. https://doi.org/10.1016/S1573-4412(01)05005-X.

———. 2019. "Bootstrap Methods in Econometrics." *Annual Review of Economics* 11 (August): 193–224. https://doi.org/10.1146/annurev-economics-080218-025651.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83, no. 5 (September): 2043–2063. https://doi.org/10.3982/ECTA11974.

Kosorok, Michael R. 2008. "Bootstrapping the Grenander Estimator." In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen,* edited by N. Balakrishnan, Edsel A. Peña, and Mervyn J. Silvapulle, 282–292. Beachwood, OH: Institute of Mathematical Statistics. https://doi.org/10.1214/193940307000000202.

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.

Mammen, Enno. 1992. *When Does Bootstrap Work?* New York, NY: Springer. https://doi.org/10.1007/978-1-4612-2950-6.

Otsu, Taisuke, and Yoshiyasu Rai. 2017. "Bootstrap Inference of Matching Estimators for Average Treatment Effects." *Journal of the American Statistical Association* 112, no. 520 (October): 1720–1732. https://doi.org/10.1080/01621459.2016.1231613.

Politis, Dimitris N., and Joseph P. Romano. 1994. "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions." *The Annals of Statistics* 22, no. 4 (December): 2031–2050. https://doi.org/10.1214/aos/1176325770.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf. 1999. *Subsampling.* New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4612-1554-7.

Romano, Joseph P., and Michael Wolf. 1999. "Subsampling Inference for the Mean in the Heavy-Tailed Case." *Metrika* 50, no. 1 (November): 55–69. https://doi.org/10.1007/s001840050035.

Sen, Bodhisattva, Moulinath Banerjee, and Michael Woodroofe. 2010. "Inconsistency of Bootstrap: The Grenander Estimator." *The Annals of Statistics* 38, no. 4 (August): 1953–1977. https://doi.org/10.1214/09-AOS777.

van der Vaart, Aad. 1998. *Asymptotic Statistics.* New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

Wu, Chien F. J. 1990. "On the Asymptotic Properties of the Jackknife Histogram." *The Annals of Statistics* 18, no. 3 (September): 1438–1452. https://doi.org/10.1214/aos/1176347759.

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.

STANDARD ERRORS WITH VERY SMALL OR VERY LARGE DATASETS

Michal Kolesár*

March 29, 2024

1. FINITE-SAMPLE ISSUES WITH STANDARD STANDARD ERRORS

As in the previous set of notes, we regress the outcome Y_i onto $X_i = (D_i, W_i)$, with $\dim(X_i) = k$. We want to do inference on the D_i regression coefficient using the "standard" standard error formulas, using either the Eicker-Huber-White (EHW) or the Liang-Zeger (LZ) formula.

When may these standard errors fail to adequately reflect the statistical uncertainty in the ordinary least squares (OLS) estimates? It is helpful to group the reasons into two broad categories: either one of the substantive assumptions fails (e.g. i.i.d. sampling), or else some "regularity conditions" fail (e.g. no fat tails). By "finite-sample issues", we mean that some of the regularity conditions that underlie the large-sample validity of the EHW and LZ standard errors fail in the sample at hand.

As we will see, this failure of asymptotics can happen for a variety of reasons. To think through them in a simple way, and to make the discussion simpler, let us suppose that the regression function $\mu(X_i) := E[Y_i \mid X_i]$ is linear, $\mu(x) = x'\theta = d\beta + w'\gamma$, and that we wish to conduct inference on β that is conditional on X. As we discussed last time, this may not be the best setup for ensuring a robust causal or descriptive interpretation for β when the linearity assumption on the regression function is violated. However, it makes it easy to think through hiccups with standard inference.

Research Question. Most of the literature takes this approach. It would be interesting to think through how one could adapt the diagnostics and solutions below to other sampling frameworks, and to allow for misspecification of μ. ⊠

Let us focus on EHW standard errors first. Let us assume that the substantive assumption—Assumption 4 in the previous set of notes that Y_i is independent across

*Email: mkolesar@princeton.edu.

i conditional on X—holds. Define $\epsilon_i := Y_i - \mu(X_i) = Y_i - X_i'\theta$. Then, for asymptotic normality, we need Assumption 5 in the previous set of notes, which imposes two regularity conditions:

NO FAT TAILS $E[\epsilon_i^{2+\eta} \mid X]$ is bounded for some $\eta > 0$. This assumption may fail because the heteroskedasticity in the data is rather extreme, or because there are outliers in ϵ_i (or equivalently Y_i).

LOW PARTIAL LEVERAGE $\max_i H_{\ddot{D},ii} \to 0$. This effectively says that there are no outliers in X_i, stated in a form that makes it easy to verify.

For inference, we also need the estimator $\hat{V}_{\text{EHW}}$ to be consistent. For this we need to strengthen the leverage condition to:

LEVERAGE FOR INFERENCE Either $k \max_i H_{X,ii} \to 0$ or $\sqrt{nk} \max_i H_{\ddot{D},ii} \to 0$.

The first part of the condition ensures that we can consistently estimate *all* regression errors. If the number of regressors is fixed, then $k \max_i H_{X,ii} \to 0$ is equivalent to $\max_i H_{X,ii} \to 0$, which we already require for asymptotic normality of the full regressor vector θ. But this condition may be violated in settings with group dummies where some of the groups may have a few observations (so we can't consistently estimate the group effect and hence the residuals for units in that group). The second condition allows for such cases at the expense of strengthening the partial leverage condition. Intuitively, this is possible since we don't actually have to estimate all individual regression errors consistently—we only need to be able to do that "on average", since they enter the EHW formula through a particular weighted average.

If we're in a "big data high-dimensional setting" in the sense that the number of controls k is large relative to the sample size, it is natural to consider asymptotics where k increases with the sample size. In this case the leverage for inference condition becomes stronger: in the best-case scenario with a balanced design, $\max_i H_{X,ii} \asymp k/n$, and $\max_i H_{\ddot{D},ii} \asymp 1/n$. The inference condition on leverage then becomes $k/n \to 0$. So in "high-dimensional designs" where k constitutes a non-negligible fraction of the sample size (say over 20%), the EHW is likely to perform poorly, even though there are no issues with asymptotic normality of $\hat{\beta}$.

Lemma 1. Suppose that the above conditions hold. Then EHW standard errors lead to asymptotically valid inference.

Proof. We need to show that the meat part of the sandwich is consistent in that

$$\frac{\sum_i \hat{\epsilon}_i^2 \ddot{D}_i^2}{\sum_i \epsilon_i^2 \ddot{D}_i^2} - 1 \xrightarrow{p} 0.$$

Since $\hat{\epsilon}_i = \epsilon_i + X_i'(\theta - \hat{\theta})$, and since the denominator is of the order $\sum_i \ddot{D}_i^2$ (by arguments as in the proof of asymptotic normality in the previous set of notes), it suffices to show

$$\frac{\sum_i 2\epsilon_i X_i'(\theta - \hat{\theta})\ddot{D}_i^2 + \sum_i (X_i'(\theta - \hat{\theta}))^2 \ddot{D}_i^2}{\sum_i \ddot{D}_i^2} \xrightarrow{p} 0.$$

Now there are two ways forward. First, we can bound the right-hand side by

$$\max_i H_{\ddot{D}_i} \cdot \left(\|\epsilon\|_2 \|X(\theta - \hat{\theta})\|_2 + \|X(\theta - \hat{\theta})\|_2^2 \right) = O_p(\max_i H_{\ddot{D}_i} \sqrt{nk}),$$

since $(\theta - \hat{\theta})' X' X (\theta - \hat{\theta}) = \epsilon H_X' \epsilon$, which is by Markov's inequality of the order k, since $E[\epsilon' H_X \epsilon] = \sum_i \sigma(X_i)^2 H_{X,ii} \leq k \max_i \sigma(x_i)^2 = O(k)$. Alternatively, we bound the right-hand side by

$$\max_i |X_i(\hat{\theta} - \theta)| \frac{\sum_i 2|\epsilon_i| \ddot{D}_i^2}{\sum_i \ddot{D}_i^2} + \max_i |X_i'(\theta - \hat{\theta})|^2. \tag{1}$$

Now, by Markov's inequality, $\frac{\sum_i 2|\epsilon_i| \ddot{D}_i^2}{\sum_i \ddot{D}_i^2} = O_p\left(\frac{\sum_i 2E[|\epsilon_i||X_i] \ddot{D}_i^2}{\sum_i \ddot{D}_i^2} \right) = O_p(1)$. Furthermore, by the Cauchy-Schwarz inequality,

$$|X_i'(\hat{\theta} - \theta)| = |e_i' H_X \epsilon| = |e_i' H_X H_X \epsilon| \leq \sqrt{H_{X,ii}} \sqrt{\epsilon' H_X \epsilon},$$

so that eq. (1) is of the order $(\max_i H_{X,ii})^{1/2} \sqrt{k} + \max_i H_{X,ii} k$. Either way, the meat part of the sandwich is consistent. □

Research Question. I think it should be possible to bound $\max_i |X_i'(\theta - \hat{\theta})|^2$ by something like $O_p(\max_i H_{X,ii} \sqrt{\log(k)})$, or perhaps $\log(n)$. In particular, if ϵ_i were Gaussian (or sub-Gaussian), then $\sum_j H_{X,ij} \epsilon_j$ are also Gaussian with variance bounded by a constant times $\max_i H_{X,ii}$. So by union and Chernoff bounds, $\max_i |\sum_j H_{X,ij} \epsilon_j| = O_p(\sqrt{(\max_i H_{X,ii})^{1/2} \log(n)})$. This similar to the rate obtained by Belloni et al. (2015) in the random regressor case using empirical process theory. ⊠

We can now see what may go wrong with inference:

1. central limit theorem (CLT) fails, either because the outcome has fat fails, or because the partial leverage is too high, so that the distribution of $\hat{\beta}$ is not close to Gaussian.

2. EHW variance estimator is not consistent: it displays finite-sample bias or substantial sampling variability, so that the t-statistic will have much fatter tails than the normal distribution that we use for critical values.

   As a result, the finite-sample coverage of confidence intervals based on EHW may be substantially below nominal coverage. (Again, it may so happen that unconditional inference is OK, we just got unlucky...).

To diagnose CLT failure, first look at outliers. If they arise due to measurement issues, consider dropping them, otherwise consider transforming the outcome if appropriate, though this of course changes the interpretation of β (taking logs, say, allows for easy interpretation of β, but winsorizing makes it tricky). Running quantile regressions may also be an appropriate alternative in certain contexts. It is also possible stick to the original OLS specification and do inference when the tails are moderately heavy using alternative inference procedures (Müller 2021).

Second, look at partial leverage. Most simply, since the maximal partial leverage is at least $1/n$, it can be high because our overall sample size is too small. More interestingly,

we may have a large overall sample size, but not that many observations actually help to pin down $\hat{\beta}$—the *effective* sample size is small. Fixing this is tricky, one may consider dropping controls, but that may then lead to violations of unconfoundedness.

Diagnosing and fixing the second issue will be our focus next. First some simple examples to fix ideas:

Example 1. Suppose that $D_1 = C\sqrt{n}$ for some constant C, while $D_i = 1$ if $i > 1$, and that $W_i = 1$. Then, one can show that $H_{X,11} = 1$, while $H_{X,ii} = 1/(n-1)$ for $i > 1$. The estimate of θ will be $\sqrt{n}$-consistent, but not asymptotically normal unless ϵ_1 happens to be normal. ⊠

Example 2 (Bo Honoré's failsafe method for detecting an outlier). Suppose we wish to check whether the first observation is an outlier, so we set $D_i = \mathbb{1}\{i = 1\}$, and let W_i be well-behaved controls. Then, as an exercise, show that $H_{X,11} = 1$. Show also that (i) $\hat{\epsilon}_1 = 0$, (ii) $X'X/n$ converges to a non-invertible limit, and (iii) $\hat{\gamma}$ will be consistent, and $\hat{\beta}$ will converge to $\beta + \epsilon_1$. Furthermore, show that the t-statistic for $\hat{\beta}$ based on EHW standard errors will converge to $\pm\infty$ irrespective of the value of β. ⊠

Example 3 (Behrens Fisher problem). Suppose that n_1 observations are treated, and there are n_0 controls. For simplicity, suppose that $W_i = 1$. The problem of inference in this context, if we assume that $\epsilon_i \mid D_i \sim \mathcal{N}(0, \sigma^2(D_i))$ is known as the Behrens-Fisher problem (Behrens 1929; Fisher 1939) (note the journal). It is clear that even if n is large, the effective number of observations is small if $\min\{n_1, n_0\}$ is small. The leverage reflects this: $H_{X,ii} = 1/n_{D_i}$.

A more complicated version of this problem arises in differences-in-differences contexts, when there are only a few treated observations. ⊠

- The takeaway message from these examples is that even if the number of observations is large, the "effective number of observations" that pins down $\hat{\beta}$ may be small. The leverage gives one sense in which we have few "effective observations" (below, we'll discuss degrees of freedom corrections which give another metric for deriving the number of effective observations).

- For a heuristic sense of whether the leverage in the sample at hand is high, consider inference on the mean. In this case, the leverage is $1/n$ for each observation. Given the rule of thumb that we need at least $n = 30$, say, for the CLT to work well in this case, this suggests that we should be careful if $\max_i H_{\ddot{D},ii} \geq 1/30$, and probably worried if $\max_i H_{\ddot{D},ii} \geq 1/10$.

CLUSTERING Similar issues arise when we cluster the standard errors. There are a few additional complications:

- The sample size is determined by the number of clusters S: the asymptotics are as $S \to \infty$.

- The rate of convergence depends on how heterogeneous the cluster sizes are, and on the within-cluster correlation structure. It'll be at most $n^{-1/2}$, but it can be even much slower than $S^{-1/2}$.

  Example 4 (Hansen and Lee 2019). Suppose we're interested in estimating the mean (so no covariates). There are two cluster sizes, $n/2$ clusters have size $n_s = 1$ and $n^{1-\alpha}/2$ clusters have size $n_s = n^\alpha$ (so $S = O(n)$). Then $\text{var}(\overline{X}) = (1 + n^\alpha)/(2n)$, so the rate of convergence is $n^{-(1-\alpha)/2}$, much slower than $S^{-1/2}$. ⊠

- We'll certainly need $\max_i H_{\ddot{D},ii} \to 0$ for the CLT to hold, though what matters will be the leverage of the whole cluster, so that a sufficient leverage condition will be substantially stronger.

2. DEGREES OF FREEDOM CORRECTION

One issue with the EHW and LZ variance estimators is that they are biased in finite samples. In particular, the bias is given by

$$B = E[\hat{V}_{\text{EHW},11} \mid X] - \mathcal{V}_{\text{cx},11} = \frac{\sum_i E[\hat{\epsilon}_i^2 - \sigma^2(X_i) \mid X_i]\ddot{D}_i^2}{(\sum_i \ddot{D}_i^2)^2},$$

Since $\hat{\epsilon}_i = \epsilon_i - e_i' H_X \epsilon$, we have $E[\hat{\epsilon}_i^2 \mid X_i] - \sigma^2(X_i) = -2H_{X,ii}\sigma^2(X_i) + \sum_j H_{X,ij}^2 \sigma^2(X_j)$, which is bounded in absolute value by a constant times $H_{X,ii}$ when the no fat tails condition holds. Thus, the order of the bias is $\frac{\sum_i H_{X,ii}\ddot{D}_i^2}{(\sum_i \ddot{D}_i^2)^2} \leq \min\{\max_i H_{X,ii}/n, k/n \max_i H_{\ddot{D},ii}\}$. Since the order of the variance is $1/n$, the bias will be asymptotically negligible if

$$\min\left\{\max_i H_{X,ii}, k \max_i H_{\ddot{D},ii}\right\} \to 0,$$

which is a slightly weaker condition than the leverage for inference condition we imposed (which also ensures that the variance of $\hat{V}_{\text{EHW},11}$ is negligible). When leverage is high, this suggests that one should be concerned with the bias of the EHW estimator. In general, the bias can be positive or negative. Under homoskedasticity,

$$\begin{aligned} E[\hat{V}_{\text{EHW}} \mid X] - \mathcal{V}_{dc} &= \sigma^2 n (X'X)^{-1} \sum_i \left[\sum_j H_{X,ij}^2 - 1\right] X_i X_i' (X'X)^{-1} \\ &= \sigma^2 n (X'X)^{-1} \sum_i (H_{X,ii} - 1) X_i X_i' (X'X)^{-1} \leq 0. \end{aligned}$$

since $H_{X,ii} \leq 1$. This expression implies that if we modify the estimator and weight the observations in inverse proportion to their leverage,

$$\hat{V}_{HC2} = n(X'X)^{-1} \sum_i \frac{\hat{\epsilon}_i^2}{1 - H_{X,ii}} X_i X_i' (X'X)^{-1},$$

we will be unbiased under the homoskedastic benchmark. This idea goes back to MacKinnon and White (1985).[1] With a single binary regressor, this estimator is unbiased even under heteroskedasticity.

While this solves the bias issue (at least if the data is close to the homoskedastic benchmark), there is still the issue of the variability of the variance estimator. If we assume that the errors are normal and homoskedastic, then we know that the appropriate distribution for the t-statistic is t_{n-k}. When we allow for heteroskedasticity, things are more complicated, since the t-statistic doesn't follow a t-distribution even under normal errors. The problem is that $\hat{V}_{HC2}$ is not a scaled χ^2_{n-k}, but instead a more complicated distribution, a weighted average of χ^2_1 (and it's also not generally independent of the numerator). Nonetheless, we may still try to approximate it by a χ^2_ν with degrees of freedom (DoF) ν chosen to match the first two moments.

If we choose ν to match the first two moments in the homoskedastic case, we arrive at the Satterthwaite (1946) DoF correction. Let G denote the $n \times n$ matrix with ith column given by $(I - H)e_i \cdot (1 - H_{X,ii})^{-1/2} X_i'(X'X)^{-1}\ell$. Here $e_i \in \mathbb{R}^n$ denotes ith unit vector. Then for inference on $\ell'\theta$ we set

$$\nu = \frac{\operatorname{tr}(G'G)^2}{\operatorname{tr}((G'G)^2)}. \tag{2}$$

Proof. We can write $\hat{\epsilon}_i = Y_i - X_i(X'X)^{-1}X'Y = e_i'(I - H)Y = e_i'(I - H)\epsilon$, so that

$$\ell'\hat{V}_{HC2}\ell = n\ell'(X'X)^{-1}\sum_i \frac{(e_i'(I-H)\epsilon)^2}{1 - H_{X,ii}} X_i X_i'(X'X)^{-1}\ell = \sum_i G_i'\epsilon\epsilon' G_i = \epsilon' GG'\epsilon$$

where G_i is the ith column of G. Let $V = \operatorname{var}(\ell'(\hat{\beta} - \beta \mid X))$. Then we can write the t-statistic as

$$\frac{\ell'(\hat{\beta} - \beta)}{\sqrt{\ell'\hat{V}_{HC2}\ell}} = \frac{Z_0}{\sqrt{\ell'\hat{V}_{HC2}\ell / V}},$$

where $Z_0 = \ell'(X'X)^{-1}X'\epsilon/\sqrt{V}$. Under the homoskedasticity benchmark, $V = \sigma^2\ell'(X'X)^{-1}\ell = \sigma^2 \operatorname{tr}(G'G)$. Also, using the spectral decomposition $GG' = P\Lambda P'$, with $\epsilon' GG'\epsilon = \sum_i \lambda_i (P\epsilon)_i^2$. The first two moments of the denominator under the homoskedastic normal benchmark are therefore $E[\ell'\hat{V}_{HC2}\ell/V \mid X] = \operatorname{tr}(G'G)/\ell'(X'X)^{-1}\ell = 1$ and $\operatorname{var}(\ell'\hat{V}_{HC2}\ell/V \mid X) = \sum_i 2\lambda_i^2\sigma^4/V^2 = \sum_i 2\lambda_i^2 / \operatorname{tr}(G'G)^2 = 2\operatorname{tr}((G'G)^2)/\operatorname{tr}(G'G)^2$. Hence, the denominator is approximately distributed χ^2_ν/ν, with ν given in eq. (2) above. □

The key point is that the adjustment depends on the distribution of the covariates, and therefore reflects any leverage issues. Even with many observations, the implied DoF may be quite small. This is most easily seen in the case with a single binary covariate. In that case the regression estimator is the difference in two means. If the sample size for one of the two means is small, then the DoF for the approximating t distribution is small, regardless of the number of observations used in calculating the other mean.

1. There are other estimators of the asymptotic variance that are sometimes used. If, for example, we normalize by $1/(1 - H_{X,ii})^2$, this is called the HC3 or jackknife variance estimator, and it is used in Young (2019).

Similar adjustments can be applied to the case with clustering. In particular, Bell and McCaffrey (2002) suggest using the variance estimator

$$\hat{V}_{BM} = n(X'X)^{-1}\sum_s X_s' A_s \hat{\epsilon}_s \hat{\epsilon}_s A_s' X_s' (X'X)^{-1}, \qquad A_s = (I - H_{ss})^{-1/2}.$$

where $H_{ss} = X_s(X'X)^{-1}X_s'$, and X_s is the submatrix of X that corresponds to cluster s. Here A_s is the symmetric square root of the inverse of $I - H_{ss}$, or of its pseudo-inverse if $I - H_{ss}$ is singular (as is the case when we include cluster fixed effects). In addition, they also propose a DoF correction so that the denominator of the t-statistic matches the first two moments of a χ^2_ν under the i.i.d. Gaussian benchmark. In particular, for inference on $\ell'\theta$, let G be an $n \times S$ matrix with columns $g_s = (I - H)_s' A_s' X_s (X'X)^{-1}\ell$, where $(I - H)_s$ is the $n_s \times n$ block of the matrix $I - H$ that corresponds to cluster s. Then compute ν as in eq. (2), but with this definition of G. One could also compute the DoF correction under other working models—see Imbens and Kolesár (2016) for details, and Hansen (2021) for a refinement of this approach.

While these DoF corrections only have a heuristic motivation, in practice they tend to significantly improve coverage relative to $\hat{V}_{\text{LZ}}$, the Stata default.

3. ALTERNATIVE ALTERNATIVES

The appeal of the HC2 estimator coupled with the DoF correction is that it's simple to compute and interpret. The downside is that we're solving the bias issue only if we're close to a homoskedastic benchmark, and that the DoF correction is a bit heuristic.

It is actually possible to construct variance estimators that are exactly unbiased. One approach is to use Hadamard products (see, for example Dobriban and Su (2018) or Cattaneo, Jansson, and Newey (2018)). However, this involves inverting large matrices, so it may not be feasible in settings with a large number of observations. Another idea proposed in Kline, Saggio, and Sølvsten (2020) and investigated in Jochmans (2022) is a leave-out approach, estimate $\sigma^2(X_i)$ in the formula not by $(Y_i - X_i'\hat{\theta})^2$ used by EHW, but by the unbiased estimator

$$\check{\sigma}_i^2 = Y_i(Y_i - X_i'\hat{\theta}_{-i}) = \frac{Y_i(Y_i - X_i'\hat{\theta})}{1 - H_{X,ii}},$$

where $\hat{\theta}_{-i}$ is the OLS estimator with observation i excluded, and the second equality uses the fact that $X_i\hat{\theta}_{-i} = X_i(X'X - X_iX_i')^{-1}(X'Y - X_iY_i) = \frac{1}{1-H_{X,ii}}X_i(X'X)^{-1}(X'Y - X_iY_i) = \frac{1}{1-H_{X,ii}}(X_i\hat{\theta} - H_{X,ii}Y_i)$ by the Woodbury formula. These papers show that under appropriate conditions, these variance estimators lead to asymptotically valid inference in even in high-dimensional settings where the number of observations is proportional to sample size, relaxing the leverage for inference condition that we needed to impose when using the EHW estimator.

Another method that tends to improve coverage in simulation is the wild bootstrap, popularized by Cameron, Gelbach, and Miller (2008). However, since it appears important to impose the null in computing the bootstrap distribution, confidence intervals have to be computed by test inversion. In particular, the procedure is as follows:

1. To test the null $\ell'\theta = c$, compute the OLS estimate θ subject to this restriction, obtaining the restricted estimate $\hat{\theta}_r$ and residuals $\hat{\epsilon}_i^r$.
2. Let $Y_i^* = X_i'\hat{\theta}^r + g_{s(i)}^* \hat{\epsilon}_i^r$, where $g_{s(i)}^* \in \{-1, 1\}$ (with equal probability), and let $X_i^* = X_i$. Compute $\hat{\theta}^*$ using OLS in this bootstrap sample.
3. As a critical value for the test statistic $|\ell'\hat{\theta} - c|$, use the $1-\alpha$ quantile of $|\ell'(\hat{\theta}^* - \hat{\theta}^r)|$

Canay, Santos, and Shaikh (2021) show formally that this method works even with a fixed number of clusters, so long as certain cluster homogeneity conditions hold on the distribution of covariates across clusters (see Ibragimov and Müller (2016) for simulation evidence on overrejection under cluster heterogeneity). This also requires the clusters to have similar sizes.

There are also two alternative methods that do not require cluster homogeneity: Canay, Romano, and Shaikh (2017) and Ibragimov and Müller (2016).

4. WEIGHTING

There are a few reasons to run weighted least squares rather than OLS:

1. Your data contains exact duplicates. This is Stata's frequency weights (`fweight`). If the weight is 5 that means there are really 5 such observations, each identical. Only integer weights are allowed. You should get numerically the same result if you expand the data using `expand pop`.
2. You weight for precision. For instance, if the data consists of cell averages, then it makes sense to weight by $1/n_i$, the inverse of the number of observations used to form the cell average. One may more generally try to estimate the heteroskedasticity function $\sigma^2(X_i)$, and weight by its inverse, as advocated for by Romano and Wolf (2017). In practice, people tend not to do this for three reasons:
   (a) The estimates of $\sigma^2(X_i)$ may be noisy, so you may actually make things worse in finite samples. If the estimates of $\sigma^2(X_i)$ are inconsistent, you may be making things worse in large samples as well. This is similar to issues that arise when using the (estimated) efficient weighting matrix in generalized method of moments (GMM)
   (b) It makes interpretation under misspecification more tricky: typically the estimates are less robust to misspecification.

3. You use sampling weights—`pweight` in Stata—because you don't sample i.i.d. from the population of interest. Note that if the sampling probability is only a function of X_i, and the regression function is correctly specified, then it is not necessary to weight. Though in practice, if we don't know how the sampling weights are constructed (and hence can't verify they are a function of X only), it may still be prudent to weight.

   Note this question of how we draw from the population of interest is moot for causal inference on the sample at hand.

Question 1. In panel data settings, where we follow states, is it a good idea to weight by the state's population?

REFERENCES

Behrens, Walter Ulrich. 1929. "Ein Beitrag Zur Fehlerberechnung Bei Wenigen Beobachtungen." *Landwirtschaftliche Jahrbücher* 68:807–837.

Bell, Robert M., and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28, no. 2 (December): 169–181. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20020029058.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. 2015. "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results." *Journal of Econometrics* 186, no. 2 (June): 345–366. https://doi.org/10.1016/j.jeconom.2015.02.014.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90, no. 3 (August): 414–427. https://doi.org/10.1162/rest.90.3.414.

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh. 2017. "Randomization Tests under an Approximate Symmetry Assumption." *Econometrica* 85, no. 3 (May): 1013–1030. https://doi.org/10.3982/ECTA13081.

Canay, Ivan Alexis, Andres Santos, and Azeem M Shaikh. 2021. "The Wild Bootstrap with a "Small" Number of "Large" Clusters." *Review of Economics and Statistics* 103, no. 2 (May): 346–363. https://doi.org/10.1162/rest_a_00887.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey. 2018. "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity." *Journal of the American Statistical Association* 113, no. 523 (July): 1350–1361. https://doi.org/10.1080/01621459.2017.1328360.

Dobriban, Edgar, and Weijie J. Su. 2018. "Robust Inference Under Heteroskedasticity via the Hadamard Estimator," July. arXiv: 1807.00347.

Fisher, Ronald Aylmer. 1939. "The Comparison of Samples with Possibly Unequal Variances." *Annals of Eugenics* 9, no. 2 (June): 174–180. https://doi.org/10.1111/j.1469-1809.1939.tb02205.x.

Hansen, Bruce E. 2021. "The Exact Distribution of the White T-Ratio." Working paper, University of Wisconsin.

Hansen, Bruce E., and Seojeong Lee. 2019. "Asymptotic Theory for Clustered Samples." *Journal of Econometrics* 210, no. 2 (June): 268–290. https://doi.org/10.1016/j.jeconom.2019.02.001.

Ibragimov, Rustam, and Ulrich K. Müller. 2016. "Inference with Few Heterogeneous Clusters." *Review of Economics and Statistics* 98, no. 1 (March): 83–96. https://doi.org/10.1162/REST_a_00545.

Imbens, Guido W., and Michal Kolesár. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98, no. 4 (October): 701–712. https://doi.org/10.1162/REST_a_00552.

Jochmans, Koen. 2022. "Heteroscedasticity-Robust Inference in Linear Regression Models With Many Covariates." *Journal of the American Statistical Association* 117, no. 538 (April): 887–896. https://doi.org/10.1080/01621459.2020.1831924.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten. 2020. "Leave-Out Estimation of Variance Components." *Econometrica* 88, no. 5 (September): 1859–1898. https://doi.org/10.3982/ECTA16410.

MacKinnon, James G., and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29, no. 3 (September): 305–325. https://doi.org/10.1016/0304-4076(85)90158-7.

Müller, Ulrich K. 2021. "A More Robust *t*-Test." Working paper, Princeton University.

Romano, Joseph P., and Michael Wolf. 2017. "Resurrecting Weighted Least Squares." *Journal of Econometrics* 197, no. 1 (March): 1–19. https://doi.org/10.1016/j.jeconom.2016.10.003.

Satterthwaite, F. E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2, no. 6 (December): 110–114. https://doi.org/10.2307/3002019.

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.

TREATMENT EFFECT HETEROGENEITY AND WEAK INSTRUMENTS

Michal Kolesár*

April 2, 2024

We can use instrumental variables (IV) regression to solve a number of issues:

1. Errors-in-variables (see, for example, Zellner 1970);
2. Deal with omitted variable bias: we'd like to recover β in the projection $E[Y_i \mid D_i, A_i] = D_i\beta + A_i'\gamma$, but A_i is not observed (see, for example, Chamberlain 2007);
3. Estimate a simultaneous equations model, such as a demand-and-supply system (see, for example, Angrist, Graddy, and Imbens 2000); or
4. Estimate treatment effects when the unconfoundedness assumption fails.[1]

Which goal we have in mind often doesn't affect the estimation and inference procedures, but it does affect the interpretation of the results—for concreteness, we will focus here on the last goal. In this lecture we'll consider:

1. Implications of treatment effect heterogeneity for estimation and inference; and
2. Weak instrument issues.

In later lectures, we'll consider two particular IV designs that have been increasingly common: "judges" designs, and the associated many instrument issues, and shift-share designs.

1. SETUP AND REVIEW OF TEXTBOOK MODEL

We'll use the usual sampling framework, assuming $\mathcal{D}_i = (Y_i, D_i, Z_i, W_i)$ are drawn i.i.d. from a large population, where D_i is the treatment, Z_i is a k-vector of instruments, and W_i is an ℓ-vector of controls. Let $X_i = (Z_i', W_i')'$. Our parameters of interest will be defined with respect to this population

*Email: mkolesar@princeton.edu.

1. This is not quite the same thing as the second issue, just because the interpretation of β in the projection $E[Y_i \mid D_i, A_i]$ is not necessarily causal.

Research Question. Think about alternative populations of interest we talked about in the context of ordinary least squares (OLS) regressions in current setting. ⊠

Define the reduced form and the first stage:

$$Y_i = Z_i'\delta + W_i'\psi_Y + u_{Y,i}, \tag{1}$$

$$D_i = Z_i'\pi + W_i'\psi_D + u_{D,i}. \tag{2}$$

The coefficients π and ψ here are defined as (unconditional) best linear predictors. Under regularity conditions, arguments in the appendix show that the OLS estimators $\hat{\delta}$ and $\hat{\pi}$ satisfy[2]

$$\sqrt{n}\left(\begin{pmatrix}\hat{\delta}\\ \hat{\pi}\\ \ddot{Z}'D/n\end{pmatrix} - \begin{pmatrix}\delta\\ \pi\\ Q\pi\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0, \mathcal{V}_0\right), \tag{3}$$

Here $\ddot{Z} = Z - H_W Z$ denotes the residual from the sample projection of Z_i onto the covariates W_i, $H_W = W(W'W)^{-1}W'$ is the hat matrix, $\tilde{Z}_i = Z_i - E[Z_i W_i'] E[W_i W_i']^{-1} W_i$ is the residual from the population projection of Z_i onto W_i, $Q = E[\tilde{Z}_i \tilde{Z}_i']$, and

$$\mathcal{V}_0 = \text{var}\begin{pmatrix} u_i \otimes Q^{-1}\tilde{Z}_i \\ (\tilde{Z}_i\tilde{Z}_i' - Q)\pi + u_{2i}\tilde{Z}_i \end{pmatrix},$$

where $u_i = (u_{Y,i}, u_{D,i})'$.

The asymptotic variance of the $\mathcal{V}_0$ can be consistently estimated using the Eicker-Huber-White (EHW) variance estimator, with the sample residual $\ddot{Z}_i$ replacing $\tilde{Z}_i$, $\hat{u}_i$ replacing u_i, and sample averages replacing population expectations. Here $\hat{u}_i = (\hat{u}_{Y,i}, \hat{u}_{D,i})$, with $\hat{u}_Y = \ddot{Y} - \ddot{Z}\hat{\delta}$ and $\hat{u}_D = \ddot{D} - \ddot{Z}\hat{\pi}$.

We'd like to use these reduced-form estimates for estimation and inference on treatment effects. To that end, we'll need to make some substantive assumptions. We'll start with the textbook assumption that the treatment effects $Y_i(1) - Y_i(0)$ are constant and linear,

Assumption 1 (Constant treatment effects). $Y_i(d) = Y_i(0) + d\beta$.

In this case, letting $\gamma = E[W_i'W_i]^{-1}E[W_i Y_i(0)]$ (notice the analogy in definition of γ for causal inference on superpopulation in the previous set of notes), we can write

$$Y_i = D_i\beta + W_i'\gamma + \epsilon_i, \tag{4}$$

where $\epsilon_i = Y_i(0) - W_i'\gamma$ is called the *structural error*. The name comes from thinking about eq. (4) as a structural equation modeling the outcomes Y_i, in which case γ would have an economic meaning—here we relax that and simply define it as the best linear predictor of $Y_i(0)$.

The key substantive assumption that we need is:

2. We include the asymptotic distribution of $\ddot{Z}'D/n$ in this result since it'll be useful when we consider asymptotics under treatment effect heterogeneity.

Assumption 2 (Random assignment). $E[Z_i \mid Y_i(d), W_i] = \Delta' W_i$ for some $\ell \times k$ matrix Δ.

This assumption entails three conditions:

RANDOM ASSIGNMENT Z is mean-independent of the potential outcomes given W

EXCLUSION RESTRICTION the potential outcomes $Y_i(d, z)$ in fact only depend on d

LINEARITY $E[Z \mid W]$ is linear in W.

The linearity condition is an analog of assuming that the propensity score is linear in controls that we imposed when analyzing OLS in previous lecture. It implies that $\tilde{Z}_i = Z_i - \Delta' W_i$. Notice this is a "design-based" identification condition, in that it only restricts the assignment of the instrument, but it makes no restrictions on the potential outcomes. One could alternatively pursue "model-based" identification.

Research Question. Can we generalize the model-based identification arguments from OLS to the present context? ⊠

Assumptions 1 and 2 deliver the moment condition

$$E[X_i \epsilon_i] = 0$$

that underlies textbook IV theory. To derive the moment condition, note that $Z_i \epsilon_i = Z_i(Y_i(0) - W_i' \gamma)$, and taking conditional expectations then, by Assumption 2, yields $\Delta' W_i (Y_i(0) - W_i' \gamma)$. This is mean zero by definition of γ. Similarly, $W_i \epsilon_i = W_i(Y_i(0) - W_i' \gamma)$ is mean zero by definition of γ.

1.1. *Estimation and Inference*

The textbook model delivers a restriction on the reduced form: substituting the first stage (2) into eq. (4) implies that the reduced-form coefficients are proportional to each other:

$$\delta = \pi \beta. \tag{5}$$

Furthermore, $\psi_Y = \gamma + \psi_D \beta$, and $u_{Yi} = u_{Di} \beta + \epsilon_i$. Therefore, the variance of the structural error is linked to $\Omega(X_i) = E[u_i u_i' \mid X_i]$:

$$\sigma^2(X_i) := \operatorname{var}(\epsilon_i \mid X_i) = \operatorname{var}(u_{Yi} - u_{Di} \beta \mid X_i) = b' \Omega(X_i) b \qquad b = \begin{pmatrix} 1 \\ -\beta \end{pmatrix}.$$

The standard approach to estimation is based on the moment condition $E[X_i \epsilon_i] = 0$. If ϵ_i is homoskedastic, the optimal weight matrix is proportional to $E[X_i X_i']^{-1}$, and using the sample analog $(X'X)^{-1}$ yields the two-stage least squares (TSLS) estimator that minimizes $(Y - D\beta - W\gamma)' H_X (Y - D\beta - W\gamma)$, where $H_X = X(X'X)^{-1}X'$ is the hat

matrix. Minimizing this over (β, θ) yields:

$$\hat{\beta}_{\text{TSLS}} = \frac{D' H_{\ddot{Z}} Y}{D' H_{\ddot{Z}} D} = \frac{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\delta}}{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\pi}}, \qquad \hat{\gamma}_{\text{TSLS}} = (W'W)^{-1} W'(Y - D\hat{\beta}_{\text{TSLS}}).$$

If $k = 1$, then the weight matrix doesn't matter, and we simply have

$$\hat{\beta}_{\text{TSLS}} = \frac{\hat{\delta}}{\hat{\pi}}.$$

By standard generalized method of moments (GMM) arguments, or by applying the delta method to eq. (3) (see Appendix), we obtain

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_1), \qquad \mathcal{V}_1 = \frac{E[\sigma^2(X_i)(\tilde{Z}_i' \pi)^2]}{(\pi' Q \pi)^2}. \tag{6}$$

This variance is usually estimated as

$$\hat{V}_1 = n \frac{\sum_i \hat{\epsilon}^2_{\text{TSLS},i} (\ddot{Z}_i' \hat{\pi})^2}{[\sum_i (\ddot{Z}_i' \hat{\pi})^2]^2}, \quad \hat{\epsilon}_{\text{TSLS}} = \ddot{Y} - \ddot{D}\hat{\beta}_{\text{TSLS}},$$

or equivalently $\hat{\epsilon}_{\text{TSLS},i} = Y_i - D_i \hat{\beta}_{\text{TSLS}} - W_i' \hat{\gamma}_{\text{TSLS}}$. Under homoskedastic errors, this simplifies to

$$\hat{V}_{1,ho} = \frac{\hat{\sigma}^2}{n^{-1} \sum_i (\ddot{Z}_i' \hat{\pi})^2} = \frac{\hat{\sigma}^2}{n^{-1} D' H_{\ddot{Z}} D},$$

where $\hat{\sigma}^2 = n^{-1} \sum_i \hat{\epsilon}^2_{\text{TSLS},i}$.

Remark 1. Note that we could alternatively estimate $\sigma^2(X_i)$ as $(\hat{u}_{Y,i} - \hat{u}_{D,i}\hat{\beta}_{\text{TSLS}})^2$. If $k = 1$, then these two approaches are equivalent, since, using $\hat{\pi}\hat{\beta}_{\text{TSLS}} = \hat{\delta}$, we get $\hat{u}_1 - \hat{u}_2 \hat{\beta}_{\text{TSLS}} = \ddot{Y} - \ddot{Z}\hat{\delta} - \ddot{D}\hat{\beta}_{\text{TSLS}} + \ddot{Z}\hat{\pi}\hat{\beta}_{\text{TSLS}} = \hat{\epsilon}$. However, if $k > 1$, this approach yields a different variance estimator, with different properties under weak instruments, many instruments, or heterogeneous treatment effects.

LIML An alternative approach to estimation that goes back to Anderson and Rubin (1949) is to assume that the structural and first-stage errors $(\epsilon_i, u_{D,i})$ are homoskedastic and jointly normal conditional on X_i, and estimate β by maximum likelihood. The resulting estimator is called limited information maximum likelihood (LIML) (viewing the problem as a simultaneous equation model, we don't use full information by not fully modeling the simultaneity). The estimator takes the form

$$\hat{\beta}_{\text{LIML}} = \underset{\beta}{\text{argmin}} \frac{(1, -\beta)\hat{\Pi}' \ddot{Z} \ddot{Z}' \hat{\Pi} (1, -\beta)}{(1, -\beta) S (1, -\beta)'} = \frac{D' H_{\ddot{Z}} Y - \kappa S_{12}}{D' H_{\ddot{Z}} D - \kappa S_{22}},$$
$$\kappa = \min \text{eig}(S^{-1} \hat{\Pi}' \ddot{Z} \ddot{Z}' \hat{\Pi}).$$

where $\hat{\Pi} = (\hat{\delta}, \hat{\pi})$, and $S = [(\ddot{Y}, \ddot{D}) - \ddot{Z}\hat{\Pi}]'[(\ddot{Y}, \ddot{D}) - \ddot{Z}\hat{\Pi}]/(n - k - \ell)$ is an estimator of $\text{var}(u_i)$ based on the reduced-form residuals. Note that TSLS can also be written in this

form, with $\kappa = 0$.

A third approach to estimation would be to base estimation directly on the restriction (5). In particular, we could form a minimum distance estimator based on this restriction, yielding the objective function

$$\begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}' W \begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}.$$

Under homoskedasticity, the variance of $(\hat{\delta}', \hat{\pi}')'$ is given by $\Omega \otimes Q^{-1}$, so that the feasible weight matrix $W = S^{-1} \otimes \ddot{Z}'\ddot{Z}/n$ will be the optimal. In this case, Goldberger and Olkin (1971) show that the minimum distance estimator of β is numerically equivalent to LIML: so we can think of LIML as an estimator that exploits the proportionality restriction (5).

Proof. Letting $a = (\beta, 1)'$, we can write the objective function as

$$\begin{aligned} \operatorname{vec}(\hat{\Pi} - \pi a')(S^{-1} \otimes (\ddot{Z}'\ddot{Z}/n)) \operatorname{vec}(\hat{\Pi} - \pi a') &= \operatorname{vec}(\hat{\Pi} - \pi a') \operatorname{vec}((\ddot{Z}'\ddot{Z}/n)(\hat{\Pi} - \pi a')S^{-1}) \\ &= \operatorname{tr}((\hat{\Pi} - \pi a')'(\ddot{Z}'\ddot{Z}/n)(\hat{\Pi} - \pi a')S^{-1}) \\ &= \operatorname{tr}(\hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\hat{\Pi}) - 2a'S^{-1}\hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\pi + \pi'(\ddot{Z}'\ddot{Z}/n)\pi a'S^{-1}a \end{aligned}$$

The first-order condition with respect to π implies $\hat{\pi}_{\text{LIML}} = \hat{\Pi}S^{-1}a/(aS^{-1}a)$, so that the objective function with π concentrated out can be written as

$$\operatorname{tr}(T) - a'S^{-1}TS^{-1}a/(aS^{-1}a)$$

where $T = \hat{\Pi}'(\ddot{Z}'\ddot{Z}/n)\hat{\Pi}$. The result then follows from the identity $a'S^{-1}TS^{-1}a/(aS^{-1}a) = b'Tb/b'Sb + \operatorname{tr}(S^{-1}T)$, where $b = (1, -\beta)'$. □

Since the minimum distance objective function doesn't rely on normality or homoskedasticity of the errors, it is clear that LIML will remain asymptotically normal and consistent even without these assumptions—in fact, one can show that it is first-order asymptotically equivalent to TSLS. One can therefore use the estimator $\hat{V}_1$ above for standard errors, with $\hat{\epsilon}_i$ replaced with $\hat{\epsilon}_{i,\text{LIML}}$.[3]

1.2. *What can go wrong?*

So in the textbook model, one could use either TSLS or LIML for estimation, and standard errors based on $\hat{V}_1$ on inference; both approaches are asymptotically equivalent. There are three main reasons why this approach may fail in finite samples:

1. The model is wrong: the proportionality restriction (5), or, equivalently, the moment condition $E[X_i\epsilon_i] = E[X_i(Y_i - D_i\beta - W_i'\gamma)] = 0$ does not hold. This can happen if Assumption 2 doesn't hold (for example, the exclusion restriction fails),

3. One could also replace $\hat{\pi}$ in the standard error formula with the maximum likelihood or minimum distance estimator of π. One could also use the information matrix of the limited information likelihood to get standard errors, although one needs to use the sandwich formula to ensure validity under heteroskedasticity.

or if there is treatment effect heterogeneity (Assumption 1 fails). Note that as far as inference is concerned, this is only an issue if $k > 1$.

Question 1. Why?

2. The delta method fails. As shown in the appendix, since we can write $\beta = \delta' Q\pi / \pi' Q\pi$, it follows that $\hat{\beta}_{\text{TSLS}} = g(\hat{\delta}, \hat{\pi}, \hat{Q})$, where $\hat{Q} = \ddot{Z}'\ddot{Z}/n$. Therefore, we can apply the delta method to eq. (3) with this function g to obtain the asymptotic distribution of TSLS. For the delta method to work, we need g to be continuously differentiable at (δ, π, Q). This will fail if $\pi = 0$. By continuity, this implies that the delta method will work poorly if π is close to zero. This is a weak instrument problem.

3. Inference on the reduced form is unreliable, in that confidence intervals (CIs) for δ or π based on the normal approximation in eq. (3) and robust standard errors are unreliable. This may happen for the usual reasons that EHW standard errors fail, as we discussed in previous lecture. Such issues may be important in practice, as shown in Young (2022). The second reason why this may be an issue is that the number of instruments k is large relative to sample size: if the number of instruments is large relative to sample size, then the reduced-form estimators $\hat{\delta}$ and $\hat{\pi}$ may not be approximately normally distributed.

We'll now explore the first two issues in detail. We'll defer the treatment of the third issue (k is large relative to sample size) to next lecture.

2. TREATMENT EFFECT HETEROGENEITY

We can't do much about failure of Assumption 2 (apart from more flexibly controlling for the covariates if we're worried about linearity of $E[D_i \mid W_i]$), so we'll focus on the implications of the failure of Assumption 1. There are two main implications:

1. TSLS will estimate a weighed average of local average treatment effects (LATEs), but LIML will not in general estimate an object that has a causal interpretation. In the words of Heckman and Vytlacil (2005):

> The relevant question regarding the choice of instrumental variables in the general class of models studied in this paper is "What parameter is being identified by the instrument?" rather than the traditional question of "What is the efficient combination of instruments for a fixed parameter?"—the question that has traditionally occupied the attention of econometricians who study instrumental variables.

2. The standard error for $\hat{\beta}_{\text{TSLS}}$ based on $\hat{V}_1$ is no longer valid.

Imbens and Angrist (1994) show that if there are no covariates, D_i is binary and Assumption 2 holds, then TSLS estimates a weighted average of LATEs, provided an additional monotonicity condition holds. As was the case with OLS under treatment effect heterogeneity, depending on the policy in question, these weights may not be particularly policy relevant, but one can defend the focus on the TSLS estimand in much the same way as when we discussed the OLS estimand. A similar result obtains under multivalued D_i, as shown in Angrist and Imbens (1995). It is straightforward to extend these results to allow for covariates (under the assumption that $E[Z_i \mid W_i]$ is linear), and you can try doing so as an exercise.

To give the result with a binary treatment, suppose that the conditional distribution of Z_i given $W_i = w$ has J_w support points. Given $W_i = w$, order the support points $\{z_{jw}\}_{j=1}^{J_w}$ so that the propensity score $E[D_i \mid Z_i = z_{jw}, W_i = w] = p(z_{j,w}, W_i) = p_{j,w}$ is increasing in j. Let $\alpha(p_{j,w}, w)$ denote the LATE for people for individuals with $W_i = w$ who get treated when the instrument they receive corresponds to propensity score $p_{j+1,w}$ or higher, but not otherwise. Then

$$\beta := \frac{\delta' Q \pi}{\pi' Q \pi} = \int \sum_{j=1}^{J_w - 1} \frac{\lambda_j(w)}{\int \sum_{j=1}^{J_w-1} \lambda_j(w)\, dF_W(w)} \alpha(p_{j,w}, w)\, dF_W(w), \tag{7}$$

where $\lambda_j(w) = (p_{j+1,w} - p_{j,w}) P(p(Z_i, W_i) > p_{j,w} \mid W_i = w) E[\tilde{Z}_i' \pi \mid W_i, p(Z_i, W_i) > p_{j,w}]$. These weights will be positive if the last term is positive. Like in the case with OLS, these weights are not particularly pretty, though I suspect they may have an efficiency interpretation.

Like with OLS, including interactions of the instrument with the covariates would make them prettier, though that means we may run into many instrument issues.

In contrast, the estimand of LIML generally no longer has a causal interpretation in the sense that it estimates a convex combination of LATEs, unless all LATEs happen to be the same, as shown in Kolesár (2013). The reason is that under treatment effect heterogeneity, the proportionality restriction (5) no longer holds. Because LIML imposes this condition with a non-diagonal weight matrix, its behavior (in terms of its probability limit and its asymptotic variance) is quite sensitive to failures of this restriction. On the other hand, TSLS constructs a single instrument $\hat{Z}_i = \ddot{Z}_i' \hat{\pi}$ based on the first stage, and estimates β using with an IV estimator using single constructed instrument, $\hat{\beta}_{\text{TSLS}} = \hat{Z}_i' Y_i / \hat{Z}_i' D_i$: this makes it much more robust to treatment effect heterogeneity. Because of this result, one should not use LIML in practice if one has doubts about eq. (5) holding.

- Interact binary instrument with covariates.

2.2. *Inference*

Since the model in eq. (4) is misspecified, it will also generally matter if one wishes to do inference conditionally on X_i, conditionally on some other variable, or unconditionally, in analogy to our discussion in the previous lecture. Here we'll focus on unconditional inference; see Evdokimov and Kolesár (2018) for conditional results, if the estimand is defined as $\delta'\ddot{Z}'\ddot{Z}\pi/\pi'\ddot{Z}'\ddot{Z}\pi$ instead.

In particular, as shown in the appendix, an application of the delta method yields

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_2), \qquad \mathcal{V}_2 = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{2i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{(\pi' Q \pi)^2}, \tag{8}$$

where

$$\epsilon_i = \tilde{Y}_i - \tilde{D}_i\beta = Y_i - D_i\beta - W_i'\gamma = \tilde{Z}_i\pi_\Delta + u_{Y,i} - u_{D,i}\beta, \tag{9}$$

$$\gamma = E[W_iW_i]^{-1}E[W_i'(Y_i - D_i\beta)] = \Delta\pi_\Delta + \psi_Y - \psi_D\beta, \tag{10}$$

and $\pi_\Delta = \delta - \pi\beta$. Note that if eq. (5) holds, then $\pi_\Delta = 0$, and $\mathcal{V}_2 = \mathcal{V}_1$. In general, however, the variances will be different, and we need to instead use the variance estimator

$$\hat{\mathcal{V}}_2 = n\frac{\sum_{i=1}^n[((\tilde{Z}_i'(\hat{\delta} - \hat{\pi}\hat{\beta}))\hat{u}_{2i} + \hat{\epsilon}_{\text{TSLS},i}(\tilde{Z}_i'\hat{\delta}))^2]}{(\sum_i \tilde{Z}_i\hat{\pi})^2},$$

Remark 2. The result that the usual standard errors are different under treatment effect heterogeneity can be found in the appendix in Imbens and Angrist (1994); it is also derived in Lee (2018) (if one simplifies the expression given in that paper, one obtains the expression above), or in Evdokimov and Kolesár (2018) (who in addition assume that $E[u_i \mid X_i] = 0$, which we don't assume here).

Remark 3 (Single instrument). If $k = 1$, then $\pi_\Delta = 0$, since δ and π are scalars, and $\beta = \delta/\pi$. Hence, $\hat{\mathcal{V}}_2 = \hat{\mathcal{V}}_1$, $\mathcal{V}_2 = \mathcal{V}_1$. Also, in this case, LIML and TSLS coincide.

Remark 4 (Known propensity score). We can interpret TSLS as an IV estimator that uses a single constructed instrument $\hat{Z}_i = Z_i'\hat{\pi}$, and controls W_i. You can think of $\hat{Z}_i$ as an estimate of $Z_i'\pi$, which under some conditions (which ones are they?) is the best single instrument. If D_i is binary, and there are no covariates, we can interpret $Z_i'\pi$ as the propensity score. In other words, β in eq. (7), and γ in eq. (10) can be thought of as solutions to the exactly identified moment condition

$$E\left[\begin{pmatrix}\pi' Z_i \\ W_i\end{pmatrix}(Y_i - D_i\beta - W_i'\gamma)\right] = 0,$$

and TSLS can also be thought of as a GMM estimator based on a sample analog of this moment condition, with $\hat{\pi} = (\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'D$ replacing the unknown nuisance parameter π. Suppose you're an oracle who knows the propensity score. Then you can use the above

moment condition, but without having to use the estimate $\hat{\pi}$. This yields the estimates $\hat{\beta}_{\text{oracle}} = (\pi'\ddot{Z}'D)^{-1}\pi'\ddot{Z}'Y$ and $\hat{\gamma}_{\text{oracle}} = (W'W)^{-1}W'(Y - D\hat{\beta}_{\text{oracle}})$. Since the model based on the above moment condition is exactly identified, by standard GMM arguments, regardless of the presence of treatment effect heterogeneity, $\sqrt{n}(\hat{\beta}_{\text{oracle}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_1)$, with

$$\mathcal{V}_1 = \frac{E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]}{E[(\tilde{Z}_i'\pi)^2]^2},$$

and ϵ_i defined in eq. (9). So under constant treatment effects, TSLS (and LIML) achieve oracle efficiency: the fact that the first-stage (i.e. the propensity score) is estimated doesn't affect the asymptotic variance (we'll explore the reason for this in later lectures). But under heterogeneous treatment effects, we do not get to this oracle: the difference between $\mathcal{V}_1$ and $\mathcal{V}_2$ exactly accounts for the fact that π is estimated.

Remark 5 (Known first stage). There is an alternative oracle estimator of β, based on the moment condition

$$E\left[\begin{pmatrix}\pi'Z_i\\ W_i\end{pmatrix}(Y_i - (Z_i'\pi)\beta - W_i'\delta)\right] = 0, \qquad \delta = \gamma + \psi_2\beta.$$

with $\hat{\beta}_{\text{oracleo}} = (\pi\ddot{Z}'\ddot{Z}\pi)^{-1}\pi'\ddot{Z}'Y$. This estimator is an analog of the TSLS estimator with a known first stage: if we know π, we don't need to run the first stage of the two-stage procedure, only the second stage. The asymptotic variance of this oracle is given by

$$\mathcal{V}_3 = \frac{E[(\epsilon_i + u_{2i}\beta)^2(\tilde{Z}_i'\pi)^2]}{E[(\tilde{Z}_i'\pi)^2]^2}.$$

In contrast to the previous case, knowing the first stage *does* have an effect on the asymptotic variance of this estimator: $\mathcal{V}_3 \neq \mathcal{V}_2$ even under homogeneous treatment effects. You may recall from your undergraduate days that it is incorrect to report standard errors from the regression of Y_i onto $\hat{Z}_i = Z_i'\hat{\pi}$ and onto W_i as the TSLS standard errors (as mentioned, e.g., on page 97 in Wooldridge 2010)—these standard errors estimate $\mathcal{V}_3$. So when person A says that (under constant treatment effects), not knowing the first stage does affect standard errors, and person B says that it does not, both are right, they just have a different oracle (a different set of moment conditions) in mind.

3. WEAK INSTRUMENTS

In many applications, the first-stage coefficients may be close to zero. In such cases, the delta method may not work well, so that CIs based on $\hat{V}_1$ or $\hat{V}_2$ may undercover, and the TSLS estimator may be biased (in the sense that its distribution will not be centered at β, even in large samples).

Example 1. As an example, consider the problem of estimating the elasticity of intertemporal substitution (EIS). One approach involves estimating the log-linearized Euler equa-

tion based on a portfolio choice problem of an agent with Epstein-Zin preferences, which is given by (see Campbell (2003) and Yogo (2004) for derivation)

$$E_t[\Delta c_{t+1} - \mu_j - \psi r_{j,t+1}] = 0$$

where $r_{j,t+1}$ is return on asset j, Δc_{t+1} is consumption growth, μ_j is a constant, and ψ is the EIS, and E_t denotes expectation conditional on the agent's information set at time t.

We could try to estimate ψ by running the regression

$$\Delta c_{t+1} = \mu_j + \psi r_{j,t+1} + e_t.$$

However, the error term in that regression, $e_t = \Delta c_{t+1} - E_t[\Delta c_{t+1}] - \psi(r_{j,t+1} - E_t[r_{j,t+1}])$ is going to be correlated with $r_{j,t+1}$. On the other hand, e_t will be by definition uncorrelated with any variables in the information set at time t, so we can use those as instruments. Alternatively, we could instrument for Δc_{t+1} using the same instruments, and estimate $1/\psi$. Instrumenting for Δc_{t+1}, computing $\hat{\psi}_{inv} = 1/\psi$ using TSLS, and then reporting $\hat{\psi}_{inv}^{-1}$ as an estimate of ψ is known as reverse TSLS.

With one instrument, the two approaches are numerically equivalent. When there is more than one instrument, both IV regressions are asymptotically equivalent under standard asymptotics, so we would expect the estimates to be approximately similar in finite samples.

The problem is that empirical estimates of both ψ and of $1/\psi$ are small. For instance, Campbell (2003, Table 9) reports a 95% CI $[-0.14, 0.28]$ for ψ, and CI $[-0.73, 2.14]$ for $1/\psi$, using quarterly U.S. data (1947–1998) on non-durable consumption and T-bill returns.

The reason is that the equivalence breaks down when the instruments are weak, and weak instruments are likely the culprit here because both consumption growth and asset returns are notoriously difficult to predict. ⊠

Remark 6 (Aside). The fact that the two approaches, estimating ψ and $1/\psi$ no longer give compatible results when instruments are weak led Hahn and Hausman (2002) to propose a test for weak instruments based on the difference between $\hat{\psi}$ and $\hat{\psi}_{inv}^{-1}$. Unfortunately, the power of the test against weak or irrelevant instruments is low and the tests are not consistent against irrelevant instruments (Hausman, Stock, and Yogo 2005).

In a recent review, Andrews, Stock, and Sun (2019) document that a substantial fraction of IV regressions in papers recently published in the *American Economic Review* have first-stage F statistics under 20, which means weak instruments are frequently encountered in practice.

Before we dive into the issues and solutions, here are the key takeaways:

- Most of the theoretical literature focused on the simplest case with constant treatment effects and homoskedastic errors for tractability. However, not all the recommendations from this literature carry over to the general case with heteroskedasticity, clustering, or serial dependence, and to heterogeneous treatment effects.

- For detecting weak instruments, it is popular to use the $F > 10$ rule of thumb. This rule relies heavily on homoskedasticity, and the rule looks quite different under heteroskedasticity. Moreover, it is unclear how one should use it in practice.

- If $k = 1$, the usual CIs work well unless the endogeneity problem is very severe. Furthermore, one can report the Anderson and Rubin (1949) CIs as a robustness check: they are robust to weak instruments, and efficient under strong instruments. The tF procedure of Lee et al. (2022) is another alternative. Since the critical values it uses are determined under extreme endogeneity, it quantifies the precision gains of the Wald test due to (implicitly or explicitly) ruling out extreme endogeneity—these precision gains can be substantial if $F \leq 20$. If one knows the sign of the first stage, the (median) bias of TSLS is minimal if one conditions on the estimated first-stage being right-signed.

- If $k > 1$, things are more complicated, and very hard if there is more than one endogenous variable. None of the existing testing procedures are robust to treatment effect heterogeneity

  Research Question. Can we develop such a procedure? ⊠

  Similarly, the finding that estimators such as LIML are more robust to the weak instrument problem is sensitive to the failure of constant treatment effects assumption.

- Of course, weak instrument and weak identification issues are present in more complicated, non-linear models. Diagnostics and solutions in these models are an active area of research.

3.1. *The weak instrument problem with a single instrument*

To see what can go wrong when the instruments are weak, let's consider the extreme case in which the instruments are irrelevant, so that $\pi = 0$. For simplicity, suppose that we have a single instrument. Then, by the central limit theorem (CLT) since $\tilde{Z}'D = \tilde{Z}'u_2$,

$$\hat{\beta}_{\text{TSLS}} = \frac{\tilde{Z}'Y}{\tilde{Z}'D} = \frac{\tilde{Z}'D\beta + \tilde{Z}'\epsilon}{\tilde{Z}'u_2} = \beta + \frac{n^{-1/2}\tilde{Z}'\epsilon}{n^{-1/2}\tilde{Z}'u_2} \Rightarrow \beta + \frac{\Sigma_{11}^{1/2}\mathcal{Z}_\epsilon}{\Sigma_{22}^{1/2}\mathcal{Z}_2},$$

where $\mathcal{Z}_\epsilon$ and $\mathcal{Z}_2$ are standard normal with covariance $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, and $\Sigma = \text{var}(\tilde{Z}_i(\epsilon_i, u_{2i})')$. We can think of ρ as measuring the endogeneity problem: under homoskedasticity, ρ measures the correlation between the structural error ϵ_i in Equation (4) and the first-stage error. If $\rho = 0$, then OLS is consistent. We therefore refer to ρ as (the degree of) endogeneity.

To simplify this a little, let's decompose $\mathcal{Z}_\epsilon$ into a part that's perfectly correlated with

$\mathcal{Z}_2$ and part that's independent letting

$$\mathcal{Z}_\perp = (1-\rho^2)^{-1/2}\mathcal{Z}_\epsilon - \rho\mathcal{Z}_2,$$

so that $\mathcal{Z}_\perp$ and $\mathcal{Z}_2$ are independent standard normal, and $\mathcal{Z}_\epsilon = \rho\mathcal{Z}_2 + \sqrt{1-\rho^2}\mathcal{Z}_\perp$. Plugging this in:

$$\hat{\beta} \Rightarrow \beta + \frac{\Sigma_{12}}{\Sigma_{22}} + \sqrt{\frac{(1-\rho^2)\Sigma_{11}}{\Sigma_{22}}}C, \tag{11}$$

where $C = \mathcal{Z}_\perp/\mathcal{Z}_2$ has a Cauchy distribution.

- If $\tilde{Z}_i$ is mean-independent of (ϵ_i, u_{2i}), then $\beta + \Sigma_{12}/\Sigma_{22} = \beta + \rho\sqrt{\Sigma_{11}/\Sigma_{22}}$ is the probability limit of OLS. So asymptotically, TSLS is median-biased, and centered around the OLS limit. Because of the Cauchy distribution, it has thick tails. So it's like taking the OLS estimator, and adding heavy-tailed noise to it.
- TSLS is inconsistent, and doesn't converge to anything even asymptotically: its distribution doesn't get less spread out even asymptotically. So there is a sharp discontinuity in the asymptotic distribution, with the rate of convergence changing, depending on whether $\pi = 0$. As a result, the bootstrap, the m-out-of-n bootstrap, and subsampling will not work either (see, for instance Andrews and Guggenberger 2010)
- The Wald statistic (i.e. the usual t-ratio) is given by

$$W = \frac{\hat{\beta}-\beta}{\sqrt{\frac{\sum_i \hat{\epsilon}^2_{\text{TSLS},i}(\tilde{Z}_i\hat{\pi})^2}{[\sum_i(\tilde{Z}_i\hat{\pi})^2]^2}}} = \frac{\hat{\beta}-\beta}{\sqrt{\frac{n^{-1}\sum_i \hat{\epsilon}^2_{\text{TSLS},i}\tilde{Z}_i^2}{[n^{-1/2}\sum_i \tilde{Z}_i u_{2i}]^2}}}$$

  Now, by CLT and the law of large numbers (LLN), $n^{-1}\sum_i \hat{\epsilon}^2_{\text{TSLS},i}\tilde{Z}_i^2 = n^{-1}\sum_i(\tilde{\epsilon}_i + \tilde{u}_i(\beta - \hat{\beta}_{\text{TSLS}}))^2\tilde{Z}_i^2 \Rightarrow \Sigma_{11}(1 - 2\rho\mathcal{Z}_\epsilon/\mathcal{Z}_2 + \mathcal{Z}_\epsilon^2/\mathcal{Z}_2^2)$, and $n^{-1/2}\sum_i \tilde{Z}_i u_{2i} \Rightarrow \Sigma_{22}^{1/2}\mathcal{Z}_2$, so by the continuous mapping theorem,

$$W \Rightarrow \frac{\mathcal{Z}_\epsilon/\mathcal{Z}_2}{\sqrt{\mathcal{Z}_2^{-2}(1 - 2\rho\mathcal{Z}_\epsilon/\mathcal{Z}_2 + \mathcal{Z}_\epsilon^2/\mathcal{Z}_2^2)}} = \frac{\mathcal{Z}_\perp/\mathcal{Z}_2 + \rho/\sqrt{1-\rho^2}}{\sqrt{1/\mathcal{Z}_2^2 + Z_\perp^2/\mathcal{Z}_2^4}}.$$

  Depending on the value of the structural correlation ρ implied by the null, the null rejection probability can be a lot lower, or a lot higher than 5%. See Figure 1.

Remark 7. Early evidence of problems with weak instruments goes back to Nelson and Startz (1990b, 1990a) who show that, in the exactly identified case, if the instruments are weak, the density of the IV estimator in finite samples may be very far away from its "asymptotic" distribution. However, their striking results appear to stem from the fact that they study the model

$$Y = D\beta + \epsilon, \qquad D = Z\pi + \underbrace{\epsilon_\perp\lambda + \epsilon}_{u_2},$$

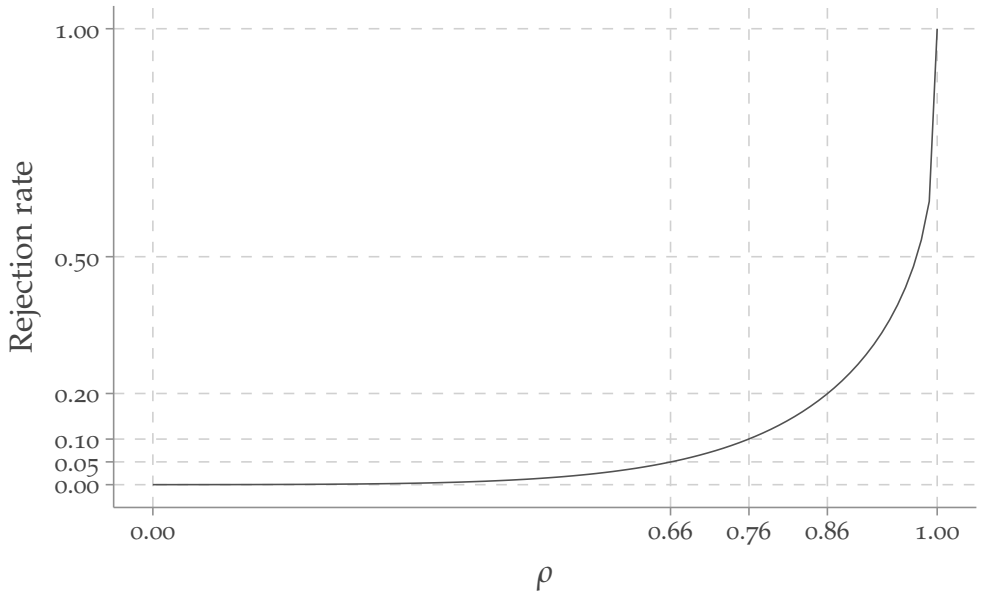


Figure 1: Rejection rate of the Wald test with nominal level 0.05 when $k = 1$, with irrelevant instruments.

and treat both Z *and* $\epsilon_\perp$ as fixed, meaning that part of the first-stage error u_2 is fixed. Bound, Jaeger, and Baker (1995) give convincing evidence of weak instrument issues based on the Angrist and Krueger (1991) study: in their Table 3, they use randomly generated information instead of the quarter of birth. The results look strikingly "reasonable".

To capture the weak instrument problem, we can think of the issue as doing inference on β given a single observation $(\hat{\delta}, \hat{\pi})$ that is distributed

$$\operatorname{vec}(\hat{\Pi}) \sim \mathcal{N}\left(\begin{pmatrix}\pi\beta\\ \pi\end{pmatrix}, \mathcal{V}\right), \qquad \mathcal{V} = E[\Omega(X_i) \otimes Q^{-1}\tilde{Z}_i\tilde{Z}_i'Q^{-1}]/n, \tag{12}$$

with $\mathcal{V}$ and Q known.

This suppresses any complications arising from issue 3, that is non-normality of the reduced form estimates, or difficulties with estimating $\mathcal{V}_0$ and focuses attention solely on the weak instruments problem. This can be formally justified in two ways:

1. Assume that $\hat{\mathcal{V}}_0 \xrightarrow{p} \mathcal{V}_0$, and that eq. (3) holds. Then (12) is the right limit experiment if we make no further assumptions, as argued in Müller (2011).

2. Assume that $\pi = C/\sqrt{n}$ for some fixed constant C. This is the idea of Staiger and Stock (1997). This is in similar spirit as the local-to-unity asymptotics in time series autoregressions: here π is local to 0. As long as we restrict ourselves to tests that are functions of $\hat{\Pi}$, then under this sequence, the problem is asymptotically equivalent the problem of inference given a single observation in the model (12).

These asymptotics are known as weak instrument asymptotics.

In the limit experiment (12), given the known reduced-form error $\mathcal{V}$, the distribution of $\hat{\beta}$, the Wald statistic, or any other object is governed by the unknown parameters (β, π). Oftentimes, it turns out to be more convenient to reparametrize the problem. If $k = 1$, in particular, it is convenient to instead use the parametrization $(E[F], \rho)$, where $E[F] = \pi_2^2/\mathcal{V}_{22} + 1$ is the expectation of the first-stage F statistic, $F = \hat{\pi}^2/\mathcal{V}_{22}$, and

$$\rho(\mathcal{V}, \beta) = \frac{\mathcal{V}_{12}/\mathcal{V}_{22} - \beta}{\sqrt{\mathcal{V}_{11}/\mathcal{V}_{22} - 2\mathcal{V}_{12}\beta/\mathcal{V}_{22} + \beta^2}} \tag{13}$$

is the degree of endogeneity (as discussed above). Then, letting $\mathcal{Z}_\epsilon$ and $\mathcal{Z}_2$ denote standard normal random variables with correlation ρ, we have

$$\hat{\beta} - \beta \sim (\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{1/2} \frac{\mathcal{Z}_\epsilon}{\sqrt{E[F] - 1} + \mathcal{Z}_2},$$

and the Wald statistic has the distribution

$$W \sim \frac{\text{sign}(\sqrt{E[F] - 1} + \mathcal{Z}_2)\mathcal{Z}_\epsilon}{\sqrt{Z_\epsilon^2/(\sqrt{E[F] - 1} + \mathcal{Z}_2)^2 - 2\rho\mathcal{Z}_\epsilon/(\sqrt{E[F] - 1} + \mathcal{Z}_2) + 1}}. \tag{14}$$

See appendix for derivation.

3.2. *Detecting weak instruments*

Let's consider the general case with potentially several instruments. To test for weak instruments, we first need to define what we mean by the term. Stock and Yogo (2005) start with eq. (12), under homoskedastic errors, so that

$$\mathcal{V} = (\Omega/n) \otimes Q^{-1}. \tag{15}$$

The parameter space is given by (β, π). They give two definitions, each of which gives a set of values $\Pi_W \subseteq \mathbb{R}^k$ for π for which the instruments are weak:

1. The bias of TSLS relative to OLS is bigger than 0.1 for some β.
2. The Wald test based on TSLS has size over 10% for some β.

We can actually read off when the instruments are weak from Table 1 in Richardson (1968), who showed that (we give a derivation based on Sawa (1972) in the appendix)

$$b_{\text{TSLS}} := \frac{E[\hat{\beta}_{\text{TSLS}} - \beta]}{\beta_{WOLS} - \beta} = 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^1 x^{k/2-1} e^{\frac{\mu^2}{2}x} dx = \\ = 1 - \frac{k(E[F] - 1)}{2} \int_0^1 x^{k/2-1} e^{(x-1)k(E[F]-1)/2} dt, \tag{16}$$

where $\beta_{WOLS} = \Omega_{12}/\Omega_{22}$ is the limit of OLS under weak-instrument asymptotics, and $E[F] = n\pi' Q\pi / k\Omega_{22} + 1$ is the expectation of the first-stage F under homoskedasticity (henceforth, we ignore the difference between Q and $\ddot{Z}'\ddot{Z}/n$ for simplicity). Sometimes, instead of $E[F]$, people use the non-centrality parameter of the F-statistic, $k(E[F]-1) = n\pi' Q\pi / k\Omega_{22}$, called the *concentration parameter*.

Some notable take-aways:

- The relative TSLS bias (relative to OLS) depends only on $E[F]$, and not the value of β or degree of endogeneity. This suggests that (at least under homoskedasticity), $E[F]$ is the right quantity to measure instrument strength
- The expression is only well-defined if $k \geq 2$. More generally, $E[\hat{\beta}^p_{\text{TSLS}}]$ exists only for $p \leq k-1$ (Richardson 1968).

  Question 2. What is the implication of this for Monte Carlos?

- For k even, we can use repeated application of integration by parts to evaluate the integral, so that, for instance,

$$b_{\text{TSLS}} = \begin{cases} e^{1-E[F]} & \text{if } k = 2, \\ \frac{1-e^{2-2E[F]}}{2E[F]-2} & \text{if } k = 4. \end{cases}$$

  For k odd, we can evaluate the integral numerically.

- By evaluating the relative bias and using the fact that $k \cdot F$ has a non-central χ^2_k distribution, we can derive critical values for testing the hypothesis $H_0 \colon b_{\text{TSLS}} \leq 0.1$. For example, for $k = 2$ we need $E[F] = 1 - \log(0.1) \approx 3.30$ for TSLS bias to be 10% of OLS bias, which leads to the critical value 7.85. For $k = 3, 4, 5$, we obtain critical values $9.18, 10.23, 10.78$, and so on. These critical values fluctuate between 10 and 11.5 for larger values of F. These roughly match Table 5.1 in Stock and Yogo (2005) for $k \geq 3$. I am not sure why they don't report the critical value for $k = 2$. For $k = 1$, there is no bias, and hence no critical value.

  You can see for yourself by running 6 lines of R code:

```
b <- function(mu, k)
    1-integrate(function(x) mu/2*x^(k/2-1)*exp((x-1)*mu/2),
                                lower=0, upper=1)$value
crit <- function(k)
    qchisq(p=0.95, df=k, ncp=uniroot(function(mu)
              b(mu, k)-0.1, lower=0.1, upper=10*k)$root)/k
```

- The $F > 10$ rule of thumb was suggested in Staiger and Stock (1997). Since the critical value is "close" to 10 regardless of the number of instruments k, this calculation "justifies" it.

For the second definition, the critical value increases in k (see Tables 5.1 and 5.2 in Stock and Yogo 2005), starting with 16.38 when $k = 1$. The fact that they increase in k is actually quite important, why?

Remark 8. A alternative second definition would be to specify a set of parameters (β, π) for which the Wald test overrejects. This would lead to a very different test, especially if we're willing to a priori restrict ρ (or equivalently β): since the least favorable value of ρ is 1, one could in such cases tolerate much smaller values of $E[F]$.

In particular, suppose that $k = 1$. Then, using Figure 2, we can compute the rejection rates of the Wald test for different values of $E[F]$ and ρ (if there is heteroskedasticity, we use the heteroskedasticity-robust version of F, we drop the subscript here). Figure 2 gives the plot, which now also appears in Angrist and Kolesár (2024).

Without restricting the value of ρ, if we want to keep the rejection rate below 10%, we need to ensure that $E[F] \geq 6.88$ according to the plot. The first-stage F statistic (the square of the t-statistic) is distributed non-central χ^2_1 with non-centrality parameter $E[F] - 1$. This distribution is stochastically increasing in the non-centrality parameter, so we can use the first-stage F statistic along with the critical value based on the 95% quantile of non-central χ^2_1 with non-centrality parameter 5.88, which yields the Stock and Yogo (2005) critical value: `qchisq(p=0.95, df = 1, ncp = 5.88)=16.56`. An $F > 10$ cutoff would apply if we want to keep the rejection rate below 13.4%.

Note, however, that the set of parameters where overrejection occurs is rather restricted: in particular, if we're willing to put a priori bounds on the value of β that would lead to $|\rho| \leq 0.76$ (for a given covariance matrix of the reduced form coefficients $\mathcal{V}$), then we never need to worry about the weak instrument problem in the sense that the Wald test will not overreject by more than 5%. Angrist and Kolesár (2024) argue that in many applications in labor economics, such large values of endogeneity are unlikely.

If $k > 1$ and the errors are not homoskedastic, $\mathcal{V}$ doesn't have the Kronecker structure in eq. (15) above, and the $F > 10$ rule of thumb should not be used, whether calculated using the robust F statistic

$$F_{\mathrm{r}} = \frac{1}{k}\hat{\pi}'\mathcal{V}_{22}^{-1}\hat{\pi} = \frac{n}{k}\hat{\pi}'QE[\Omega_{22}(X_i)\tilde{Z}_i\tilde{Z}_i']Q\hat{\pi}$$

or the homoskedastic version, F_{h}

As an alternative, Montiel Olea and Pflueger (2013) instead suggest a pre-test based on what they call an effective first-stage F,

$$F_{\mathrm{eff}} = \frac{\hat{\pi}'Q\hat{\pi}}{\mathrm{tr}(\mathcal{V}_{22}Q)} = \frac{k/n \cdot \Omega_{22}F_{\mathrm{h}}}{\mathrm{tr}(Q^{-1}E[\Omega_{22}(X_i)\tilde{Z}_i\tilde{Z}_i']/n)}$$

Note if $k = 1$, then $F_{\mathrm{eff}} = F_{\mathrm{r}}$.

The reason for suggesting this statistic is that the bad behavior of the TSLS estimator is determined whether its denominator, given by $\hat{\pi}'Q\hat{\pi}$ in the limit experiment (12), is close to zero. F_{h} measures this object, while F_{r} measures the wrong object: its non-

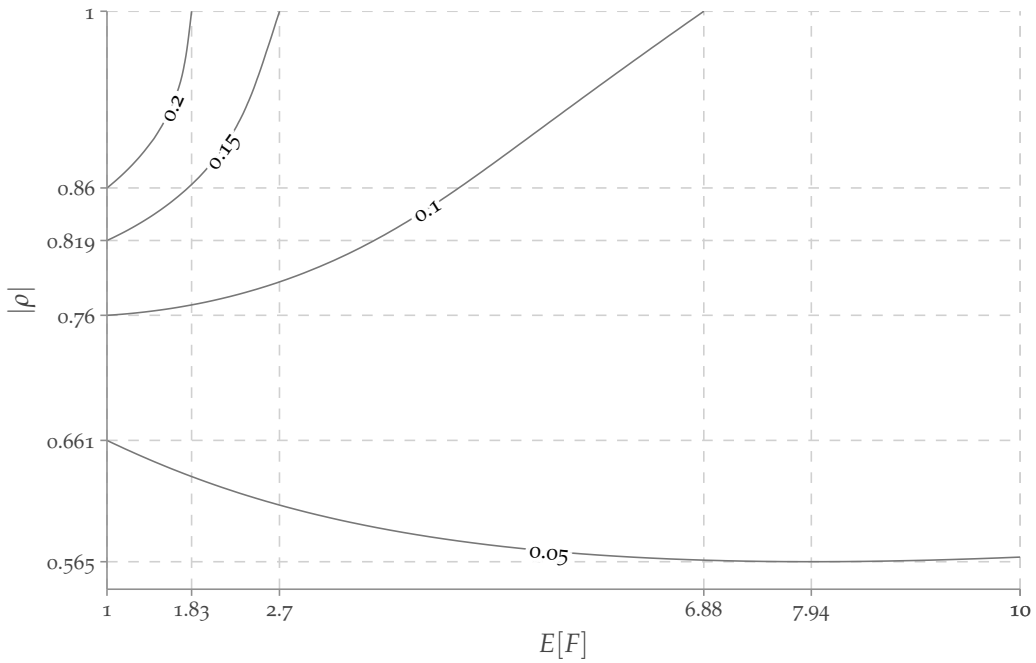


Figure 2: Contour plot of rejection rates of the Wald test with nominal level 0.05 when $k = 1$ as a function of $E[F]$ and ρ. Figure 1 gives a cross-section of this plot at $E[F] = 1$.

centrality parameter is not proportional to $\pi' Q \pi$. F_{h} gets the standard errors (scaling) wrong, however; F_{eff} gets them right "on average". Montiel Olea and Pflueger (2013) compute the critical values needed to ensure that an approximation to the bias of TSLS relative to the OLS bias doesn't exceed 10%.

To summarize:

- If $k = 1$, $F_{\text{eff}} = F_{\text{r}}$, and one can use the Stock and Yogo (2005) cutoff for ensuring that the size of the Wald test doesn't exceed 10%, that is $F_{\text{r}} \geq 16.6$
- If $k > 1$, use the Montiel Olea and Pflueger (2013) test, that ensures small TSLS bias

Note the discrepancy in what the rule is based upon: it's not possible to state the rule in terms of TSLS bias if $k = 1$, and, at the same time, a rule using the size of the Wald test based on TSLS when $k > 1$ would be very conservative.

However, screening on the first-stage F-statistic appears to compound, rather than reduce, inferential problems arising from weak instruments (see Andrews, Stock, and Sun (2019) for discussion and suggestive evidence), so one should use F-statistic as a diagnostic, not a decision rule.

When $k = 1$, then eq. (12) simplifies to a ratio of means problem,

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \mathcal{N}_2 \left(\pi \begin{pmatrix} \beta \\ 1 \end{pmatrix}, \mathcal{V} \right), \qquad \mathcal{V} = \frac{E[\Omega(X_i)\tilde{Z}_i^2]}{nQ^2}$$

Doing inference about β is therefore equivalent to the problem of doing inference about ratio of means of a bivariate Gaussian variable. This connection has been pointed out by Zellner (1978), and Mariano and McDonald (1979).

The ratio of normal means is an old problem in statistics, dating back to at least Fieller (1932). Fieller (1940, 1954) proposed the following solution: under the null $H_0 \colon \beta = \beta_0$, the distribution of the statistic $\hat{\delta} - \hat{\pi}\beta_0$ does not depend on the nuisance parameter π: it's pivotal, so that

$$S(\beta_0) = \frac{(\hat{\delta} - \beta_0\hat{\pi})^2}{b_0'\mathcal{V}b_0} \sim_{H_0} \chi_1^2, \qquad b_0 = \begin{pmatrix} 1 \\ -\beta_0 \end{pmatrix}.$$

Therefore, the CI

$$\begin{aligned} &\left\{\beta_0 \in \mathbb{R} \colon S(\beta_0) \leq z_{1-\alpha/2}^2\right\} = \left\{\beta_0 \colon \frac{(\hat{\beta} - \beta_0)^2 F}{\mathcal{V}_{11}/\mathcal{V}_{22} - 2\beta_0\mathcal{V}_{12}/\mathcal{V}_{22} + \beta_0^2} \leq z_{1-\alpha/2}^2\right\} \\ &= \left\{\beta_0 \colon (F - z_{1-\alpha/2}^2)\beta_0^2 + 2(z_{1-\alpha/2}^2\mathcal{V}_{12}/\mathcal{V}_{22} - \hat{\beta}F)\beta_0 + \hat{\beta}^2 F - z_{1-\alpha/2}^2\mathcal{V}_{11}/\mathcal{V}_{22} \leq 0\right\}. \end{aligned}$$

will have coverage $1 - \alpha$ independently of the parameter π. Here $F = \hat{\pi}^2/\mathcal{V}_{22}$ denotes the first-stage F statistic.

Let $\bar{F} = F\mathcal{V}_{22}\hat{b}'\mathcal{V}\hat{b}/|\mathcal{V}|$ denote the F-statistic for testing $(\delta, \pi) = 0$, where $\hat{b} = (1, -\hat{\beta})'$. Note that $\bar{F} - F = F(\hat{\beta}\mathcal{V}_{22} - \mathcal{V}_{12})^2/|\mathcal{V}| \geq 0$.

Since the above display is a quadratic equation, there are three forms the CI can take.

1. If $F \geq z_{1-\alpha/2}^2$ (that is, we reject the null that $\pi = 0$), then the CI takes the form of an interval $[C_1, C_2]$, with endpoints given by

$$C_j = \hat{\beta} + z_{1-\alpha/2}^2 \frac{\hat{\beta} - \mathcal{V}_{12}/\mathcal{V}_{22}}{F - z_{1-\alpha/2}^2} \pm z_{1-\alpha/2} \frac{\sqrt{(\bar{F} - z_{1-\alpha/2}^2)|\mathcal{V}|}}{(F - z_{1-\alpha/2}^2)\mathcal{V}_{22}}$$

2. If $\bar{F} \geq z_{1-\alpha/2}^2 \geq F$, then it takes the form $(-\infty, C_2] \cup [C_1, \infty)$, with C_j given in the previous display

3. If $\bar{F} \leq z_{1-\alpha/2}^2$ (we don't reject the null that $\Pi = 0$, with critical value based on χ_1^2 rather than χ_2^2), then it is given by $\mathbb{R}$.

This CI is a special case of the CI proposed by Anderson and Rubin (1949) (discussed below), so it is often referred to as the Anderson-Rubin (AR) CI.

In contrast, the usual Wald CI is given by

$$\hat{\beta} \pm z_{1-\alpha/2} \frac{\sqrt{\hat{b}' \mathcal{V} \hat{b}}}{|\hat{\pi}|} = \hat{\beta} \pm z_{1-\alpha/2} \frac{\sqrt{\bar{F} |\mathcal{V}|}}{F \mathcal{V}_{22}}.$$

Comparing the two intervals, it follows that

- The AR CI is always longer than the usual Wald CI. It's obviously longer if $F \leq z^2_{1-\alpha/2}$. Otherwise, if $F \geq z^2_{1-\alpha/2}$, then it is longer so long as $(\bar{F} - z^2_{1-\alpha/2})/(F - z^2_{1-\alpha/2})^2 \geq \bar{F}/F^2 \iff 0 \geq F(F - \bar{F}) + \bar{F}(z^2_{1-\alpha/2} - F)$. But the right-hand side is always negative since $z^2_{1-\alpha/2} \leq F \leq \bar{F}$.
- Under standard asymptotics, $F = O_p(n)$, and $\bar{F} = O_p(n)$, so that

  $$C_j = \hat{\beta} + O_p(1/n) + z_{1-\alpha/2} \frac{\sqrt{\hat{b}' \mathcal{V} \hat{b}}}{|\hat{\pi}|} \sqrt{1 + O_p(1/n)},$$

  so that the AR and Wald CIs are asymptotically equivalent.
- This suggests that one should always use the AR CI: it works under weak instruments, and it is as efficient as the Wald CI under strong instruments. One problem with the CI, however, is that it is not bet-proof, as discussed in Müller and Norets (2016): because its coverage is exactly 95%, and because we know that if the CI is given by $\mathbb{R}$, its conditional coverage is 100%, this means that its conditional coverage (conditional on knowing the shape of the CI) when it's not the whole real line is lower than 95%: therefore one can make money by betting against it. The formal argument is given by Theorem 1 in Müller and Norets (2016). To make it bet-proof, we'd need to enlarge it when it doesn't consist of the whole real line.[4]
- As a test, the AR test has an appealing optimality property: it is uniformly most powerful (UMP) among all unbiased tests. This was shown in Moreira (2009); I give a self-contained proof in Section B.
- The AR test is used little in practice. Motivated by this, Lee et al. (2022) propose an alternative procedure called tF that is based on the observation that the worst-case rejection of the Wald test occurs at $\rho = 1$. When $\rho = 1$, the t-statistic depends on the data only through the first-stage F, and they use this observation to derive critical values $c_\alpha(F)$ that depend on it. These critical values tend to infinity as $F \to z^2_{1-\alpha/2}$.

ESTIMATION AND SIGN-SCREENING No estimator can be fully immune to bias since it is impossible to construct a consistent or at least median unbiased estimator when the instruments are irrelevant.

4. The bet-proofness concept is related to the Cox (1958) problem: the alternative, shorter CI we considered there is not bet-proof.

However, we can construct an unbiased estimator if the sign of the first stage is known, as pointed out in Andrews and Armstrong (2017). In particular, they point out that we can construct a unique unbiased estimator of $1/\pi$ if $\pi \geq 0$ as

$$\frac{1}{\mathcal{V}_{22}^{1/2}} m(t_1),$$

where $m(x) = (1 - \Phi(x))/\phi(x)$ is the Mills' ratio, and $t_1 = \hat{\pi}/\mathcal{V}_{22}^{1/2}$ is the first-stage t-statistic (so $F = t_1^2$). Then $\hat{\pi}_\perp = \hat{\delta} - \mathcal{V}_{12}/\mathcal{V}_{22} \cdot \hat{\pi}$ (the part of $\hat{\delta}$ that's independent of $\hat{\pi}$) has expectation $\pi\beta - \mathcal{V}_{12}/\mathcal{V}_{22}\pi$, an unbiased estimator of β can be constructed as

$$\hat{\beta}_U = t_1 m(t_1)\hat{\beta}_{\text{TSLS}} + (1 - t_1 m(t_1))\beta_{WOLS},$$

where $\beta_{WOLS} = \mathcal{V}_{12}/\mathcal{V}_{22}$ is the weak-instrument limit of OLS. The curious thing is that if $t_1 \geq 0$—that is, the first-stage is right-signed—then the estimator shrinks TSLS towards OLS. The shrinkage interpretation of $\hat{\beta}_U$ seems surprising: since $\hat{\beta}_{\text{TSLS}}$ is biased towards OLS, shrinkage towards OLS increases bias. This counterintuitive fact arises because the estimator $\hat{\beta}_U$ is unbiased by virtue of averaging a conditional positive bias when $t_1 > 0$ and with conditional negative bias when $t_1 < 0$.

It is hard to imagine an analyst who is prepared to sign the population first stage while ignoring the sign of the estimated first stage. Such conditioning, however, strips $\hat{\beta}_U$ of its appeal. What about $\hat{\beta}_{\text{TSLS}}$? Angrist and Kolesár (2024) show that sign-screening actually halves the median bias of TSLS, without reducing coverage: in contrast with procedures that screen on the *magnitude* of the first-stage F statistic, screening on the sign of the corresponding t-statistic has to have little effect on rejection rates for a conventional Wald test.

Question 3. What is the practical takeaway?

3.4. *Overidentified case*

This case is much more complicated:

- We can generalize Fieller's idea: Since

  $$\hat{\delta} - \beta\hat{\pi} \sim \mathcal{N}(0, Q^{-1}E[b'\Omega(X_i)b\tilde{Z}_i\tilde{Z}_i']Q^{-1}/n),$$

  the test that rejects the null $H_0\colon \beta = \beta_0$ whenever

  $$AR = n(\hat{\delta} - \beta_0\hat{\beta}_2)' \left(QE[b'\Omega(X_i)b\tilde{Z}_i\tilde{Z}_i']^{-1}Q\right)(\hat{\delta} - \beta_0\hat{\beta}_2)$$

  is greater than the 95th quantile of χ_k^2 will have size 5%. This test is called the Anderson and Rubin (1949) test. The idea generalizes naturally to GMM models.

  However, the bet-proofness problem becomes even more severe, since the resulting

CI will be empty with positive probability. Furthermore, the CI is no longer efficient under standard asymptotics: it is longer than the Wald CI.

The reason that the CI may be empty is that it tests the joint null the TSLS estimand being equal to β_0, and δ being proportional to π.

- The conditional likelihood ratio (CLR) test suggested by Moreira (2003) is more powerful than AR, and enjoys some optimality properties under homoskedastic errors (Andrews, Moreira, and Stock 2006) (that do not, however, carry over to the heteroskedastic case). It is available in Stata.
- In the just identified case, we need values of $|\rho|$ that are unreasonably high in cross-section applications (very close to 1) to generate severe overrejection of the Wald test: it is hard to come up with empirical examples where Wald CIs are substantively misleading. This changes when $k > 1$, as we've seen from the intertemporal elasticitity of substitution example.
- There is no procedure when $k > 1$ that allows for treatment effect heterogeneity.

APPENDICES

A. DERIVATIONS

Proof of Equation (3). Let $\Delta = E[W_i W_i']^{-1} E[W_i Z_i']$. We can think of $\hat{\delta}, \hat{\pi}$ and $\ddot{Z}'D/n$ as method of moments estimators based on the moment condition $E[g(\mathcal{D}_i, \theta)]$, where $\theta = \text{vec}(\delta, \pi, Q\pi, \text{vec}(\Delta'))$, and

$$g(\mathcal{D}_i, \theta) = \begin{pmatrix} \text{vec}((Z_i - \Delta' W_i)(Y_i - Z_i'\delta, D_i - Z_i'\pi)) \\ (Z_i - \Delta' W_i) D_i - Q\pi \\ \text{vec}((Z_i - \Delta' W_i) W_i') \end{pmatrix} = \begin{pmatrix} \tilde{u}_i \otimes \tilde{Z}_i \\ \tilde{Z}_i D_i - Q\pi \\ W_i \otimes \tilde{Z}_i \end{pmatrix}, \tag{17}$$

where $\tilde{u}_i = u_i + \Psi' W_i$. The derivative of this moment condition is

$$\Gamma = -\begin{pmatrix} I_2 \otimes Q & 0 & \Psi' E[W_i W_i'] \otimes I_K \\ 0 & I_K & E[D_i W_i'] \otimes I_K \\ 0 & 0 & E[W_i W_i'] \otimes I_K \end{pmatrix}, \quad \Gamma^{-1} = \begin{pmatrix} I_2 \otimes Q^{-1} & 0 & -\Psi' \otimes Q^{-1} \\ 0 & I_K & -(\pi'\Delta + \psi') \otimes I_K \\ 0 & 0 & E[W_i W_i']^{-1} \otimes I_K \end{pmatrix},$$

while its variance is

$$\Sigma = E\begin{pmatrix} \tilde{u}_i \tilde{u}_i' \otimes \tilde{Z}_i \tilde{Z}_i' & \tilde{u}_i \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - \tilde{Z}_i \pi' Q) & \tilde{u}_i W_i' \otimes \tilde{Z}_i \tilde{Z}_i' \\ \tilde{u}_i' \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - Q\pi \tilde{Z}_i') & (\tilde{Z}_i D_i - Q\pi)(\tilde{Z}_i D_i - Q\pi)' & W_i' \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - Q\pi \tilde{Z}_i') \\ W_i \tilde{u}_i' \otimes \tilde{Z}_i \tilde{Z}_i' & W_i \otimes (\tilde{Z}_i \tilde{Z}_i' D_i - \tilde{Z}_i \pi' Q) & W_i W_i' \otimes \tilde{Z}_i \tilde{Z}_i' \end{pmatrix}.$$

The upper block of $\Gamma^{-1}\Sigma\Gamma^{-1'}$ is given by

$$\mathcal{V}_0 = E\begin{pmatrix} \Omega(X_i) \otimes Q^{-1} \tilde{Z}_i \tilde{Z}_i' Q^{-1} & u_i \otimes (Q^{-1} \tilde{Z}_i (\tilde{Z}_i' \overline{D}_i - \pi' Q)) \\ u_i' \otimes ((\overline{D}_i \tilde{Z}_i - Q\pi) \tilde{Z}_i' Q^{-1}) & (\tilde{Z}_i \overline{D}_i - Q\pi)(\tilde{Z}_i \overline{D}_i - Q\pi)' \end{pmatrix} = \text{var}\begin{pmatrix} u_i \otimes Q^{-1} \tilde{Z}_i \\ \tilde{Z}_i \overline{D}_i - Q\pi \end{pmatrix},$$

where $\overline{D}_i = D_i - E[D_i W_i'] E[W_i W_i']^{-1} W_i = \tilde{Z}_i' \pi + u_{2i}$, and $\Omega(X_i) = E[u_i u_i' \mid X_i]$ is the conditional variance of the reduced-form errors. Since $\tilde{Z}_i \overline{D}_i - Q\pi = (\tilde{Z}_i \tilde{Z}_i' - Q)\pi + u_{2i} \tilde{Z}_i$, this yields the

result. □

Proof of Equation (6) and Equation (8). TSLS can be thought of as an estimator of

$$\beta = g(\theta) = \delta' Q\pi / \pi' Q\pi,$$

with θ defined as in eq. (17). The derivative of this function is given by

$$G = \frac{1}{\pi' Q\pi} \begin{pmatrix} b \otimes Q\pi \\ \pi_\Delta, \end{pmatrix}$$

where $\pi_\Delta = \delta - \beta\pi$. Hence, using the fact that $\pi' Q\pi_\Delta = 0$ by definition of β,

$$G'\mathcal{V}_0 G = E\left[u_{\Delta,i}^2(\pi'\tilde{Z}_i)^2 + 2u_{\Delta,i}\pi'\tilde{Z}_i(\overline{D}_i\tilde{Z}_i'\pi_\Delta) + (\overline{D}_i\tilde{Z}_i'\pi_\Delta)^2\right] = E\left[\left(\epsilon_i \cdot \pi'\tilde{Z}_i + u_{2i} \cdot \tilde{Z}_i'\pi_\Delta\right)^2\right]$$

where $b'u_i = u_{\Delta,i}$ and $\epsilon_i = \tilde{Z}_i\pi_\Delta + u_{\Delta,i} = (Y_i - D_i\beta) - W_i' E[W_i W_i]^{-1} E[W_i(Y_i - D_i\beta)]$. This yields eq. (8). If $\pi_\Delta = 0$, then one obtains eq. (6). □

Proof of Equation (16). In a model with normal homoskedastic reduced form errors

$$\begin{pmatrix} (\ddot{Z}'\ddot{Z})^{1/2}(\hat{\delta} - \hat{\pi}\beta)/\sqrt{b'\Omega b} \\ (\ddot{Z}'\ddot{Z})^{1/2}\hat{\pi}/\Omega_{22}^{1/2} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes \lambda \sim \begin{pmatrix} \sqrt{1-\rho^2} & \rho \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \end{pmatrix},$$

where $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are standard normal k-vectors, and $\lambda = (\ddot{Z}'\ddot{Z})^{1/2}\pi/\Omega_{22}^{1/2}$. Recall also that $\beta_{WOLS} = \Omega_{12}/\Omega_{22} = \rho\sqrt{b'\Omega b/\Omega_{22}} + \beta$. Therefore,

$$\begin{aligned} \hat{\beta}_{\text{TSLS}} - \beta &= \frac{\hat{\pi}'\ddot{Z}'\ddot{Z}'(\hat{\delta} - \hat{\pi}\beta)}{\hat{\pi}'\ddot{Z}'\ddot{Z}'\hat{\pi}} \\ &= \sqrt{\frac{b'\Omega b}{\Omega_{22}}(1-\rho^2)}\frac{\mathcal{Z}_1'(\mathcal{Z}_2 + \lambda)}{(\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)} + (\beta_{WOLS} - \beta)\frac{(\mathcal{Z}_2+\lambda)'\mathcal{Z}_2}{(\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)}. \end{aligned}$$

By Proposition 1(2) in Bao and Kan (2013), the expectation of both terms exists. The expectation of the first term is zero by iterated expectations. As noted in the proof of Lemma 4 in Magnus (1986), for $x > 0$, it follows by setting $z = tx$ in the gamma function identity $(s-1)! = \Gamma(s) = \int_0^\infty z^{s-1}e^{-z}dz$ that $1/x^s = \frac{1}{(s-1)!}\int_0^\infty t^{s-1}e^{-tx}dt$. With $s = 1$, we therefore get, by Fubini's theorem, that for any X_1, X_2 with $X_2 > 0$,

$$E[X_1/X_2] = \int_0^\infty E[X_1 e^{-tX_2}]dt$$

Let $X_1 = (\mathcal{Z}_2+\lambda)'\lambda$, $X_2 = (\mathcal{Z}_2+\lambda)'(\mathcal{Z}_2+\lambda)$, and $\mu^2 = \lambda'\lambda$. We have

$$\begin{aligned} E[X_1^s e^{-tX_2}] &= \int \lambda'(z+\lambda)e^{-t(z+\lambda)'(z+\lambda)}\frac{1}{(2\pi)^{k/2}}e^{-z'z/2}dz \\ &= e^{-\frac{2t}{2t+1}\frac{\mu^2}{2}}\frac{1}{(2t+1)^{k/2}}E_{Z\sim\mathcal{N}_k(-\frac{2t}{2t+1}\lambda, I_k/(2t+1))}[\lambda'(Z+\lambda)] = e^{-\mu^2/2}e^{\frac{1}{2t+1}\frac{\mu^2}{2}}\frac{1}{(2t+1)^{k/2+1}}\mu^2, \end{aligned}$$

where the second line follows by "completing the square". Hence,

$$\frac{E[\hat{\beta}_{\text{TSLS}} - \beta]}{\beta_{WOLS} - \beta} = 1 - E[X_1/X_2] = 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^\infty \frac{2}{(2t+1)^{k/2+1}} e^{\frac{1}{2t+1}\frac{\mu^2}{2}} dt$$
$$= 1 - \frac{\mu^2}{2} e^{-\mu^2/2} \int_0^1 x^{k/2-1} e^{x\mu^2/2} dx$$

which yields the result. Recall that

$$M(a,b,z) = \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zu} u^{a-1} (1-u)^{b-a-1} \, du \qquad \square$$

is Kummer's confluent hypergeometric function (sometimes denoted ${}_1F_1(a;b;z)$). We can therefore equivalently write the integral as $\frac{2}{k} M(k/2, k/2+1, \mu^2/2)$.

Proof of Equation (14). Write the limit experiment as

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \begin{pmatrix} \beta\pi \\ \pi \end{pmatrix} + \begin{pmatrix} \sqrt{\mathcal{V}_{11} - \mathcal{V}_{12}^2/\mathcal{V}_{22}} & \mathcal{V}_{12}/\mathcal{V}_{22}^{1/2} \\ 0 & \mathcal{V}_{22}^{1/2} \end{pmatrix} \begin{pmatrix} \mathcal{Z}_\perp \\ \mathcal{Z}_2 \end{pmatrix},$$

where $\mathcal{Z}_\perp$ and $\mathcal{Z}_2$ are independent standard normal. The asymptotic variance of $\hat{\beta}$ is given by $E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]/E[(\tilde{Z}_i'\pi)^2]^2 = E[\epsilon_i^2\tilde{Z}_i^2]/(E[\tilde{Z}_i^2]^2\pi^2) = (\mathcal{V}_{11} - 2\mathcal{V}_{12}\beta + \mathcal{V}_{22}\beta^2)/\pi^2$, and the asymptotic variance estimator is given by $(\mathcal{V}_{11} - 2\mathcal{V}_{12}\hat{\beta} + \mathcal{V}_{22}\hat{\beta}^2)/\hat{\pi}^2$. Thus, the Wald statistic can then be written as

$$W = \frac{\hat{\beta} - \beta}{\sqrt{(\hat{\pi}/\mathcal{V}_{22}^{1/2})^{-2}(\hat{\beta}^2 - 2\hat{\beta}\mathcal{V}_{12}/\mathcal{V}_{22} + \mathcal{V}_{11}/\mathcal{V}_{22})}}.$$

Let $\mathcal{Z}_\epsilon = (\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{-1/2}(\sqrt{\mathcal{V}_{11}/\mathcal{V}_{22} - \mathcal{V}_{12}^2/\mathcal{V}_{22}^2}\mathcal{Z}_\perp + (\mathcal{V}_{12}/\mathcal{V}_{22} - \beta)\mathcal{Z}_2)$. Note that $\mathcal{Z}_\epsilon$ is standard normal, with $E[\mathcal{Z}_\epsilon\mathcal{Z}_2] = \rho$. Then

$$\hat{\beta} - \beta = \frac{(\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22})^{1/2}\mathcal{Z}_\epsilon}{\lambda + \mathcal{Z}_2}, \qquad \hat{\pi}/\mathcal{V}_{22}^{1/2} = \lambda + \mathcal{Z}_2,$$

Hence,

$$\hat{\beta}^2 - 2\hat{\beta}\mathcal{V}_{12}/\mathcal{V}_{22} + \mathcal{V}_{11}/\mathcal{V}_{22} = \left(\mathcal{V}_{11}/\mathcal{V}_{22} + \beta^2 - 2\beta\mathcal{V}_{12}/\mathcal{V}_{22}\right)\left(\frac{Z_\epsilon^2}{(\lambda+\mathcal{Z}_2)^2} - 2\rho\frac{\mathcal{Z}_\epsilon}{\lambda+\mathcal{Z}_2} + 1\right)$$

and the Wald statistic has the distribution

$$W = \frac{\mathcal{Z}_\epsilon/(\lambda+\mathcal{Z}_2)}{\sqrt{\frac{1}{(\lambda+\mathcal{Z}_2)^2}\left(\frac{Z_\epsilon^2}{(\lambda+\mathcal{Z}_2)^2} - 2\rho\frac{\mathcal{Z}_\epsilon}{\lambda+\mathcal{Z}_2} + 1\right)}} = \frac{\text{sign}(\lambda+\mathcal{Z}_2)\mathcal{Z}_\epsilon}{\sqrt{Z_\epsilon^2/(\lambda+\mathcal{Z}_2)^2 - 2\rho\mathcal{Z}_\epsilon/(\lambda+\mathcal{Z}_2) + 1}}.$$

For numerical results, observe that the rejection region of the Wald statistic is quadratic in $\mathcal{Z}_\epsilon$:

$$[(\lambda+\mathcal{Z}_2)^2 - z_{1-\alpha/2}^2]\mathcal{Z}_\epsilon^2 + 2\rho z_{1-\alpha/2}^2(\lambda+\mathcal{Z}_2)\mathcal{Z}_\epsilon - z_{1-\alpha/2}^2(\lambda+\mathcal{Z}_2)^2 \geq 0$$

If $(\lambda+\mathcal{Z}_2)^2 \geq z_{1-\alpha/2}^2$, we reject for large values of $\mathcal{Z}_\epsilon$. If $(\lambda+\mathcal{Z}_2)^2 \leq (1-\rho^2)z_{1-\alpha/2}^2$, then determinant, $4(\lambda+\mathcal{Z}_2)^2 z_{1-\alpha/2}^2\left[(\lambda+\mathcal{Z}_2)^2 - (1-\rho^2)z_{1-\alpha/2}^2\right]$, is negative, and we never reject. Otherwise, we reject for $\mathcal{Z}_\epsilon$ in an interval. To compute the rejection probabilities, condition on $\mathcal{Z}_2$, and use the fact that $P(\mathcal{Z}_\epsilon < x \mid \mathcal{Z}_2) = \Phi((x - \rho\mathcal{Z}_2)/\sqrt{1-\rho^2})$. $\square$

B. OPTIMALITY OF ANDERSON-RUBIN

We use the following result:

Theorem 9 (Lehmann and Romano 2005, Theorem 4.4.1). Suppose the statistic (U, T) has distribution $C(\theta, \vartheta)e^{\theta u + \vartheta' t}d\nu(u, t)$ with respect to some measure ν. Then a test that rejects if $U \notin [C_1(T), C_2(T)]$, where the cutoffs are chosen to ensure that the test is unbiased and has size α conditional on T is UMP unbiased unconditionally.

Proof. The proof uses the fact that the conditional distribution $U \mid T$ is $C_t(\theta)e^{\theta u}d\nu_t(u)$, which is an exponential family with no nuisance parameters, and T is sufficient for ϑ if we fix θ. Therefore, there exists a UMP unbiased test, that rejects if U is outside an interval, where the cutoffs are chosen to satisfy a size and an unbiasedness restriction. We then need to argue these conditional results imply UMP unbiasedness unconditionally. □

In our case, it follows from eq. (12), that the data are given by $Y := (\hat{\delta}, \hat{\pi})' \sim \mathcal{N}(\pi a, \Omega)$, where we define $\pi := \pi$, $a := (\beta, 1)'$, and $\Omega := \mathcal{V}$. Hence, the density is proportional to

$$e^{\pi a'\Omega^{-1}Y} = e^{\pi a_0'\Omega^{-1}Y + \pi(a-a_0)'\Omega^{-1}Y} = e^{\pi a_0'\Omega^{-1}Y + \theta e_1'\Omega^{-1}Y},$$

where we let $a_0 = (\beta_0, 1)$, $\theta = (\beta - \beta_0)\pi$, and π is the nuisance parameter ϑ. To get the density of $U = e_1'\Omega^{-1}Y$ conditional on $T = a_0'\Omega^{-1}Y$, observe that since $\left(b, e_2\right)^{-1} = \left(e_1, a\right)'$, where $b = (1, -\beta)'$, we have

$$\begin{pmatrix} e_1 & a \end{pmatrix}' \Omega^{-1} \begin{pmatrix} e_1 & a \end{pmatrix} = \left(\begin{pmatrix} b & e_2 \end{pmatrix}' \Omega \begin{pmatrix} b & e_2 \end{pmatrix} \right)^{-1} = \frac{1}{|\Omega|} \begin{pmatrix} \Omega_{22} & -e_2'\Omega b \\ -e_2'\Omega b & b'\Omega b \end{pmatrix}. \tag{18}$$

Post-multiplying both sides by $\left(b, e_2\right)' = \left(\begin{smallmatrix} 1 & -\beta \\ 0 & 1 \end{smallmatrix}\right)$ then yields

$$\begin{pmatrix} e_1 & a \end{pmatrix}' \Omega^{-1} = \frac{1}{|\Omega|} \begin{pmatrix} \Omega_{22} & -e_2'\Omega b - \Omega_{22}\beta \\ -e_2'\Omega b & b'\Omega b + e_2'\Omega b\beta \end{pmatrix}$$

Pre-multiplying both sides by $(b'\Omega b, e_2'\Omega b)$ yields

$$b'\Omega b e_1'\Omega^{-1} + e_2'\Omega b a'\Omega^{-1} = b, \tag{19}$$

where we use the fact that

$$\left| \begin{pmatrix} b & e_2 \end{pmatrix}' \Omega \begin{pmatrix} b & e_2 \end{pmatrix} \right| = |\Omega|.$$

Now,

$$\begin{pmatrix} e_1'\Omega^{-1}Y \\ a_0'\Omega^{-1}Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} e_1'\Omega^{-1}a\pi \\ a_0'\Omega^{-1}a\pi \end{pmatrix}, \begin{pmatrix} e_1 & a_0 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} e_1 & a_0 \end{pmatrix} \right).$$

Thus, by the formula for the conditional normal distribution,

$$e_1'\Omega^{-1}Y \mid a_0'\Omega^{-1}Y \sim \mathcal{N}\left(m, \frac{1}{|\Omega| \cdot a_0'\Omega^{-1}a_0}\right),$$

where

$$\begin{aligned} m &= e_1'\Omega^{-1}a\pi + \frac{e_1'\Omega^{-1}a_0}{a_0'\Omega^{-1}a_0}(a_0'\Omega^{-1}Y - a_0'\Omega^{-1}a\pi) \\ &= e_1'\Omega^{-1}a\pi - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}(a_0'\Omega^{-1}Y - a_0'\Omega^{-1}a\pi) = \frac{1}{b_0'\Omega b_0}b_0'a\pi - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}Y \\ &= \frac{\theta}{b_0'\Omega b_0} - \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}Y. \end{aligned}$$

Here the second equality uses eq. (18), and the third equality uses eq. (19). Since $\theta = 0$ under the null, for both a one- and a two-sided test, we reject for large values of the z-statistic

$$\frac{\left[e_1'\Omega^{-1} + \frac{e_2'\Omega b_0}{b_0'\Omega b_0}a_0'\Omega^{-1}\right]Y}{1/\sqrt{|\Omega|a_0'\Omega^{-1}a_0}} = \frac{\frac{1}{b_0'\Omega b_0}b_0'Y}{1/\sqrt{|\Omega|a_0'\Omega^{-1}a_0}} = \frac{b_0'Y}{\sqrt{b_0'\Omega b_0}},$$

where the first equality uses eq. (19), and the second one uses eq. (18).

REFERENCES

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20, no. 1 (March): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Donald W. K., and Patrik Guggenberger. 2010. "Asymptotic Size and a Problem with Subsampling and with the m Out Of n Bootstrap." *Econometric Theory* 26, no. 02 (April): 426. https://doi.org/10.1017/S0266466609100051.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74, no. 3 (May): 715–752. https://doi.org/10.1111/j.1468-0262.2006.00680.x.

Andrews, Isaiah, and Timothy B. Armstrong. 2017. "Unbiased Instrumental Variables Estimation under Known First-Stage Sign." *Quantitative Economics* 8, no. 2 (July): 479–503. https://doi.org/10.3982/QE700.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11, no. 1 (August): 727–753. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240, no. 2 (March): 105398. https://doi.org/10.1016/j.jeconom.2022.12.012.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67, no. 3 (July): 499–527. https://doi.org/10.1111/1467-937X.00141.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity." *Journal of the American Statistical Association* 90, no. 430 (June): 431–442. https://doi.org/10.1080/01621459.1995.10476535.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Bao, Yong, and Raymond Kan. 2013. "On the Moments of Ratios of Quadratic Forms in Normal Random Variables." *Journal of Multivariate Analysis* 117 (May): 229–245. https://doi.org/10.1016/j.jmva.2013.03.002.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90, no. 430 (June): 443–450. https://doi.org/10.1080/01621459.1995.10476536.

Campbell, John Y. 2003. "Consumption-Based Asset Pricing." Chap. 13 in *Handbook of the Economics of Finance*, edited by G. M. Constantinides, M. Harris, and R. Stulz, vol. 1B, 803–887. Amsterdam: Elsevier. https://doi.org/10.1016/S1574-0102(03)01022-7.

Chamberlain, Gary. 2007. "Decision Theory Applied to an Instrumental Variables Model." *Econometrica* 75, no. 3 (May): 609–652. https://doi.org/10.1111/j.1468-0262.2007.00764.x.

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.

Evdokimov, Kirill, and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." January. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf.

Fieller, Edgar C. 1932. "The Distribution of the Index in a Normal Bivariate Population." *Biometrika* 24, nos. 3/4 (November): 428–440. https://doi.org/10.1093/biomet/24.3-4.428.

Fieller, Edgar C. 1940. "The Biological Standardization of Insulin." *Supplement to the Journal of the Royal Statistical Society* 7 (1): 1–64. https://doi.org/10.2307/2983630.

———. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society. Series B (Methodological)* 16, no. 2 (July): 175–185. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x.

Goldberger, Arthur S., and Ingram Olkin. 1971. "A Minimum-Distance Interpretation of Limited-Information Estimation." *Econometrica* 39, no. 3 (May): 635–639. https://doi.org/10.2307/1913273.

Hahn, Jinyong, and Jerry A. Hausman. 2002. "A New Specification Test for the Validity of Instrumental Variables." *Econometrica* 70, no. 1 (January): 163–189. https://doi.org/10.1111/1468-0262.00272.

Hausman, Jerry A., James H. Stock, and Motohiro Yogo. 2005. "Asymptotic Properties of the Hahn-Hausman Test for Weak-Instruments." *Economics Letters* 89, no. 3 (December): 333–342. https://doi.org/10.1016/j.econlet.2005.06.007.

Heckman, James J., and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica* 73, no. 3 (May): 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity." Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2022. "Valid t-Ratio Inference for IV." *American Economic Review* 112, no. 10 (October): 3260–3290. https://doi.org/10.1257/aer.20211063.

Lee, Seojeong. 2018. "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs." *Journal of Business & Economic Statistics* 36, no. 3 (July): 400–410. https://doi.org/10.1080/07350015.2016.1186555.

Lehmann, Erich L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.

Magnus, Jan R. 1986. "The Exact Moments of a Ratio of Quadratic Forms in Normal Variables." *Annales d'Économie et de Statistique*, no. 4, 95–109. https://doi.org/10.2307/20075629.

Mariano, Roberto S., and James B. McDonald. 1979. "A Note on the Distribution Functions of LIML and 2SLS Structural Coefficient in the Exactly Identified Case." *Journal of the American Statistical Association* 74, no. 368 (December): 847–848. https://doi.org/10.1080/01621459.1979.10481040.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics* 31, no. 3 (July): 358–369. https://doi.org/10.1080/00401706.2013.806694.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71, no. 4 (July): 1027–1048. https://doi.org/10.1111/1468-0262.00438.

———. 2009. "Tests with Correct Size When Instruments Can Be Arbitrarily Weak." *Journal of Econometrics* 152, no. 2 (October): 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012.

Müller, Ulrich K. 2011. "Efficient Tests under a Weak Convergence Assumption." *Econometrica* 79, no. 2 (March): 395–435. https://doi.org/10.3982/ECTA7793.

Müller, Ulrich K., and Andriy Norets. 2016. "Credibility of Confidence Sets in Nonstandard Econometric Problems." *Econometrica* 84, no. 6 (November): 2183–2213. https://doi.org/10.3982/ECTA14023.

Nelson, Charles R., and Richard Startz. 1990a. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58, no. 4 (July): 967–976. https://doi.org/10.2307/2938359.

———. 1990b. "The Distribution of the Instrumental Variables Estimator and Its *t*-Ratio When the Instrument Is a Poor More." *The Journal of Business,* A Conference in Honor of Merton H. Miller's Contributions to Finance and Economics, 63, no. 1 (January): S125–S140. https://doi.org/10.1086/296497.

Richardson, David H. 1968. "The Exact Distribution of a Structural Coefficient Estimator." *Journal of the American Statistical Association* 63, no. 324 (December): 1214–1226. https://doi.org/10.1080/01621459.1968.10480921.

Sawa, Takamitsu. 1972. "Finite-Sample Properties of the k-Class Estimators." *Econometrica* 40, no. 4 (July): 653–680. https://doi.org/10.2307/1912960.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65, no. 3 (May): 557–586. https://doi.org/10.2307/2171753.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." Chap. 5 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006.

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

Yogo, Motohiro. 2004. "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak." *Review of Economics and Statistics* 86, no. 3 (August): 797–810. https://doi.org/10.1162/0034653041811770.

Young, Alwyn. 2022. "Consistency without Inference: Instrumental Variables in Practical Application." *European Economic Review* 147 (August): 104112. https://doi.org/10.1016/j.euroecorev.2022.104112.

Zellner, Arnold. 1970. "Estimation of Regression Relationships Containing Unobservable Independent Variables." *International Economic Review* 11, no. 3 (October): 441–454. https://doi.org/10.2307/2525323.

———. 1978. "Estimation of Functions of Population Mean and Regression Coefficients Including Structural Coefficients: A Minimum Expected Loss (MELO) Approach." *Journal of Econometrics* 8, no. 2 (October): 127–158. https://doi.org/10.1016/0304-4076(78)90024-6.

MANY INSTRUMENTS AND JUDGES

Michal Kolesár*

April 3, 2024

In this lecture, we'll consider some issues that arise in the linear instrumental variables (IV) model when the number of instruments K is large. We may find ourselves in a setting with many instruments for two main reasons. First, when we take an original, low-dimensional instrument, and interact it with controls. A classic example is Angrist and Krueger (1991). More generally, we could take an underlying low-dimensional set of instruments and covariates, and take a series expansion of it in an effort to find a transformation that leads to a strong first stage—see Belloni et al. (2012).

The second setting in which this problem arises is when we use fixed effects (group indicators) as instruments: if there are many effects, we have many instruments. This setting has been quite common in recent empirical work, and is sometimes referred to as "judges" or "examiners" designs. The basic idea is as follows. Suppose we're interested in the effect of incarceration on some economic outcome, as in Kling (2006). We can't just regress the outcome of interest on the length of sentence, since it's unlikely, even if we control for observables, that the length of sentence is as good as randomly assigned. However, we can exploit the fact that (perhaps conditional on time and location dummies) the judge assigned to a defendant is as good as random. If some judges tend to hand out longer sentences than others, this will generate variation in the sentencing length that's as good as random, and we can use we can use judge indicators as instruments. If there are many judges, then we have many instruments.

Apart from studying the effects of incarceration on economic outcomes (e.g. Kling 2006; Aizer and Doyle 2015), this design has been used in a number of other settings: we may exploit random assignment of judges to bankruptcy cases, criminal cases, or patent cases, doctors to shifts, or social workers to cases—see the introduction to Frandsen, Lefgren, and Leslie (2023) for a partial list of applications.

We'll consider the effects of many instruments on:

1. the bias of the two-stage least squares (TSLS) estimator, and consider alternatives to TSLS
2. standard errors
3. the interpretation of the first-stage F statistic

*Email: mkolesar@princeton.edu.

We'll see that the conclusion in Bound, Jaeger, and Baker (1995) critique of Angrist and Krueger (1991) that "the natural experiment afforded by the interaction between compulsory school attendance laws and quarter of birth does not give much usable information regarding the causal effect of education on earnings" is not quite correct: with the right estimator and right standard errors, we can still extract useful information from the data and report a reliable assessment of its accuracy.

1. SETUP AND ESTIMATION

We use the same notation as in the previous lecture, with the reduced form and the first stage given by

$$Y_i = Z_i'\delta + W_i'\psi_Y + u_{Yi}, \tag{1}$$
$$D_i = Z_i'\pi + W_i'\psi_D + u_{Di}. \tag{2}$$

The parameter of interest is given by

$$\beta := \frac{\delta' Q \pi}{\pi' Q \pi},$$

where $Q = E[\tilde{Z}_i \tilde{Z}_i']$, with $\tilde{Z}_i = Z_i - E[Z_i W_i'] E[W_i W_i']^{-1} W_i$. To measure the overall strength of the instruments, define

$$r_n = n\pi' Q \pi = nE[(\tilde{Z}_i'\pi)^2].$$

You can think of r_n as the effective sample size, in the sense that it scales the sample size n by instrument strength, the denominator in the definition of β. As discussed in the previous lecture, under homogeneous treatment effects, β corresponds to the causal effect of D_i on Y_i, and the reduced-form coefficients are proportional to each other, $\delta = \pi\beta$. Under heterogeneous treatment effects, β has the interpretation as a weighted average of local average treatment effects (LATEs) if we make a monotonicity assumption.

Remark 1 (Notation). For any matrix A, let $H_A = A(A'A)^{-1}A'$ denote the hat matrix (also called a projection matrix), and let $\ddot{A} = A - H_W A$ denote the residuals after projecting A onto the covariates. Thus, for instance, $\ddot{Z}_i$ is the sample analog of $\tilde{Z}_i$. Let $X_i = (Z_i', W_i')'$ collect the right-hand side (RHS) (or "exogenous") variables.

Example 1 (Judges design). To make the above setup concrete, let us consider a setting in which individual i is assigned judge Q_i, which is random conditional on i's geographic location G_i. Then the covariates and instruments are both indicators: $W_{i\ell} = \mathbb{1}\{G_i = \ell\}$, and $Z_{ik} = \mathbb{1}\{Q_i = k\}$. Let us assume that each judge only works in one location, and that the treatment of interest D_i is an indicator for incarceration. We're interested in the effect of incarceration on some economic outcome. There are L locations and $K + L$

judges; in each location we drop one judge to avoid collinearity—call this judge the reference judge. Then $\psi_{D,\ell}$ measures the average sentencing rate of the reference judge in location ℓ, and π_k measures the sentencing rate of judge k relative to the reference judge in the same location. ⊠

Remark 2. The monotonicity assumption in Example 1 requires that judges agree on the ranking of the defendants, they just disagree on the cutoff at which they start sentencing them. Such an assumption can be problematic. For example, Chan, Gentzkow, and Yu (2019) point out that decisions of physicians differ due to both skill and preferences. They look at radiologists who diagnose whether a patient has pneumonia, and define a type II error as a patient who was not diagnosed, but who have a subsequent pneumonia diagnosis in the next 10 days. They show that radiologists who diagnose at higher rates actually have higher rather than lower type II error rates. Similarly, Kleinberg et al. (2018) find that the increase in crime associated with judges who are more likely to release defendants on bail is about the same as if these more lenient judges randomly picked the extra defendants to release on bail. See also Frandsen, Lefgren, and Leslie (2023). However, since failure of monotonicity only affects the interpretation of β, to keep the statistical issues separate from identification issues, we put these issues aside here.

To capture the effect of the potentially large number of instruments, $K = \dim(Z_i)$, or covariates $L = \dim(W_i)$ on the finite-sample behavior of estimators in the asymptotic approximation, we consider asymptotics in which both K and L are allowed to grow with sample size. Under such asymptotics, we obtain the following result:

Lemma 3. *The* TSLS *estimator suffers from own observation bias towards ordinary least squares (OLS). In particular, suppose* $E[u_i \mid X_i] = 0$, *with* $\Omega(X_i) = E[u_i u_i' \mid X_i]$ *bounded,* $E[\tilde{Z}_i \mid W_i] = 0$, *and* $L/n \to 0$.

Then consistency of TSLS *is in general requires* $K/r_n \to 0$. *Under homoskedastic errors,*

$$\hat{\beta}_{TSLS} = \beta + \frac{(\Omega_{YD} - \Omega_{DD}\beta)K}{r_n + \Omega_{DD}K} + o_p(1) = (1-w)\beta + w\beta_{OLS} + o_p(1), \quad w = \frac{K}{r_n/\Omega_{DD} + K},$$

where $\beta_{OLS} = (\Omega_{YD} - \Omega_{DD}\beta)/\Omega_{DD} + \beta$ *is the probability limit of OLS.*

Proof. We sketch the argument, see Evdokimov and Kolesár (2018) for a precise proof. Recall from extremum estimation theory that typically, under regularity conditions, an estimator $\hat{\beta}$ that minimizes a sample objective function $\hat{Q}_n(\beta)$ converges in probability to the minimizer of $\beta_n = \operatorname{argmin}_b Q_n(b) = E[\hat{Q}_n(b)]$ in the sense that $\hat{\beta}_n - \beta_n \xrightarrow{p} 0$, provided that (uniformly) $\hat{Q}_n(b) - Q_n(b) \xrightarrow{p} 0$. In our case, the TSLS objective function is

$$\begin{aligned}\hat{Q}_n(b) &= (Y - Db)' H_{\ddot{Z}} (Y - Db) \\ &= (\delta - \pi b)' \ddot{Z}' \ddot{Z} (\delta - \pi b) + 2(\delta - \pi b) H_{\ddot{Z}} (u_Y - u_D b) + (u_Y - u_D b) H_{\ddot{Z}} (u_Y - u_D b),\end{aligned}$$

where the second line follows from $H_{\ddot{Z}}(Y - Db) = \ddot{Z}(\delta - \pi b) + H_{\ddot{Z}}(u_Y - u_D b)$. Using the assump-

tion $E[u_i \mid Z_i, W_i] = 0$, we have[1]

$$Q_n(b) = (\delta - \pi b)' E[\ddot{Z}'\ddot{Z}](\delta - \pi b) + \sum_i E[(u_Y - u_D b)^2 H_{\ddot{Z},ii}].$$

This is minimized at

$$\beta_n = \frac{\pi' E[\ddot{Z}'\ddot{Z}]\delta + \sum_i E[u_{Yi} u_{Di} H_{\ddot{Z},ii}]}{\pi' E[\ddot{Z}'\ddot{Z}]\pi + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]} = \beta + \frac{\pi' E[\ddot{Z}'\ddot{Z}](\delta - \pi\beta) + \sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]}{\pi' E[\ddot{Z}'\ddot{Z}]\pi + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]}$$

Under regularity conditions, replacing $E[\ddot{Z}'\ddot{Z}]$ in this expression by nQ will have a negligible effect, so that we may write

$$\beta_n = \beta + \frac{\sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]}{r_n + \sum_i E[u_{Di}^2 H_{\ddot{Z},ii}]} + o_p(1),$$

By boundedness of $\Omega(X_i)$, $\sum_i E[(u_{Yi} - u_{Di}\beta) u_{Di} H_{\ddot{Z},ii}]$ is bounded by a constant times $\sum_i H_{\ddot{Z},ii} = K$, so the bias is of the order $K/(r_n + K)$, which gives the rate for the consistency result. □

The key thing here is that the TSLS bias scales with K/r_n, rather than K/n: thus, the TSLS bias can be substantial even if K/n is very small if the instruments are not very strong.

The bias comes from the fact that the single constructed instrument $\hat{Z}_{\text{TSLS},i} = Z_i'\hat{\pi} + W_i'\hat{\psi}_D$ used by TSLS puts positive weight on own treatment status D_i, which means that the constructed instrument is slightly endogenous. The total net weight, holding overall instrument strength constant, scales linearly with K, so that the TSLS bias scales with the number of instruments.

Remark 4. Recall from the previous lecture note that the oracle who knows the first stage would use the single instrument $\hat{Z}_i^* = Z_i'\pi + W_i'\psi_D$, estimating β as

$$\hat{\beta}^* = \frac{\hat{Z}^{*\prime}\ddot{Y}}{\hat{Z}^{*\prime}\ddot{D}} = \frac{\hat{D}^{*\prime}Y}{\hat{D}^{*\prime}D}, \qquad \hat{D}^* = \ddot{Z}\pi.$$

Here $\hat{D}^*$ is the single instrument $\hat{Z}^*$ with the covariates partialled out. We also discussed that $\hat{Z}^*$ (or equivalently $\hat{D}^*$) has the interpretation as the optimal instrument under some conditions. The TSLS estimator replaces the unknown $\hat{D}^*$ by the estimate $H_{\ddot{Z}}D = \ddot{Z}\hat{\pi}$, $\hat{\pi} = (\ddot{Z}'\ddot{Z})^{-1}\ddot{Z}'D$. Under standard asymptotics that hold π and K fixed, so that $r_n \asymp n$, this has no effect on consistency of the estimator. However, when K is large, the noise in $\hat{\pi}$ may become non-negligible relative to the signal r_n. Lemma 3 makes it precise when this happens.

Remark 5 (Partial solutions). The first solution TSLS inconsistency, dating back to Bekker (1994), the first paper to analyze the many instrument problem, is to use limited information maximum likelihood (LIML), or variants thereof. Indeed, one can show that under homogeneous treatment effect and homoskedastic errors, we only need $\sqrt{K}/r_n \to 0$ for consistency, a substantially weaker requirement. To ensure consistency under heteroskedastic errors, further adjustments are needed. However, as discussed in the pre-

1. The expectation of a quadratic form $Y'HY$ is $E[Y'HY] = E[\text{tr}(HYY')] = \text{tr}(HE[YY'])$, since trace is a linear operator.

vious lecture, an issue with LIML-like estimators is that they are not robust to heterogeneous treatment effects.

The second solution is to directly correct for the TSLS bias by subtracting an estimate of its bias. This is easy to do under homoskedastic errors, and leads to variants of the bias-corrected TSLS estimator that dates back to Nagar (1959). However, its consistency does depend on homoskedasticity. We can think of the jackknife estimators, discussed next, as heteroskedasticity-robust analogs.

If the bias is caused by using D_i in constructing the single instrument $\hat{Z}_{\text{TSLS},i}$, then an obvious solution is to not use it. There are multiple ways of implementing this idea. The first option is to construct predictors $\hat{\pi}_{-i}$ and $\hat{\psi}_{D,-i}$ of π and ψ_D based on a regression of D onto (Z, W), but with the ith observation removed, to construct a single instrument

$$\hat{Z}_{\text{JIVE1},i} = Z_i'\hat{\pi}_{-i} + W_i'\hat{\psi}_{D,-i}.$$

In the judges design, $\hat{Z}_{\text{JIVE1},i}$ is just the average sentencing rate of the judge assigned to me, using all observations except myself,

$$\hat{Z}_{\text{JIVE1},i} = \frac{\sum_{j\neq i} D_i \mathbb{1}\{Q_j = Q_i\}}{\sum_{j\neq i} \mathbb{1}\{Q_j = Q_i\}}.$$

Then run an IV regression of Y_i onto D_i and W_i, using $\hat{Z}_{\text{JIVE1},i}$ as an instrument for D_i. The resulting estimator is known as jackknife instrumental variables estimator (JIVE1), and can be written as

$$\hat{\beta}_{\text{JIVE1}} = \frac{\hat{Z}_{\text{JIVE1}}'\ddot{Y}}{\hat{Z}_{\text{JIVE1}}'\ddot{D}} = \frac{\hat{D}_{\text{JIVE1}}'Y}{\hat{D}_{\text{JIVE1}}'D}, \quad \hat{Z}_{\text{JIVE1}} = \ddot{H}D, \quad \hat{D}_{\text{JIVE1}} = (I - H_W)\hat{Z},$$
$$\ddot{H} = (I - \text{diag}(H_X))^{-1}(H_X - \text{diag}(H_X)),$$

In particular, we don't need to run n first-stage regressions to construct $\hat{Z}_{\text{JIVE1},i}$ for each i; we can do everything in one step using matrix algebra.[2] This is the version of jackknife discussed in Angrist, Imbens, and Krueger (1999), and Blomquist and Dahlberg (1999), and goes back to Phillips and Hale (1977). However, a problem with this implementation is that when we project out the covariates from $\hat{Z}_i$, we re-introduce the own observation bias: the instrument $\hat{D}_i$ with the covariates projected out does depend on D_i. In the judges design, we adjust $\hat{Z}_{\text{JIVE1},i}$ by the average sentencing rate in the location of i,

$$\hat{D}_{\text{JIVE1},i} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_j \hat{Z}_j \mathbb{1}\{G_j = G_i\}}{\sum_j \mathbb{1}\{G_j = G_i\}} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_j D_j \mathbb{1}\{G_j = G_i\}}{\sum_j \mathbb{1}\{G_j = G_i\}},$$

since $\sum_{j:\, Q_j=Q_i} \hat{Z}_j = \sum_{j:\, Q_j=Q_i} D_j$. Because the average sentencing rate in location i depends on D_i, this re-introduces the own observation bias. Indeed, we generally need

2. Though implementing this estimator (and the estimators below) in practice does require computing the diagonal of the projection matrix H_X, which may challenging when the dimensionality of the data is very large.

$L/r_n \to 0$ for consistency (see Evdokimov and Kolesár (2018) for a formal result), which can be a stringent requirement if there are many covariates. Furthermore, under homoskedastic errors, the bias goes in the opposite direction than the TSLS bias:

$$\hat{\beta}_{\text{JIVE1}} = \beta - \frac{(\Omega_{YD} - \Omega_{DD}\beta)L}{r_n - \Omega_{DD}L} + o_p(1) = (1+\lambda)\beta - \lambda\beta_{OLS} + o_p(1),\ \lambda = \frac{L}{r_n/\Omega_{DD} - L}.$$

In many applications the JIVE1 bias can be *bigger* in magnitude than that of TSLS—see Section 4.

There are two solutions to this problem. The first is to *first* partial out the covariates, and then do the leave-one-out prediction, leading to the improved improved jackknife instrumental variables estimator (IJIVE1) estimator proposed in Ackerberg and Devereux (2009). That is: (i) compute the residuals $\ddot{Z}, \ddot{D}, \ddot{Y}$ from regressing Z, D, Y onto W, and then (ii) compute $\hat{D}_{\text{IJIVE1},i} = \ddot{Z}_i\hat{\pi}_{-i}$, where $\hat{\pi}_{D,-i}$ is the estimate from the regression of $\ddot{D}$ onto $\ddot{Z}$, with observation i excluded. In one step, this can be written as:

$$\hat{\beta}_{\text{IJIVE1}} = \frac{\ddot{D}'\ddot{H}'\ddot{Y}}{\ddot{D}'\ddot{H}'\ddot{D}}, \qquad \ddot{H} = (I - \text{diag}(H_{\ddot{Z}}))^{-1}(H_{\ddot{Z}} - \text{diag}(H_{\ddot{Z}})).$$

The second solution is to exclude own observation both when calculating the severity of the judge assigned to i, and also when calculating the average sentencing rate in the location of i,

$$\hat{D}_{\text{UJIVE},i} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_{j\neq i} D_j \mathbb{1}\{G_j = G_i\}}{\sum_{j\neq i}\mathbb{1}\{G_j = G_i\}}$$

Kolesár (2013) calls this estimator unbiased jackknife instrumental variables estimator (UJIVE). With general instruments and covariates, the estimator takes the form

$$\hat{\beta}_{\text{UJIVE}} = \frac{D'\ddot{H}'Y}{D'\ddot{H}'D},$$
$$\ddot{H} = (I - \text{diag}(H_X))^{-1}(H_X - \text{diag}(H_X)) - (I - \text{diag}(H_W))^{-1}(H_W - \text{diag}(H_W)).$$

This implements the following procedure in one step: (i) run the regression of D onto (W, Z) with ith observation removed to compute $\hat{\pi}_{-i}$ and $\hat{\psi}_{D,-i}$. Compute $\hat{Z}_{\text{UJIVE},i} = Z_i'\hat{\pi}_{-i} + W_i'\hat{\psi}_{D,-i}$. (ii) adjust $\hat{Z}_{\text{UJIVE},i}$ for covariates by regressing D onto W, with ith observation removed to compute $\hat{\tau}_{-i}$, constructing $\hat{D}_{\text{UJIVE},i} = \hat{Z}_{\text{UJIVE},i} - W_i\hat{\tau}_{-i}$. Then use this as a single instrument, $\hat{\beta}_{\text{UJIVE}} = \hat{D}'_{\text{UJIVE}}Y/\hat{D}'_{\text{UJIVE}}D$.

Both jackknife estimators, UJIVE and IJIVE1 are consistent so long as:

1. The instruments are not too weak, in the sense that $\sqrt{K}/r_n \to 0$ (without this condition, no estimator can be consistent).

2. There aren't too many covariates: IJIVE1 requires $LK/r_n n \to 0$, while consistency of UJIVE only requires $\sqrt{L}/r_n \to 0$.

2. INFERENCE

Let us define $\pi_\Delta = \delta - \pi\beta$, $u_{\Delta,i} = u_{Yi} - u_{Di}\beta$, $\gamma = E[W_i W_i']^{-1} E[W_i(Y_i - D_i\beta)] = E[W_i W_i']^{-1} E[W_i Z_i]\pi_\Delta + \psi_Y - \psi_D\beta$, and $\epsilon_i = \tilde{Z}_i'\pi_\Delta + u_{\Delta,i}$. Then we can write the "structural" equation as

$$Y_i = D_i\beta + W_i'\gamma + \epsilon_i.$$

We saw in the previous lecture that the oracle estimator satisfied

$$\mathcal{V}_{1,n}^{-1/2}(\hat{\beta}^* - \beta) \Rightarrow \mathcal{N}(0,1), \qquad \mathcal{V}_{1,n} = \frac{E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

The variance $\mathcal{V}_{1,n}$ is the variance estimated by the conventional robust standard errors, such as those in Stata. These are also the correct standard errors for TSLS, UJIVE, or IJIVE1 if (i) there is no treatment effect heterogeneity, and (ii) $K/r_n \to 0$ for UJIVE, or IJIVE1, and $K^2/r_n \to 0$ for TSLS. If (i) fails, then, as discussed last time, we don't achieve the oracle variance, but instead the correct asymptotic variance is given by

$$\mathcal{V}_{2,n} = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{D,i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

What about (ii)? If $K/r_n \to 0$, but $K^2/r_n \not\to 0$, then TSLS will be consistent, but asymptotically biased, and inference based on TSLS will be difficult. However, inference based on UJIVE or IJIVE1 is simple in this case: use an estimate of $\mathcal{V}_{2,n}$, or, if we want to impose homogeneity of treatment effects, $\mathcal{V}_{1,n}$. If K/r_n doesn't converge to zero, then we need to account for the presence of many instruments in the asymptotic variance formula (Evdokimov and Kolesár 2018, Theorem 5.4)

$$(\mathcal{V}_{2,n} + \mathcal{V}_{MI,n})^{-1/2}(\hat{\beta}_{\text{UJIVE}} - \beta) \Rightarrow \mathcal{N}(0,1),$$
$$\mathcal{V}_{MI,n} = \frac{1}{r_n^2}\sum_{i\neq j}(H_{\tilde{Z},ij}^2 u_{\Delta,i}^2 u_{D,j}^2 + H_{\tilde{Z},ij}^2 u_{\Delta,i}u_{D,i}\cdot u_{\Delta,j}u_{D,j}), \quad (3)$$

The same result holds with $\hat{\beta}_{\text{UJIVE}}$ replaced with $\hat{\beta}_{\text{IJIVE1}}$. This result generalizes that in Chao et al. (2012), who give a formula for JIVE1, assuming constant treatment effects and assuming that L is fixed with sample size.

Under homoskedastic errors, $\Omega(X_i) = \Omega = E[u_i u_i']$, this additional many instrument term in the variance formula simplifies to[3]

$$\mathcal{V}_{MI,n} = \frac{K}{r_n^2}(E[(u_{Yi} - u_{Di}\beta)^2]\cdot E[u_{Di}^2] + E[(u_{Yi} - u_{Di}\beta)u_{Di}]^2)(1 + o_p(1)).$$

3. This follows since $\sum_{i,j} H_{\tilde{Z},ij}^2 = \text{tr}(H_{\tilde{Z}}) = K$, and $\sum_i H_{\tilde{Z},ii}^2 \leq \max_i H_{\tilde{Z},ii}\sum_i H_{\tilde{Z},ii} = o(1)K$, provided that $\max_i H_{\tilde{Z},ii} \to 0$ (this is a familiar leverage condition).

2.1. *Estimating the standard errors*

For correct inference, in addition for accounting for the additional $\mathcal{V}_{MI,n}$ term in the asymptotic variance, we need to ensure that our estimate of $\mathcal{V}_{1,n} + \mathcal{V}_{MI,n}$, or, preferably, $\mathcal{V}_{2,n} + \mathcal{V}_{MI,n}$ is consistent even under the many instrument asymptotics. The issue is that under standard asymptotics, the usual estimates of π, β, and γ are all consistent, so we can just use plug-in estimators. However, simple plugin estimators are not consistent once we let K and L increase with the sample size.

To see the issue, suppose, for simplicity, that the reduced-form errors are homoskedastic, and that there is no treatment effect heterogeneity. Then

$$\mathcal{V}_{1,n} = \mathcal{V}_{2,n} = \frac{E[\epsilon_i^2]}{r_n} \approx \frac{E[\epsilon_i^2]}{\pi' \ddot{Z}' \ddot{Z} \pi}.$$

The natural estimator, used by default in Stata for inference based on TSLS is given by

$$\frac{\frac{1}{n}\sum_i \hat{\epsilon}_{i,\text{TSLS}}^2}{D' H_{\ddot{Z}} D}.$$

Using footnote 1, it follows that $E[D' H_{\ddot{Z}} D] = E[\pi \ddot{Z}' \ddot{Z} \pi] + KE[u_{Di}^2] \approx r_n + K\Omega_{DD}$. This leads to a downward bias in the standard errors. See Section 4 for an illustration.

The remedy to the bias in the denominator is simple: use the denominator of IJIVE1 or UJIVE to estimate r_n. Getting the numerator right is trickier—see Evdokimov and Kolesár (2018).

[[TODO: describe estimator of the many instrument term]].

3. FIRST STAGE F

In some papers, due to the dimensionality of the problem, researchers calculate the instrument $\hat{Z}_i$ manually, often as a leave-one-out prediction, effectively computing JIVE1 by hand. Often, they then forget that $\hat{Z}_i$ is a constructed instrument, and compute the first stage F statistic, as well as all other results, as if $K = 1$. This will of course overstate the actual instrument strength.

When using the (correctly computed) first-stage F statistic for diagnostics, remember that the $F > 10$ rule of thumb is a test of the hypothesis that the TSLS bias, relative to the bias of OLS exceeds 0.1. So if F is small, this is an indication that TSLS is biased. Since we have argued that it is good practice to use jackknife estimators precisely to eliminate the TSLS bias, a small F statistic is not necessarily a concern when UJIVE or IJIVE1 are used.

In particular, with homoskedastic errors,

$$E[F] = \frac{E[\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\pi}]}{KE[u_{Di}^2]} = \frac{\pi E[\ddot{Z}' \ddot{Z}] \pi}{KE[u_{Di}^2]} + 1 \approx \frac{r_n}{KE[u_{Di}^2]} + 1.$$

Indeed, we have seen that if $r_n/K \to 0$, TSLS will be inconsistent; but jackknife estimators will remain consistent so long as $r_n/\sqrt{K} \to \infty$, and we can still do inference using jackknife estimators, provided we account for the additional many instrument term in the asymptotic variance.

4. SUMMARY AND ILLUSTRATION

If there are many instruments, in the sense that the number of instruments K is non-negligible relative to the sample size scaled by instrument strength r_n, the TSLS estimator will be biased. Instead, use IJIVE1 or UJIVE. To ensure reliable inference, the standard errors need to account for an additional many instrument term in the asymptotic variance, and avoid the downward bias that's present in the default standard errors estimator based on TSLS.

To illustrate these takeaways, we consider the dataset from Angrist and Krueger (1991), who use quarter of birth as in instrument in a regression of log earnings on education. Let us focus on the subsample of men born in 1930–39 from 1980 census. We consider four specifications. The first three correspond to a version of Table III, and to Tables V and VII in Angrist and Krueger (1991). In the first specification, we just use quarter of birth as an instrument. In the second, we interact it with 10 year of birth indicators, for a total of 30 instruments. In the third, we also interact it with state of birth indicators, for a total of $30 + 50 \times 3 = 180$ instruments. Finally, we consider a specification in which we use a triple interaction of year with state and with quarter of birth. We drop Alaska and Hawaii, which have some empty cells, or cells with only one observation (for which the jackknifing would be impossible). This reduces the number of observations by 324, and we're left with 329,185 observations. This leaves us with a total of $49 \times 10 \times 3 = 1470$ instruments.

The ratio K/n is very small in all of these specifications. However, the ratio of r_n to K is fairly small, as is the first-stage F, indicating that there will likely be a problem with TSLS. Indeed, we see that

1. The estimate gets close to OLS in Panel 4, in line with the theory above
2. The estimate of $\mathcal{V}_{1,n}$ and $\mathcal{V}_{2,n}$ appears to be biased downward in multiple panels, in line with the theory above.

Question 1. Do these two observations explain the findings of Bound, Jaeger, and Baker (1995)?

We also see that JIVE1 is rather erratic. This stems from the issue that the dimension of controls L is high relative to the strength of identification. In Panel 4, the JIVE1-based estimate of r_n is in fact negative—while, as discussed in Section 2.1, TSLS overestimates r_n, and hence underestimates the standard error, JIVE1 underestimates it, and hence

Table 1: Application to Angrist and Krueger (1991)

Estimator	Estimate	$\hat{V}_1^{1/2}$	$\hat{V}_2^{1/2}$	$\sqrt{\hat{V}_2 + \hat{V}_{MI}}$	$\hat{r}_n/K$
Panel A: OLS					
OLS	0.0670	0.0004			
Panel B: Instrument is QOB. $F = 34.0$					
TSLS	0.1026	0.0195	0.0198		366.0
JIVE1	0.1039	0.0203	0.0206	0.0209	351.6
UJIVE	0.1036	0.0201	0.0204	0.0207	355.2
Panel C: Instrument is $QOB \times YOB$. $F = 4.9$					
TSLS	0.0891	0.0162	0.0176		52.6
JIVE1	0.0959	0.0224	0.0244	0.0273	38.3
UJIVE	0.0938	0.0204	0.0222	0.0211	41.9
Panel D: Instrument is $QOB \times YOB + QOB \times SOB$, $F = 2.6$					
TSLS	0.0928	0.0097	0.0112		26.2
JIVE1	0.1211	0.0205	0.0243	0.0273	12.7
UJIVE	0.1096	0.0160	0.0187	0.0211	16.1
Panel E: Instrument is $QOB \times YOB \times SOB$, $F = 1.1$					
TSLS	0.0721	0.0049	0.0067		11.6
JIVE1	0.0320	0.0307	0.0425	0.0515	-1.9
UJIVE	0.1110	0.0397	0.0548	0.0663	1.4

tends to overestimate the standard errors. This explains why the JIVE1–based standard errors tend to be the largest.

In contrast, UJIVE is quite stable. The many-instrument term in the standard error formula doesn't matter in the first two specification, but increases the standard errors by 12% and 21%, respectively, in the last two specifications. Allowing for heterogeneity in the treatment effects has a similar effect.

REFERENCES

Ackerberg, Daniel A., and Paul J. Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *Review of Economics and Statistics* 91, no. 2 (May): 351–362. https://doi.org/10.1162/rest.91.2.351.

Aizer, Anna, and Joseph J. Doyle Jr. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." *The Quarterly Journal of Economics* 130, no. 2 (May): 759–803. https://doi.org/10.1093/qje/qjv003.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14 (1): 57–67. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62, no. 3 (May): 657–681. https://doi.org/10.2307/2951662.

Belloni, Alexandre, Daniel L. Chen, Victor Chernozhukov, and Christian B. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80, no. 6 (November): 2369–2429. https://doi.org/10.3982/ECTA9626.

Blomquist, Soren, and Matz Dahlberg. 1999. "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments." *Journal of Applied Econometrics* 14, no. 1 (January): 69–88. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<69::AID-JAE521>3.0.CO;2-7.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90, no. 430 (June): 443–450. https://doi.org/10.1080/01621459.1995.10476536.

Chan, David, Matthew Gentzkow, and Chuan Yu. 2019. *Selection with Variation in Diagnostic Skill: Evidence from Radiologists.* Working Paper 26467. Cambridge, MA: National Bureau of Economic Research, November. https://doi.org/10.3386/w26467.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. 2012. "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments." *Econometric Theory* 12, no. 1 (February): 42–86. https://doi.org/10.1017/S0266466611000120.

Evdokimov, Kirill, and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." January. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113, no. 1 (January): 253–277. https://doi.org/10.1257/aer.20201860.

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.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96, no. 3 (May): 863–876. https://doi.org/10.1257/aer.96.3.863.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity." Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Nagar, Anirudh Lal. 1959. "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations." *Econometrica* 27, no. 4 (October): 575–595. https://doi.org/10.2307/1909352.

Phillips, G. D. A., and C. Hale. 1977. "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems." *International Economic Review* 18, no. 1 (February): 219–228. https://doi.org/10.2307/2525779.

MANY INSTRUMENTS AND JUDGES

Michal Kolesár

EC0539B, Fall 2022

April 3, 2024

- Consider linear instrumental variables (IV) model when number of instruments K large. Arises in 2 ways:
    1. Interact low-dim instrument with controls (Angrist and Krueger 1991). Or use approach of Belloni et al. (2012).
    2. Use fixed effects (group indicators) as instruments: "judge" or "examiner" designs. Studies of effects of incarceration on economic outcomes (Kling 2006; Aizer and Doyle 2015), studies that exploit random assignment of judges to bankruptcy cases, criminal cases, or patent cases, doctors to shifts…

Key issues

1. two-stage least squares (TSLS) biased with many instruments. Alternatives?
2. standard errors
3. Interpretation of first-stage F statistic

Setup and Estimation

Inference

First stage F

Summary and illustration

- Same notation as in previous lecture, with reduced form and first stage

$$Y_i = Z_i'\delta + W_i'\psi_Y + u_{Yi}, \tag{1}$$

$$D_i = Z_i'\pi + W_i'\psi_D + u_{Di}. \tag{2}$$

  The parameter of interest is $\beta := \frac{\delta' Q \pi}{\pi' Q \pi}$, where $Q = E[\tilde{Z}_i \tilde{Z}_i']$, and $\tilde{Z}_i = Z_i - E[Z_i W_i'] E[W_i W_i']^{-1} W_i$.

- To measure overall strength of instruments, define measure of "effective sample size"

$$r_n = n\pi' Q \pi = nE[(\tilde{Z}_i'\pi)^2].$$

- $X_i = (Z_i', W_i')'$ collects the right-hand side (RHS) (or "exogenous") variables.

- For any matrix A, let $H_A = A(A'A)^{-1}A'$ denote hat matrix (also called a projection matrix), and let $\ddot{A} = A - H_W A$ denote residuals after projecting A onto the covariates.
  - Thus $\tilde{Z}_i$ is the population analog of $\ddot{Z}_i$

Judges design

- Individual i assigned judge Q_i, random conditional on i's geographic location G_i.
- Covariates and instruments both indicators: $W_{i\ell} = \mathbb{1}\{G_i = \ell\}$, and $Z_{ik} = \mathbb{1}\{Q_i = k\}$.
- There are L locations and $K + L$ judges; in each location we drop one judge to avoid collinearity—call this judge reference judge.
- $\psi_{D,\ell}$ is average sentencing rate of the reference judge in location ℓ, and π_k is sentencing rate of judge k relative to the reference judge in the same location.

- Monotonicity assumption requires that judges agree on the ranking of the defendants, they just disagree on the cutoff at which they start sentencing them.
- Can be problematic: Chan, Gentzkow, and Yu (2019) point out that decisions of physicians differ due to both skill and preferences.
- Similarly, Kleinberg et al. (2018) find that the increase in crime associated with judges who are more likely to release defendants on bail is about the same as if these more lenient judges randomly picked the extra defendants to release on bail.
- Failure of monotonicity only affects the interpretation of β; to keep statistical issues separate from identification issues, we put these issues aside here.

Consider asymptotics in which $K = \dim(Z_i)$ and $L = \dim(W_i)$ can to grow with sample size

Lemma

The TSLS estimator suffers from own observation bias towards ordinary least squares (OLS). In particular, suppose $E[u_i \mid X_i] = 0$, with $\Omega(X_i) = E[u_i u_i' \mid X_i]$ bounded, $E[\tilde{Z}_i \mid W_i] = 0$, and $L/n \to 0$.

Then consistency of TSLS is in general requires $K/r_n \to 0$. Under homoskedastic errors,

$$\hat{\beta}_{\text{TSLS}} = \beta + \frac{(\Omega_{YD} - \Omega_{DD}\beta)K}{r_n + \Omega_{DD}K} + o_p(1) = (1-w)\beta + w\beta_{\text{OLS}} + o_p(1), \quad w = \frac{K}{r_n/\Omega_{DD} + K},$$

where $\beta_{\text{OLS}} = (\Omega_{YD} - \Omega_{DD}\beta)/\Omega_{DD} + \beta$ is the probability limit of OLS.

TSLS bias scales with K/r_n, rather than K/n!

- Bias arises because single constructed instrument $\hat{Z}_{\text{TSLS},i} = Z_i'\hat{\pi} + W_i'\hat{\psi}_D$ used by TSLS puts positive weight on own treatment status D_i
- Total net weight, holding overall instrument strength constant, scales linearly with K, so TSLS bias scales with the number of instruments.
- Solution 1 (Bekker 1994): use limited information maximum likelihood (LIML), or variants thereof. Only need $\sqrt{K}/r_n \to 0$ for consistency, substantially weaker requirement. But LIML-like estimators not robust to heterogeneous treatment effects.
- Solution 2: subtract estimate of bias. Easy to do under homoskedastic errors: leads to variants of the bias-corrected TSLS estimator that dates back to Nagar (1959). But consistency does depend on homoskedasticity.

- Bias caused by using D_i in constructing the single instrument $\hat{Z}_{\text{TSLS},i}$: obvious solution is to not use it!
- Option 1: construct predictors $\hat{\pi}_{-i}$ and $\hat{\psi}_{D,-i}$ of π and ψ_D based on a regression of D onto (Z, W), but with the ith observation removed, to construct a single instrument $\hat{Z}_{\text{JIVE1},i} = Z_i'\hat{\pi}_{-i} + W_i'\hat{\psi}_{D,-i}$
  - What is this in judges design?
- Then run an IV regression of Y_i onto D_i and W_i, using $\hat{Z}_{\text{JIVE1},i}$ as an instrument for D_i. Resulting estimator known as jackknife instrumental variables estimator (JIVE1):

$$\hat{\beta}_{\text{JIVE1}} = \frac{\hat{Z}'_{\text{JIVE1}}\ddot{Y}}{\hat{Z}'_{\text{JIVE1}}\ddot{D}}, \quad \hat{Z}_{\text{JIVE1}} = \ddot{H}D, \quad \ddot{H} = (I - \text{diag}(H_X))^{-1}(H_X - \text{diag}(H_X)),$$

- Don't need to run n first-stage regressions
- Problem: when we project out the covariates from $\hat{Z}_i$, we re-introduce the own observation bias
  - In judges design, adjust instrument by average sentencing rate in the location of i
- Generally need $L/r_n \to 0$ for consistency (see Evdokimov and Kolesár (2018)). Under homo, bias goes in opposite direction to TSLS bias:

$$\hat{\beta}_{\text{JIVE1}} = \beta - \frac{(\Omega_{YD} - \Omega_{DD}\beta)L}{r_n - \Omega_{DD}L} + o_p(1) = (1+\lambda)\beta - \lambda\beta_{OLS} + o_p(1), \;\; \lambda = \frac{L}{r_n/\Omega_{DD} - L}.$$

- Leads to option 2: *first* partial out the covariates, and then do the leave-one-out prediction. Called improved improved jackknife instrumental variables estimator (IJIVE1) estimator, proposed in Ackerberg and Devereux (2009).

$$\hat{\beta}_{\text{IJIVE1}} = \frac{\ddot{D}'\ddot{H}'\ddot{Y}}{\ddot{D}'\ddot{H}'\ddot{D}}, \qquad \ddot{H} = (I - \text{diag}(H_{\ddot{Z}}))^{-1}(H_{\ddot{Z}} - \text{diag}(H_{\ddot{Z}})).$$

- Option 3: to exclude own observation both when calculating the severity of the judge assigned to i, and also when calculating the average sentencing rate in the location of i.

Kolesár (2013) calls this estimator unbiased jackknife instrumental variables estimator (UJIVE).

$$\hat{D}_{\text{UJIVE},i} = \hat{Z}_{\text{JIVE1},i} - \frac{\sum_{j \neq i} D_j \mathbb{1}\{G_j = G_i\}}{\sum_{j \neq i} \mathbb{1}\{G_j = G_i\}}$$

That is, (i) run jackknife the regression of D onto (W, Z) to compute $\hat{Z}_{\text{UJIVE},i} = Z_i' \hat{\pi}_{-i} + W_i' \hat{\psi}_{D,-i}$. (ii) adjust $\hat{Z}_{\text{UJIVE},i}$ for covariates by jackknife regression of D onto W. Then use this as a single instrument, $\hat{\beta}_{\text{UJIVE}} = \hat{D}_{\text{UJIVE}}' Y / \hat{D}_{\text{UJIVE}}' D$.

- Both UJIVE and IJIVE1 consistent so long as:
  1. $\sqrt{K}/r_n \to 0$ (without this condition, no estimator can be consistent).
  2. There aren't too many covariates: IJIVE1 requires $LK/r_n n \to 0$, while consistency of UJIVE only requires $\sqrt{L}/r_n \to 0$.

Setup and Estimation

Inference

First stage F

Summary and illustration

- Define $\pi_\Delta = \delta - \pi\beta$, $u_{\Delta,i} = u_{Yi} - u_{Di}\beta$,
  $\gamma = E[W_i W_i']^{-1} E[W_i(Y_i - D_i\beta)] = E[W_i W_i']^{-1} E[W_i Z_i]\pi_\Delta + \psi_Y - \psi_D\beta$, and $\epsilon_i = \tilde{Z}_i'\pi_\Delta + u_{\Delta,i}$.
- Write the "structural" equation as

$$Y_i = D_i\beta + W_i'\gamma + \epsilon_i.$$

- saw in the previous lecture that the oracle estimator satisfied

$$\mathcal{V}_{1,n}^{-1/2}(\hat{\beta}^* - \beta) \Rightarrow \mathcal{N}(0,1), \qquad \mathcal{V}_{1,n} = \frac{E[\epsilon_i^2(\tilde{Z}_i'\pi)^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

  Variance $\mathcal{V}_{1,n}$ is estimated by the conventional robust standard errors, such as those in Stata. Also correct standard errors for TSLS, UJIVE, or IJIVE1 if (i) there is no treatment effect heterogeneity, and (ii) $K/r_n \to 0$ for UJIVE, or IJIVE1, and $K^2/r_n \to 0$ for TSLS.

- If (i) fails, then, as discussed last time, don't achieve oracle variance, but instead the correct asymptotic variance is given by

$$\mathcal{V}_{2,n} = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{D,i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{r_n E[(\tilde{Z}_i'\pi)^2]}.$$

- What if (ii) fails? What if $K/r_n \to 0$, but $K^2/r_n \not\to 0$ and we use TSLS?

- If $K/r_n \not\to 0$, and use UJIVE or IJIVE1, we need to account for the presence of many instruments in the asymptotic variance formula (Evdokimov and Kolesár 2018, Theorem 5.4)

$$(\mathcal{V}_{2,n} + \mathcal{V}_{MI,n})^{-1/2}(\hat{\beta}_{\text{UJIVE}} - \beta) \Rightarrow \mathcal{N}(0,1),$$

$$\mathcal{V}_{MI,n} = \frac{1}{r_n^2}\sum_{i\neq j}(H^2_{\tilde{Z},ij}u^2_{\Delta,i}u^2_{2,j} + H^2_{\tilde{Z},ij}u_{\Delta,i}u_{2,i}\cdot u_{\Delta,j}u_{2,j}), \quad (3)$$

- Under homoskedasticity, $\Omega(X_i) = \Omega = E[u_i u_i']$, additional many instrument term simplifies to

$$\mathcal{V}_{MI,n} = \frac{K}{r_n^2}(E[(u_{i1} - u_{2i}\beta)^2]\cdot E[u_{2i}^2] + E[(u_{i1} - u_{2i}\beta)u_{2i}]^2)(1 + o_p(1)).$$

Setup and Estimation

Inference

First stage F

Summary and illustration

- Some papers calculate the instrument $\hat{Z}_i$ manually, often as a leave-one-out prediction, effectively computing JIVE1 by hand. But this does not mean that $K = 1$! That will overstate actual instrument strength.
- When using the (correctly computed) first-stage F for diagnostics, remember that the $F > 10$ rule of thumb tests hypothesis that TSLS bias, relative to the bias of OLS exceeds 0.1.
- But small F statistic not necessarily a concern when UJIVE or IJIVE1 are used. Under homoskedasticity,

$$E[F] = \frac{E[\hat{\pi}_2 \tilde{Z}' \tilde{Z} \hat{\pi}_2]}{KE[u_{2i}^2]} = \frac{\pi_2 E[\tilde{Z}' \tilde{Z}] \pi_2}{KE[u_{2i}^2]} + 1 \approx \frac{r_n}{KE[u_{2i}^2]} + 1.$$

- If $r_n/K \to 0$, TSLS will be inconsistent; but jackknife estimators will remain consistent so long as $r_n/\sqrt{K} \to \infty$

Setup and Estimation

Inference

First stage F

Summary and illustration

- If K non-negligible relative to effective sample size r_n, TSLS biased. Instead, use IJIVE1 or UJIVE.
  - K/r_n may be large even if K/n small!
  - JIVE1 not a good solution
- To ensure reliable inference, standard errors need to account for additional many instrument term in the asymptotic variance
  - Even more important is to avoid downward bias that's present in the default standard errors estimator based on TSLS—see notes.
- Will now illustrate in application to Angrist and Krueger (1991)

Estimator	Estimate	$\hat{V}_1^{1/2}$	$\hat{V}_2^{1/2}$	$\sqrt{\hat{V}_2 + \hat{V}_{MI}}$	$\hat{r}_n/K$
Panel A: OLS					
OLS	0.0670	0.0004			
Panel B: Instrument is QOB. $F = 34.0$					
TSLS	0.1026	0.0195	0.0198		366.0
JIVE1	0.1039	0.0203	0.0206	0.0209	351.6
UJIVE	0.1036	0.0201	0.0204	0.0207	355.2
Panel C: Instrument is $QOB \times YOB$. $F = 4.9$					
TSLS	0.0891	0.0162	0.0176		52.6
JIVE1	0.0959	0.0224	0.0244	0.0273	38.3
UJIVE	0.0938	0.0204	0.0222	0.0211	41.9
Panel D: Instrument is $QOB \times YOB + QOB \times SOB$, $F = 2.6$					
TSLS	0.0928	0.0097	0.0112		26.2
JIVE1	0.1211	0.0205	0.0243	0.0273	12.7
UJIVE	0.1096	0.0160	0.0187	0.0211	16.1
Panel E: Instrument is $QOB \times YOB \times SOB$, $F = 1.1$					
TSLS	0.0721	0.0049	0.0067		11.6
JIVE1	0.0320	0.0307	0.0425	0.0515	-1.9
UJIVE	0.1110	0.0397	0.0548	0.0663	1.4

REFERENCES I

Ackerberg, Daniel A., and Paul J. Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *Review of Economics and Statistics* 91, no. 2 (May): 351–362. https://doi.org/10.1162/rest.91.2.351.

Aizer, Anna, and Joseph J. Doyle Jr. 2015. "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges." *The Quarterly Journal of Economics* 130, no. 2 (May): 759–803. https://doi.org/10.1093/qje/qjv003.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62, no. 3 (May): 657–681. https://doi.org/10.2307/2951662.

Belloni, Alexandre, Daniel L. Chen, Victor Chernozhukov, and Christian B. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica* 80, no. 6 (November): 2369–2429. https://doi.org/10.3982/ECTA9626.

Chan, David, Matthew Gentzkow, and Chuan Yu. 2019. *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*. Working Paper 26467. Cambridge, MA: National Bureau of Economic Research, November. https://doi.org/10.3386/w26467.

Evdokimov, Kirill, and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." January. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf.

REFERENCES II

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.

Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96, no. 3 (May): 863–876. https://doi.org/10.1257/aer.96.3.863.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity." Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Nagar, Anirudh Lal. 1959. "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations." *Econometrica* 27, no. 4 (October): 575–595. https://doi.org/10.2307/1909352.

SHIFT-SHARE DESIGNS

Michal Kolesár

ECO539B, Fall 2022

April 11, 2024

- Observations $i = 1, \ldots, N$ corresponding to regions (commuting zones (CZs), counties etc). Interested in effect of some treatment D_i on outcome Y_i.
- Construct regional instruments X_i by combining national-level shifters (or shocks) $\mathcal{X}_s$ to different sectors s (e.g. industries) with regional shares (or weights) w_{is},

$$X_i = \sum_{s=1}^{S} w_{is} \mathcal{X}_s, \qquad w_{is} \geq 0.$$

So X_i is a shift-share instrument.

Key issues

How do we think about: (i) identification and (ii) inference?

EXAMPLES

- Bartik (1991) or Blanchard and Katz (1992), want to estimate inverse labor supply elasticity
  - Y_i: log change in wages; D_i: log change in employment. Need instrument for labor demand. $\mathcal{X}_s$: employment growth of industry s; w_{is}: lagged employment shares.
- Autor, Dorn, and Hanson (2013, ADH) want effect of Chinese imports on US labor markets.
  - Y_i: local labor market outcome (employment, wages, …); D_i measures rise of Chinese imports to location i. Weak labor markets may be more likely to import. $\mathcal{X}_s$: growth of Chinese exports to non-US countries in industry s; w_{is}: lagged industry shares.
- Effects of immigration (e.g. Card 2001) on natives, or other outcomes
  - D_i: local immigration. i can be region-skill group cell, national-level education-experience cell, or simply region. w_{is}: share of immigrants from country s; $\mathcal{X}_s$: (normalized) change or growth in # of immigrants from country s. Very similar to the logic of Bartik (1991).
- Also: effects on local labor markets of trade competition, technological change, credit supply; impact of sectoral shocks on marriage patters, crime levels, innovation…See, e.g., Adão, Kolesár, and Morales (2019) for partial list.

Identification

Estimation and Inference

Empirical Applications

Need a sense in which X_i is as good as randomly assigned. Two approaches:

1. View w_{is} as randomly assigned. Explored in Goldsmith-Pinkham, Sorkin, and Swift (2020). Can then think of setup as overidentified instrumental variables (IV) model, with X_i as one of many possible ways of combining "instruments" w_{is}.
    - If we believe exogeneity of w_{is}, no issues with estimation of inference, so long as (i) $n \to \infty$, and (ii) we can form groups of regions across which w_{is} are independent.
    - Can try to "open the black box" of the constructed instrument by looking at the implicit weights X_i puts on each share, called "Rotemberg weights" in this context; IV analog to leverage analysis in ordinary least squares (OLS).
    - But shares w_{is} not plausibly exogenous in many applications: worry there are shift-share terms with structure similar to X_i in the structural residual (will explore below).
2. View $\mathcal{X}_s$ as randomly assigned. First proposed by a working-paper version of Borusyak, Hull, and Jaravel (2022); advocated for by Adão, Kolesár, and Morales (2019). Focus on here.

To serve as background to help think through econometric issues, useful to consider stylized model (taken from Adão, Kolesár, and Morales 2019).

- Economy with multiple sectors $s = 1, \ldots, S$ and regions $i = 1, \ldots, J$
- Labor demand and supply depend on wage ω_i:

$$\begin{aligned} &\text{Demand of } s \text{ in } i: && \log L_{is} = -\sigma_s \log \omega_i + \log D_{is}, && \sigma_s > 0, \\ &\text{Supply in } i: && \log L_i = \phi \log \omega_i + \log S_i, && \phi > 0, \end{aligned}$$

  σ_s and ϕ: labor demand and supply elasticities; ω_i: wage; S_i: supply shifter;
- Decompose labor demand shifter D_{is} into observed shifter of interest χ_s and potentially unobserved components: $\log D_{is} = \rho_s \log \chi_s + \log u_s^D + \log u_{is}^D$.

- Also decompose labor supply shocks into group-specific shifters:

$$\log S_i = \sum_{g=1}^{G} \tilde{w}_{ig} \log u_g^S + \log u_i^S,$$

- Workers assumed immobile across regions, freely mobile across sectors $\implies$ market clearing: $L_i = \sum_{s=1}^{S} L_{is}$ for $i = 1, \ldots, J$.
- Assume $\{\hat{\chi}_s, \hat{u}_s^D, \hat{u}_{is}^D, \hat{u}_g^S, \hat{u}_i^S\}_{i,s,g} \sim F$, where $\hat{z} = \log(z^t / z^0)$.

- Up to first-order approximation around initial equilibrium

$$\hat{L}_i = \sum_{s=1}^{S} l_{is}^0 (\gamma_{is}\hat{\chi}_s + \lambda_i \hat{u}_s^D + \lambda_i \hat{u}_{is}^D) + (1-\lambda_i)\left(\sum_{g=1}^{G} \tilde{w}_{ig}\hat{u}_g^S + \hat{u}_i^S\right),$$

  l_{is}^0: initial employment share, $\lambda_i \equiv \phi\left[\phi + \sum_s l_{is}^0 \sigma_s\right]^{-1}$, and $\gamma_{is} \equiv \rho_s \lambda_i$ (recall ρ_s is elasticity of demand wrt χ_s). Similar to a reduced-form regression considered in Autor, Dorn, and Hanson (2013).

- Solving for wages yields regression like that in Bartik (1991) or Blanchard and Katz (1992)

$$\hat{\omega}_i = \frac{1}{\phi}\hat{L}_i - \frac{1}{\phi}\left(\sum_{g=1}^{G} \tilde{w}_{ig} \log u_g^S + \log u_i^S\right)$$

  Can use demand shifters $\hat{\chi}_s$ to estimate $1/\phi$ with IV, since $\phi^{-1} = (\partial\hat{\omega}_i/\partial\hat{\chi}_s)/(\partial\hat{L}_i/\partial\hat{\chi}_s)$

KEY TAKEAWAYS

1. Change in regional employment and wage changes both combine multiple shift-share terms; complicated correlation structure in residuals
2. Shifter effects depend on parameters that are heterogeneous across i and s.

Focus on reduced-form effect of X_i onto Y_i for clarity of argument.

- As always, first ask: what do we mean by "effect"? What's the ideal experiment here?
- Use potential outcomes, $Y_i(\chi_1, \ldots, \chi_s) = Y_i(0) + \sum_s w_{is}\chi_s\beta_{is}$. Observed outcome is $Y_i(\mathcal{X}_1, \ldots, \mathcal{X}_s)$.
  - Exercise: how would one define potential outcomes in Goldsmith-Pinkham, Sorkin, and Swift (2020)? What's the idealized experiment? Can we think of an economic model with such a structure?
  - ADH example: What would regional outcomes be if we assign different shocks $\mathcal{X}_s$ to Chinese export growth? Posit effect proportional to regional employment exposure to industry s, w_{is}.
  - For studying effect of demand shocks $\hat{\chi}_s$ in stylized model,

$$Y_i = \hat{L}_i,\ w_{is} = l_{is}^0,\ \chi_s = \hat{\chi}_s,\ \ \beta_{is} = \gamma_{is},$$

    and $Y_i(0) = \lambda_i \sum_{s=1}^S w_{is}(\hat{u}_s^D + \hat{u}_{is}^D) + (1-\lambda_i)\left(\sum_{g=1}^G \tilde{w}_{ig}\hat{u}_g^S + \hat{v}_i\right)$ aggregates all shifters other than $\hat{\chi}_s$.
- w_{is} are equilibrium objects, condition on them throughout.

- With no controls, we need $\mathcal{X}$ to be as good as randomly assigned, in line with discussion in OLS section of the course: $E[\mathcal{X}_s \mid \mathcal{F}_0] = 0$, $\mathcal{F}_0 = (Y(0), B, W)$. We use matrix notation: A has N rows, $\mathcal{A}$ has S rows, B has elements β_{is}, W has elements w_{is}. OLS estimand:

$$\beta = \frac{\sum_i E[X_i Y_i \mid \mathcal{F}_0]}{\sum_i E[X_i^2 \mid \mathcal{F}_0]} = \frac{\sum_{i=1}^N \sum_{s=1}^S \pi_{is}\beta_{is}}{\sum_{i=1}^N \sum_{s=1}^S \pi_{is}}, \tag{1}$$

  where $\pi_{is} = w_{is}^2 \operatorname{var}(\mathcal{X}_s \mid W)$.
- Treats population of interest to be sample at hand. Answers what we'd be estimating in idealized experiment where $\mathcal{X}_s$ randomly assigned.
- Although shares endogenous, we exploit that variation in shares w_{is} generates variation in exposure to exogenous shocks $\mathcal{X}_s$: this is basis for identification.

- If have controls with exact shift-share structure, $Z_i = \sum_s w_{is} \mathcal{Z}_s$, then need to assume (why?)

$$E[\mathcal{X}_s \mid \mathcal{Z}, Y(0), B, W] = \mathcal{Z}_s' \gamma.$$

  Important: implies that if $\sum_s w_{is} \neq 1$, need to include $\sum_s w_{is}$ as control (why?)
  With controls, weights in (1) become $\pi_{is} = w_{is}^2 \operatorname{var}(\mathcal{X}_s \mid W, \mathcal{Z})$, and X_i is replaced by residual after projecting it off Z_i.

- What if controls don't have exact shift-share structure? One option is to assume that such controls "proxy" for unobserved sectoral shocks, see Adão, Kolesár, and Morales (2019) for formalization of this idea.

- **Research question:** is this the best way of thinking about the issue? Can we give some other explanation for why not run shift-share regressions at sectoral level?

1. Is β an interesting object? Is it policy relevant?
   - Analogous to policy relevance of OLS estimand in OLS lecture.
2. What are we estimating in the presence of cross-regional spillovers?
   - Analogous to relevance of comparing treated and control outcomes in an experiment where we worry about peer effects or other types of spillovers. See discussion of Stable unit treatment value assumption (SUTVA) in lecture on OLS.

We will not address these concerns here

Identification

Estimation and Inference

Empirical Applications

- In stylized economic model, $Y_i(0)$ (and hence regression residual) incorporates terms that have shift-share structure, with shares that could be identical to (other demand shocks u_s^D) or different from (supply shocks to occupational groups, u_g^S) w_{is}. Correct inference requires taking into account potential cross-regional correlation in residuals across observations with similar values of w_{is}.
- Regions with similar sectoral employment shares $\{w_{is}\}_{s=1}^S$ also tend to have similar regression residuals and hence similar $X_i\epsilon_i$. Correlation independent of regions' geographic location $\implies$ not captured by clustering on geographic units.

- Use 2000–2007 observed changes in labor market outcomes $\{Y_i\}_{i=1}^{n}$ and 1990 employment shares $\{w_{is}\}_{i=1,s=1}^{N,S}$ for US commuting zones.
- $N = 722$ regions; $S = 397$ sectors corresponding to 4-digit Standard Industrial Classification (SIC) codes
- For each simulation draw m, generate $\mathcal{X}_s^m \sim \mathcal{N}(0, 5)$, and estimate

$$Y_i = \delta + \beta \sum_{s=1}^{S} w_{is} \mathcal{X}_s^m + \epsilon_i.$$

  Computer-generated shocks cannot have impacted US labor mkt outcomes $\implies \beta = 0$.
- Consider Eicker-Huber-White (EHW) (*Robust*) and state-clustered standard errors (*Cluster*), commonly used in practice.

Standard errors and rejection rate of $H_0: \beta = 0$ at 5% level, from Adão, Kolesár, and Morales (2019).

	Estimate		Median std. error		Rejection rate	
	Mean	SD	Robust	Cluster	Robust	Cluster
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Change in the share of working-age population						
Employed	−0.01	2.00	0.73	0.92	48.5%	38.1%
Employed in mfg	−0.01	1.88	0.60	0.76	55.7%	44.8%
Employed in non-mfg	0.00	0.94	0.58	0.67	23.2%	17.6%
Panel B: Change in average log weekly wage						
Employed	−0.03	2.66	1.01	1.33	47.3%	34.2%
Employed in mfg	−0.03	2.92	1.68	2.11	26.7%	16.8%
Employed in non-mfg	−0.02	2.64	1.05	1.33	45.4%	33.7%

- To help focus on key issues, assume $\beta_{is} = \beta$, and abstract from controls. See Adão, Kolesár, and Morales (2019) for results with controls. Then

$$\hat{\beta} - \beta = \frac{\sum_i X_i Y_i(0)}{\sum_i X_i^2} = \frac{\sum_s \mathcal{X}_s \sum_i w_{is} Y_i(0)}{\sum_{s,t} \mathcal{X}_s \mathcal{X}_t \sum_i w_{is} w_{it}}.$$

- Use design-based framework where we condition on $\mathcal{F}_0 = (Y(0), W)$, and consider repeated sampling of $\mathcal{X}_s$. This is implicit in placebo. Would want only 5 of 100 researchers find a non-zero treatment effect if each of them comes up with a different set of irrelevant shifters.
- Key substantive assumption: $\mathcal{X}_s$ independent across s (but perhaps not identically distributed). No restriction on correlation structure of objects in $\mathcal{F}_0$.

- Need $S \to \infty$ for consistency.
- To ensure low "leverage", also need sectors to be small in the sense that $\max_s n_s/N \to 0$, where $n_s = \sum_i w_{is}$ is total share of sector s (and also regularity conditions on $Y_i(0)$)
  Intuition: consider "concentrated sectors", $w_{is} = \mathbb{1}\{s = s(i)\}$, where $s(i)$ is concentration of region i. Then setup equivalent to OLS in randomized controlled trials with cluster-level treatment, assumption equivalent to the largest cluster asymptotically negligible.
- In contrast, Goldsmith-Pinkham, Sorkin, and Swift (2020) can live with small, finite S.

- Sandwich form with bread as usual, but meat is different today:

$$\widehat{se}(\hat{\beta}) = \frac{\sqrt{\sum_{s=1}^{S} X_s^2 \hat{R}_s^2}}{\sum_{i=1}^{N} X_i^2}, \qquad \hat{R}_s = \sum_{i=1}^{N} w_{is} \hat{\epsilon}_i.$$

  With "concentrated sectors", $w_{is} = \mathbb{1}\{s = s(i)\}$, $\widehat{se}(\hat{\beta})$ equivalent to clustering regions that specialize in same sector—very different from clustering on state!

- In line with rule of thumb that one should "cluster" at level of variation of regressor of interest.
- Formula essentially forms sectoral clusters, using w_{is} to weight residuals.
- When effective # of clusters/sectors small, can impose H_0 when estimating the residual (call it AKMo). AKM vs AKMo analogous to Wald vs LM tests in likelihood models. In IV context, AKMo generalizes Anderson and Rubin (1949) CI to allow structural errors to be correlated. Robust to weak instruments.

Identification

Estimation and Inference

Empirical Applications

- Interested in effect of Chinese exports on US labor mkt outcomes. i: CZ; s: 4-digit SIC code, as in placebo. Use 1990–2007 data (2-period panel). $N = 1,444$ (722 CZs $\times$ 2 time periods).
- Estimate shift-share IV regression with:
  - Outcome: 10-year equivalent change in employment share
  - Endogenous variable: shift-share regressor, Chinese exports to US aggregated using beginning-of-period employment shares
  - Shift-share instrument: Chinese exports to rest of world, aggregated using beginning-of-period employment shares
  - Largest set of controls used in ADH; AKM and AKMo methods cluster at 3-digit SIC level.
- CIs on average 23% (AKM) or 66% (AKMo) wider, significance not affected

	All	Manuf.	Non-Manuf.
	(1)	(2)	(3)
Panel A: IV Regression			
$\hat{\beta}$	-0.77	-0.60	-0.18
Robust	$[-1.10, -0.45]$	$[-0.78, -0.41]$	$[-0.47, 0.12]$
Cluster	$[-1.12, -0.42]$	$[-0.79, -0.40]$	$[-0.45, 0.10]$
AKM	$[-1.25, -0.30]$	$[-0.84, -0.35]$	$[-0.54, 0.18]$
AKM0	$[-1.69, -0.39]$	$[-1.01, -0.36]$	$[-0.84, 0.14]$
Panel B: Reduced-Form Regression			
$\hat{\beta}$	-0.49	-0.38	-0.11
Robust	$[-0.71, -0.27]$	$[-0.48, -0.28]$	$[-0.31, 0.08]$
Cluster	$[-0.64, -0.34]$	$[-0.45, -0.30]$	$[-0.27, 0.05]$
AKM	$[-0.81, -0.17]$	$[-0.52, -0.23]$	$[-0.35, 0.12]$
AKM0	$[-1.24, -0.24]$	$[-0.67, -0.25]$	$[-0.64, 0.08]$

- Estimate $1/\phi$ in

$$\hat{\omega}_i = \frac{1}{\phi}\hat{L}_i + Z_i\delta + \epsilon_i$$

  Z_i: same vector of controls as in ADH. $N = 1,444$ (722 CZs $\times$ 2 time periods);

- To instrument for $\hat{L}_i$, use
  1. National employment growth, as in Bartik (1991) (actually, a leave-one-out version of it); or
  2. Increase Chinese imports to high-income countries excl. US, as in ADH
- Two approaches give similar point estimates, but while AKM CIs broadly similar to usual ones with Bartik IV, 20% and 250% wider for AKM and AKM0 with ADH IV (AKM and AKMo methods cluster at 3-digit SIC level).
  - National employment growth absorbs most sector-level shocks, not much shift-share structure left in residual

	First-Stage	Reduced-Form	2SLS
	(1)	(2)	(3)
Panel A: Bartik IV—Leave-one-out estimator			
$\hat{\beta}$	0.87	0.71	0.82
Robust	[0.68, 1.06]	[0.53, 0.89]	[0.65, 0.98]
Cluster	[0.62, 1.12]	[0.46, 0.96]	[0.60, 1.03]
AKM (*leave-one-out*)	[0.59, 1.15]	[0.47, 0.94]	[0.61, 1.02]
AKM0 (*leave-one-out*)	[0.53, 1.15]	[0.42, 0.94]	[0.59, 1.09]
Panel B: ADH IV			
$\hat{\beta}$	−0.72	−0.48	0.67
Robust	[−1.04, −0.39]	[−0.80, −0.16]	[0.36, 0.98]
Cluster	[−0.93, −0.50]	[−0.78, −0.18]	[0.35, 0.99]
AKM	[−1.19, −0.24]	[−0.88, −0.07]	[0.27, 1.07]
AKM0	[−1.83, −0.35]	[−1.27, −0.10]	[0.18, 1.14]

I am not clear on the econometrics in the following settings:

- Some papers use IV regressions where the instrument is an interaction term, $X_{it} = X_t W_i$. See, for instance Nunn and Qian (2014) (interact US wheat production with propensity to receive aid from US as instrument for US food aid), or Kearney and Levine (2015) (instrument to "16 and Pregnant" popularity in the area with predicted popularity: W_i MTV rating; X_t: indicator for show running)
  - Some discussion: Jaeger, Joyce, and Kaestner (2020) with reply, Christian and Barrett (2024), Kahn-Lang and Lang (2020).
- In some settings, the number of sectors is as small as $S = 2$ (Nakamura and Steinsson 2014). How to do inference? Limited progress made in https://arxiv.org/abs/1905.13660.
- Number of structural papers use shift-share instruments (e.g. Diamond 2016), many with non-linear structure. Some progress thinking through these in Borusyak and Hull (2023), but many open questions (inference, identification with controls etc).

REFERENCES I

Adão, Rodrigo, Michal Kolesár, and Eduardo Morales. 2019. "Shift-Share Designs: Theory and Inference." *The Quarterly Journal of Economics* 134, no. 4 (November): 1949–2010. https://doi.org/10.1093/qje/qjz025.

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20, no. 1 (March): 46–63. https://doi.org/10.1214/aoms/1177730090.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review* 103, no. 6 (October): 2121–2168. https://doi.org/10.1257/aer.103.6.2121.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/9780585223940.

Blanchard, Olivier J., and Lawrence F. Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 1992 (1): 1–75. https://doi.org/10.2307/2534556.

Borusyak, Kirill, and Peter Hull. 2023. "Nonrandom Exposure to Exogenous Shocks." *Econometrica* 91, no. 6 (November): 2155–2185. https://doi.org/10.3982/ECTA19367.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-Experimental Shift-Share Research Designs." *The Review of Economic Studies* 89, no. 1 (January): 181–213. https://doi.org/10.1093/restud/rdab030.

REFERENCES II

Card, David. 2001. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration." *Journal of Labor Economics* 19, no. 1 (January): 22–64. https://doi.org/10.1086/209979.

Christian, Paul, and Christopher B Barrett. 2024. "Spurious Regressions and Panel IV Estimation: Revisiting the Causes of Conflict." *The Economic Journal* 134, no. 659 (March): 1069–1099. https://doi.org/10.1093/ej/uead091.

Diamond, Rebecca. 2016. "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000." *American Economic Review* 106, no. 3 (March): 479–524. https://doi.org/10.1257/aer.20131706.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik Instruments: What, When, Why, and How." *American Economic Review* 110, no. 8 (August): 2586–2624. https://doi.org/10.1257/aer.20181047.

Jaeger, David A., Theodore J. Joyce, and Robert Kaestner. 2020. "A Cautionary Tale of Evaluating Identifying Assumptions: Did Reality TV Really Cause a Decline in Teenage Childbearing?" *Journal of Business & Economic Statistics* 38, no. 2 (April): 317–326. https://doi.org/10.1080/07350015.2018.1497510.

Kahn-Lang, Ariella, and Kevin Lang. 2020. "The Promise and Pitfalls of Differences-in-Differences: Reflections on *16 and Pregnant* and Other Applications." *Journal of Business & Economic Statistics* 38, no. 3 (July): 613–620. https://doi.org/10.1080/07350015.2018.1546591.

Kearney, Melissa S., and Phillip B. Levine. 2015. "Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing." *American Economic Review* 105, no. 12 (December): 3597–3632. https://doi.org/10.1257/aer.20140012.

REFERENCES III

Nakamura, Emi, and Jón Steinsson. 2014. "Fiscal Stimulus in a Monetary Union: Evidence from US Regions." *American Economic Review* 104, no. 3 (March): 753–792. https://doi.org/10.1257/aer.104.3.753.

Nunn, Nathan, and Nancy Qian. 2014. "US Food Aid and Civil Conflict." *American Economic Review* 104, no. 6 (June): 1630–1666. https://doi.org/10.1257/aer.104.6.1630.

REGRESSION DISCONTINUITY

Michal Kolesár*

April 11, 2024

1. IDENTIFICATION

We're interested in the effect of a treatment D_i on an outcome Y_i. In a regression discontinuity (RD) design, the treatment is determined, either fully or partially, by the value of some variable X_i, called a running variable, crossing a threshold. Without loss of generality, normalize the threshold to 0. In the sharp RD design,

$$D_i = \mathbb{1}\{X_i \geq 0\}, \tag{1}$$

so that all units with X_i exceeding 0 are treated. For example, D_i may be an indicator for being elected, and X_i may be the margin of victory, as in Lee (2008). In the fuzzy RD, the running variable induces a jump in the treatment probability,

$$\lim_{x \downarrow 0} P(D_i = 1 \mid X_i = x) - \lim_{x \uparrow 0} P(D_i = 1 \mid X_i = x) > 0. \tag{2}$$

For instance, van der Klaauw (2002) is interested in the effect of financial aid on attending college. Since students are put into "financial aid groups", having a numerical score X_i based on the objective part of the application (SAT scores, grades) over some cutoff 0 discontinuously increases the chances of receiving aid. D_i is an indicator for receiving aid. Another nice example comes from Bleemer and Mehta (2022). Here D_i is an indicator for majoring in economics, and Y_i is earnings. The paper exploits the fact that UC Santa Cruz students can't declare an econ major if their GPA in intro econ courses (X_i) is below 2.8.

We observe $\{(Y_i, X_i, D_i)\}_{i=1}^n$. Let us assume this triple is drawn i.i.d. from some well-defined population. For any variable A_i, let $\mu_A(x) = E[A_i \mid X_i = x]$.

*Email: mkolesar@princeton.edu.

1.1. *Sharp RD*

Here the parameter of interest is the discontinuity in the regression function $\mu_Y(x) := E[Y_i \mid X_i = x]$ at the cutoff:

$$\tau_Y = \lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x).$$

The variable X_i may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity of the conditional distribution (or of a feature of this conditional distribution such as the conditional expectation) of the outcome as a function of this covariate at the cutoff value can be interpreted as evidence of a causal effect of the treatment. In particular, assume

Assumption 1 (Continuity). $\mu_{Y(0)}(x) := E[Y_i(0) \mid X_i = x]$ and $\mu_{Y(1)}(x) := E[Y_i(1) \mid X_i = x]$ are both continuous in x at 0.

Although in theory, we only need continuity at $x = 0$, it is rare that such assumption is reasonable without having continuity at all values of x. Indeed, examining continuity of the regression function away from the cutoff is a good way of checking whether Assumption 1 is reasonable in practice, as we discuss in Section 2 below.

Under eq. (1) and Assumption 1, τ_Y identifies the treatment effect for individuals at the cutoff:

$$\begin{aligned}\tau_Y &\overset{(i)}{=} \lim_{x \downarrow 0} E[Y_i(1) \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i(0) \mid X_i = x] \\ &\overset{(ii)}{=} E[Y_i(1) - Y_i(0) \mid X_i = 0].\end{aligned}$$

where (i) follows from eq. (1) and (ii) follows from Assumption 1. See Figure 1.

Remark 1 (Practical implication). In practice, Assumption 1 requires:

1. No perfect manipulation: individuals cannot perfectly manipulate their running variable. Imperfect manipulation is allowed, although it may create problems with estimation and inference, as we'll discuss in Remark 5. So in the Lee (2008) application, it is fine if Mark Harris hires McCready to tamper with votes, so long as McCready can't do it in a way that ensures victory.

2. Nothing else happens at the cutoff except for a change in treatment status. This is a strong assumption in geography-based or spatial RDs. Age-based cutoffs also need to be treated with care: if say the cutoff is retirement age (as in, e.g., Battistin et al. 2009), one needs to be careful when say estimating the effect of retirement on an outcome of interest, as other things may also change one an individual reaches retirement age (they become eligible for Medicare, they downsize and children move out etc.).

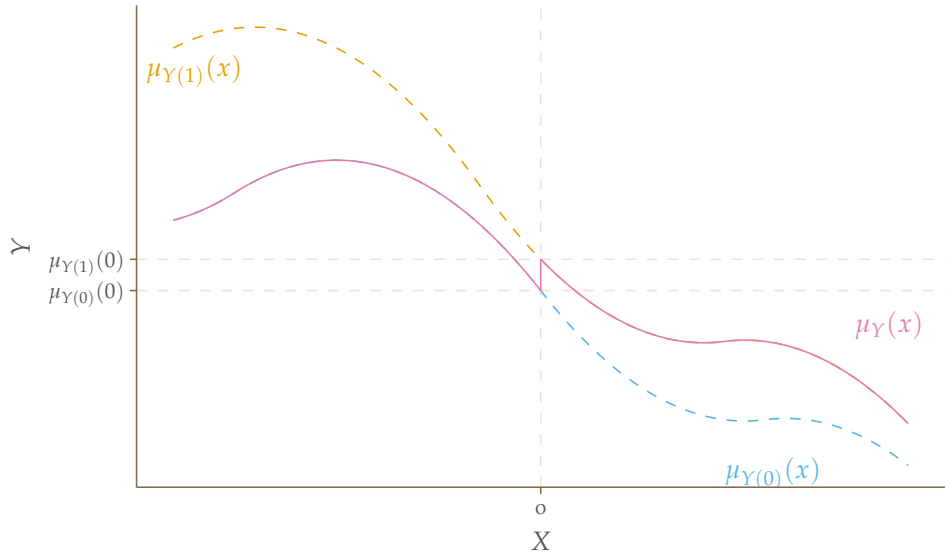


Figure 1: Regression functions for observed outcome (solid) and potential outcomes (dashed). The parameter of interest is the jump in the observed regression function at zero, which corresponds to $\tau_Y = \mu_{Y(1)}(0) - \mu_{Y(0)}(0)$.

1.2. *Fuzzy RD*

In many cases, "treatment eligibility" $\mathbb{1}\{X_i \geq 0\}$ does not perfectly determine the actual treatment: the jump in the treatment probability in eq. (2) is positive, but doesn't go all the way from zero to one. In such fuzzy cases, we scale the jump τ_Y by the size of the jump in treatment probability:

$$\theta = \frac{\lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x)}{\lim_{x \downarrow 0} \mu_D(x) - \lim_{x \uparrow 0} \mu_D(x)} =: \frac{\tau_Y}{\tau_D}.$$

where $\mu_D(x) := P(D_i = 1 \mid X_i = x)$ is the propensity score. To interpret this ratio, let $D_i(1)$ denote the potential treatment status of an individual if we were to make them eligible, and let $D_i(0)$ denote their treatment status if we were to make them ineligible. This may require the cutoff to be in principle manipulable, but that is often the case in administrative settings. The observed treatment corresponds to $D_i = D_i(\mathbb{1}\{X_i \geq 0\})$.

As in Imbens and Angrist (1994), assume a version of monotonicity:

Assumption 2 (Monotonicity). $P(D_i(1) \geq D_i(0) \mid X_i) = 1$

This says that if I were to make the individual eligible (by, say, moving the cutoff), it either has no effect on their treatment, or else induces them to take the treatment; nobody selects out of treatment.

Assumption 3 (Continuity). $\mu_{Y(d)}(x)$, $\mu_{D(d)}(x)$, and $\mu_{D(d)Y(d')}(x)$ are continuous at $x = 0$ for $d, d' \in \{0, 1\}$.

This reduces to Assumption 1 if $D_i(d) = d$.

Under eq. (2) and Assumptions 2 and 3,

$$\theta = E[Y_i(1) - Y_i(0) \mid X_i = 0, D_i(1) > D_i(0)],$$

the local average treatment effect (LATE) for individuals who are at the cutoff, and who are compliers: they select into treatment if I move the cutoff so they clear it, out of treatment if I move the cutoff so they don't clear it. This is a very local treatment effect!

Proof. Letting $\tau_i = Y_i(1) - Y_i(0)$, we have

$$\begin{aligned}
\lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x) &\overset{(1)}{=} \lim_{x \downarrow 0} E[Y_i(0) + D_i(1)\tau_i \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i(0) + D_i(0)\tau_i \mid X_i = x] \\
&\overset{(2)}{=} E[(D_i(1) - D_i(0))\tau_i \mid X_i = 0] \\
&\overset{(3)}{=} E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0), X_i = 0] P(D_i(1) > D_i(0) \mid X_i = 0),
\end{aligned}$$

where (1) follows since $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$, (2) follows from Assumption 3, and (3) follows by Assumption 2. Finally, by Assumption 3, $\lim_{x \downarrow 0} \mu_D(x) - \lim_{x \uparrow 0} \mu_D(x) = P(D(1) = 1 \mid X = 0) - P(D(0) = 1 \mid X = 0)$, which equals $P(D(1) > D(0) \mid X = 0)$ by Assumption 2. Equation (2) ensures that the denominator is non-zero. □

Note that this setup is a bit different from Hahn, Todd, and van der Klaauw (2001), who were the first to give a LATE interpretation to θ. In particular, they define potential treatments $D_i(x)$ as treatment status that would obtain if $X_i = x$, which requires the running variable to be manipulable. That is typically more restrictive than requiring the cutoff to be manipulable. They also assume that $(Y_i(1) - Y_i(0))$ and $D_i(x)$ are jointly independent of X_i near 0, which effectively requires that the running variable is as good as randomly assigned near the cutoff. In contrast, the setup here only requires continuity of the conditional distribution in X_i, not full independence.

Fuzzy RD thus is like a local instrumental variables (IV) model: eligibility $\mathbb{1}\{X_i \geq 0\}$ is an instrument for treatment. Implicit in Assumption 3 is an exclusion restriction that eligibility itself doesn't affect potential outcomes. In fact, as we'll see, estimation is exactly the same as in an IV model. Inference will be different because we'll want to account for functional form bias.

1.3. *Alternative frameworks*

The framework we describe above was based on sampling form a large population, and assuming continuity of potential outcome and treatment distribution at the cutoff. There are two other frameworks that have been proposed based on randomization.

LOCAL RANDOMIZATION FRAMEWORK The first is based on the idea that close enough to the cutoff, the treatment in sharp RD can be thought of "as good as randomly assigned": that is, not only do the potential outcomes change smoothly as a function

of X, they are independent of X in a small neighborhood of the cutoff. The heuristic justification is that if units either have imperfect knowledge of the cutoff or have no ability to precisely manipulate their own score, units with scores close enough to the cutoff will have the same chance of being barely above the cutoff as barely below it.

Cattaneo, Frandsen, and Titiunik (2015) formalize this idea by letting $Y_i(D_i, X_i)$ denote potential outcomes, and assuming that for some small window $\mathcal{W}$ around the cutoff, $Y_i(D_i, X_i)$ doesn't depend on X_i (exclusion restriction), and that $Y_i(D_i) \perp\!\!\!\perp X_i \mid X_i \in \mathcal{W}$ (random assignment). Note that this sort of exclusion restriction is not necessarily satisfied even if we assumed random assignment of treatment near the cutoff. In contrast, our framework in Section 1.1 doesn't require an exclusion restriction on X_i: $Y_i(D_i)$ may depend on X_i, as long as it does so smoothly. We do not require that $\mu_{Y(0)}$ and $\mu_{Y(1)}$ are exactly flat near the cutoff.

Under this setup, one can treat the potential outcomes as fixed, and conduct randomization inference based on repeated sampling of X_i: either by using randomization tests, or by other methods for analyzing randomized experiments that may have an asymptotic justification. The key issue is how to choose the window $\mathcal{W}$: choosing it too small loses a lot of power, and if we choose it larger, the exclusion restriction will be questionable. Cattaneo, Frandsen, and Titiunik (2015) propose a heuristic window selection mechanism based on the idea that in a randomized experiment, the distribution of observed covariates has to be equal between the treated and the controls. However, formalizing this trade-off is harder than the analogous bias-variance trade-off that arises when selecting bandwidths under the framework in Section 1.1.

RANDOM CUTOFF ASSIGNMENT Another approach is to think of the cutoff c as being as good as randomly assigned, drawn from a known distribution P, with the realized cutoff equal to $c = 0$. This approach has been proposed in Ganong and Jäger (2018). We can then use finite-sample randomization tests to test the sharp null that the policy has no effect on the outcomes.

In particular, let $\hat{\tau}_c$ denote the statistic of interest, computed under the assumption that the cutoff is at c. For example, it may correspond to the estimate of a treatment effect. The actual estimate is given by $\hat{\tau}_0$. Now, under the sharp null that the treatment has no effect, the distribution of $\hat{\tau}_0$ is given, under repeated sampling of the cutoff (holding everything else fixed) by the distribution of $\hat{\tau}_C$ with $C \sim P$. Therefore, can reject the null of no effect, if, say $|\tau_0|$ exceeds the 95% quantile of this distribution, which we can simulate if we know P. The key issue is how to pick P. Ganong and Jäger (2018) propose using a uniform distribution within a small window of the cutoff (which as the issue that the realized value of c is always in the middle), a uniform distribution over $\{X_i\}_{i=1}^n$, or using institutional knowledge of how c was determined.

2. FALSIFICATION TESTS

Assumption 1 (or Assumption 3 in fuzzy RD) is not testable directly. We can, however, test the two conditions in Remark 1 that imply it.

MANIPULATION OF THE RUNNING VARIABLE To test for presence of manipulation, we can check the continuity of the density of the running variable, that is, whether $\lim_{x\downarrow 0} f(x) = \lim_{x\uparrow 0} f(x)$. One often accompanies a formal test with a plot of the density on either side of the cutoff. While continuity of the density is, strictly speaking, neither necessary nor sufficient for manipulation, it makes intuitive sense.

The original test, developed by McCrary (2008) used local polynomial estimators to estimate the density to the right and to the left of the cutoff, based on pre-binned data. Cattaneo, Jansson, and Ma (2020) modify the test to reduce the number of tuning parameters: instead of pre-binning the density, they propose running a local polynomial regression of the empirical cumulative distribution function (CDF) onto X_i. The difference in estimated slopes at the cutoff is then the estimate of the jump in the density. Otsu, Xu, and Matsushita (2013) develop a test using a local likelihood approach.

An alternative approach, studied in Bugni and Canay (2021), is to formalize the idea that under no manipulation, the running variable should be effectively randomized close to the cutoff. In particular, Bugni and Canay (2021) suggest picking q values of the running variable closest (in absolute value) to the cutoff, and count the number of positive values. Under the null, the density should be locally constant near the cutoff, so that the number of positive values should be distributed $\text{Binom}(q, 1/2)$. Under Lispchitz smoothness of the density f, if $q = o(n^{2/3})$, the bias from assuming that the density is exactly locally constant will be asymptotically negligible, and the test will control size.

Example 1. In one of the first applications of fuzzy RD in economics, Angrist and Lavy (1999) exploit the fact that following the advice of a rabbinic scholar Maimonides, class sizes in Israel are capped at 40 students, so that a grade cohort with 41 students is supposed to be split into 2 classes, while a cohort with 39 students remains in one large class. In a follow-up analysis using a more recent sample (2002–2011 vs 1991 in the original paper), Angrist et al. (2019) document clear enrollment manipulation at the Maimonides cutoffs—see Figure 2. This leads them to use predicted enrollment, rather than actual enrollment, as a running variable in their analysis. The original sample also displays similar, albeit slightly less striking evidence of sorting, as reported by Otsu, Xu, and Matsushita (2013). ⊠

COVARIATE BALANCE CHECKS This idea goes back to Lee (2008): the treatment should have no effect on pre-determined covariates. Therefore, if one estimates the effect of the treatment on some pre-determined variable Z_i (using the same estimator and confidence interval as used for estimating the effect of the treatment on the outcome of interest Y_i), the estimated treatment effect should not be significantly different from

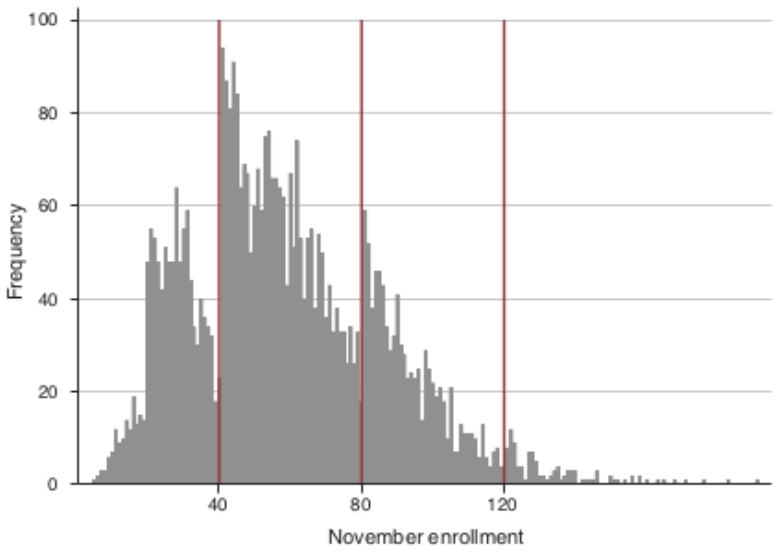


Figure 2: Figure 1 from Angrist et al. (2019). The fifth-grade enrollment distribution, as reported by school headmasters in November. Red reference lines indicate Maimonides' rule cutoffs at which an additional class is added.

zero. This test is an analog of a covariance balance check in randomized experiments.

Again, one could alternatively use the local randomization approach. In particular, Canay and Kamat (2018) propose a permutation test based on the q closest observations to the cutoff. Order the covariates W_i according to the running variable, obtaining $S = (W^-_{(q)}, \ldots, W^-_{(1)}, W^+_{(1)}, \ldots, W^+_{(q)})$. Compare their empirical CDFs $\hat{F}^+(w) = \frac{1}{q}\sum_j \mathbb{1}\{W^+_{(q)} \leq w\}$ using the Cramér-von Mises test statistic

$$T(S) = \frac{1}{2q}\sum_{j=1}^{2q}\left[\hat{F}^-(S_j) - \hat{F}^+(S_j)\right]^2,$$

Compute the critical value using a permutation test by permuting the elements of S. Note that the rule of thumb for picking q proposed in the published version of the paper is not correct, one needs to pick q that grows more slowly to control the bias. Note also that although the intuition behind this test is based on a local randomization framework, the justification is asymptotic, and under the usual framework.

DISCONTINUITY AWAY FROM CUTOFF Similar to testing for discontinuity in the density, it's also a good idea to at least visually inspect that we don't observe jumps in μ_Y away from the cutoff.

3. ESTIMATION AND INFERENCE

Let us focus on sharp RD for concreteness. Statistically, the problem of estimating τ_Y just amounts to estimating a conditional mean at 0 separately for the treated and untreated subpopulations, and taking a difference.

The key issue is that since 0 is a boundary point in both regression problems, there is an unavoidable need for extrapolation, because by design there are no units with $X_i = 0$ for which we observe $Y_i(0)$, and also because typically there are too few units that are very close to the cutoff.

This is why using parametric methods, such as specifying that $\mu_{Y(1)}$ and $\mu_{Y(0)}$ are exactly polynomial of order p, or using global nonparametric methods, such as using series estimators (where, if the basis is polynomial, the only difference is that p grows with sample size), is unattractive. Global estimators in general, and global polynomial estimators in particular, may place large weight on observations far away from the cutoff in constructing the estimate $\hat{\mu}_{Y(1)}(0)$ ($\hat{\mu}_{Y(0)}(0)$), which corresponds to the intercept in the regression of Y_i on powers of X_i for the treated (untreated) units. That is, polynomial estimators can be written as

$$\hat{\tau}_Y = \sum_i w(X_i) Y_i, \tag{3}$$

where the weight $w(X_i)$ on Y_i depends on the value of the running variable X_i. The weights sum to one for observations above the cutoff, and sum to minus one for observations below the cutoff. Some of these weights are negative, unless the order of the polynomial is 0. Furthermore, the average magnitude of the weight tends to increase with the order of the polynomial, so that if p is large, some observations will receive very large weights. As a result, in such cases, small amounts of misspecification (true regression function that's not exactly polynomial, small measurement error in the outcomes) may generate large biases in the intercept estimates. This leads to estimators with large mean squared error, and confidence intervals with poor coverage. See Gelman and Imbens (2019) for a thorough discussion of this point.

A better alternative is to use local methods, which by design only place non-zero weight on observations that are near the cutoff. This can be seen as the key distinction between "parametric" and "nonparametric" thinking: while in "parametric" models, we don't worry about extrapolation bias, in "nonparametric" models, we both (i) take into account the potential extrapolation bias when choosing between different estimators; we don't just minimize variance, and (ii) we should try to account for the potential bias when conducting inference.

3.1. Standard approach to estimation

We now discuss the standard approach to estimation, based on local polynomial regression. In local polynomial regression, one picks a bandwidth h and a polynomial order

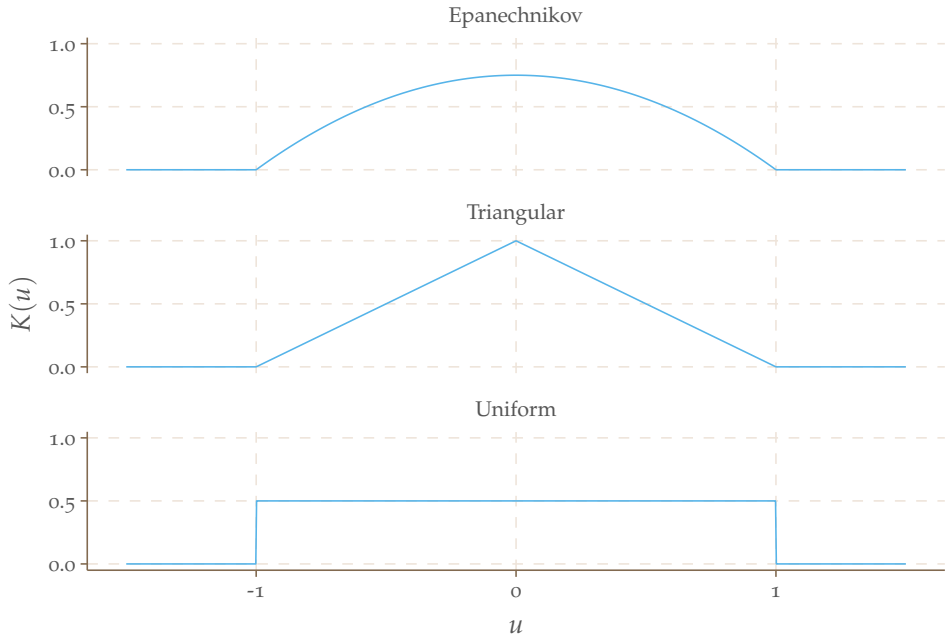


Figure 3: Uniform, Triangular, and Epanechnikov kernels

p. All observations with distance further away from the cutoff than h are discarded, and the estimates $\hat{\mu}_{Y(1)}(0)$ and $\hat{\mu}_{Y(0)}(0)$ correspond to the intercept in the regression of Y_i on powers of X_i for the treated (untreated) units, using only observations within distance h of the cutoff. If we set $h = \infty$, we get back to global polynomial regression. More generally, one may want to downweight observations that are relatively further away from the cutoff, putting weight $K(x/h)$ on observation i if $X_i = x$. Here K is a kernel function. If we use the uniform kernel $K(u) = \mathbb{1}\{|u| \leq 1\}$, then we get back to the unweighted case, placing equal weight on observations within h of the cutoff, and 0 weight on observations farther away than h. Other popular choices of kernel include the triangular kernel $K(u) = (1 - |u|)_+$, or the Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)_+$. See Figure 3.

The intercept estimate is then obtained by a weighted least squares regression of Y onto powers of X:

$$\hat{\mu}_{Y(1)}(0) = e_1' \left(\sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) Y_i$$

where $m(X) = (1, x, \ldots, x^p)$, and $e_1 = (1, 0, \ldots, 0)'$ is the first unit vector. With $\hat{\mu}_{Y(0)}(0)$ defined analogously, the estimate is given by

$$\hat{\tau}_{Y,h} = \hat{\mu}_{Y(1)}(0) - \hat{\mu}_{Y(0)}(0).$$

We can compute $\hat{\tau}_{Y,h}$ all in one step as the coefficient on $D_i = 1\{X_i \geq 0\}$ in a weighted

regression of Y_i onto $D_i, m(X_i)$, and the interaction $D_i m(X_i)_{-1} = D_i(X_i, \ldots, X_i^p)$, with weights $K(X_i/h)$.

To implement this method, one needs to pick K, p, and h. The first two are relatively easy:

1. The order of the polynomial depends on the amount of smoothness we assume. In particular, suppose that $\mu_{Y(1)}$ and $\mu_{Y(0)}$ both belong to the Hölder class of order $p+1$, that is, they are (almost everywhere) $p+1$ times differentiable, with a bounded $(p+1)$th derivative. Then it is optimal to use a polynomial of order p. In practice, $p = 1$ or $p = 2$ are typically used.

2. The kernel choice typically matters less. The triangular and Epanechnikov kernel are both slightly more efficient choices than the uniform (see Cheng, Fan, and Marron (1997) and Armstrong and Kolesár (2020) for formal results).

BANDWIDTH SELECTION The choice of bandwidth, in contrast, is more complicated, and more consequential. The key tradeoff is between bias and variance: while larger h leads to a lower variance of the estimate, since we're using more data, it also leads to a larger bias: unless the true regression function is exactly polynomial of order p inside the estimation window, putting weight on observations away from the cutoff will bias the estimate.

Observe that the local polynomial estimate has the form of a linear estimator, as in eq. (3), with the weights given by

$$\begin{aligned} w(x;h) = e_1' \left(\sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \mathbb{1}\{x \geq 0\} K(x/h) m(x) \\ - e_1' \left(\sum_i \mathbb{1}\{X_i < 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \mathbb{1}\{x < 0\} K(x/h) m(x) \end{aligned}$$

The finite-sample conditional bias is therefore given by

$$\begin{aligned} \text{bias}(\hat{\tau}_{Y,h}) &= \sum_i E[w(X_i;h) Y_i \mid X] - (\mu_{Y(1)}(0) - \mu_{Y(0)}(0)) \\ &= \sum_i \mathbb{1}\{X_i \geq 0\} w(X_i)(\mu_{Y(1)}(X_i) - \mu_{Y(0)}(0)) \\ &\qquad + \sum_i \mathbb{1}\{X_i < 0\} w(X_i)(\mu_{Y(1)}(X_i) - \mu_{Y(0)}(0)), \end{aligned}$$

where the second equality follows since for the local polynomial estimator, $\sum_i \mathbb{1}\{X_i \geq 0\} w(X_i) = -\sum_i \mathbb{1}\{X_i < 0\} w(X_i) = 1$. The variance is given by

$$\text{var}(\hat{\tau}_{Y,h} \mid X) = \sum_i w(X_i;h)^2 \sigma^2(X_i),$$

where $\sigma^2(x) = \text{var}(Y_i \mid X_i = x)$. One could get a consistent estimate of the variance

(discussed below) that's consistent uniformly over all bandwidth choices considered. Therefore, if one could also get an estimate of the bias, one could estimate the mean squared error (MSE) for each bandwidth choice and pick the bandwidth that minimizes it. Estimating the bias directly is difficult, however, so the classic approach is to use a Taylor approximation to the bias. In particular, if we Taylor-expand $\mu_{Y(d)}(x)$ around 0, then as $h \to 0$ and $n \to \infty$ (see Theorem 3.2 in Fan and Gijbels 1996):

$$\begin{aligned}\text{bias}(\hat{\tau}_{Y,h}) &= \left[C_B(p,K)\mu_{Y(1)}^{(p+1)}(0)h^{p+1} - C_B(p,K)\mu_{Y(0)}^{(p+1)}(0)h^{p+1}\right](1+o(1)),\\ &= C_B(p,K)h^{p+1}\left[\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0)\right](1+o(1)),\end{aligned}$$

where $C_B(p,K)$ is a constant that depends only on the order of the polynomial and the kernel. One could similarly approximate the variance as $nh \to \infty$ (again, see Theorem 3.2 in Fan and Gijbels 1996):

$$\begin{aligned}\text{var}(\hat{\tau}_{Y,h} \mid X) &= \left[C_V(p,K)\frac{\sigma^2(0_+)}{f_X(0_+)nh} + C_V(p,K)\frac{\sigma^2(0_+)}{f_X(0_+)nh}\right](1+o(1))\\ &= C_V(p,K)\left[\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)nh}\right](1+o(1)),\end{aligned}$$

where $C_V(p,K)$ is a constant that depends only on the order of the polynomial and the kernel, $\sigma^2(0_+) = \lim_{x \downarrow 0}\text{var}(Y_i \mid X_i = x)$, $\sigma^2(0_-)$ is defined similarly, $f_X(0_+)$ and $f_X(0_-)$ are the densities at 0 of the running variable for the treated and untreated, respectively, and the second line assumes that there is no jump in the density $f_X(x)$ of the running variable at 0, so that $f_X(0_+) = f_X(0_-) = 2f_X(0)$.

Then the asymptotic approximation to the MSE is given by

$$\text{AMSE}(h) = C_B(p,K)^2 h^{2(p+1)}(\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0))^2 + C_V(p,K)\left[\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)nh}\right],$$

which we can minimize analytically over h to yield the (pointwise) optimal bandwidth

$$h_{\text{PT}}^* = \left(\frac{C_V(p,K)}{2(p+1)C_B(p,K)^2}\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)(\mu_{Y(1)}^{(p+1)}(0) - \mu_{Y(0)}^{(p+1)}(0))^2 \cdot n}\right)^{\frac{1}{2p+3}}. \tag{4}$$

Of course, this bandwidth is not feasible, because we do not know the variances $\sigma^2(0_+)$, $\sigma^2(0_-)$, the derivatives $\mu_{Y(1)}^{(p+1)}(0)$, $\mu_{Y(0)}^{(p+1)}(0)$, or the density $f_X(0)$. Imbens and Kalyanaraman (2012) propose a feasible version of this bandwidth based on plugging in estimates of these unknown quantities.

Notice that, so long as $\mu_{Y(1)}^{(p+1)}(0) \neq \mu_{Y(0)}^{(p+1)}(0)$, the optimal bandwidth shrinks at the rate $O(n^{-\frac{1}{2p+3}})$. This is optimal rate: it ensures that the squared bias is of the same order $O(n^{-\frac{2p+2}{2p+3}})$ as the variance. As long as we choose the bandwidth to be of this order, the resulting convergence rate of $\hat{\tau}_Y$ will be $O_p(n^{-\frac{p+1}{2p+3}})$. So, for example, if $p = 1$ (local linear regression), the estimator converges at the rate $n^{-2/5}$, slower than the parametric rate

$n^{-1/2}$. But one could get closer to the optimal rate by assuming more smoothness: if one assumes a third-order Hölder class, then one can run a local quadratic regression ($p = 2$), with a faster convergence rate equal to $n^{-3/7}$ if the bandwidth is picked optimally.

3.2. *Problems with the standard approach*

There are two issues with this approach. First, its performance can be arbitrarily bad even if we use the infeasible bandwidth choice h^*_{PT}. This is because the Taylor-expansion method effectively assumes that we can approximate $\mu_{Y(d)}$ locally around zero by a polynomial of order $p+1$. This is fine as long as the bandwidth h^*_{PT} we end up choosing is not too large. But if in this approximation the $p+1$th derivative to the right and to the left of the cutoff are similar, so that $\mu^{(p+1)}_{Y(1)}(0) \approx \mu^{(p+1)}_{Y(0)}(0)$, the implied optimal bandwidth choice h^*_{PT} will be large, at which point the local polynomial approximation may become very misleading. In this case, the Taylor approximation effectively decides that close to zero, the bias of $\hat{\mu}_{Y(1)}(0)$ is similar to that of $\hat{\mu}_{Y(0)}(0)$, so that the biases cancel out, implying that we should use a large bandwidth. But while such conclusion may be accurate for bandwidths close zero, it may be quite inaccurate for large bandwidths.

To illustrate this point, consider the following example from Armstrong and Kolesár (2020). Suppose that $\mu_{Y(1)}(x) = -\mu_{Y(0)}(x) = x^{p+2}$ if p is even; if p is odd, suppose $\mu_{Y(1)}(x) = -\mu_{Y(0)}(x) = x^{p+3}$. Then the $(p+1)$th derivative of both $\mu_{Y(0)}$ and $\mu_{Y(1)}$ at zero is zero, implying that the optimal bandwidth is infinite. The resulting estimator is therefore a global pth order polynomial least squares estimator. Its mean squared error will be large, since this estimator is not even consistent.[1]

To address this problem, plug-in bandwidths such as the Imbens and Kalyanaraman (2012) bandwidth selector that estimate h^*_{PT} include tuning parameters to prevent the bandwidth from getting too large. However, the MSE of the resulting estimator at such functions is then determined almost entirely by these tuning parameters.

The second problem with the standard approach is that, in order to implement the plug-in method, one needs to estimate the $(p+1)$th order derivatives $\mu^{(p+1)}_{Y(1)}(0)$ and $\mu^{(p+1)}_{Y(0)}$. This is a harder problem than the original problem of estimating the intercepts $\mu_{Y(1)}(0)$ and $\mu_{Y(0)}(0)$. Formally, this shows up in the smoothness requirement on $\mu_d(x)$: we need these functions to be in the Hölder class of order $p+2$ or higher. But if we are willing to assume this higher order of smoothness, it is no longer optimal to use local polynomial regression of order p: we should be using local polynomial regression of order $p+1$! So the resulting estimator is optimal in the class of estimators (local polynomial estimators of order p), that is itself suboptimal.

1. To ensure consistency and finiteness of h^*_{PT}, one therefore needs to assume that $\mu^{(p+1)}_{Y(1)}(0) \neq \mu^{(p+1)}_{Y(0)}(0)$. However, the MSE at h^*_{PT} can still be arbitrarily poor whenever $\mu^{(p+1)}_{Y(1)}(x)$, and $\mu^{(p+1)}_{Y(0)}(x)$ are similar near zero, but not so globally.

To prevent these issues, one can instead adapt a minimax approach: choose the bandwidth to minimize the *worst-case* mean squared error of $\hat{\tau}_Y$: this is the estimation approach proposed in Armstrong and Kolesár (2018). That is, minimize

$$\begin{aligned}
&\sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} (\operatorname{bias}(\hat{\tau}_{Y,h})^2 + \operatorname{var}(\hat{\tau}_{Y,h})) \\
&\quad = \operatorname{var}(\hat{\tau}_{Y,h}) + \sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} \operatorname{bias}(\hat{\tau}_{Y,h})^2 = \sum_i w(X_i;h)\sigma^2(X_i) + \\
&\qquad \sup_{\mu_{Y(1)},\mu_{Y(0)}\in\mathcal{F}_{\mathrm{H},p+1}(M)} \Bigg[\sum_{i:\, X_i\geq 0} w(X_i;h)(\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0)) \\
&\qquad\qquad + \sum_{i:\, X_i<0} w(X_i;h)(\mu_{Y(0)}(X_i) - \mu_{Y(0)}(0)) \Bigg]^2,
\end{aligned}$$

where the equality follows since the variance of the estimator doesn't depend on μ, and $\mathcal{F}_{\mathrm{H},p+1}(M)$ is the Hölder class, the class of $p+1$ times differentiable functions, with the $(p+1)$th derivative bounded by a constant M (we'll discuss the choice of the curvature parameter M below). While the sup in the above display is an infinite-dimensional optimization problem, it turns out that one can solve it in closed form: see Armstrong and Kolesár (2020, Theorem B.3).

Example 2 ($p = 0$). For local constant regression ($p = 0$), the bias-maximizing function takes the form Mx. This is easy to see: the weights are simply given by $w(X_i;h) = K(X_i/h)$ for $X_i \geq 0$; since the weights are positive, we need to make $\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0)$ as large as possible. Now, $\mu_{Y(1)} \in \mathcal{F}_{\mathrm{H},0}(M)$ if and only if $|\mu_{Y(1)}(x) - \mu_{Y(1)}(x')| \leq M|x - x'|$. Therefore, $\mu_{Y(1)}(X_i) - \mu_{Y(1)}(0) \leq MX_i$. The equality is sharp for all X_i if and only if $\mu_{Y(1)}(x) = a + Mx$, for an arbitrary intercept a, showing that $\mu_{Y(1)}(x) = Mx$ is indeed least favorable. The proof for $\mu_{Y(0)}(x)$ is similar. ⊠

If $p = 1$ (local linear regression), the bias-maximizing function takes the form $\mu_{Y(1)}(x) = -Mx^2/2$ and $\mu_{Y(0)}(x) = Mx^2/2$. So for $p = 1$, the worst-case MSE is given by

$$\sum_i w(X_i;h)\sigma^2(X_i) + B_{M,h}^2, \qquad B_{M,h} = -\frac{M}{2}\sum_{i=1}^n w(X_i;h)X_i^2 \operatorname{sign}(X_i). \tag{5}$$

We denote the minimizer by h^*_{MSE}. Computing it is not feasible, since we do not know the variance function σ^2. In practice, one can assume homoskedastic errors (in analogy to ordinary least squares (OLS)), and use a preliminary variance estimator $\hat{\sigma}^2$ to obtain a feasible version of this bandwidth.

- In contrast with the classic approach of minimizing the asymptotic approximation to the MSE, this approach doesn't rely on any asymptotic approximation (apart from the variance estimation), and doesn't require any regularization to prevent

the bandwidth from getting too large.

- The approach doesn't require any assumptions on the distribution of X_i: in particular, nothing changes if the distribution of the running variable is discrete, as is often the case in practice.

To compare this resulting bandwidth to that based on the classic approach, it is useful to consider a large-sample version of the worst-case MSE criterion. If the density of the running variable f is well-behaved (which rules out discrete running variables), and $\sigma^2(X_i)$ is continuous, then (see Equation (19) in Armstrong and Kolesár 2020)

$$h^*_{\text{MSE}} = \left(\frac{C_V(p,K)}{2(p+1)\tilde{C}_B(p,K)^2} \cdot \frac{\sigma^2_+(0) + \sigma^2_-(0)}{2f_X(0) \cdot 4M^2 \cdot n} \right)^{\frac{1}{2p+3}} (1 + o_p(1)),$$

where $\tilde{C}_B(p,K)$ is a kernel constant that's slightly larger than the constant $C_B(p,K)$ in eq. (4). Comparing this expression with eq. (4), the key difference is in the term $4M^2$: rather than plugging in an estimate of the difference between the $(p+1)$th derivatives of $\mu_{Y(1)}$ and $\mu_{Y(0)}$ at 0, the bandwidth uses the priori worst-case difference, equal to $2M$ under $\mathcal{F}_{\text{H},p}(M)$. This ensures good performance of the resulting estimator simultaneously for all possible conditional means $\mu_{Y(1)}$ and $\mu_{Y(0)}$ in the parameter space $\mathcal{F}_{\text{H},p}(M)$ (i.e. uniformly over $\mathcal{F}_{\text{H},p}(M)$). Thus, unlike h^*_{PT}, this bandwidth doesn't yield poor performance in cases where the $(p+1)$th derivatives of $\mu_{Y(0)}$ and $\mu_{Y(1)}$ are similar locally to 0, but not so globally.

CHOICE OF M The key question is how to pick the curvature parameter M, as the optimal bandwidth depends on this choice. The issue is that if we pick M too conservatively, in the sense that, say, in the $p = 1$ case, the second derivatives of the functions $\mu_{Y(1)}$ and $\mu_{Y(0)}$ are much smaller than M over the support of X_i, the resulting bandwidth will be too conservative: it would be nice to be able to estimate the bound on the second derivative from the data, and use this estimate $\hat{M}$.

If one wants to conduct inference based on the estimator $\hat{\tau}_{Y,h^*_{\text{MSE}}}$, it turns out that such a data-driven rule for choosing M (or any data-driven rule) is not possible without further restrictions (see Armstrong and Kolesár 2018).

Research Question. It remains an open question how one could construct a data-driven rule if one was only interested in estimation. ⊠

This result is essentially an instance of the general issue with using pre-testing or using model selection rules, such as using cross-validation or information criteria like AIC or BIC to pick which controls to include in a regression: doing so leads to distorted confidence intervals. Here the curvature parameter M indexes the size of the model: a large M is the analog of saying that all available covariates need to be included in the model to purge omitted variables bias; a small M is the analog of saying that a small subset of them will do. Just like one needs to use institutional knowledge of

the problem at hand to decide which covariates to include in a regression, ideally one uses problem-specific knowledge to select M. Analogous to reporting results based on different subsets of controls in columns of a table with regression results, one can vary the choice of M by way of sensitivity analysis.

Depending on the problem at hand, it may be difficult to translate problem-specific intuition about how close we think the regression function is to a linear function into a statement about the curvature parameter M. In such cases, it is convenient to have a rule of thumb for selecting M using the data. Armstrong and Kolesár (2020) suggest the following rule of thumb for calibrating M, based on formalizing the heuristic that the local smoothness of μ_d is no smaller than its smoothness at large scales:

- Fit a global polynomial on either side of the cutoff, and calculate the largest second derivative of the fitted polynomial. Set M to this value.

Armstrong and Kolesár (2020) show that if the second derivative of $\mu_{Y(1)}$ and $\mu_{Y(0)}$ near zero is indeed bounded by the largest second derivative of a global polynomial approximation to $\mu_{Y(0)}$ and $\mu_{Y(1)}$, the rule of thumb will indeed consistently estimate an upper bound on the true M (other reasonable rules of thumb have been proposed—see, for instance, Imbens and Wager (2019)).

Research Question. The additional restriction justifying the above rule of thumb is a little hard to interpret—is it possible to come up with a better data-driven rule for the choice of M, based on a more interpretable restriction on μ? ⊠

Since these methods for choosing M are meant just as a rule of thumb for start of the analysis, it is a good idea to consider alternative choices by way of sensitivity analysis. Also, plotting an approximation of μ that imposes this value of M (by, say, fitting a spline inside the estimation window) can help visually assess whether this choice is reasonable.

3.4. *Inference*

Since $\hat{\tau}_Y$ is a weighted average of the outcomes (in the sense of eq. (3)), by the Lindeberg central limit theorem, the t-statistic will satisfy

$$\frac{\hat{\tau}_{Y,h} - \tau_Y}{\operatorname{var}(\hat{\tau}_{Y,h})^{1/2}} \approx \mathcal{N}\left(\frac{\operatorname{bias}(\hat{\tau}_{Y,h})}{\operatorname{var}(\hat{\tau}_{Y,h})^{1/2}}, 1\right) + o_p(1), \tag{6}$$

so long as the weights $w(X_i)$ are not too large for any given observation.

The key issue in inference is that if the bandwidth h was chosen to optimally trade of bias against variance, the bias-standard deviation ratio $b = \operatorname{bias}(\hat{\tau}_{Y,h}) / \operatorname{var}(\hat{\tau}_{Y,h})^{1/2}$ will not be approximately zero. There are two standard approaches to handle this issue.

The first is undersmoothing: calculate the optimal bandwidth (either the bandwidth h^*_{PT} with regularization to prevent the bandwidth from getting too large, or else the bandwidth h^*_{MSE}). Then use a bandwidth that is smaller than this optimal bandwidth in

the sense that it shrinks to zero at a faster rate. Of course, the issue in finite samples is how to define "smaller": one can in practice justify any bandwidth choice.

The second is bias correction: try to estimate the bias of $\hat{\tau}_{Y,h}$ and subtract it off. For this to be feasible, one again needs to assume more smoothness than was initially assumed to make the local polynomial estimator of order p optimal. Furthermore, even if one had enough smoothness, the bias estimate is often noisy, and the coverage of the resulting confidence intervals is often poor (Hall 1992). To ameliorate this issue, Calonico, Cattaneo, and Titiunik (2014) propose adjusting the variance estimator to take into account the variability of the bias estimate, which they call robust bias correction (RBC). In an important special case, when the pilot bandwidth used to estimate the bias is the same as the main bandwidth h for the local linear estimator that estimates τ_Y, this procedure amounts to running a local *quadratic* regression, but with the original bandwidth h (that was picked as to be optimal for local linear regression). As a result, one can think of this procedure as a particular way of implementing undersmoothing, with a more principled stance on the amount of undersmoothing.

Both of these approaches have the disadvantage that the resulting confidence intervals (CIs) will not be optimal under the smoothness assumptions originally used to justify the choice of the initial estimator $\hat{\tau}_{Y,h}$. Armstrong and Kolesár (2018, 2020) suggest an alternative approach that doesn't have this problem based on bounding the bias in eq. (6). In particular, although we do not know the ratio of the bias to the standard deviation, we can bound it by the ratio of the worst-case bias over $\mathcal{F}_{\mathrm{H},p}(M)$ to the standard deviation—we already calculated this worst-case bias in eq. (5): it is given by $B_{M,h}$, so that $b \leq B_{M,h}/\operatorname{var}(\hat{\tau}_{Y,h})^{1/2} =: \bar{B}$.

Thus, the t-statistic is asymptotically $\mathcal{N}(b,1)$, with $|b| \leq \bar{B}$. Since the quantiles of the absolute value $\mathcal{N}(b,1)^2$ are increasing in b, we can use the 95% percent quantile of the $|\mathcal{N}(\bar{B},1)|$ distribution as our critical value, which leads to the CI

$$\hat{\tau}_{Y,h} \pm \mathrm{cv}_\alpha(\bar{B}) \operatorname{var}(\hat{\tau}_{Y,h})^{1/2},$$

where $\mathrm{cv}_\alpha(b)$ is the α quantile of the $|\mathcal{N}(b,1)|$ distribution (equivalently, the square root of the $1-\alpha$ quantile of a χ^2 distribution with 1 degree of freedom, and non-centrality parameter b^2, which is readily available in statistical software). This CI is honest in the sense that its validity doesn't rely on undersmoothing, or any other asymptotic promises about how the bandwidth would shrink with the sample size, and it is valid uniformly over the whole parameter space $\mathcal{F}_{\mathrm{H},p}(M)$. It is also *bias-aware* in the sense that its length reflects the potential finite-sample bias of the estimator.

Remark 2 (Variance estimation). Since the estimator $\hat{\tau}_{Y,h}$ is just a weighted least squares estimator, one can use the Eicker-Huber-White (EHW) asymptotic variance estimator to estimate $\operatorname{var}(\tau_{Y,h})$: $\hat{V}_{EHW,h} = \sum_i w(X_i)^2(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the fitted value based on the local polynomial regression. This variance estimate is conservative in finite samples, for the same reasons that the EHW estimator is conservative for inference on the conditional best linear predictor (as discussed in the OLS lecture). Alternatively, one could

use a nearest-neighbor variance estimator (Abadie and Imbens 2006; Abadie, Imbens, and Zheng 2014), replacing $(Y_i - \hat{Y}_i)^2$ with $\frac{J}{J+1}(Y_i - J^{-1}\sum_{j=1}^{J} Y_{j(i)})^2$, where $j(i)$ is the jth closest observation to Y_i (in terms of the distance $|X_{j(i)} - X_i|$) on the same side of the cutoff as i. Here J is a fixed small number, such as $J = 3$.

Remark 3 (Discrete running variable). This bias-aware CIs also have the advantage that they allow the running variable to be discrete, which is formally ruled out in the undersmoothing and RBC approaches. In contrast, the other popular proposal for handling discrete covariates, to cluster the errors by the running variable (Lee and Card 2008), has a serious deficiency: it may lead to confidence intervals that are *shorter* than unclustered CIs. See Kolesár and Rothe (2018) for a detailed discussion of this point.

Of course, if the *sampling design* is clustered, then it is appropriate to use clustered standard errors to estimate $\text{var}(\hat{\tau}_{Y,h})$, as discussed the lecture on OLS.

Remark 4 (Leverage). The main condition to deliver the asymptotic normality result in eq. (6) is that the maximal partial leverage of the estimator goes to zero,

$$\max_i w(X_i;h)^2 / \sum_{j=1}^{n} w(X_j;h) \to 0.$$

Since our estimator is just a weighted least squares estimator, this is just a weighted least squares version of the partial leverage condition we encountered when discussing asymptotic normality in the OLS lecture. You can verify that the uniform kernel is used, then this condition is equivalent to $\max_i \max_i H_{\ddot{D},ii} \to 0$, where $\ddot{D}_i$ is a residual from projecting $D_i = \{X_i \geq 0\}$ onto $m(X_i)$ and $D_i(X_i, \ldots, X_i^p)$ for observations inside the estimation window. Computing this leverage as a routine diagnostic is a good idea. If the leverage is too high (say bigger than 0.1), one can use a bigger bandwidth to ensure that the central limit theorem (CLT) approximation is accurate. This will yield greater bias of the estimator, but any bias will be reflected in the CI through a larger critical value.

Remark 5 (Imperfect manipulation). Though imperfect manipulation doesn't mess up identification, it will impact inference because it will typically lead to much larger curvature M of the conditional mean function. Furthermore, manipulation will typically result in the regression function to be less smooth at the cutoff than away from the cutoff, creating potential problems with the rule of thumb calibration of M, and the issues discussed with standard inference will be particularly salient.

3.5. *Fuzzy RD*

In fuzzy RD, we can estimate τ_Y just like in the sharp case—this is our estimate of the reduced form effect of eligibility on the outcome. We can estimate the first-stage effect of eligibility on treatment in the same way, replacing the outcome Y_i with D_i, yielding

the estimator $\hat{\tau}_{D,h}$. Their ratio will then yield an estimate of θ, $\hat{\theta}_h = \hat{\tau}_{Y,h}/\hat{\psi}_{D,h}$. This is equivalent to a weighted IV regression, using weights $K(X_i/h)$, with $\mathbb{1}\{X_i \geq 0\}$ as an instrument for D_i, and using $m(X_i)$ and the interaction $D_i m(X_i)_{-1} = D_i(X_i, \ldots, X_i^p)$ as covariates.

The variance of the estimator can error can be computed by the delta method, yielding the IV variance formula

$$\operatorname{var}(\hat{\theta}_h) = \frac{\operatorname{var}(\tau_{Y,h}) + \theta^2 \operatorname{var}(\tau_{D,h}) - 2\operatorname{cov}(\tau_{D,h}, \tau_{Y,h})\theta}{\tau_D^2},$$

where $\operatorname{cov}(\tau_{D,h}, \tau_{Y,h}) = \sum_i w(X_i, h)^2 \operatorname{cov}(Y_i, D_i \mid X_i)$ is the covariance of the estimators.

To construct confidence intervals, we also need the worst-case bias of the estimator. For this we need to use a linearization argument (Armstrong and Kolesár 2020, Section 3.2.2), yielding the bias bound $B_{M,h}$ as in the sharp case, except now $M = (M_Y + |\theta| M_D)/|\tau_D|$, which depends on θ itself, the second derivative bound M_Y on the curvature in the reduced form regression, and the second derivative bound M_D on the curvature in the first stage regression. Therefore, optimal bandwidth calculations will require a preliminary estimate of $|\theta|$. See Noack and Rothe for Anderson-Rubin style inference that doesn't require linearization for validity.

4. EMPIRICAL ILLUSTRATION

We use the dataset from Lee (2008). The dataset contains 6,558 observations on elections to the US House of Representatives between 1946 and 1998. The running variable $X_i \in [-100, 100]$ is the Democratic margin of victory (in percentages) in election i. The outcome variable $Y_i \in [0, 100]$ is the Democratic vote share (in percentages) in the next election. Given the inherent uncertainty in final vote counts, the party that wins is essentially randomized in elections that are decided by a narrow margin, so that the RD parameter τ_Y measures the incumbency advantage for Democrats for elections decided by a narrow margin—the impact of being the current incumbent party in a congressional district on the vote share in the next election. Figure 4 plots the averages of the raw data.

For estimation, we use $p = 1$ (local linear regression), and the triangular kernel. To determine the bandwidth, we use the Armstrong and Kolesár (2020) rule of thumb, which yields $M = 0.14$, which is driven by observatinons with $X \leq -50$ (You can see from Figure 4 that there is a lot of curvature when $X \leq -05$). If (somewhat arbitrarily) we restrict attention to the 4,900 observations within distance 50 of the cutoff, we obtain $M = 0.04$. Table 1 shows the estimation results. Note the fairly small value of the bandwidth, in spite of the low value of the second derivative, M, selected. In contrast, the Imbens and Kalyanaraman (2012) bandwidth selector picks $h = 29.4$, due to small estimates of the second derivatives near the cutoff.

- Graphical analysis is both very useful and popular, and potentially misleading.

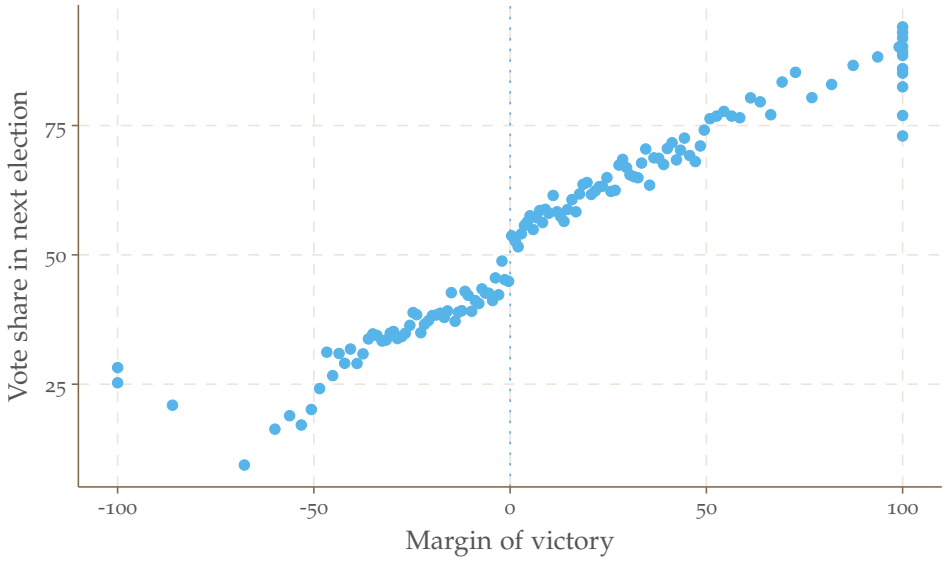


Figure 4: Lee (2008) RD example. Points correspond to 20-observation averages.

M	Estimate	Bias	SE	95% bias-aware CI	Effective obs.	h	$\bar{L}$
0.14	5.85	0.89	1.37	$(2.69, 9.01)$	764	7.7	0.01
0.04	6.24	0.71	1.12	$(3.66, 8.81)$	1250	12.8	0.01

Table 1: Lee (2008) RD example: estimation results. Bias refers to the worst-case bias under the assumed value of M. h refers to the estimate of the optimal bandwidth that minimizes the worst-case MSE given in eq. (5). $\bar{L}$ refers to maximum leverage.

Useful for detecting outliers or potential issues, and as a sanity check.

5. EXTENSIONS

5.1. *Covariate adjustments*

In practice, we commonly have available a vector of L pre-treatment covariates W_i. We can use such covariates in a balance check as discussed in Section 2, but are there any advantages to incorporating them directly into the local polynomial regression? In standard linear regression there are two reasons to include covariates: (i) we think that it helps to make the assumption that the treatment is as good as randomly assigned more plausible, or (ii) including covariates may help precision if they soak up variation in the regression residual. What about in RD?

Under the standard RD framework, where identification is achieved via Assumption 1, including covariates is hard to justify on "identification" grounds. That is, if we think that $E[Y_i(d) \mid X_i, W_i = w]$ is continuous in X_i, then, by iterated expectations, $\mu_{Y(d)}(x) = E[Y_i(d) \mid X_i = x] = \int E[Y_i(d) \mid X_i = x, W_i = w] f(w \mid x) dw$ will be continuous so long as the conditional density f of W_i is continuous. So, in contrast to worries about omitted variable bias in linear regression, the reason to include covariates in RD cannot be that we are worried about failure of Assumption 1, and think that it only holds conditional on covariates. In fact, since conditional expectations "smooth" non-linearities, we expect $\mu_{Y(d)}$ to be smoother than if we also condition on the covariates.

This leaves us with precision as the main reason to include covariates (actually, perhaps one could also argue that covariates help reduce the bias of the estimator). One option, explored in Frölich and Huber (2019) is to estimate the conditional treatment effects $E[Y_i(1) \mid X_i = 0, W_i] - E[Y_i(0) \mid X_i = 0, W_i]$, and then average them using the conditional distribution of the covariates given $X_i = 0$. Such strategy may be hard to implement, however, unless the covariate dimension L is very low, since it involves running kernel regression with $L + 1$ right-hand side variables. A simpler approach is to add the covariates linearly, regressing Y_i onto $m(X_i)$ and W_i for observations within an estimation window. This leads to the estimator

$$\tilde{\tau}_{Y,h} = \tilde{\beta}_{Y,h,1}, \quad \tilde{\beta}_{Y,h} = \left(\sum_{i=1}^{n} K(X_i/h) \tilde{m}(X_i, W_i) \tilde{m}(X_i, W_i)' \right)^{-1} \sum_{i=1}^{n} K(X_i/h) \tilde{m}(X_i, W_i) Y_i, \quad (7)$$

where $\tilde{m}(x, w) = (I\{x \geq 0\}, I\{x \geq 0\}x, 1, x, w)'$. Denote the coefficient on W_i in this regression by $\tilde{\gamma}_{Y,h}$; this corresponds to the last L elements of $\tilde{\beta}_{Y,h}$. As in the case without covariates, we first take the bandwidth h as given, and defer bandwidth selection choice to the end of this subsection.

Remark 6 (Interactions). It is tempting to interact the covariates with the treatment, in analogy to how covariates are included when estimating the average treatment effect

(ATE) under unconfoundedness. That is, we regress Y_i onto an intercept, $m(X_i)$, $(1 - D_i)W_i$, and $D_i W_i$ for observations within an estimation window. As the window shrinks to zero, this is equivalent to the difference in intercepts when projecting Y_i onto a constant and W_i for units just above vs just below the cutoff. By standard regression results, the intercepts in these regressions are given by $E[Y_i(1) \mid X_i = 0] - \mu_W(0)\gamma_{I,+}$ and $E[Y_i(0) \mid X_i = 0] - \mu_W(0)\gamma_{I,-}$, where $\gamma_{I,+} = E[Y_i(1)(W_i - \mu_W(0)) \mid X_i = 0] / \operatorname{var}(W_i \mid X_i = 0)$ and $\gamma_{I,-} = E[Y_i(1)(W_i - \mu_W(0)) \mid X_i = 0] / \operatorname{var}(W_i \mid X_i = 0)$ are projections of the outcome on demeaned covariates. If these projections are different for $Y_i(1)$ and $Y_i(0)$, then this strategy won't yield a consistent estimator of τ_Y, as pointed out in Calonico et al. (2019). One solution to this issue is to demean the covariates, but Calonico et al. (2019) argue that because the demeaning has to be local this has a negative effect on the large-sample variance and convergence rates of the estimator.

To motivate the estimator in eq. (7), we need to formalize the assumption that the covariates are predetermined (without any assumptions on the covariates, it is optimal to ignore the covariates and use the unadjusted estimator $\hat{\tau}_{Y,h}$). Let $\mu_W(x) = E[W_i \mid X_i = x]$ denote the regression function from regressing the covariates on the running variable, and let

$$\Sigma_{WW}(x) = \operatorname{var}(W_i \mid X_i = x), \qquad \Sigma_{WY}(x) = \operatorname{cov}(W_i, Y_i \mid X_i = x).$$

We assume that the variance and covariance functions are continuous, except possibly at zero. Let $\gamma_Y = (\Sigma_{WW}(0_+) + \Sigma_{WW}(0_-))^{-1}(\Sigma_{WY}(0_+) + \Sigma_{WY}(0_-))$ denote the coefficient on W_i when we regress Y_i onto W_i for observations at the cutoff. Let $\tilde{Y}_i := Y_i - W_i'\gamma_Y$ denote the covariate-adjusted outcome. To formalize the assumption that the covariates are pre-determined, we assume that $\tau_W = \lim_{x \downarrow 0} f_W(0) - \lim_{x \uparrow 0} f_W(0) = 0$, which implies that τ_Y can be identified as the jump in the covariate-adjusted outcome $\tilde{Y}_i$ at 0. Following Appendix B.1 in Armstrong and Kolesár (2018), we also assume that the covariate-adjusted outcome varies smoothly with the running variable (except for a possible jump at the cutoff), in that the second derivative of

$$\tilde{\mu}(x) := \mu_Y(x) - \mu_W(x)'\gamma_Y$$

is bounded by a known constant $\tilde{M}$. In addition, we assume μ_W has bounded second derivatives.

Under these assumptions, if γ_Y was known and hence $\tilde{Y}_i$ was directly observable, we could estimate τ as in the case without covariates, replacing M with $\tilde{M}$ and Y_i with $\tilde{Y}_i$. Furthermore, such approach would be optimal under homoskedasticity assumptions. Although γ_Y is unknown, it turns out that the estimator $\tilde{\tau}_{Y,h}$ has the same large sample behavior as the infeasible estimator $\hat{\tau}_{\tilde{Y},h}$.

Proof. To show this, note that by standard regression algebra, $\tilde{\tau}_{Y,h}$ can equivalently be written as

$$\tilde{\tau}_{Y,h} = \hat{\tau}_{Y - W'\tilde{\gamma}_{Y,h},h} = \hat{\tau}_{\tilde{Y},h} - \sum_{k=1}^{K} \hat{\tau}_{W_k,h}(\tilde{\gamma}_{Y,h,k} - \gamma_{Y,k}).$$

The first equality says that covariate-adjusted estimate is the same as an unadjusted estimate that replaces the original outcome Y_i with the covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_{Y,h}$. The second equality uses the decomposition $Y_i - W_i'\tilde{\gamma}_{Y,h} = \tilde{Y}_i - W_i'(\tilde{\gamma}_{Y,h} - \gamma_Y)$ to write the estimator as a sum of the infeasible estimator and a linear combination of placebo RD estimators $\hat{\tau}_{W_k,h}$ that replace Y_i in the outcome equation with the kth element of W_i.

Since μ_W has bounded second derivatives, these placebo estimators converge to zero, with rate that is at least as fast as the rate of convergence of the infeasible estimator $\hat{\tau}_{\tilde{Y},h}$: $\hat{\tau}_{W_k,h} = O_p(B_{\tilde{M},h} + \text{sd}(\hat{\tau}_{\tilde{Y},h}))$. Furthermore, under regularity conditions, $\tilde{\gamma}_{Y,h}$ converges to γ_Y, so that the second term in the previous display is asymptotically negligible relative to the first. □

Consequently, we can form bias-aware CIs based on $\tilde{\tau}_{Y,h}$ as in the case without covariates, treating the covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_Y$ as the outcome,

$$\tilde{\tau}_{Y,h} \pm \text{cv}_{1-\alpha}(B_{\tilde{M},h}/\operatorname{var}(\hat{\tau}_{\tilde{Y},h})^{1/2})\operatorname{var}(\hat{\tau}_{\tilde{Y},h})^{1/2}, \operatorname{var}(\hat{\tau}_{\tilde{Y},h}) = \sum_{i=1}^{n} w(X_i,h)^2\sigma^2_{\tilde{Y}}(x_i),$$

where $\sigma^2_{\tilde{Y}}(x_i) = \sigma^2_Y(x_i) + \gamma_Y'\Sigma_{WW}(x_i)\gamma_Y - 2\gamma_Y'\Sigma_{WY}(x_i)$. If the covariates are effective at explaining variation in the outcomes, then $\sum_i w(X_i,h)^2 \cdot (\gamma_Y'\Sigma_{WW}(x_i)\gamma_Y - 2\gamma_Y'\Sigma_{WY}(x_i))$ will be negative, and $\text{sd}(\hat{\tau}_{\tilde{Y},h}) \leq \text{sd}(\hat{\tau}_{Y,h})$. If the smoothness of the covariate-adjusted conditional mean function $\mu_Y - \mu_W'\gamma_Y$ is greater than the smoothness of the unadjusted conditional mean function μ_Y, so that $\tilde{M} \leq M$, then using the covariates will help tighten the confidence intervals.

Implementation of covariate-adjustment requires a choice of $\tilde{M}$. We can first estimate the model without covariates (using a rule of thumb to calibrate M, the bound on the second derivative of μ_Y), and compute the bandwidth $\check{h}$ that's MSE optimal without covariates. Based on this bandwidth, we compute a preliminary estimate $\tilde{\gamma}_{Y,\check{h}}$ of γ_Y, and use this preliminary estimate to compute a preliminary covariate-adjusted outcome $Y_i - W_i'\tilde{\gamma}_{Y,\check{h}}$. We then calibrate $\tilde{M}$ using the rule of thumb, using this preliminary covariate-adjusted outcome as the outcome.

5.2. *Other extensions*

There are a number of extensions that we don't have time to talk about. Here are some of these extensions with relevant references:

- Sharp and fuzzy regression kink designs: Card et al. (2015).
- Bunching designs: Blomquist et al. (2021) and Bertanha, McCallum, and Seegert (2023)
- Multiple cutoffs. See Bertanha (2020), Cattaneo et al. (2016).
- Cutoffs based on multidimensional running variable.
- Extrapolating treatment effects away from the cutoff. See Angrist and Rokkanen (2015), Dong and Lewbel (2015), Bertanha and Imbens (2020), Cattaneo et al. (2021).

REFERENCES

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74, no. 1 (January): 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

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., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114, no. 2 (May): 533–575. https://doi.org/10.1162/003355399556061.

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." *American Economic Review: Insights* 1, no. 3 (December): 309–324. https://doi.org/10.1257/aeri.20180120.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110, no. 512 (October): 1331–1344. https://doi.org/10.1080/01621459.2015.1012259.

Armstrong, Timothy B., and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86, no. 2 (March): 655–683. https://doi.org/10.3982/ECTA14434.

———. 2020. "Simple and Honest Confidence Intervals in Nonparametric Regression." *Quantitative Economics* 11, no. 1 (January): 1–39. https://doi.org/10.3982/QE1199.

Battistin, Erich, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. 2009. "The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach." *American Economic Review* 99, no. 5 (December): 2209–2226. https://doi.org/10.1257/aer.99.5.2209.

Bertanha, Marinho. 2020. "Regression Discontinuity Design with Many Thresholds." *Journal of Econometrics* 218, no. 1 (September): 216–241. https://doi.org/10.1016/j.jeconom.2019.09.010.

Bertanha, Marinho, and Guido W. Imbens. 2020. "External Validity in Fuzzy Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 38, no. 3 (July): 593–612. https://doi.org/10.1080/07350015.2018.1546590.

Bertanha, Marinho, Andrew H. McCallum, and Nathan Seegert. 2023. "Better Bunching, Nicer Notching." *Journal of Econometrics* 237, no. 2 (December): 105512. https://doi.org/10.1016/j.jeconom.2023.105512.

Bleemer, Zachary, and Aashish Mehta. 2022. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major." *American Economic Journal: Applied Economics* 14, no. 2 (April): 1–22. https://doi.org/10.1257/app.20200447.

Blomquist, Sören, Whitney K. Newey, Anil Kumar, and Che-Yuan Liang. 2021. "On Bunching and Identification of the Taxable Income Elasticity." *Journal of Political Economy* 129, no. 8 (August): 2320–2343. https://doi.org/10.1086/714446.

Bugni, Federico A., and Ivan Alexis Canay. 2021. "Testing Continuity of a Density via G-Order Statistics in the Regression Discontinuity Design." *Journal of Econometrics* 221, no. 1 (March): 138–159. https://doi.org/10.1016/j.jeconom.2020.02.004.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs Using Covariates." *The Review of Economics and Statistics* 101, no. 3 (July): 442–451. https://doi.org/10.1162/rest_a_00760.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82, no. 6 (November): 2295–2326. https://doi.org/10.3982/ECTA11757.

Canay, Ivan Alexis, and Vishal Kamat. 2018. "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design." *The Review of Economic Studies* 85, no. 3 (July): 1577–1608. https://doi.org/10.1093/restud/rdx062.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83, no. 6 (November): 2453–2483. https://doi.org/10.3982/ECTA11224.

Cattaneo, Matias D., Brigham R. Frandsen, and Rocío Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate." *Journal of Causal Inference* 3, no. 1 (January): 1–24. https://doi.org/10.1515/jci-2013-0010.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115, no. 531 (September): 1449–1455. https://doi.org/10.1080/01621459.2019.1635480.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. 2016. "Interpreting Regression Discontinuity Designs with Multiple Cutoffs." *The Journal of Politics* 78, no. 4 (October): 1229–1248. https://doi.org/10.1086/686802.

———. 2021. "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs." *Journal of the American Statistical Association* 116, no. 536 (October): 1941–1952. https://doi.org/10.1080/01621459.2020.1751646.

Cheng, Ming-Yen, Jianqing Fan, and J. S. Marron. 1997. "On Automatic Boundary Corrections." *The Annals of Statistics* 25, no. 4 (August): 1691–1708. https://doi.org/10.1214/aos/1031594737.

Dong, Yingying, and Arthur Lewbel. 2015. "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models." *Review of Economics and Statistics* 97, no. 5 (December): 1081–1092. https://doi.org/10.1162/REST_a_00510.

Fan, Jianqing, and Irène Gijbels. 1996. *Local Polynomial Modelling and Its Applications.* Monographs on Statistics and Applied Probability 66. New York, NY: Chapman & Hall/CRC. https://doi.org/10.1201/9780203748725.

Frölich, Markus, and Martin Huber. 2019. "Including Covariates in the Regression Discontinuity Design." *Journal of Business & Economic Statistics* 37, no. 4 (October): 736–748. https://doi.org/10.1080/07350015.2017.1421544.

Ganong, Peter, and Simon Jäger. 2018. "A Permutation Test for the Regression Kink Design." *Journal of the American Statistical Association* 113, no. 522 (April): 494–504. https://doi.org/10.1080/01621459.2017.1328356.

Gelman, Andrew, and Guido W. Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37, no. 3 (July): 447–456. https://doi.org/10.1080/07350015.2017.1366909.

Hahn, Jinyong, Petra Elisabeth Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69, no. 1 (January): 201–209. https://doi.org/doi.org/10.1111/1468-0262.00183.

Hall, Peter. 1992. "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density." *The Annals of Statistics* 20, no. 2 (June): 675–694. https://doi.org/10.1214/aos/1176348651.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79, no. 3 (July): 933–959. https://doi.org/10.1093/restud/rdr043.

Imbens, Guido W., and Stefan Wager. 2019. "Optimized Regression Discontinuity Designs." *The Review of Economics and Statistics* 101, no. 2 (May): 264–278. https://doi.org/10.1162/rest_a_00793.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review* 108, no. 8 (August): 2277–2304. https://doi.org/10.1257/aer.20160945.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142, no. 2 (February): 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142, no. 2 (February): 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142, no. 2 (February): 698–714. https://doi.org/10.1016/j.jeconom.2007.05.005.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita. 2013. "Estimation and Inference of Discontinuity in Density." *Journal of Business & Economic Statistics* 31, no. 4 (October): 507–524. https://doi.org/10.1080/07350015.2013.818007.

van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43, no. 4 (November): 1249–1287. https://doi.org/10.1111/1468-2354.t01-1-00055.

REGRESSION DISCONTINUITY

Michal Kolesár

ECO539B, Fall 2022

April 11, 2024

SETTING

- Observations $i = 1, \ldots, n$. Interested in effect of some treatment D_i on outcome Y_i.
- Key feature of regression discontinuity (RD) is that treatment is fully or partially determined by whether running variable X_i crosses a threshold.
- Normalizing threshold to zero:

$$D_i = \mathbb{1}\{X_i \geq 0\}, \qquad \text{(sharp RD)}$$

$$\lim_{x \downarrow 0} P(D_i = 1 \mid X_i = x) - \lim_{x \uparrow 0} P(D_i = 1 \mid X_i = x) > 0. \qquad \text{(fuzzy RD)}$$

- Running examples: Lee (2008), Haggag and Paci (2014), van der Klaauw (2002), and Bleemer and Mehta (2022).

Identification

Falsification

Estimation and Inference in sharp RD

Empirical illustration

Extensions

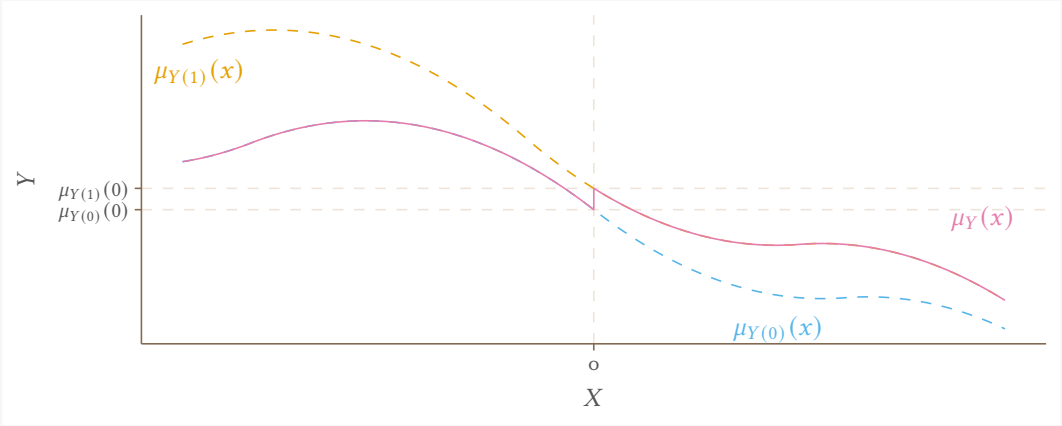
- Parameter of interest is jump in observed conditional mean function $\mu_Y(x) := E[Y_i \mid X_i = x]$ at cutoff:

$$\tau_Y = \lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x).$$

- Key assumption is continuity: $\mu_{Y(0)}(x) := E[Y_i(0) \mid X_i = x]$ and $\mu_{Y(1)}(x) := E[Y_i(1) \mid X_i = x]$ are both continuous in x at 0.
  - Rules out perfect manipulation of X_i (imperfect manipulation OK in theory). Taxi drivers in Haggag and Paci (2014) cannot keep driving until the fare is over \$15. But it is fine if Mark Harris hires McCready to tamper with votes in Lee (2008), since McCready can't do it in a way that ensures Harris' victory.
  - Nothing else happens at the cutoff except for a change in treatment status. Strong assumption in geography-based or spatial RDs. Age-based cutoffs also need to be treated with care. In Battistin et al. (2009), things other than retirement may happen at retirement age cutoff.

Under continuity, τ_Y identifies average treatment effect (ATE) at the cutoff:

$$\tau_Y = \lim_{x \downarrow 0} E[Y_i(1) \mid X_i = x] - \lim_{x \uparrow 0} E[Y_i(0) \mid X_i = x] = E[Y_i(1) - Y_i(0) \mid X_i = 0].$$



- If jump in treatment probability at cutoff is smaller than 1, scale jump τ_Y by the size of the jump in treatment probability:

$$\theta = \frac{\lim_{x \downarrow 0} \mu_Y(x) - \lim_{x \uparrow 0} \mu_Y(x)}{\lim_{x \downarrow 0} \mu_D(x) - \lim_{x \uparrow 0} \mu_D(x)} = \frac{\tau_Y}{\tau_D}$$

- To interpret this, let $D_i(1), D_i(0)$ denote potential treatment if we make individual (in)eligible, by say, making exception to GPA cutoff requirement, or by changing the cutoff. Then $D_i = D_i(\mathbb{1}\{X_i \geq 0\})$
- Need two assumptions:

monotonicity $P(D_i(1) \geq D_i(0) \mid X_i) = 1$. Like in Imbens and Angrist (1994).

continuity $\mu_{Y(d)}(x)$, $\mu_{D(d)}(x)$, and $\mu_{D(d)Y(d')}(x)$ are continuous at $x = 0$ for $d, d' \in \{0, 1\}$. Again, allows for imperfect manipulation, like re-taking intro econ courses to improve GPA in Bleemer and Mehta (2022).

Under monotonicity and continuity, θ identifies local average treatment effect (LATE) at the cutoff

$$\theta = E[Y_i(1) - Y_i(0) \mid X_i = 0, D_i(1) > D_i(0)],$$

- Fuzzy RD is a local instrumental variables (IV) model: eligibility $\mathbb{1}\{X_i \geq 0\}$ is an instrument for D_i. Implicit in continuity assumption is exclusion restriction that eligibility itself doesn't affect potential outcomes.

- Some papers propose a local randomization framework to formalize idea that sharp RD is like a localized random experiment, and fuzzy RD is like a localized experiment with imperfect compliance. But this would require $\mu_{Y(1)}$ and $\mu_{Y(0)}$ to be flat close to cutoff, which is not typically true. Also not clear how to pick right neighborhood.
- Ganong and Jäger (2018) propose randomization approach, thinking of cutoff as random. Allows for simple randomization inference, but would need to specify counterfactual cutoff distribution.
- See notes for more discussion

Identification

Falsification

Estimation and Inference in sharp RD

Empirical illustration

Extensions

Continuity assumption "questionable" if density of running variable not smooth around cutoff.

Formal tests described in the notes, in practice graphical evidence often convincing.

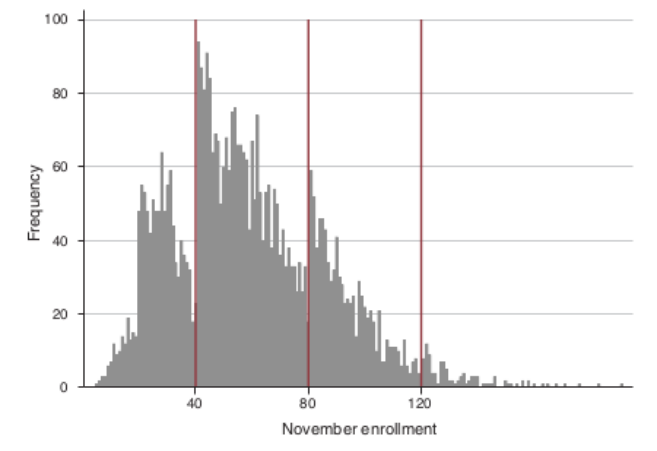


Figure 1 from Angrist et al. (2019). Fifth-grade enrollment distribution, as reported by school headmasters in November. Red reference lines indicate Maimonides' rule cutoffs at which an additional class is added.

- Idea similar to "placebo" tests in other contexts: treatment should have no effect on pre-determined covariates
- Run RD, but with pre-determined covariate W_i as outcome. Can actually test the whole distribution, not just the mean, by comparing distribution of W_i just below and just above cutoff: Canay and Kamat (2018) propose a permutation test based on q closest observations to the cutoff.
  - Order the covariates W_i according to the running variable, obtaining $S = (W^-_{(q)}, \ldots, W^-_{(1)}, W^+_{(1)}, \ldots, W^+_{(q)})$.
  - Compare their empirical cumulative distribution functions (CDFs) $\hat{F}^+(w) = \frac{1}{q}\sum_j \mathbb{1}\{W^+_{(q)} \leq w\}$ using the Cramér-von Mises test statistic

$$T(S) = \frac{1}{2q}\sum_{j=1}^{2q}\left[\hat{F}^-(S_j) - \hat{F}^+(S_j)\right]^2,$$

  - Compute the critical value using a permutation test by permuting the elements of S.

Identification

Falsification

Estimation and Inference in sharp RD

Empirical illustration

Extensions

- Just need to estimate conditional mean at 0 separately for the treated and untreated subpopulations
- Key issue is that 0 is a boundary point in both regression problems, so extrapolation unavoidable
  - Parametric methods, such as specifying that $\mu_{Y(1)}$ and $\mu_{Y(0)}$ are exactly polynomial of order p, or using global nonparametric methods unattractive: observations far away from cutoff receive large weight
- Most estimators, including polynomial estimators can be written

$$\hat{\tau}_Y = \sum_i w(X_i) Y_i, \tag{1}$$

with $\sum_{i:\, X_i \geq 0} w(X_i) = -\sum_{i:\, X_i < 0} w(X_i) = 1$.

- average magnitude $|w(\cdot)|$ tends to increase with the order of polynomial p: if p large, some observations very influential.
- small misspecification can translate into large bias (Gelman and Imbens 2019)
- better to use local polynomial regression with $p = 1$ or 2.

- Local methods (e.g. local linear or local quadratic regression) only place weigh on obs near cutoff
- Key distinction between "parametric" and "nonparametric" thinking: In "parametric" models, we don't worry about extrapolation bias. In "nonparametric" models, we both
    1. take into account the potential extrapolation bias when choosing between different estimators; don't just minimize variance
    2. should try to account for potential bias when conducting inference.
- How to operationalize this?

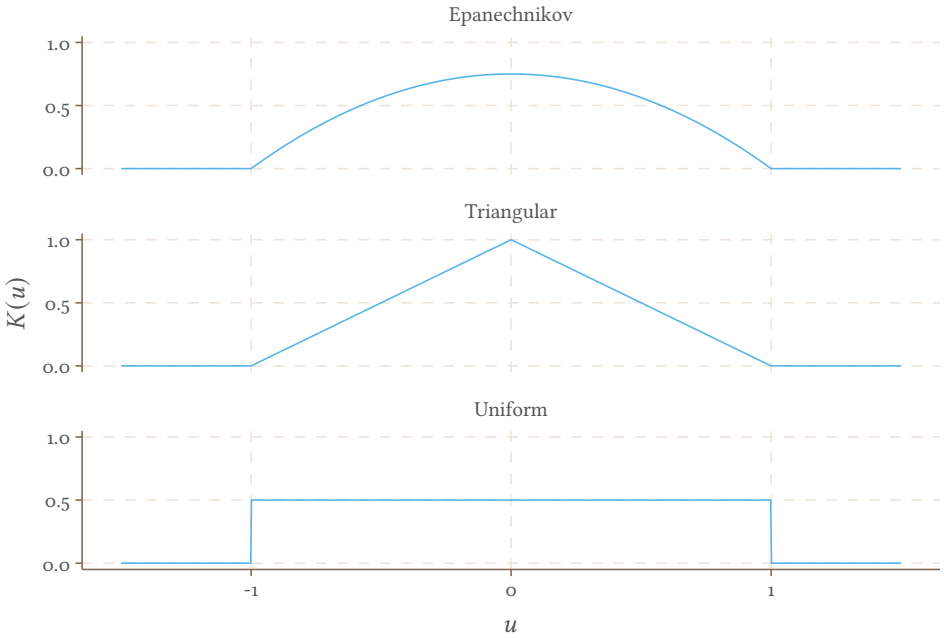
- Pick bandwidth h and polynomial order p. Keep only obs with distance h of cutoff
- Regress Y_i on powers of X_i above an below cutoff, difference in estimates is an estimate of τ_Y
- Can further downweight obs relatively further away from cutoff using kernel weights $K(x_i/h)$. Same as difference between weighted and ordinary least squares (OLS):

$$\hat{\mu}_{Y(1)}(0) = e_1' \left(\sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) m(X_i)' \right)^{-1} \sum_i \mathbb{1}\{X_i \geq 0\} K(X_i/h) m(X_i) Y_i$$

$m(X) = (1, x, \ldots, x^p)$, and $e_1 = (1, 0, \ldots, 0)'$. Then

$$\hat{\tau}_{Y,h} = \hat{\mu}_{Y(1)}(0) - \hat{\mu}_{Y(0)}(0).$$

Can compute in one step by regressing Y_i onto D_i interacted with $m(X_i)$, with weights $K(X_i/h)$.


Epanechnikov
Triangular
Uniform
1.0
0.5
0.0
$K(u)$
-1
0
1
u

- How to pick polynomial order p? Formally depends on the amount of smoothness we assume. $p = 1$ optimal if we assume μ_Y twice differentiable with bounded second derivatives (Hölder class of order 2)
- How to pick K? Doesn't matter much, can use uniform for simplicity. triangular and Epanechnikov slightly more efficient (Cheng, Fan, and Marron 1997; Armstrong and Kolesár 2020).
- How to pick h? More tricky and more consequential

- key tradeoff is between bias and variance:
  1. larger h lowers variance (we use more data)
  2. but also tends to increase bias, unless true regression function exactly polynomial of order p inside estimation window
- Estimator just weighted average of outcomes as in (1), so bias

$$\sum_i w(x_i; h)\mu_Y(x_i) - \tau_Y$$

  and variance as in regression conditional on X (remember OLS lecture!),
  $\text{var}(\hat{\tau}_{Y,h} \mid X) = \sum_i w(X_i; h)^2 \sigma^2(X_i)$.
- Variance estimation easy (doesn't depend on μ_Y), but bias estimation tricky.

- Classic approach: approximate $\mu_{Y(d)}$ locally by Taylor expansion As $h \to 0$ and $n \to \infty$ (see Theorem 3.2 in Fan and Gijbels 1996):

$$\begin{aligned}\text{bias}(\hat{\tau}_{Y,h}) &= \left[C_B(p,K)\mu^{(p+1)}_{Y(1)}(0)h^{p+1} - C_B(p,K)\mu^{(p+1)}_{Y(0)}(0)h^{p+1}\right](1+o(1)),\\ &= C_B(p,K)h^{p+1}\left[\mu^{(p+1)}_{Y(1)}(0) - \mu^{(p+1)}_{Y(0)}(0)\right](1+o(1)),\end{aligned}$$

where $C_B(p,K)$ is a constant that depends only on the order of the polynomial and the kernel. One could similarly approximate the variance as $nh \to \infty$, and hence the mean squared error (MSE), which yields (pointwise) optimal bandwidth

$$h^*_{\text{PT}} = \left(\frac{C_V(p,K)}{2(p+1)C_B(p,K)^2}\frac{\sigma^2(0_+)+\sigma^2(0_-)}{2f_X(0)(\mu^{(p+1)}_{Y(1)}(0)-\mu^{(p+1)}_{Y(0)}(0))^2\cdot n}\right)^{\frac{1}{2p+3}}. \tag{2}$$

- this bandwidth is not feasible, because we do not know the variances $\sigma^2(0_+)$, $\sigma^2(0_-)$, the derivatives $\mu^{(p+1)}_{Y(1)}(0)$, $\mu^{(p+1)}_{Y(0)}(0)$, or the density $f_X(0)$.
- Imbens and Kalyanaraman (2012) propose a feasible version of this bandwidth based on plugging in estimates of these unknown quantities: very popular in practice.
- So long as $\mu^{(p+1)}_{Y(1)}(0) \neq \mu^{(p+1)}_{Y(0)}(0)$, optimal bandwidth shrinks at rate $O(n^{-\frac{1}{2p+3}})$.
  - optimal if we assume $p+1$ derivative.
  - resulting convergence rate of $\hat{\tau}_Y$ is $O_p(n^{-\frac{p+1}{2p+3}})$
  - Can get arbitrarily close to parametric rate by assuming enough derivatives…

1. Arbitrarily bad performance, even if we use infeasible h^*_{PT}.
   - Taylor-expansion method effectively assumes that we can approximate $\mu_{Y(d)}$ locally around zero by a polynomial of order $p+1$.
   - Fine if h^*_{PT} ends up small. But if $\mu^{(p+1)}_{Y(1)}(0) \approx \mu^{(p+1)}_{Y(0)}(0)$ h^*_{PT} large, and Taylor approximation can be very poor
   - Consider local linear regression and $-\mu_{Y(0)}(x) = \mu_{Y(1)}(x) = x^3$. $h^*_{\text{PT}} = \infty$, and we're not even consistent!
   - To address this problem, plug-in bandwidths such as the Imbens and Kalyanaraman (2012) bandwidth selector that estimate h^*_{PT} include tuning parameters to prevent bandwidth from getting too large. But method then driven by tuning parameter choice
2. To implement h^*_{PT}, need to estimate derivatives of order $p+1$:
   - much harder than our initial problem of estimating intercept
   - requires derivatives of order $p+2$ exist. But if that's the case, could have used polynomial of order $p+1$ instead!
   - estimator optimal in class of estimators (local polynomial estimators of order p), that is itself suboptimal.

- Choose bandwidth to minimize worst-case MSE of $\hat{\tau}_Y$ over all possible μ_Y that have second derivatives bounded by M.
- If we use local linear regression, least favorable function has closed form:$\mu_{Y(1)}(x) = -Mx^2/2$ and $\mu_{Y(0)}(x) = Mx^2/2$
  - Intuition?
- Closed-form expression for worst-case MSE: no Taylor approximation!

$$\sum_i w(X_i;h)\sigma^2(X_i) - M\left[\sum_{i:\,X_i\geq 0} w(X_i;h)X_i^2 - \sum_{i:\,X_i<0} w(X_i;h)X_i^2\right]^2. \tag{3}$$

Can minimize numerically to obtain finite-sample optimal bandwidth h^*_{MSE}. In practice, need to estimate variance—can assume homoskedasticity to make that part easy.

- No assumptions on distribution of X_i: in particular, nothing changes if the distribution of the running variable is discrete
- Compare bandwidths in large samples:

$$h^*_{\text{PT}} = \left(\frac{C_V(p,K)}{2(p+1)C_B(p,K)^2} \frac{\sigma^2(0_+) + \sigma^2(0_-)}{2f_X(0)(\mu^{(p+1)}_{Y(1)}(0) - \mu^{(p+1)}_{Y(0)}(0))^2 \cdot n} \right)^{\frac{1}{2p+3}}$$

$$h^*_{\text{MSE}} = \left(\frac{C_V(p,K)}{2(p+1)\tilde{C}_B(p,K)^2} \cdot \frac{\sigma^2_+(0) + \sigma^2_-(0)}{2f_X(0) \cdot 4M^2 \cdot n} \right)^{\frac{1}{2p+3}} (1 + o_p(1)),$$

- To implement, need to figure out second derivative bound, curvature M (global problem) instead of second derivative at zero.

- If M too large, we'll be unnecessarily conservative. Can we use data to estimate it well?
- No possible without further restrictions if goal is inference (Armstrong and Kolesár 2018).
  - Instance of the general issue with using pre-testing or using model selection rules: model selection distorts infernece
  - Here curvature parameter M indexes model size. Large M like saying we more of available covariates possible confounders in OLS, small M like saying we don't need to include them. Without restrictions on OLS coeffs, best we can do is include all of them! Otherwise need to use institutional knowledge
  - Ideally would use institutional knowledge to pick M: hard sell!
  - Analogous to reporting results based on different subsets of controls in columns of a table with regression results, vary choice of M by way of sensitivity analysis.
- Armstrong and Kolesár (2020) suggest rule of thumb for calibrating M, based on heuristic that local smoothness of μ_d is no smaller than its smoothness at large scales:

- Fit a global polynomial on either side of the cutoff, and calculate the largest second derivative of the fitted polynomial. Set M to this value.
- Formally "works" if second derivative of $\mu_{Y(1)}$ and $\mu_{Y(0)}$ near zero indeed bounded by the largest second derivative of a global polynomial approximation to $\mu_{Y(0)}$ and $\mu_{Y(1)}$
- Question whether "better" calibration possible (alternatives have been proposed, e.g. Imbens and Wager (2019), but suffer form same issues)

- Good idea to plot approximation to μ that imposes this rule of thumb, by, say, fitting splines.

- Estimator just weights average of outcomes, so asymptotically normal under minimal assumptions, so long as weights $w(X_i)$ not too large:

$$\frac{\hat{\tau}_{Y,h} - \tau}{\text{var}(\hat{\tau}_{Y,h})^{1/2}} \approx \mathcal{N}\left(\frac{\text{bias}(\hat{\tau}_{Y,h})}{\text{var}(\hat{\tau}_{Y.h})^{1/2}}, 1\right) + o_p(1), \tag{4}$$

- But if h optimally chosen, $b = \text{bias}(\hat{\tau}_{Y,h})/\text{var}(\hat{\tau}_{Y.h})^{1/2}$ not close to zero!

1. Undersmooth: choose h smaller than optimal. But how small? In practice anything goes.
2. Bias-correct: try to estimate bias and subtract it off.
   - Like with h^*_{PT}, can only do this if have more smoothness than optimal for local linear to be optimal.
   - Even if feasible, bias estimate noise, and resulting confidence intervals (CIs) poor (Hall 1992)
   - Calonico, Cattaneo, and Titiunik (2014) propose adjusting variance estimator to take into account the variability of bias estimate, which they call robust bias correction (RBC).
   - Important special case: if bandwidth for bias estimation equals h, this reduces to local quadratic regression (but with original bandwidth, calibrated for local linear)
   - I think of RBC as particular (more principled) way of implementing undersmoothing.

- t-stat asymptotically normal, but don't know mean b. But we have bound on bias—already calculated it to compute optimal bandwidth, so use it to b
- leads to CI

$$\hat{\tau}_{Y,h} \pm \text{cv}_\alpha(\bar{B}) \, \text{var}(\hat{\tau}_{Y,h})^{1/2},$$

  where $\text{cv}_\alpha(b)$ is the α quantile of the $|\mathcal{N}(b,1)|$ and $\bar{B}$ is given by ratio of bias bound to standard error.
- Advantages:
  1. honest: validity doesn't rely on undersmoothing, or any other asymptotic promises about how the bandwidth would shrink with the sample size
  2. valid uniformly over the whole parameter space of all functions μ_Y with second derivative bounded by M
  3. *bias-aware*: length reflects the potential finite-sample bias of the estimator.

1. Can estimate variance using Eicker-Huber-White (EHW). But we're doing inference conditional on X_i, so this is conservative by lecture on OLS. Using nearest-neighbor variance estimator better
   - This is a case where we are misspecified and willing to admit it
2. Bias-aware inference works with discrete running variable. Discrete X_i formally ruled out in the undersmoothing and RBC approaches.
   Another popular proposal for handling discrete covariates: cluster the errors by the running variable (Lee and Card 2008).
   - Has a serious deficiency: it may lead to confidence intervals that are *shorter* than unclustered CIs. See Kolesár and Rothe (2018) for a detailed discussion of this point. Intuition: creates inference with a small number of clusters problem, but doesn't solve bias issue.
3. Since we're just doing OLS, main regularity check is to check leverage not too high

Identification

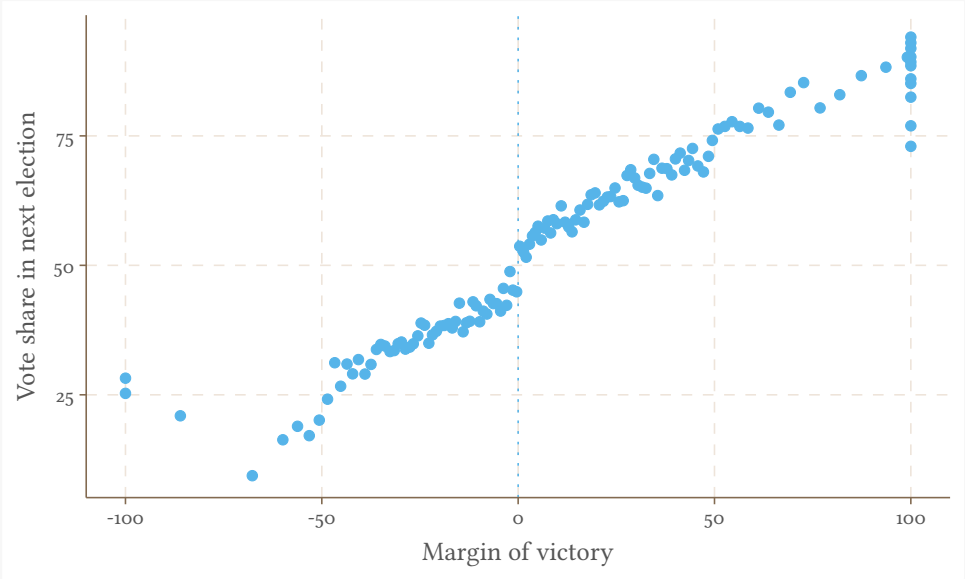
Falsification

Estimation and Inference in sharp RD

Empirical illustration

Extensions

- Use the dataset from Lee (2008) on 6,558 observations on elections to US House of Representatives between 1946 and 1998
- Running variable $X_i \in [-100, 100]$ is the Democratic margin of victory (in percentages) in election i. The outcome variable $y_i \in [0, 100]$ is the Democratic vote share (in percentages) in the next election.


Vote share in next election
75
50
25
-100
-50
0
50
100
Margin of victory

- For estimation, we use $p = 1$ (local linear regression), and the triangular kernel.
- Armstrong and Kolesár (2020) rule of thumb yields $M = 0.14$, which is driven by observatinons with $X \leq -50$ (can see from graph). If (somewhat arbitrarily) restrict attention to the 4,900 observations within distance 50 of the cutoff, we obtain $M = 0.04$.
- For comparison, IK bandwidth is about 30, due to small curvature near cutoff.

M	Estimate	Bias	SE	95% bias-aware CI	Effective obs.	h	$\bar{L}$
0.14	5.85	0.89	1.37	(2.69, 9.01)	764	7.7	0.01
0.04	6.24	0.71	1.12	(3.66, 8.81)	1250	12.8	0.01

Identification

Falsification

Estimation and Inference in sharp RD

Empirical illustration

Extensions

See notes for discussion of first two extensions and references for the other extensions

- Fuzzy RD
- Incorporating covariates
- kink designs
- bunching designs
- multiple cutoffs, or multi-dimensional running variables
- extrapolating treatment effects away from cutoff

REFERENCES I

Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides' Rule Redux." *American Economic Review: Insights* 1, no. 3 (December): 309–324. https://doi.org/10.1257/aeri.20180120.

Armstrong, Timothy B., and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86, no. 2 (March): 655–683. https://doi.org/10.3982/ECTA14434.

———. 2020. "Simple and Honest Confidence Intervals in Nonparametric Regression." *Quantitative Economics* 11, no. 1 (January): 1–39. https://doi.org/10.3982/QE1199.

Battistin, Erich, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. 2009. "The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach." *American Economic Review* 99, no. 5 (December): 2209–2226. https://doi.org/10.1257/aer.99.5.2209.

Bleemer, Zachary, and Aashish Mehta. 2022. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major." *American Economic Journal: Applied Economics* 14, no. 2 (April): 1–22. https://doi.org/10.1257/app.20200447.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82, no. 6 (November): 2295–2326. https://doi.org/10.3982/ECTA11757.

Canay, Ivan Alexis, and Vishal Kamat. 2018. "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design." *The Review of Economic Studies* 85, no. 3 (July): 1577–1608. https://doi.org/10.1093/restud/rdx062.

REFERENCES II

Cheng, Ming-Yen, Jianqing Fan, and J. S. Marron. 1997. "On Automatic Boundary Corrections." *The Annals of Statistics* 25, no. 4 (August): 1691–1708. https://doi.org/10.1214/aos/1031594737.

Fan, Jianqing, and Irène Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability 66. New York, NY: Chapman & Hall/CRC. https://doi.org/10.1201/9780203748725.

Ganong, Peter, and Simon Jäger. 2018. "A Permutation Test for the Regression Kink Design." *Journal of the American Statistical Association* 113, no. 522 (April): 494–504. https://doi.org/10.1080/01621459.2017.1328356.

Gelman, Andrew, and Guido W. Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37, no. 3 (July): 447–456. https://doi.org/10.1080/07350015.2017.1366909.

Haggag, Kareem, and Giovanni Paci. 2014. "Default Tips." *American Economic Journal: Applied Economics* 6, no. 3 (July): 1–19. https://doi.org/10.1257/app.6.3.1.

Hall, Peter. 1992. "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density." *The Annals of Statistics* 20, no. 2 (June): 675–694. https://doi.org/10.1214/aos/1176348651.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

REFERENCES III

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* 79, no. 3 (July): 933–959. https://doi.org/10.1093/restud/rdr043.

Imbens, Guido W., and Stefan Wager. 2019. "Optimized Regression Discontinuity Designs." *The Review of Economics and Statistics* 101, no. 2 (May): 264–278. https://doi.org/10.1162/rest_a_00793.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review* 108, no. 8 (August): 2277–2304. https://doi.org/10.1257/aer.20160945.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142, no. 2 (February): 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142, no. 2 (February): 655–674. https://doi.org/10.1016/j.jeconom.2007.05.003.

van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43, no. 4 (November): 1249–1287. https://doi.org/10.1111/1468-2354.t01-1-00055.

DIFFERENCE IN DIFFERENCES

Michal Kolesár*

April 25, 2024

1. 2 BY 2 DIFFERENCE-IN-DIFFERENCES

Let us first briefly review difference-in-differences (DiD) designs with 2 groups $g \in \{0,1\}$ and 2 time periods $t \in \{0,1\}$. Treatment is a deterministic function of time and group membership. We denote by $D_{gt} = \mathbb{1}\{g = 1, t = 1\}$ the treatment of group g at time t. Let F_{gt} denote the distribution of observed outcomes in group g at time t. Let $Y_{gt} \sim F_{gt}$ denote a random variable with this distribution. We focus on identification, and therefore omit the unit-level subscripts. Let $Y_{gt}(0)$ and $Y_{gt}(1)$ denote potential outcomes, so that $Y_{gt} = Y_{gt}(D_{gt})$. For simplicity, there are no covariates. We may have panel data (we draw samples from the joint distribution of (Y_{g0}, Y_{g1})), or a repeated cross-section (we draw samples from the marginal distribution F_{gt}).

The DiD estimand is given by

$$\beta = E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}].$$

In the panel case, this can be estimated as a regression of the difference $Y_{g1} - Y_{g0}$ onto a constant and $D_{g1} = D_{g1} - D_{g0}$. Since there are only two periods, this is equivalent to running a fixed effects regression with unit and time fixed effects. In the repeated cross-section case, we regress the outcome on the group dummy interacted with a time dummy.

The estimand can be written as the average treatment effect for the treated (ATT) plus a selection bias term coming from differential trends among the treated and control units:

$$\beta = E[Y_{11}(1) - Y_{11}(0)] + \underbrace{E[Y_{11}(0) - Y_{10}(0)] - E[Y_{01}(0) - Y_{00}(0)]}_{\text{Differential trend}}. \tag{1}$$

This follows directly from the definition of the estimand and the fact that $Y_{gt} = Y_{gt}(1)$ if $g = t = 1$, and $Y_{gt} = Y_{gt}(0)$ otherwise.

Therefore, the crucial assumption underlying DiD designs is that the treated and

*Email: mkolesar@princeton.edu.

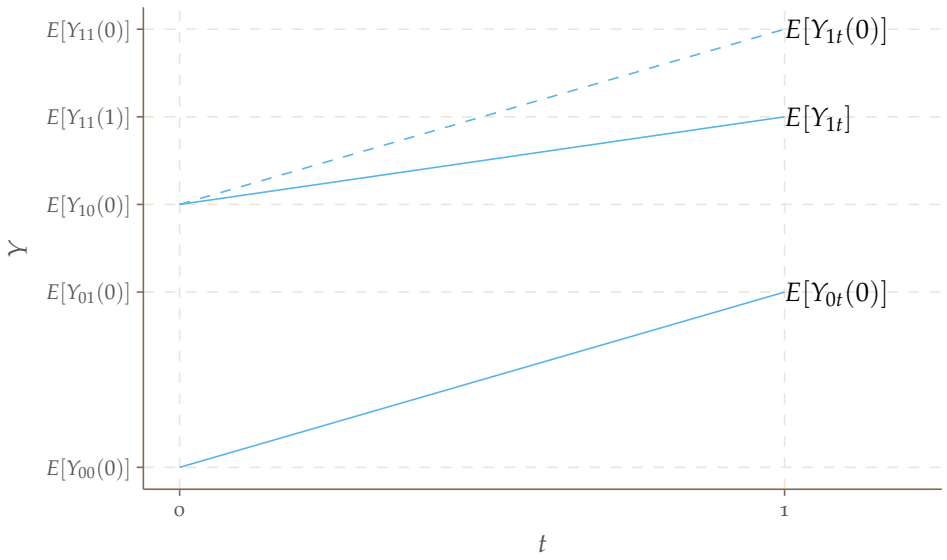


Figure 1: Common trends assumption in DiD designs. The solid lines are observed, the dotted line is imputed by shifting $E[Y_{10}]$ by $E[Y_{01}] - E[Y_{00}]$.

control units have *common trends* (or *parallel trends*):

$$E[Y_{1t}(0) - Y_{10}(0)] = E[Y_{0t}(0) - Y_{00}(0)] \tag{2}$$

for $t = 1$. This allows us to impute the counterfactual mean $E[Y_{11}(0)]$, as illustrated in Figure 1.

Remark 1. For the DiD method to make sense, we only need a notion of a potential outcome $Y_{11}(0)$. It is not necessary to define potential outcomes $Y_{0t}(1)$. In some examples, conceptualizing a treated outcome for the control units can be difficult—but we do not need to do so.

Remark 2 (Ashenfelter's dip). The parallel trends assumption is fragile—it allows selection on levels, but not on differences. It rules out the "dip" observed in Ashenfelter (1978, p. 51): "earnings of trainees [in a labor market program] tend to fall, both absolutely and relative to the comparison group, in the year prior to training". In other words, the treated group experiences a downward "dip" relative to the control group in earnings prior to treatment. In such case DiD methods will overstate the treatment effect if there is reversion to the mean.

Remark 3 (No anticipation). By indexing the potential outcomes in terms of the current treatment status only, we are implicitly making two assumptions, in addition to the Stable unit treatment value assumption (SUTVA) assumption. The first one is *no anticipation*: future treatment status doesn't influence current outcomes (for this reason, perhaps it may make sense to index the treatment by when a law was passed rather than by when

it was implemented). The second one is *no dynamic effects*: past treatment effects don't influence present outcomes.

1.1. *Falsification tests*

There are two falsification tests commonly used in practice to check the common trends assumption: estimate the treatment for groups not affected, and, with multiple pre-treatment periods, estimate pretrends.

EFFECTS ON GROUPS NOT AFFECTED For example, Gruber (1994) uses DiD to evaluate the effect of the passage of mandatory maternity benefits in some US states on the wages of married women of childbearing age. In this context, the common trends assumption implies that in the absence of the intervention, married women of childbearing age would have experienced the same increase in log wages in states that adopted mandated maternity benefits and states that did not. To evaluate this assumption, Gruber (1994) compares the changes in log wages in adopting and non-adopting states for single men and for women over 40 years old.

PRETRENDS TEST Suppose we have data on multiple periods, running from $-T_0 \leq 0$ to $T_1 \geq 1$, but there are still just two groups, with the treatment path given by $D_{gt} = \mathbb{1}\{g = 1, t \geq 0\}$. Write

$$E[Y_{gt}] = \mathbb{1}\{t \neq 0\}\lambda_t + \alpha_g + \mathbb{1}\{g = 1, t \neq 0\}\beta_t, \tag{3}$$

where $\alpha_g = E[Y_{g0}]$, $\lambda_t = E[Y_{0t} - Y_{00}]$, and $\beta_t = E[Y_{1t} - Y_{10}] - E[Y_{0t} - Y_{00}]$. We normalize $\lambda_0 = \beta_0 = 0$ (with two groups, the number of coefficients we can identify is two times the number of time periods, so if we keep the group effects α_g, we need to drop on β_t and one λ_t). Under this normalization, β_t for $t < 0$ measures violations from common trends, since $\beta_t = E[Y_{1t}(0) - Y_{10}(0)] - E[Y_{0t}(0) - Y_{00}(0)]$, which equals zero under eq. (2). For $t \geq 1$, β_t measures the dynamic effect of the treatment.

We can implement eq. (3) by simply running a regression of the outcome on group fixed effects (with repeated cross-sections, with panel data we can use unit fixed effects), time fixed effects with $\mathbb{1}\{t = 0\}$ excluded, and their interaction. We then test the joint significance of the coefficients on the leads to treatment adoption, β_t for $t < 0$, using an F-test. It is also common to show the estimates graphically to visually assess the common trends assumption.

In some cases, we may want to exclude or collapse some of the pre-treatment indicators $\mathbb{1}\{g = 1, t \neq 0\}$. Say with $T_0 = -3$, we may include $\mathbb{1}\{g = 1, t < -1\}$ and $\mathbb{1}\{g = 1, t = 1\}$, or perhaps only include $\mathbb{1}\{g = 1, t = 1\}$. Similarly, if we think that the dynamic effect of the treatment is constant after some time period, we can collapse the post-treatment indicators. Say with $T_1 = 3$, we may include $\mathbb{1}\{g = 1, t = 1\}$ and

Table 1: Table 12 from Snow (1855, p. 90).

	Deaths from Cholera	
Water supply	1849	1854
Southwark & Vauxhall	2261	2458
Lamberth	162	37

Total deaths in sub-districts served by different water companies during two Cholera outbreaks in London, in 1849 and 1854.

$\mathbb{1}\{g = 1, t > 1\}$.

The nice thing about the pretrends test is that it can be done using pre-treatment data alone.

1.2. *Examples*

Example 1. The first DiD design appears as Table 12 in Snow (1855), reproduced in Table 1. Snow challenged the conventional wisdom that cholera spreads by "bad air": he noted that Lamberth changed its water source away from the Thames river in 1852, while Southwark & Vauxhall did not. Unfortunately, there are no standard errors. ☒

Example 2 (Card and Krueger 1994). On April 1, 1992, New Jersey (NJ) raised the state minimum wage from \$4.25 to \$5.05 per hour (the law was passed in 1989). In the meantime, in Pennsylvania (PA) the minimum wage remained constant and equal to \$4.25. To study the effect of the minimum wage increase, Card and Krueger (1994) surveyed 410 fast-food restaurants in NJ and 7 counties in eastern PA (Burger King, Wendy's, KFC, and Roy Rogers), in February 1992, and again in November 1992. Employment contracted in NJ, but so it did in PA, due to an upward trend in unemployment in 1991–1993 in the mid-Atlantic. Table 2 shows the results from regressing various outcomes of interest on a NJ dummy, with the primary outcome being full-time equivalent (FTE) employment. One can tell a coherent story based on these results. However, these results were subsequently questioned in a comment by Neumark and Wascher (2000), who find the opposite effect on employment using administrative payroll data obtained from a sample of 235 restaurants in NJ and PA, drawn from the same geographic areas and the same chains (and hence most overlapping substantially with the Card and Krueger sample). In a response to this comment, Card and Krueger (2000), use Bureau of Labor Statistics data to plot employment in fast-food restaurants in NJ and PA relative to February 1992 levels, reproduced in Figure 2. More pre-intervention data would have been useful, but, nonetheless, the figure is not encouraging. ☒

Example 3 (Meyer, Viscusi, and Durbin 1995). On July 15, 1980, Kentucky (KY) raised the maximum benefit paid by workers' compensation (medical and cash benefits due to

Table 2: Replication of Card and Krueger (1994).

	FTE (1)		Months to raise (2)		Price of entrée (3)	
NJ	2.75	(1.34)	2.51	(2.15)	0.08	(0.04)
Constant	−2.28	(1.25)	1.26	(1.96)	−0.03	(0.04)
Observations	384		321		375	
R^2	0.015		0.005		0.007	

Notes: FTE: Change in FTE employment. Months to raise: Months to first salary raise.

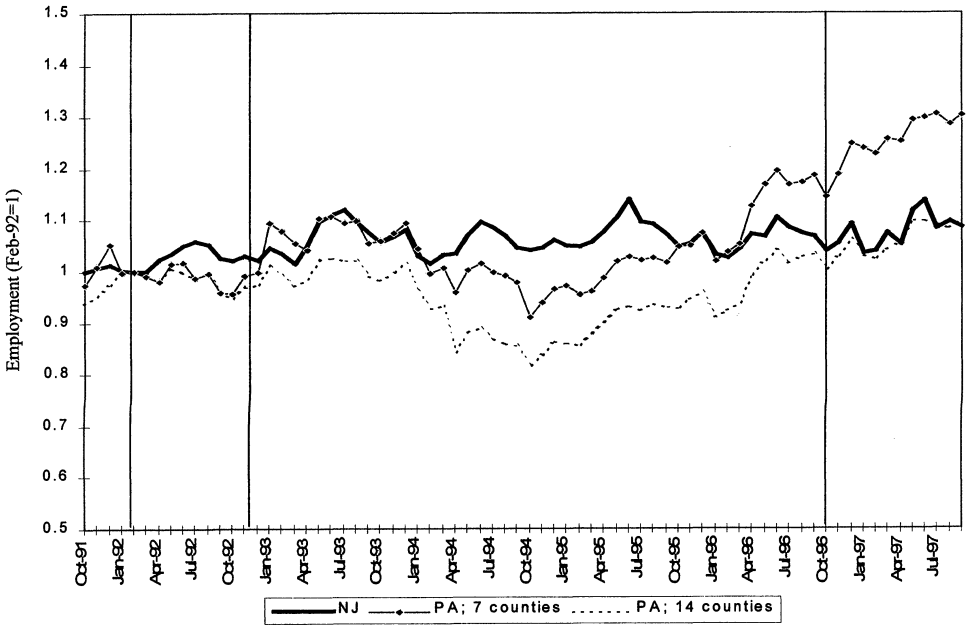


Figure 2: Figure 2 from Card and Krueger (2000). Vertical lines indicate dates of original Card and Krueger survey and another minimum wage increase in October 1996

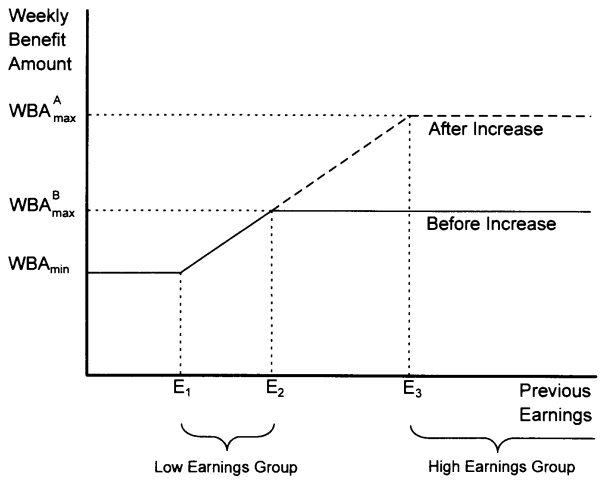


Figure 3: Figure 1 from Meyer, Viscusi, and Durbin (1995). Temporary total benefit schedule before and after an increase in the maximum weekly benefit.

temporary total disability as a result of on-the-job injury) from $131 to $217 per week. A similar change was passed in Michigan (MI) on January 1, 1982, with the maximum benefit going from $181 to $307 per week. Low earners, who were not affected by this change, provide a natural control group for high earners, who were (see Figure 3). Meyer, Viscusi, and Durbin (1995) use a repeated cross-section to look at the effect of this change on the duration of injury claims benefit. The key results are replicated in Table 3. We see that there is a significant effect on log duration in the larger KY sample, but not much action in terms of effect of duration. ⊠

Example 4 (Donohue and Wolfers 2005). Compelling DiD analyses show observations for a period long enough to discern the underlying trends, with attention focused on how deviations from trend relate to changes in policy. A telling graph in this vein is Figure 4, reproduced from Donohue and Wolfers (2005), showing that changes in death penalty have little effect on trends in homicide rates. ⊠

1.3. *Changes-in-changes*

The common trend assumption not invariant to nonlinear transformations of the dependent variable: if it holds in levels, it doesn't hold in logs[1]. This makes it important that we choose the outcome variable carefully.

To address this issue, Athey and Imbens (2006) provide an alternative to the basic DiD model, called the changes-in-changes (CiC) model. Suppose that $Y_{gt}(0) = h_t(U_g)$, with

1. This is not quite right: the common trends is insensitive to functional form assumptions if we can partition the population into a fraction θ for whom the untreated potential outcome depends only on group membership, but not time, and a fraction $1-\theta$ for whom it depends only on time, but not group membership. Arguably, this is a rather special case. See Roth and Sant'Anna (2023).

Table 3: Replication of Meyer, Viscusi, and Durbin (1995).

	Mean duration (weeks)				Mean of log duration			
	KY		MI		KY		MI	
	(1)		(2)		(3)		(4)	
Panel A: DiD								
High × After	0.95	(1.28)	1.96	(3.97)	0.19	(0.07)	0.19	(0.16)
High	4.91	(0.88)	3.82	(2.50)	0.26	(0.05)	0.17	(0.11)
After	0.77	(0.51)	2.69	(1.90)	0.01	(0.04)	0.10	(0.08)
Constant	6.27	(0.30)	10.96	(1.09)	1.13	(0.03)	1.41	(0.06)
Panel B: CiC								
ATT (upper bound)	1.08	(1.61)	2.37	(4.43)	0.58	(0.16)	0.34	(0.16)
ATT (lower bound)	0.07	(1.60)	1.30	(4.48)	0.14	(0.13)	0.02	(0.16)
Observations	5626		1524		5626		1524	

Notes: FTE: Change in FTE employment. Months to raise: Months to first salary raise. CiC refers to the changes-in-changes estimator discussed in Section 1.3.

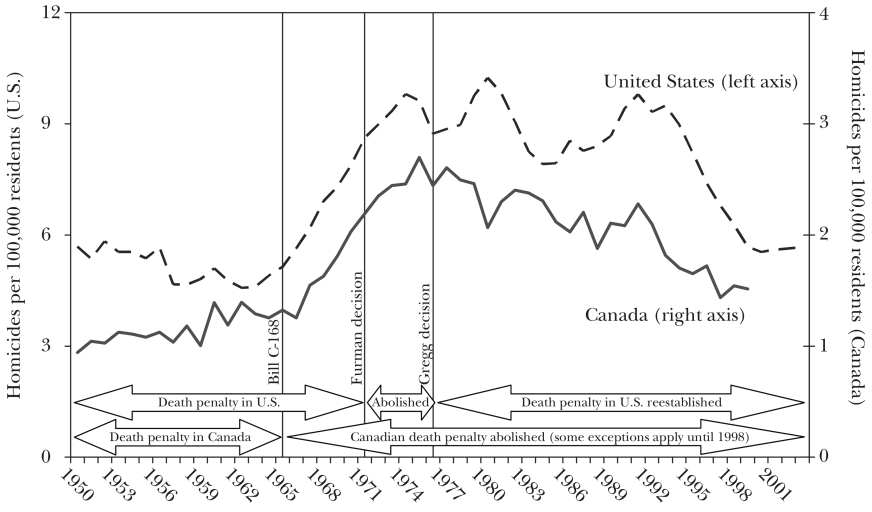


Figure 4: Figure 2 from Donohue and Wolfers (2005). Homicide rates and the death penalty in the United States and Canada.

(i) h strictly increasing in the unobservable u; (ii) the distribution of the unobservable doesn't vary over time within groups $U_g \mid t = 0 \sim U_g \mid t = 1$; an (iii) and overlap condition holds: the support of U_1 is contained in its support given U_0. Together, these assumptions imply that the relative ranking of treatment and control units is stable over time: if control units account for 95% of top earners in period 0, they would also have accounted for 95% of them in period 1 in absence of the intervention (this is similar to the rank invariance assumption that Chernozhukov and Hansen (2005) make in a different context).

The advantage of this setup is that we can now devise an identification strategy that's invariant to transformations of the outcome; the cost is that we're restricting the whole distribution of $Y_{gt}(0)$, not just its mean. The CiC model is useful as a robustness check in cases where the means of the treated and control groups are very different, when we worry about the fragility of the mean restriction imposed by the DiD model.

Theorem 4 (Athey and Imbens 2006). Let F_{gt} denote the cumulative distribution function (CDF) of Y_{gt}. Then, under assumptions (i), (ii), and (iii) above,

$$E[Y_{11}(0)] = E[F_{01}^{-1}(F_{00}(Y_{10}))], \tag{4}$$

Proof. We have

$$F_{0t}(h_t(u)) = P(h_t(U_0) \leq h_t(u)) = P(U_0 \leq u),$$

where the second equality uses the strict monotonicity of u. Thus, $F_{01}(h_1(u)) = F_{00}(h_0(u))$, so that

$$h_1(u) = F_{01}^{-1}(F_{0,0}(h_0(u))), \qquad h_0(u) \in \text{support}(Y_{00}).$$

This is the period 1 outcome for someone with period 0 outcome equal to $y = h_0(u)$. Since the distribution of U_g is time-invariant, this implies (4) as claimed, provided that $\text{support}(Y_{10}) \subseteq \text{support}(Y_{00})$, which holds by assumption (iii). □

We can estimate eq. (4) using empirical distributions and sample averages. In particular, let $\mathbf{Y}_{gt}$ denote the vector of outcomes for group g and time t, and let $\hat{F}_{gt}$ denote its empirical CDF. Define

$$\hat{F}_{01}^{-1}(q) = \min\{y \in \text{support}(\mathbf{Y}_{01}) \colon \hat{F}_{01}(y) \geq q\}.$$

(I am giving an explicit formula here, because this is not the default way of computing the sample quantile function in many software packages). Then we can estimate the ATT as

$$\frac{1}{n_{11}} \sum_{y \in \mathbf{Y}_{11}} y - \frac{1}{n_{10}} \sum_{y \in \mathbf{Y}_{10}} \hat{F}_{01}^{-1}(\hat{F}_{00}(y)), \tag{5}$$

where n_{gt} is the length of the $\mathbf{Y}_{gt}$ vector.

Here the transform $F_{01}^{-1}(F_{00}(y))$ gives the second-period outcome for an individual with an unobserved component u such that $h_0(u) = y$. The logic is as follows. Take an individual from group 1 in period 0 with outcome equal to y. Compute their quantile $q = F_{00}(y)$ if they were in the control group. This quantile wouldn't change in period

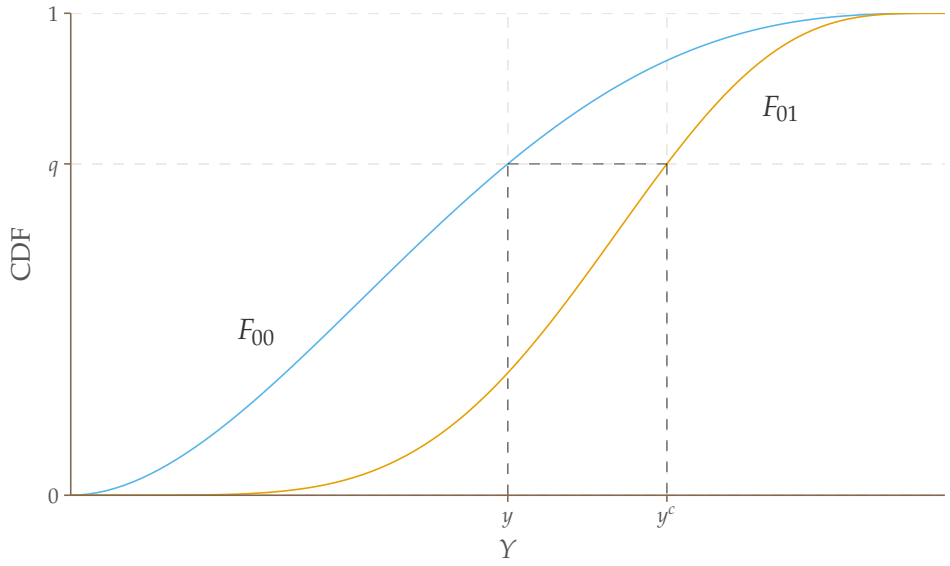


Figure 5: Computation of the counterfactual outcome y^c in period 1 for an individual in group 1 with outcome y in period 0.

1, so their counterfactual outcome in period 1 would need to correspond to this outcome: $y^c = F_{01}^{-1}(F_{00}(y))$. See Figure 5. Since we're matching quantiles, the exercise is invariant to monotone transformations of the outcome. In contrast, in a DiD model, the counterfactual outcome is estimated as $y^c = y + E[Y_{01}] - E[Y_{00}]$, which is not invariant to monotone transformations of Y.

Things to note:

- See the paper for inference results, for an extension to a multi-period and multi-group case (here the paper suggests estimating the potentially many treatment effects $E[Y_{gt}(1) - Y_{gt}(0)]$).
- Since we only need the distribution of U_g to be time-invariant, with panel data, this accommodates a fixed effect structure, $Y_{it}(0) = h_t(\eta_i + \epsilon_{it})$, where ϵ_{it} is an idiosyncratic shock with the same distribution across periods. Estimation is the same as in the repeated cross-section case, inference will be different.
- If the outcome is discrete, then the model doesn't quite make sense, since h can't be strictly increasing in u, if u is continuously distributed. If we weaken assumption (i) to only require h to be weakly increasing, then it turns out that the ATT is only partially identified, and eq. (5) estimates the lower endpoint of the identified set. To estimate the upper endpoint, replace $\hat{F}_{00}(y)$ in eq. (5) with $1 - \hat{F}_{-00}(-y)$, where $\hat{F}_{-00}$ is the empirical CDF of $-\mathbf{Y}_{00}$.

  Question 1. Can you see this result from Figure 5? (imagine discrete F_{00} that jumps at y)

- The support condition ensures that the quantile $q' = F_{00}(y)$ is well-defined. If $\mathbf{Y}_{01}$ is not in the support of Y_{00}, one option is to trim, similarly to trimming in average treatment effect (ATE) estimation under unconfoundedness, or to bound the ATT (**Homework question:** what are the bounds if the outcome is not bounded?)
- Athey and Imbens (2006) propose to accommodate covariates via the model $y = h_t(u) + x'\beta$.

Example 3 (continued). Panel B in Table 3 applies the estimator to the Meyer, Viscusi, and Durbin (1995) data, accounting for the discreteness in the outcome variable. Even though there are over 130 support points, the bounds are quite wide, and the DiD estimates from both a levels and a logs specification are included in the bounds. ⊠

1.4. *Conditional common trends*

In some cases, it may be more plausible to assume conditional common trends,

$$E[Y_{11}(0) - Y_{10}(0) \mid X_1 = x] = E[Y_{01}(0) - Y_{00}(0) \mid X_0 = x],$$

conditional on some time-invariant covariate X (Abadie 2005, about 3k citations). Here X just needs to be known at time 0; it could, in principle, correspond to past outcomes. Note this assumption does not imply the common trends assumption unless the distribution of X is the same between the treated and untreated (i.e. $X_1 \sim X_0$); and common trends obviously doesn't imply conditional common trends. So it's neither a weaker nor a stronger assumption. Under this assumption, one could estimate DiD conditional on X, and average the estimates, similarly to a regression approach to estimating the ATT under unconfoundedness.

Remark 5 (Treatment effects under unconfoundedness). Recall that, with cross-section data, if we assume a version of unconfoundedness, $E[Y(0) \mid D, X] = E[Y(0) \mid X] =: \mu_0(X)$, and let $p(X) = P(D = 1 \mid X)$ denote the propensity score, then

$$\begin{aligned} E[Y(1) - Y(0) \mid D = 1] &= E[Y(1) \mid D = 1] - E[E[Y \mid D = 0, X] \mid D = 1] \\ &= E[Y(1) \mid D = 1] - \frac{1-p}{p} E\left[\frac{p(X)Y}{1-p(X)} \mid D = 0\right]. \end{aligned}$$

So we can either use regression or propensity score weighting to estimate the ATT.

Proof. Here the first equality uses iterated expectations, and the fact that by unconfoundedness, $E[Y(0) \mid D = 1, X] = E[Y(0) \mid D = 0, X] = E[Y \mid D = 1, X] = \mu_0(X)$. The second uses the fact that by applying Bayes formula twice, $f_{X|D_1=1}(x) = p(x)f_X(x)/P(D_1 = 1) = \frac{1-p}{p}\frac{p(x)}{1-p(x)} f_{X|D_1=0}(x)$, so that for any function g, we have $E[g(X) \mid D = 1] = \frac{1-p}{p} E[g(X)p(X)/(1-p(X)) \mid D = 0]$. Thus,

letting $g(x) = E[Y \mid X = x, D = 0]$, we have

$$E[Y(0) \mid D = 1] = E[E[Y \mid X, D = 0] \mid D = 1]$$
$$= \frac{1-p}{p} E\left[E[Y \mid X, D = 0] \frac{p(X)}{1-p(X)} \mid D = 0\right] = \frac{1-p}{p} E\left[\frac{p(X)Y}{1-p(X)} \mid D = 0\right]. \quad \square$$

In our setting, we can apply Remark 5 with $Y_{g1} - Y_{g0}$ playing the role of Y, and with $p(X) = P(G = 1 \mid X)$. This delivers an analogous result:

$$E[Y_{11}(1) - Y_{11}(0)] = E[Y_{11} - Y_{10}] - E[E[Y_{01} - Y_{00} \mid X_0 = X_1]]$$
$$= E[Y_{11} - Y_{10}] - E\left[\frac{(Y_{01} - Y_{00})p(X_0)/p}{(1-p(X_0))/(1-p)}\right].$$

Intuitively, the propensity score weighting gives more weight to observations in the control group that look more like treatment group units. To implement it, we need to either estimate $p(X)$ nonparametrically, or estimate the conditional mean $E[Y_{01} - Y_{00} \mid X_0]$ nonparametrically.

- Note that adding the covariates linearly in the regression does not estimate the ATT in general (the equation in the display after eq. (8) in Abadie (2005)). **Homework question:** what do we estimate? What assumptions ensure the estimate has a causal interpretation? (Hint: think back to notes on ordinary least squares (OLS)).
- By the analogy with estimation of the ATT under unconfoundedness, one could also use other estimation strategies, such as matching.

1.5. *Fuzzy designs*

In many applications (about 10% of DiD papers according to a count in de Chaisemartin and D'Haultfœuille 2018), the treatment can be thought of as an encouragement or a subsidy. One could of course estimate the effect of the encouragement, which corresponds to the intent-to-treat (ITT) effect. Can we divide the ITT estimate by the first stage as in fuzzy regression discontinuity (RD), or instrumental variables (IV) to estimate the causal effect of a treatment of interest?

Example 5 (Duflo 2001). In 1973–74, the Indonesian government launched a major primary school construction program. Because the program intensity was related to 1972 enrollment rates, which vary across regions, we classify individuals into two groups $g = 0, 1$ depending on whether the individual was born in a region with high construction. There are two cohorts: cohort $t = 0$ consists of men aged 12–17 in 1974, who were out of primary school by the time the program launched. Cohort $t = 1$ consists of men aged 2–6 in 1974. Simple DiD estimates suggest an effect of 0.12 on years of education, and 0.026 on log wages (Table 3). Using the $\{t = 1\} \times \{g = 1\}$ interaction as an instrument for education, the implied effect on the return to education is quite high, about

19.5% (In Table 7 in the paper, it's quite a bit lower, 0.0752, I am not sure why). Assuming school construction affects wages only through years of schooling, under what conditions can we interpret these estimates as causal effects? ⊠

Denote the treatment of interest by D, and denote the encouragement design by $Z_{gt} = \mathbb{1}\{g = t = 1\}$. Let $Y_{gt}(d)$ denote the potential outcomes, and $D_{gt}(z)$ the potential treatments.[2] The IV estimand is given by

$$\beta_W = \frac{E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}]}{\pi}, \qquad \pi = E[D_{11} - D_{10}] - E[D_{01} - D_{00}].$$

What is this estimating? de Chaisemartin and D'Haultfœuille (2018) make the common trends assumption (2), in analogy to the sharp case where $D = Z$. Let $\tau_{gt} = Y_{gt}(1) - Y_{gt}(0)$, so that $Y_{gt} = Y_{gt}(0) + D_{gt}\tau_{gt}$. The $Y_{gt}(0)$ intercepts then cancel out by the common trends assumption, and we obtain

$$\beta_W = \frac{E[\tau_{11}D_{11}] - E[\tau_{10}D_{10}] - (E[\tau_{01}D_{01}] - E[\tau_{00}D_{00}])}{\pi}$$

Since $E[\tau_{gt}D_{dt}] = E[\tau_{gt} \mid D_{gt} = 1]E[D_{gt}]$, this is a non-convex combination of ATTs for different groups: the weights sum to one by definition of π, but some of them are negative. Bad news if there is heterogeneity in treatment effects. de Chaisemartin and D'Haultfœuille (2018) try to save the day by restricting heterogeneity in treatment effects, and by making no-defier type assumptions $D_{11}(1) \geq D_{10}(0)$, but we have to be able to conceptualize moving individuals through time to make sense of such assumptions. Regardless of the interpretation issues, they have limited success with this strategy.

de Chaisemartin and D'Haultfœuille (2018) propose an alternative approach based on the assumption that $E[Y_{g1}(d) - Y_{g0}(d) \mid D_{g0}(0) = d]$ doesn't depend on g. However, this approach only leads to bounds when the share of treated in the control group changes. They also consider an approach based on adapting the CiC model.

Remark 6. One could reach a very different conclusion about the attractiveness of the β_W estimand if we set things up differently. Instead of assuming (2), let us impose common trends on the treatment, $E[D_{11}(0) - D_{10}(0)] = E[D_{01}(0) - D_{00}(0)]$ Then $\pi = E[D_{11}(1) - D_{11}(0)]$ by arguments as in eq. (1). Also, make a common trends assumption on the outcomes in absence of the subsidy,

$$E[Y_{11}(D_{11}(0)) - Y_{10}(D_{10}(0))] = E[Y_{01}(D_{01}(0)) - Y_{00}(D_{00}(0))].$$

This allows us to interpret the ITT DiD regression as a causal effect. By arguments analogous to (1),

$$E[Y_{11} - Y_{10}] - E[Y_{01} - Y_{00}] = E[Y_{11}(D_{11}(1)) - Y_{11}(D_{11}(0))].$$

Make the monotonicity assumption that $D_{11}(1) \geq D_{11}(0)$ (now it's an assumption about

2. The setup in de Chaisemartin and D'Haultfœuille (2018) instead works with potential treatments that consider moving an individual to period t. That does not seem feasible...

what would happen if we canceled the subsidy, rather than a no defiers assumption across time). Then $E[Y_{11}(D_{11}(1)) - Y_{11}(D_{11}(0))] = E[(D_{11}(1) - D_{11}(0))\tau_{11}] = E[\tau_{11} \mid D_{11}(1) > D_{11}(0)]E[D_{11}(1) - D_{11}(0)]$, so that

$$\beta_W = E[\tau_{11} \mid D_{11}(1) > D_{11}(0)].$$

So there is no issue...

Question 2. What is the takeaway from all this?

Research Question. This setup is a special case of a model-based approach to IV. There are presumably things we can say in general here. What are they? ⊠

2. MULTIPLE PERIODS (AND GROUPS)

Suppose now that there are more than two periods. To simplify the discussion, we analyze the data at the individual level, $i = 1, \ldots, n$. Let $t = 1, \ldots, T$ index time. We abstract from individual-level controls. The notation $Y_{it}(d)$ and D_{it} is as before. Unlike the previous section, we don't make any restrictions on when the D_{it} may be zero or non-zero. Later, we'll consider particular designs for D_{it}. As in the case with two groups, we consider D_{it} to be a deterministic function of the individual and time. Equivalently, we can think of the exercise as conditioning on D_{it}. Make the common trends assumption:

$$E[Y_{it}(0) - Y_{i0}(0)] = \lambda_t, \qquad t \geq 1, i = 1, \ldots, n. \tag{6}$$

Remark 7. Since the treatment is non-random, the treatment assignment is exogenous, in the sense discussed in our OLS lecture. To see the connection, let us map the notation to the usual notation for estimating treatment effects under unconfoundedness. Let Y_j denote the outcome of a randomly picked observation, and let $i(j)$ and $t(j)$ be the identity of the unit associated with observation j, and $t(j)$ be the time index associated with this observations. Let $Y_j(d) = Y_{i(j),t(j)}$ denote the potential outcomes. Our covariates are $W_j = (i(j), t(j))$. What makes this setup somewhat special is that the treatment D_j is a deterministic function of the covariates. Using this notation, eq. (6) is equivalent to the model-based assumption (Assumption 6 in the lecture note on OLS)

$$E[Y_j(0) \mid D_j, W_j] = E[Y(0) \mid W_j] = \alpha_{i(j)} + \lambda_{t(j)}. \tag{7}$$

In our setup, the first equality ("exogeneity") is trivial (why?), and the content comes from the second equality, that it suffices to control for the individual effect and the time effect in an additive manner using two-way fixed effects (2WFE): we don't need to control for a more complicated, non-linear function of the unit and time dummies.

Under this setup, a DiD comparison using any two groups and two time periods comparison yields a causal comparison. In particular, since $Y_{it} = \tau_{it}D_{it} + Y_{it}(0)$, where

$\tau_{it} = Y_{it}(1) - Y_{it}(0)$, the $Y_{it}(0)$ terms cancel, and we obtain:

$$E[Y_{it} - Y_{is} - (Y_{jt} - Y_{jt})] = D_{it}E[\tau_{it}] - D_{is}E[\tau_{is}] - (D_{jt}E[\tau_{jt}] - D_{js}E[\tau_{js}]), \tag{8}$$

So if $D_{it} = 1$ and $D_{is} = D_{jt} = D_{js} = 0$, then this comparison estimates an ATT. The question is how we should aggregate these ATTs. The standard approach is to realize that eq. (6) implies

$$E[Y_{it}] = \alpha_i + \lambda_t + D_{it}E[\tau_{it}], \quad \alpha_i = E[Y_{i0}(0)], \tag{9}$$

$\lambda_0 = 0$, and $\lambda_t = E[Y_{0t} - Y_{00}]$. If the treatment effect $\tau_{it} = \tau$ is constant, we can estimate τ using a 2WFE regression. If the variance of the residual $Y_{it} - \alpha_i - \lambda_t - D_{it}\tau$ is homoskedastic, then this estimator has the usual optimality properties by the Gauss-Markov theorem.

2.1. *Heterogeneous treatment effects*

Suppose now that τ_{it} is not constant. What is the 2WFE regression estimating? By Remark 7, we can think of the problem as a problem of inference under heterogeneous treatment effects using the model-based approach to identification. By the results from the lecture on OLS, we estimate a weighted average of treatment effects $E[\tau_{it}]$, with weights

$$\lambda_{it} = \frac{\ddot{D}_{it}D_{it}}{\sum_{i,t}\ddot{D}_{it}D_{it}}, \tag{10}$$

where

$$\ddot{D}_{it} = D_{it} - \overline{D}_i - \frac{1}{n}\sum_j (D_{jt} - \overline{D}_j), \tag{11}$$

is the residual from regressing D_{it} onto unit and time fixed effects. Here where $\overline{D}_i = \frac{1}{T}\sum_{t=1}^{T} D_{it}$. This is Theorem 1 in de Chaisemartin and D'Haultfœuille (2020).

If $T = 2$, and no units are treated initially, we get $\ddot{D}_{i2}D_{i2} = D_{it}(1 - 1/2 - \pi_1 + \pi_1/2) = D_{i2}(1 - \pi_1)/2$, where π_1 is the fraction of units treated in period 2, so we estimate the ATT. The trouble is that, as discussed in the OLS lecture, the weights λ_{it} are in general negative for some groups and time periods. So we are estimating a linear, but not a convex combination of the (i,t)-level ATTs. The first differences estimator suffers from a similar problem. Again, the issue stems from the fact that the model-based assumption in eq. (7) does not generally guarantee that controlling for the covariates linearly in a regression will yield an estimate of a convex combination of treatment effects. This also implies, for instance, that using, say an interactive fixed effects specification will not fix the problem.

To provide additional intuition, Goodman-Bacon (2021, Theorem 1) decomposes the 2WFE estimator $\hat{\beta}_{\text{2WFE}}$ into a weighted average of different 2×2 DiD estimands, assuming

an event study design, or, equivalently a staggered adoption design (i.e. the treatments never switch off). The next result generalizes this decomposition to a setting where groups may drop the treatment, and considerably simplifies resulting expression.

Lemma 8. Let $\pi_{ij}^{ab} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a\}\,\mathbb{1}\{D_{jt} = b\}$ *denote the fraction of time* $D_{it} = a$ *and* $D_{jt} = b$*, and let* $\Delta_{ij}^{ab} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a\}\,\mathbb{1}\{D_{jt} = b\}(Y_{it} - Y_{jt})/\pi_{ij}^{ab}$*. Then*

$$\hat{\beta}_{2WFE} = \frac{\sum_{ij}(\pi_{ij}^{10}\pi_{ij}^{11}(\Delta_{ij}^{10} - \Delta_{ij}^{11}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00}))}{\sum_{ij}(\pi_{ij}^{01}\pi_{ij}^{11} + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10})}. \tag{12}$$

Proof. By the Frisch–Waugh–Lovell (FWL) theorem, the 2WFE estimator may be written as

$$\hat{\beta}_{\text{2WFE}} = \frac{\sum_{it}\ddot{D}_{it}Y_{it}}{\sum_{it}\ddot{D}_{it}^2} = \frac{\frac{1}{N}\sum_{ijt}(D_{it} - \overline{D}_i)(Y_{it} - Y_{jt})}{\sum_{it}\ddot{D}_{it}^2},$$

where the second equality follows by switching the order of summation in the identity

$$\sum_{it}\ddot{D}_{it}Y_{it} = \sum_{it}(D_{it} - \overline{D}_i)Y_{it} - \frac{1}{N}\sum_{ijt}(D_{jt} - \overline{D}_j)Y_{it}.$$

Since $\overline{D}_i = \pi_{ij}^{10} + \pi_{ij}^{11}$, the numerator of $\hat{\beta}_{\text{2WFE}}$ may be written as

$$\begin{aligned}
\sum_{it}\ddot{D}_{it}Y_{it} &= \frac{1}{N}\sum_{ijt}(D_{it} - \pi_{ij}^{10} - \pi_{ij}^{11})(Y_{it} - Y_{jt}) \\
&= \frac{1}{N}\sum_{ijt}(D_{it}(\pi_{ij}^{00} + \pi_{ij}^{01}) - (1 - D_{it})(\pi_{ij}^{10} + \pi_{ij}^{11}))(Y_{it} - Y_{jt}) \\
&= \frac{T}{N}\sum_{ij}((\pi_{ij}^{00} + \pi_{ij}^{01})(\pi_{ij}^{11}\Delta_{ij}^{11} + \pi_{ij}^{10}\Delta_{ij}^{10}) - (\pi_{ij}^{10} + \pi_{ij}^{11})(\pi_{ij}^{01}\Delta_{ij}^{01} + \pi_{ij}^{00}\Delta_{ij}^{00})),
\end{aligned}$$

where the second line uses $1 = \pi_{ij}^{00} + \pi_{ij}^{01} + \pi_{ij}^{10} + \pi_{ij}^{11}$. Rearranging the expression yields

$$\begin{aligned}
\sum_{it}\ddot{D}_{it}Y_{it} &= \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01})) \\
&\quad + \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{01})), \\
&= \frac{T}{N}\sum_{ij}(\pi_{ij}^{00}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{00}) + \pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00})) \\
&= \frac{T}{N}\sum_{ij}(\pi_{ij}^{01}\pi_{ij}^{11}(\Delta_{ij}^{11} - \Delta_{ij}^{01}) + (\pi_{ij}^{00} + 2\pi_{ij}^{01})\pi_{ij}^{10}(\Delta_{ij}^{10} - \Delta_{ij}^{00})),
\end{aligned}$$

where the second line uses $\Delta_{ij}^{10} - \Delta_{ij}^{01} = \Delta_{ij}^{10} - \Delta_{ij}^{00} + (\Delta_{ji}^{10} - \Delta_{ji}^{00})$, and the third line uses and uses the symmetry property $\Delta_{ij}^{11} - \Delta_{ij}^{00} = -(\Delta_{ji}^{11} - \Delta_{ji}^{00})$. Applying the same arguments to the denominator of $\hat{\beta}_{\text{2WFE}}$ and switching the i and j index in the first sum yields the result. □

The lemma says that the estimator is a weighted average of two types of DiD comparisons:

1. $\Delta_{ij}^{10} - \Delta_{ij}^{00}$. This is the usual DiD comparison, comparing i and j over periods where

i is treated and j is not, vs neither of them is treated. By eq. (8), is an unbiased estimate of the ATT for group g over the periods $D_{gt} = 1, D_{ht} = 0$.

2. $\Delta_{ij}^{10} - \Delta_{ij}^{11}$. This is a "forbidden comparison": j serves as a control for i, but we're not comparing to an untreated period, but to a period in which both are treated. Two ways of thinking about this. To parse through them, let $\tau_{i|j}^{ab} = \frac{1}{T}\sum_t \mathbb{1}\{D_{it} = a, D_{jt} = b\}E[\tau_{it}]/\pi_{ij}^{ab}$. First, if we only assume eq. (6), then by eq. (8), this estimates $\Delta_{ij}^{10} - \Delta_{ij}^{11} = \tau_{i|j}^{10} - \tau_{i|j}^{11} + \tau_{j|i}^{11}$, and this is how we wind up with negative weights on treatment effects. Alternatively, if we assume that eq. (6) also holds for $Y(1)$, then this estimates the ATE for the untreated, $E[\tau_{j|i}^{01}]$ (show this!). However, assuming common trends for both $Y(0)$ and $Y(1)$ restricts treatment effect heterogeneity, since it implies (why?) $E[\tau_{it}] = \kappa_i + \gamma_t$.

Thus, in DiDs designs, to interpret β_{2WFE} as a weighted average of treatment effects, we need to assume that eq. (6) holds for $Y(1)$ as well. Otherwise, the 2WFE estimand doesn't have a causal interpretation whenever the fitted values from the propensity score regression exceed one.

Example 6. Suppose that there are three time periods, $t = 1, 2, 3$, and two equal-sized groups. Group 1 starts being treated at $t = 2$, but group 0 only in the last period, $t = 3$. Then $\ddot{D}_{03} = 1/6$, while $\ddot{D}_{12} = 1/3$, and $\ddot{D}_{13} = -1/6$. So by eq. (10), we're estimating $\frac{1}{2}\tau_{03} + \tau_{12} - \frac{1}{2}\tau_{13}$, where $\tau_{gt} = E[\tau_{it} \mid g(i) = g]$ are group-specific treatment effects.

Alternatively, using Lemma 8, we're estimating the average of two DiD estimators, one comparing $\overline{Y}_{12} - \overline{Y}_{02} - (\overline{Y}_{11} - \overline{Y}_{01})$, where $\overline{Y}_{gt}$ are group means, which is unbiased for τ_{12}, with weight $1/2$, and the other one comparing $\overline{Y}_{12} - \overline{Y}_{02} - (\overline{Y}_{13} - \overline{Y}_{03})$ with weight $1/2$, which is unbiased for $\tau_{12} - \tau_{13} + \tau_{03}$. If we assume parallel trends on the treatment effects, so that $\tau_{g2} - \tau_{g3}$ is constant, this second term equals τ_{02}. ⊠

What to do? There have been many proposals (e.g. Callaway and Sant'Anna 2021; Sun and Abraham 2021; Imai and Kim 2021; Borusyak, Jaravel, and Spiess 2024).

1. We can compute these weights: a good idea to always to this as a robustness check!

2. Either estimate the ATT for the largest subpopulation that we can, or else pick the weights to combine the 2×2 DiD estimates efficiently.

3. If we worry about dynamic treatment effects, we can estimate the (weighted) ATT for units who just switch into treatment (suggested by Imai and Kim 2021), switched two periods ago, etc.

4. de Chaisemartin and D'Haultfœuille (2020) suggest imposing common trends on $Y_{gt}(1)$ as well, and estimate the average treatment effect for "switchers" (ATT for units just entering treatment, plus the ATE for the untreated units who just dropped treatment). But by the discussion above, if we do make this additional assumption, then there is no need to change the estimator! This point was made in Fabre (2023).

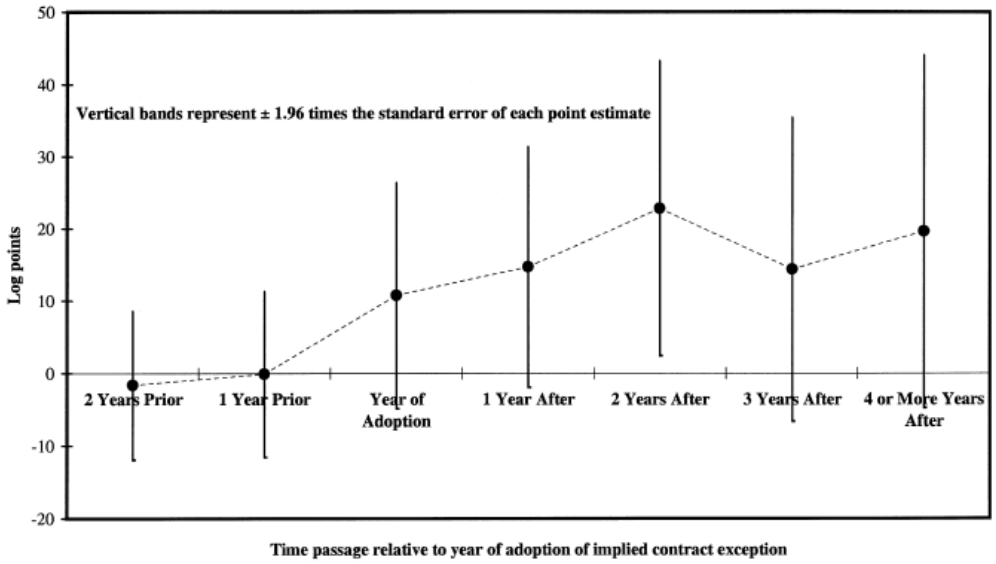


Figure 6: Figure 3 from Autor (2003), plotting coefficients from a regression of log temporary employment on a state-specific linear trend, and leads and lags of the adoption of implied contract exception by a state.

Research Question. Is there some attractive default? Part of the issue is that some of the proposed solutions require the treatment to be binary, but it's a bit awkward to have completely different methods for continuous versus binary treatments. ⊠

2.2. *Pretrends in event study designs*

As we discussed in Section 1, in staggered adoption designs (or equivalently, event study designs), it is common to regress Y_{it} onto leads and lags of when the treatment was adopted, and plot the coefficients to assess pretrends. Figure 6 reproduces such a plot from Autor (2003), who is interested in the effect of passing an implied contract exception to the employment at will doctrine on temporary help services.[3] Angrist and Pischke (2009) refer to this as a "Granger causality test". As discussed there, the coefficient on the period just before adoption is normalized to zero (if no units are untreated, we need to drop another leads or lag to avoid multicollinearity).

Sun and Abraham (2021) show that this test only works under constant treatment effects—we may then have non-zero pretrends even if the common trends assumption holds (unless we only have two groups as in eq. (3)). The issue is that we're trying

3. As a rule, US labor law allows "employment at will", which means that workers can be fired for just cause or no cause, at the employer's whim. But beginning in 1967, state courts have allowed a number of exceptions to the employment-at-will doctrine, which raised the chance that the employer may face an "unjust dismissal" lawsuit. The implied contract exception prohibits the firing of a worker after an "implied contract" has been established. Such a contract is an expectation of continued employment, which can be created in the form of oral assurances or expectations created by employer handbooks or policies.

to estimate multiple treatments, and doing this using regression where we control for the covariates linearly. But such regressions are not generally robust to treatment effect heterogeneity (Goldsmith-Pinkham, Hull, and Kolesár 2024). How to best conduct such a test in a manner that is robust to heterogeneity in treatment effects is an open question.

In addition, one may worry about the usual power issues with such a pretrend check, and about what to do if the test rejects. Freyaldenhoven, Hansen, and Shapiro (2019) propose to exploit the presence of a covariate that is affected by a confounder we worry about, but not by the policy. This allows for an IV solution to the pretrend problem, under the (strong) assumption that the dynamic relationship between the treatment and the instrument is the same as that between the confounder and the treatment. Rambachan and Roth (2023) propose conducting sensitivity analysis to deviations from the parallel trends assumption.

REFERENCES

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *The Review of Economic Studies* 72, no. 1 (January): 1–19. https://doi.org/10.1111/0034-6527.00321.

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.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics* 60, no. 1 (February): 47–57. https://doi.org/10.2307/1924332.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74, no. 2 (March): 431–497. https://doi.org/10.1111/j.1468-0262.2006.00668.x.

Autor, David H. 2003. "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labor Economics* 21, no. 1 (January): 1–42. https://doi.org/10.1086/344122.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." Forthcoming, *Review of Economic Studies* (February). https://doi.org/10.1093/restud/rdae007.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225, no. 2 (December): 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84, no. 4 (September): 772–793. https://www.jstor.org/stable/2118030.

———. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply." *American Economic Review* 90, no. 5 (December): 1397–1420. https://doi.org/10.1257/aer.90.5.1397.

Chernozhukov, Victor, and Christian B. Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73, no. 1 (January): 245–261. https://doi.org/10.1111/j.1468-0262.2005.00570.x.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2018. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85, no. 2 (April): 999–1028. https://doi.org/10.1093/restud/rdx049.

———. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110, no. 9 (September): 2964–2996. https://doi.org/10.1257/aer.20181169.

Donohue, John J., III, and Justin Wolfers. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58, no. 3 (December): 791–845.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *The American Economic Review* 91, no. 4 (September): 975–813. https://doi.org/10.1257/aer.91.4.795.

Fabre, Anaïs. 2023. "Robustness of Two-Way Fixed Effects Estimators to Heterogeneous Treatment Effects." Working paper, Toulouse School of Economics, https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2022/wp_tse_1362.pdf.

Freyaldenhoven, Simon, Christian B. Hansen, and Jesse M. Shapiro. 2019. "Pre-Event Trends in the Panel Event-Study Design." *American Economic Review* 109, no. 9 (September): 3307–3338. https://doi.org/10.1257/aer.20180609.

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

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225, no. 2 (December): 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014.

Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review* 84, no. 3 (June): 622–641. https://www.jstor.org/stable/2118071.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29, no. 3 (July): 405–415. https://doi.org/10.1017/pan.2020.33.

Meyer, Bruce, W. Kip Viscusi, and David Durbin. 1995. "Workers' Compensation and Injury Duration." *American Economic Review* 85, no. 3 (June): 322–340. https://www.jstor.org/stable/2118177.

Neumark, David, and William Wascher. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment." *American Economic Review* 90, no. 5 (December): 1362–1396. https://doi.org/10.1257/aer.90.5.1362.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A More Credible Approach to Parallel Trends." *Review of Economic Studies* 90, no. 5 (September): 2555–2591. https://doi.org/10.1093/restud/rdad018.

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2023. "When Is Parallel Trends Sensitive to Functional Form?" *Econometrica* 91, no. 2 (March): 737–747. https://doi.org/10.3982/ECTA19402.

Snow, John. 1855. *On the Mode of Communication of Cholera.* 2nd ed. London: John Churchhill.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225, no. 2 (December): 175–199. https://doi.org/10.1016/j.jeconom.2020.09.006.

SIMULATION-BASED INFERENCE

Michal Kolesár*

April 25, 2024

In 2024, Princeton flew out 14 junior candidates. Five papers were mostly concerned with theory (economic or econometric), and of the remaining nine papers, 5 featured simulation-based inference.[1] In previous years, the fraction of papers using the methods we discuss in this note is similarly around a third.

1. SIMULATION APPROACHES

Suppose we have i.i.d. data $Z_i = (Y_i, X_i)$, $i = 1, \dots, n$, where X_i are exogenous covariates, and the conditional distribution of the outcome Y_i given X_i is known up to a finite-dimensional parameter θ. For concreteness, suppose that this model has the form

$$Y_i = m(X_i, \epsilon_i, \theta), \tag{1}$$

where the distribution F of the vector ϵ_i is known. The outcome Y_i may be vector-valued, so that the setup is general enough so that Y_i may be a vector of equilibrium outcomes, or a time series path of outcomes in a dynamic panel model.

In principle, we should be able to write down the (conditional on X_i) likelihood, and estimate the model using maximum likelihood estimator (MLE). But in models of the form (1), evaluating the likelihood involves integration. If we can't evaluate the integral analytically, doing so numerically is hard if ϵ_i is high-dimensional (say $\dim(\epsilon_i) \geq 5$). There are a lot of examples: models with missing data, models with latent (unobserved) variables, or complicated discrete choice models.

Sometimes even simple-looking discrete choice models end up being surprisingly hard to estimate by likelihood methods:

Example 1 (Multinomial discrete choice). Suppose that the utility from choice j is given by $Y_{ij}^* = X_{ij}'\beta + U_{ij}$, $j = 0, \dots, J$, with the distribution of $U_i = (U_{i0}, \dots, U_{iJ})$ known up to its

*Email: mkolesar@princeton.edu.

1. These were Tim de Silva (finance) "Insurance versus Moral Hazard in Income-Contingent Student Loan Repayment", Aleksei Oskolkov (international finance) "Heterogeneous Impact of the Global Financial Cycle", Anna Russo (environmental, with K. M. Aspelund) "Additionality and Asymmetric Information in Environmental Markets", Charlie Rafkin (public, with Evan Soltas) "Eviction as Bargaining Failure: Hostility and Misperceptions in the Rental Housing Market", and Evan Soltas (public) "Tax Incentives and the Supply of Low-Income Housing".

covariance Σ. We can write this as $U_i = A\epsilon_i$, $\epsilon_i \sim F$, and $\Sigma = AA'$. Let a_j' denote the jth row of A, so that we can equivalently write $U_{ij} = a_j'\epsilon_i$. Let Y_i denote the $(J+1)$-vector of zeros with one in the position of the observed choice. That is, $Y_{ij} = 1$ if $Y_{ij}^* \geq \max_\ell Y_{i\ell}^*$ (assuming no ties), and $Y_{ij} = 0$ otherwise. This is a special case of (1), with vector outcome Y_i, and

$$m_j(X_i, \epsilon_i, \theta) = \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta + a_j'\epsilon_i \geq a_\ell'\epsilon_i\}.$$

The log-likelihood is given by $\sum_{i=1}^n \sum_{j=0}^J \mathbb{1}\{Y_{ij} = 1\} \log P_\theta(Y_{ij} = 1 \mid X_i)$. To build the likelihood, we therefore need to calculate all conditional choice probabilitys (CCPs), $P_\theta(Y_{ij} = 1 \mid X_i)$. Unfortunately,

$$P_\theta(Y_{ij} = 1 \mid X_i) = \int \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\epsilon\} f(\epsilon)\, \mathrm{d}\epsilon$$

is a J-dimensional integral that you don't want to evaluate numerically unless J is very small.

In some special cases, we do have analytic solutions. For example, consider the multinomial logit model, in which $\epsilon_{ij} \sim F(u)$, with $F(u) = e^{-e^{-u}}$. This distribution is called *Gumbel distribution*, or *Type-I extreme value distribution*[2]. Thus, $\Sigma = I$, so that a_i is the unit vector, ϵ_{ij} has density $e^{-u}e^{-e^{-u}}$. Then

$$\begin{aligned}
P_\theta(Y_{i0} = 1 \mid X_i) &= \int_{-\infty}^{\infty} \prod_{\ell=1}^{J} P(\epsilon_\ell \leq (X_{i0} - X_{i\ell})'\beta + \epsilon_0 \mid \epsilon_0) e^{-\epsilon_0} e^{-e^{-\epsilon_0}}\, \mathrm{d}\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-e^{-(X_{i0}-X_{i1})'\beta - \epsilon_0}} \cdots e^{-e^{-(X_{i0}-X_{iJ})'\beta - \epsilon_0}} e^{-\epsilon_0} e^{-e^{-\epsilon_0}}\, \mathrm{d}\epsilon_0 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\epsilon_0}[1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{iJ})'\beta}]\right) e^{-\epsilon_0}\, \mathrm{d}\epsilon_0 \\
&= \frac{1}{1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{iJ})'\beta}} = \frac{e^{X_{i0}'\beta}}{\sum_{j=0}^J e^{X_{ij}'\beta}},
\end{aligned}$$

where the last line follows from

$$\int_{-\infty}^{\infty} e^{-\epsilon} e^{-e^{-\epsilon-c}}\, \mathrm{d}\epsilon = \int_{-\infty}^{\infty} e^{-u+c} e^{-e^{-u}}\, \mathrm{d}u = e^c \int_{-\infty}^{\infty} e^{-u} e^{-e^{-u}} = e^c.$$

by setting $c = -\log(1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{i1})'\beta})$.

But what if $\epsilon \sim \mathcal{N}_{J+1}(0, I)$ and we don't restrict Σ?[3] The likelihood for this multinomial probit model is hard to evaluate unless J is very small. ⊠

2. The names come from the fact that if X_i are i.i.d. exponential (so that the cumulative distribution function (CDF) is $1 - e^{-x}$), and we let $M_n = \max_{i \leq n} X_i$ denote the maximum, then Emil Julius Gumbel showed that $M_n - \log(n) \Rightarrow F(u)$. Indeed, $P(M_n - \log(n) \leq u) = (1 - e^{-u}/n)^n \to e^{-e^{-u}}$.

3. The covariance matrix Σ needs some normalization beyond requiring that it be symmetric positive semi-definite; we'll come back to this issue in Section 2.3.

Example 2 (Random coefficients logit). The multinomial logit model is restrictive: it implies independence of irrelevant alternatives, and it may be able to match the observed CCPs. McFadden and Train (2000) showed that if we allow the coefficients β to be random, the model can match the CCPs generated by *any* discrete choice random utility model. Such a model is called a random coefficients multinomial logit, or sometimes mixed logit, since the choice probabilities are mixtures of logit probabilities. In other words, suppose that each individual draws their own coefficients from some distribution G_θ $\beta_i \sim G_\theta$. Then $P_\theta(Y_{ik} = 1 \mid X_i) = \int \frac{e^{X_{ik}'\beta}}{\sum_{j=0}^J e^{X_{ij}'\beta}} dG_\theta(\beta)$. If G_θ and X_i is sufficiently flexible, we can match any observed choice patterns. A tractable choice for the mixing distribution G_θ is $\beta \sim \mu + \Lambda\epsilon$, where Λ is an $K \times L$ matrix of "factor loadings" onto factors ϵ, distributed $\mathcal{N}(0, I_K)$, say. Then the CCP becomes an L-dimensional integral,

$$P_\theta(Y_{ik} = 1 \mid X_i) = \int \frac{e^{X_{ik}'\mu + X_{ik}'\Lambda\epsilon}}{\sum_{j=0}^J e^{X_{ij}'\mu + X_{ij}'\Lambda\epsilon}} dF(\epsilon). \tag{2}$$

⊠

Example 3 (Panel probit). Another simple model that's non-trivial to estimate is the panel Probit model. For simplicity, suppose there are only two alternatives, so that $Y_{it}^* = X_{ij}'\beta + \epsilon_{it}$, and we observe $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, with $Y_{it} = \mathbb{1}\{Y_{it}^* \geq 0\}$. In general, we'd expect ϵ_{it} to be correlated over time, since factors that are not observed by the researcher can persist over time. Then with the autocovariance structure unrestricted, the likelihood for observing the sequence Y_i is a T-dimensional integral. ⊠

1.1. *Simulated method of moments*

The classic reference is McFadden (1989), see also Pakes and Pollard (1989). Notice that (1) and iterated expectations imply that for any function $s(z)$, we have the moment condition

$$E\left[g(Z_i, \theta_0)\right] = 0, \quad g(Z_i, \theta) = s(Z_i) - E_{\theta_0}[s(Z_i) \mid X_i] = s(Z_i) - \int s(X_i, m(X_i, \epsilon, \theta_0))\, \mathrm{d}F(\epsilon).$$

For example, we may take $s(z) = yg(x)$, so that $g(Z_i, \theta) = (Y_i - E[Y_i \mid X_i])g(X_i)$. The integral may be hard to evaluate. If we could evaluate it, we would be able to use the generalized method of moments (GMM) estimator

$$\hat{\theta}_{\text{GMM}} = \operatorname*{argmin}_\theta \frac{1}{n}\sum_i g(Z_i, \theta)' \hat{W}_n \frac{1}{n}\sum_i g(Z_i, \theta),$$

with asymptotic distribution (under regularity conditions)

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}),$$

where $G = E[\frac{\partial}{\partial\theta} g(Z_i, \theta_0)]$ and

$$\Omega = \text{var}(g(Z_i, \theta_0)) = E[\text{var}_{\theta_0}(s(Z_i) \mid X_i)].$$

The key idea of simulated method of moments (SMM) is that we can replace the hard-to-evaluate integral with an unbiased estimate of it. In particular, simulate M samples of $\{\tilde{\epsilon}_1^m, \ldots, \tilde{\epsilon}_n^m\}_{m=1}^M$ from the known distribution of ϵ, and replace the integral (which is an expectation) by the sample average based on the simulated samples, generating the moment

$$\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta) = s(Z_i) - \frac{1}{M} \sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta)), \tag{3}$$

where $\tilde{\epsilon}_i = (\epsilon_i^1, \ldots, \tilde{\epsilon}_i^M)$ is a vector of simulation samples for the ith observation. We then apply GMM to this moment condition,

$$\hat{\theta}_{\text{SMM}} = \frac{1}{n} \sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)' \hat{W} \frac{1}{n} \sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta).$$

One issue (that we'll come back to) is that if Y_i is discrete as in discrete choice models, $\tilde{g}$ is not differentiable in θ, which creates issues both for minimizing the sample GMM objective function[4], for proving normality, and for estimating the asymptotic variance. Using empirical process theory, it is possible to nonetheless show:

Proposition 1. Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{SMM} - \theta) \Rightarrow \mathcal{N}(0, (G'WG)^{-1} G'W\Omega_{SMM} WG(G'WG)^{-1}),$$

where (using differentiation under the integral sign)

$$G = \frac{\partial}{\partial\theta} E[\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)] = E\left[\frac{\partial}{\partial\theta} g(Z_i, \theta_0)\right].$$

and

$$\Omega_{SMM} = \text{var}(\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)) = \left(1 + \frac{1}{M}\right) \Omega.$$

So the gradient G in the asymptotic variance is the same as in the GMM case, but simulation increases the variance of the moment condition.[5] The formula for Ω_{SMM} obtains

4. One needs to use non-derivative based methods, such as Nelder-Mead, but minimization can be challenging unless $\dim(\theta)$ is small.

5. Because the simulation draws are ancillary (like the precision of the scale in the example in Cox (1958)), one can make the argument that we should do inference conditional them. In that case, there is no noise, but there is bias! Will need $M \to \infty$ to get consistency.

because the simulation draws are independent of each other and of the data, so that

$$\begin{aligned}
&\operatorname{var}(\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)) \\
&\quad = E\left[\operatorname{var}\left(g(Z_i, \theta_0) + E_{\theta_0}(s(Z_i) \mid X_i) - \frac{1}{M}\sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta_0))\right) \;\middle|\; X_i\right] \\
&\quad = E[\operatorname{var}(g(Z_i, \theta_0)) \mid X_i] + E\left[\operatorname{var}(\frac{1}{M}\sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta_0)) \;\middle|\; X_i)\right] \\
&\quad = E[\operatorname{var}(g(Z_i, \theta_0)) \mid X_i] + \frac{1}{M} E\left[\operatorname{var}(g(Z_i, \theta_0) \mid X_i)\right].
\end{aligned}$$

You can see that more generally, if say the simulation draws are correlated with each other, the asymptotic variance formula is the sum of the variance of the moment condition, and the simulation variance.

Remark 2 (Aside). The key new regularity condition (relative to the usual regularity conditions for GMM with smooth moment conditions) for Proposition 1 is a stochastic equicontinuity condition allowing us to use the differentiation under the integral sign, that if $\delta_n \to 0$, then

$$\sup_{\|\theta_n - \theta_0\| \le \delta_n} \frac{\sqrt{n}\|\tilde{g}_n(\theta_n) - \tilde{g}_n(\theta_0) - E[g(Z_i, \theta_n)]\|}{1 + \sqrt{n}\|\theta_n - \theta_0\|} \xrightarrow{p} 0, \tag{4}$$

where $\tilde{g}_n(\theta_0) = n^{-1}\sum_{i=1}^n \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)$. Stochastic equicontinuity is a topic for 519 (check Ch 7 in Newey and McFadden 1994, if you're interested). The intuition is that we already know $\tilde{g}_n(\theta) \xrightarrow{p} E[g(Z_i, \theta)]$ by the law of large numbers (LLN), so that eq. (4) holds already pointwise (i.e. without the sup) for any $\theta_n \neq \theta_0$. The condition strengthens the pointwise convergence to hold uniformly.

Remark 3 (Estimating the asymptotic variance). Ω in the asymptotic variance formula is easy to estimate. On the other hand, since the moment condition $g(Z_i, \theta_0)$ is hard to evaluate, estimating G can be challenging. The problem is that in discrete choice models, $\tilde{g}_n(\theta) = n^{-1}\sum_{i=1}^n \tilde{g}(Z_i, \epsilon_i, \theta)$ is not typically differentiable in θ. One option is to use importance sampling to make it smooth (see Section 2.2 below). Another option is to take a numerical derivative, with a large enough step size s_n, as discussed in Newey and McFadden (1994, Chapter 7.3). In particular, the estimator

$$\hat{G}_j = \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta})}{s_n} \quad \text{or} \quad \hat{G}_j = \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta} - s_n e_j)}{2s_n}$$

of Ge_j (and similarly for other columns of G) will satisfy

$$\hat{G}_j - Ge_j \xrightarrow{p} 0, \tag{5}$$

if the step size s_n satisfies $s_n \to 0$ and $\sqrt{n}s_n \to \infty$. The idea is similar to kernel smoothing. To pick s_n in practice, one possibility is to plot $\hat{G}_j$ as a function of s_n, and then choose

s_n, small, but not in a region where the function is very choppy.

In some cases, one may obtain consistency under even weaker conditions on s_n. See Hong, Mahajan, and Nekipelov (2015) for a thorough treatment of numerical estimation of derivatives.

Proof of eq. (5). By triangle inequality, letting $g(\theta) = E[g(Z_i, \theta)]$ and $\tilde{g}_n(\theta) = n^{-1}\sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)$,

$$\left\| \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta})}{s_n} - G e_j \right\| \\ \leq \frac{\left\| \tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\theta_0) - g(\hat{\theta} + s_n e_j) \right\|}{s_n} + \left\| \frac{g(\hat{\theta} + s_n e_j)}{s_n} - G e_j \right\| + \frac{\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0)\|}{s_n}$$

Now, by eq. (4) and the fact that $\hat{\theta} - \theta = O_p(n^{-1/2})$, the first term is bounded by

$$o_p(1)\frac{n^{-1/2} + \|\hat{\theta} + s_n e_j - \theta\|}{s_n} \leq o_p(1/s_n\sqrt{n}) + o_p(1) + o_p(1)\|\hat{\theta} - \theta\|/s_n = o_p(1/\epsilon\sqrt{n}).$$

By Taylor's theorem and the fact that g is differentiable at θ_0, $g(\hat{\theta} + s_n e_j) = g(\theta_0) + G(\hat{\theta} + s_n e_j - \theta_0) + o(\|\hat{\theta} + s_n e_j - \theta_0\|)$, the second term is bounded by

$$\left\| \frac{g(\hat{\theta} + s_n e_j)}{s_n} - G e_j \right\| \leq \|G(\hat{\theta} - \theta_0)/s_n\| + o(\|\hat{\theta} + s_n e_j - \theta_0\|/s_n) \\ \leq (\|G\| + o(1))\|\hat{\theta} - \theta_0\|/s_n + o(1) \leq O_p(1/\epsilon\sqrt{n}).$$

Finally, again using the triangle inequality and eq. (4),

$$\frac{\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0)\|}{s_n} \leq s_n^{-1}\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0) - g(\hat{\theta})\| + s_n^{-1}\|g(\hat{\theta})\| = o_p(1/s_n\sqrt{n}) + O_p(1/s_n\sqrt{n}). \quad \square$$

Example 4. First we for simplicity consider a binomial choice example. We have $Y_i = \mathbb{1}\{X_i'\theta_0 + \epsilon_i \geq 0\}$, with distribution of $\epsilon_i \sim F$ known and independent of X_i. This model delivers the moment (by setting $s(Z_i) = Y_i X_i$)

$$E[(Y_i - P_{\theta_0}(Y_i = 1 \mid X_i))X_i] = 0.$$

Since $P_\theta(Y_i = 1 \mid X_i) = F(X_i'\theta)$, we can estimate θ by GMM, using the moment $g(Z_i, \theta) = (Y_i - F(X_i'\theta))X_i$. Since we're exactly identified, the weight matrix doesn't matter, and the elements of the asymptotic variance would be given by

$$G = E[f(X_i'\theta_0)X_i X_i'],$$
$$\Omega = E[\mathrm{var}(Y_i \mid X_i)X_i X_i'] = E[F(X_i'\theta_0)(1 - F(X_i'\theta_0))X_i X_i'],$$

with $f = \frac{\mathrm{d}}{\mathrm{d}u}F(u)$ denoting the density. Suppose we don't know how to calculate the CDF $F(u)$, but we do know how to draw from the distribution. We could then use a

SMM estimator, with the moment (3) now given by

$$0 = E[\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)] \qquad \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta) = \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{X_i'\theta + \tilde{\epsilon}_i^m \geq 0\} \right) X_i,$$

replacing the unknown CCP with a sample average based on the M simulated samples. Note that $\tilde{g}$ is a step function as a function of θ. ⊠

Example 1 (continued). In the multinomial model, we could try to match the CCPs,

$$g_j(Z_i, \theta) = (\mathbb{1}\{Y_{ij} = 1\} - P_\theta(Y_{ij} = 1 \mid X_i))X_i \tag{6}$$

(so that $s(Z_i) = X_i \otimes Y_i$). Since the event $\{Y_{ij} = 1\}$ is equivalent to $\bigcap_{\ell=1}^{J} \{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\epsilon_i\}$, we can estimate the CCP by

$$\frac{1}{M} \sum_{m=1}^{M} \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\tilde{\epsilon}_i^m\},$$

where $\tilde{\epsilon}_i^m$ is drawn from F. In other words, replace $P_\theta(Y_{ij} = 1 \mid X_i)$ with sample frequency with which j is the choice with the highest utility in the simulated data. ⊠

Example 2 (continued). The CCP in eq. (2) can be estimated as

$$\hat{P}_\theta(Y_{ik} = 1 \mid X_i) = \frac{1}{M} \sum_{m=1}^{M} \frac{e^{X_{ik}'\mu + X_{ik}'\Lambda\tilde{\epsilon}_i^m}}{\sum_{j=0}^{J} e^{X_{ij}'\mu + X_{ij}'\Lambda\tilde{\epsilon}_i^m}},$$

which is smooth in the parameters. See McFadden and Train (2000) for a discussion of how to pick the moments. ⊠

Remark 4. Because the moment condition (3) is mean zero even with a single simulation draw per observation ($M = 1$), using the number of simulation draws only affects the asymptotic variance of the estimator. The estimator's consistency or asymptotic normality is not affected by the simulation. Furthermore, the standard errors are only inflated by a factor of $\sqrt{(1 + 1/M)}$, which means that for practical purposes, $M = 10$ (say) is sufficient (leading to 5% larger standard errors).

1.2. *Simulated maximum likelihood*

This idea goes back to Lerman and Manski (1981): if the likelihood contains an integral that's hard to evaluate, replace it with a sample average based on simulated samples.

Example 4 (continued). We can also estimate θ by MLE,

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmin}_\theta \frac{1}{n} \sum_i q(Y_i, F(X_i, \theta)) \quad q(Y_i, F) = -\mathbb{1}\{Y_i = 1\} \log F - \mathbb{1}\{Y_i = 0\} \log(1 - F).$$

If F is hard to compute, we replace it with a simulated version,

$$\hat{\theta}_{\text{SMLE}} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_i q(Y_i, \hat{F}(X_i, \theta)), \qquad \hat{F}(X_i, \theta) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{X_i'\theta + \tilde{\epsilon}_i^m \geq 0\}. \qquad \boxtimes$$

More generally, for whatever discrete choice model, we can approximate the log likelihood as

$$\sum_{i=1}^{n} \sum_{j=0}^{J} \mathbb{1}\{Y_{ij} = 1\} \log \hat{P}_\theta(Y_{ij} = 1 \mid X_i).$$

This is particularly popular for estimating the random coefficient logit, see, for example Huber and Train (2001).

The main problem with this method is that because the log-likelihood is non-linear, even if $\hat{F}$ is an unbiased estimate of F, (i.e. $E[\hat{F}(X_i, \theta) \mid X_i] = F(X_i'\theta)$) it's still the case that $q(Y_i, E[\hat{F}(X_i, \theta) \mid X_i]) \neq E[q(Y_i, \hat{F}_i(\theta)) \mid X_i]$, so that the limiting objective function is not correct (and the SMLE is inconsistent) unless $M \to \infty$ as $n \to \infty$. For SMLE to be asymptotically equivalent to MLE, we need $M^2/n \to \infty$. For these reasons, it's a good idea to use the sandwich variance estimator when estimating the standard errors, as discussed in McFadden and Train (2000).

The methods of simulated moments and that of simulated scores, discussed next, were initially motivated by the desire for a simulation-based estimator that is consistent even for a fixed number of draws.

1.3. *Method of simulated scores*

This method is due to Hajivassiliou and McFadden (1998). One issue with SMM relative to MLE is that it is less efficient, even if we didn't incur a simulation error (because we have a parametric model, unless there is endogeneity in X_i). Another is that it is unclear how to pick the moments. If, as the moments, we used the score, we would solve both issues.

For concreteness, let us consider a discrete choice model, with the CCP given by $P(Y_{ij} = 1 \mid X_i) = p_j(\theta, X_i)$, and $\sum_{j=0}^{J} p_j(\theta, X_i) = 1$. Then the likelihood is given by $\sum_{i,j} Y_{ij} \log p_j(\theta, X_i)$, with the score given by

$$\sum_{i,j} Y_{ij} \frac{1}{p_j(\theta, X_i)} \frac{\partial p_j(\theta, X_i)}{\partial \theta}.$$

Since the expected value of the score is zero, this gives a moment condition. If the CCPs are hard to evaluate, we could instead try to get independent unbiased estimates of $\partial p_j(\theta, X_i)/\partial\theta$ and of $1/p_j(\theta, X_i)$. The first one is easy, since we can get unbiased estimates of $p_j(\theta, X_i)$ and the derivative is a linear operator. The trouble comes in getting reliable unbiased estimates of $1/p_j(\theta, X_i)$. Since $1/p$ is the mean of a geometric distribution (number of Bernoulli trials needed to get $Y_{ij} = 1$), one estimator is the number

of trials needed to generate $Y_{ij} = 1$. The disadvantage of this approach is that it can be slow, especially if some CCPs are small.

1.4. *Indirect inference*

This method is due to Smith (1993) (part of his 1990 PhD thesis), further developed in Gourieroux, Monfort, and Renault (1993). It was originally developed in a time-series context; we focus here on the cross-section and panel case.

The intuition for the SMM method is that the sample moments $\frac{1}{n}\sum_{i=1}^{n} s(Z_i)$ from the real data should match sample moments from the simulated data, if the simulated data is generated using the true θ_0. The insight of indirect inference is that we don't need to restrict ourselves to moments: if we have the right model generating the data, then *any feature* of the simulated data should match the real data. In particular, we can focus on matching features of the real data that we think are important.

To make this concrete, let $\hat{\pi}$ be some easily-computable reduced-form parameter $\pi_0 = h(\theta_0)$, which satisfies

$$\sqrt{n}(\hat{\pi} - \pi_0) \Rightarrow \mathcal{N}(0, \Omega). \tag{7}$$

For example, π could be a reduced-form regression, quantile regression parameters, reduced-form choice probabilities, or even maximum likelihood estimates based on a simpler (and therefore potentially misspecified) model (the case studied in Gallant and Tauchen 1996), or GMM estimates based on such model. The only thing to be careful about is that in deriving Ω, we don't want to assume that the model for π is correctly specified. So if π is, say, the MLE estimand, we want to compute Ω using the sandwich formula.

The model defining the reduced-form parameter is called an *auxiliary model*. To fix ideas, suppose $\hat{\pi}$ is an M-estimator

$$\hat{\pi} = \operatorname*{argmin}_{\pi} \frac{1}{n}\sum_{i=1}^{n} q(Z_i, \pi), \tag{8}$$

which defines $\pi = h(\theta) = \operatorname{argmin}_{\varpi} \int q(z, \varpi)\, \mathrm{d}P_\theta(z)$. This map $h\colon \theta \to \pi$ is called the *binding function* by Gourieroux, Monfort, and Renault (1993). If it were a known function, then we could use minimum distance to get an estimate of θ.

Typically, however, it is hard to figure out the form of h, for the same reason that it is hard to evaluate the likelihood in the first place. On the other hand, it's easy to simulate the data: generate draws $\{\tilde{\epsilon}_i^m\}_{m=1}^{M}$ from the (known) distribution of ϵ_i, and set

$$\tilde{Y}_i^m(\theta) = m(X_i, \tilde{\epsilon}_i^m, \theta)$$

and $\tilde{X}_i^m = X_i$.[6] This gives us M simulated samples $\{\tilde{Z}_i^m(\theta) = (\tilde{Y}_i^m(\theta), X_i)\}_{i=1}^n$. On each of these simulated samples, we can estimate π using the same estimator that we used to get $\hat{\pi}$ on the real data: $\tilde{\pi}^m(\theta) = \operatorname{argmin}_\pi n^{-1} \sum_{i=1}^n q(\tilde{Z}_i^m(\theta), \pi)$. We then estimate θ by minimizing the distance between the reduced-form estimate from the data and those from the simulated samples,

$$\hat{\theta}_{\text{II}} = \underset{\theta}{\operatorname{argmin}} \left(\hat{\pi} - \tilde{\pi}(\theta)\right)' \hat{W} \left(\hat{\pi} - \tilde{\pi}(\theta)\right), \qquad \tilde{\pi}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \tilde{\pi}^m(\theta).$$

An alternative (and asymptotically equivalent) approach is to pool all the simulated data together and estimate $\tilde{\pi}$ based on one big sample of size Mn.

Proposition 5. Suppose that as $n \to \infty$,

1. $\pi(\theta) \neq \pi(\theta_0)$ *if* $\theta \neq \theta_0$
2. $\pi(\theta)$ *is continuous*
3. Θ *is compact*
4. $\sup_\theta |\tilde{\pi}(\theta) - \pi(\theta)| \xrightarrow{p} 0$
5. $\hat{\pi} \xrightarrow{p} h(\theta_0)$, *and* $\hat{W} \xrightarrow{p} W$ *positive definite.*

and that $M \geq 1$ is fixed. Then $\hat{\theta}_{II} \xrightarrow{p} \theta_0$.

Proof. This follows by verifying conditions UC and ID in the theorem for consistency of extremum estimators that is discussed in 519. □

Using empirical process methods (see Theorem 7.2 in Newey and McFadden 1994), and assuming (7), it follows that

$$\sqrt{n}(\tilde{\pi}^m(\theta) - \pi_0) \Rightarrow \mathcal{N}(0, \Omega),$$

with the limiting distribution

$$\sqrt{n}(\hat{\theta}_{\text{II}} - \theta_0) \Rightarrow \mathcal{N}\left(0, \left(1 + \frac{1}{M}\right)(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}\right).$$

Similar to the SMM case, we would need $M \to \infty$ to be as efficient as a minimum distance estimator.

Remark 6 (Picking the reduced form). If the auxiliary model is the true model (that is, if $q(\cdot)$ in (8) is the true log-likelihood), with $\dim(\pi) = \dim(\theta)$, then solving $\hat{\theta} = h(\hat{\pi})$ gives the MLE of θ. This suggests that it should be a good idea to pick the auxiliary model to be close to the true model. However, so long as $h(\cdot)$ is invertible (which may be in practice hard to check!), the indirect inference estimator will be consistent—it doesn't

6. One could also make the simulated samples have sample size different from n, in which case X_i^m would be drawn from the empirical distribution of X

matter if the auxiliary model is misspecified. The choice of a particular auxiliary model only affects efficiency of the estimator. The choice is often driven by what aspect of the data you want to fit.

Remark 7. Indirect inference is similar to calibration in macro—except that we do get standard errors as well. See the working paper https://arxiv.org/abs/2109.08109 for how to get the standard errors in calibration exercises.

Remark 8. Notice that the efficient weight matrix is proportional to Ω^{-1}, which, since Ω is just the variance of the reduced-form estimator, can be estimated directly from the data—we do not need a two-step estimator, in which we first get a consistent estimate of θ, and then use the estimate to form an efficient weight matrix.

Often, people set $\hat{W} = \text{diag}(\hat{\Omega})^{-1}$. It may have something to do with misspecification of the model (1).

Research Question. How should one deal with such potential misspecification? ⊠

2. PRACTICAL ISSUES

2.1. *An important implementation note*

Remark 9. For the simulation methods to work, it is important to draw the sequence $\{\tilde{\epsilon}_i^1, \ldots, \tilde{\epsilon}_i^M\}_{i=1}^n$ once, and hold it fixed as we search over β. Otherwise, the objective function will jump wildly as we move over the Θ space, and consistency will fail (and, on a more basic level, the estimator is not well-defined). See Figure 1. In fact, in the derivation of the asymptotics, we have implicitly treated $(Y_i, X_i, \tilde{\epsilon}_i)$ as the data.

One way of ensuring the draws are fixed is to set the seed to a particular value. A better way is to draw the sequence of epsilons once at the beginning, store them, and then use the set each time you need to evaluate the objective function. In other words, treat the simulation draws as part of the data after you generate them. It's more transparent, marginally quicker (since you don't need to re-generate the draws), and it also makes it easy to share the draws along with the data with another person or another programming language if the whole estimation uses more than one language.

2.2. *Improving on frequency simulators*

In the description of the simulation methods above, we have used a simple frequency simulator to estimate the CCPs $p_j(\theta, X_i) := P_\theta(Y_{ij} = 1 \mid X_i)$,

$$\hat{p}_j(\theta, X_i) = \frac{1}{M} \sum_{m=1}^{M} \tilde{Y}_{ij}^m(\theta).$$

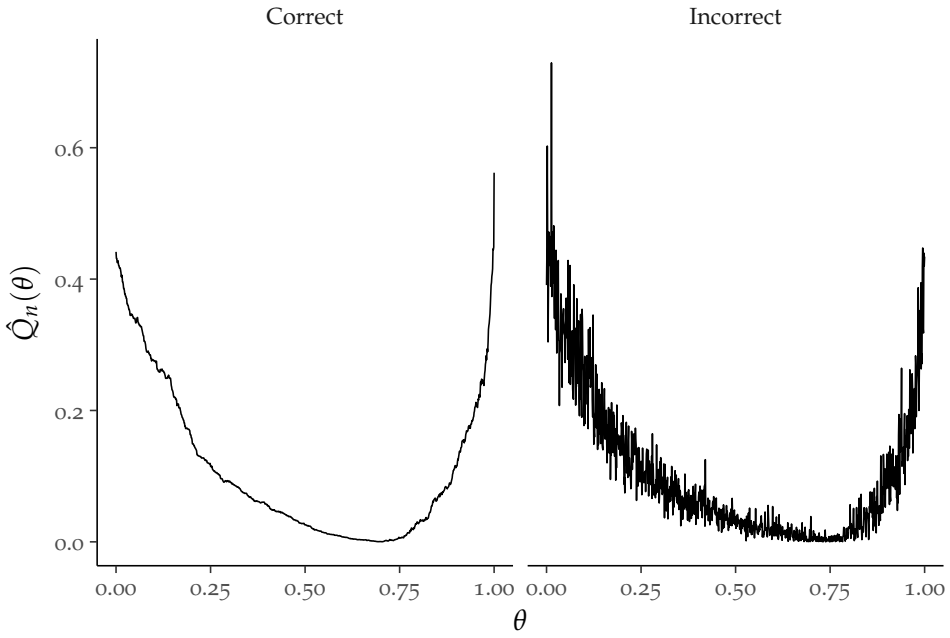


Figure 1: If $\tilde{\epsilon}$ is not held fixed, the objective function oscillates wildly ("Incorrect" panel). When $\tilde{\epsilon}$ is held fixed ("Correct" panel), the objective function is much better behaved, although it is still not differentiable.

This has multiple problems:

1. Slow if J is large
2. Sample objective function is discontinuous in θ, since $\tilde{Y}_{ij}^m(\theta)$ will be a step function. Apart from complications with asymptotic theory, this makes the sample objective function hard to optimize.
3. There is a positive probability that $\hat{p}_j(\theta, X_i) = 0$, which creates issues with some estimators (such as simulated maximum likelihood). A related problem is that the frequency sampler will be imprecise if the CCP is close to zero.

We'll briefly discuss two approaches to ameliorate this problem: kernel smoothing and importance sampling (or, more precisely, importance sampling coupled with a change of variables), both of which were suggested in McFadden (1989).

Remark 10. Besides these two methods, there have been other proposals for implementing simulation-based methods. For example, the frequency simulator uses $n \times M$ independent draws $\tilde{\epsilon}_i^m$. If n is very large, this may be computationally prohibitive. An alternative is to use the same set of M draws $\{\tilde{\epsilon}^m\}_{m=1}^M$ for each observation. See, for example, Hong, Li, and Li (2021) for an implementation of this idea in the context of the Berry, Levinsohn, and Pakes (1995) model, and Armstrong et al. (2017) for a general theory.

KERNEL SMOOTHING The idea of kernel smoothing is to replace the discrete choice indicators by a smooth function of the underlying continuous latent variables that determine the model's discrete outcomes. An obvious problem is how to choose the amount of smoothing, and that the smoothing induces finite-sample bias. See Bruins et al. (2018) for an application of this idea in the context of indirect inference.

IMPORTANCE SAMPLING This idea was imported to econometrics by Kloek and van Dijk (1978). To explain it, consider Example 4, and use a change of variables to rewrite the CCP as

$$P_\theta(Y_i = 1 \mid X_i = x) = E[\mathbb{1}\{X_i'\theta + \epsilon_i \geq 0\} \mid X_i = x] = \int \mathbb{1}\{u \geq 0\} f(u - x'\theta)\,\mathrm{d}u$$
$$= \int \mathbb{1}\{u \geq 0\} \frac{f(u - x'\theta)}{g(u \mid x)} g(u \mid x)\,\mathrm{d}u,$$

where f is the density of ϵ_i, and $g(u \mid x)$ is some other density. If we draw samples $\{\tilde{u}_i^m\}_{m=1}^M$ this density, we can estimate the CCP by

$$\frac{1}{M}\sum_m \mathbb{1}\{\tilde{u}_i^m \geq 0\} \frac{f(\tilde{u}_i^m - X_i'\theta)}{g(\tilde{u}_i^m \mid X_i)},$$

which is differentiable in θ so long as f is differentiable. Instead of holding the draws $\tilde{\epsilon}_i^m$ (and their implicit weights, $1/M$) constant as we change θ, we're holding the utility $\tilde{u}_i^m$ constant, but vary the importance weights $\frac{1}{M}\frac{f(\tilde{u}_i^m - X_i'\theta)}{g(\tilde{u}_i^m \mid X_i)}$ that we put on each simulated observation as we search over θ.

- Since the integrand is only non-zero if u is positive, the support of g only needs to be on $\mathbb{R}_+$: for instance, we can set it to a truncated normal, or an exponential distribution. Then we can get rid of the indicator in the preceding display.
- A possible choice for $g(u \mid x)$ is to set it to $f(u - x'\tilde{\theta})$, truncated to the positive part of the real line, where $\tilde{\theta}$ is an initial guess of θ.
- Since $\mathbb{1}\{\tilde{u}_i^m \geq 0\}/g(\tilde{u}_i^m \mid X_i)$ doesn't vary with θ, we can store it along with the simulation draws as we search over θ.

Example 1 (continued). Suppose $F = \Phi$, so we're in the multinomial probit case. Let $V_{ij} = (Y_{i0}^* - Y_{ij}^*, \ldots, Y_{i,j-1}^* - Y_{ij}^*, Y_{i,j+1}^* - Y_{ij}^*, \ldots, Y_{iJ}^* - Y_{ij}^*)$ denote the vector of utility differences relative to choice j. Then V_{ij} conditional on X_i is multivariate normal with mean and variance depending on $\theta = (\beta, \Sigma)$. Let $f(V_{ij} \mid X_i, \theta)$ denote its pdf. Then

$$P_\theta(Y_{ij} = 1 \mid X_i = x) = \int \prod_{\ell=1}^J \mathbb{1}\{v_\ell \leq 0\} f(v \mid x, \theta)\,\mathrm{d}v$$
$$= \int \prod_{\ell=1}^J \mathbb{1}\{v_\ell \leq 0\} \frac{f(v \mid x, \theta)}{g(v \mid x)} g(v \mid x)\,\mathrm{d}v,$$

where g is a density supported on $(\mathbb{R}_-)^J$, such as the product of truncated normal variables or a product of exponentials. If we can draw $\{\tilde{v}_i^m\}_{m=1}^M$ from it, then we can estimate the CCP by

$$\frac{1}{M}\sum_{m=1}^{M}\frac{f(\tilde{v}_i^m \mid x,\theta)}{g(\tilde{v}_i^m \mid x)}.$$

Again, this will yield a smooth moment condition. ⊠

Remark 11. In some cases, it may be hard to compute $m(X_i,\epsilon_i,\theta)$. This is the case when Y_i is the equilibrium of a game, or solution to a dynamic program. It would then be very useful if we only had to compute the equilibrium once, and not every time we change θ as we're optimizing. See Ackerberg (2009) for how one can use change of variables coupled with importance sampling to avoid having to recompute the equilibrium each time.

Remark 12. Importance sampling is also a variance reduction method. Consider the problem of calculating

$$p(\theta) = \int h(\epsilon,\theta)f(\epsilon)\,\mathrm{d}\epsilon,$$

where f is a density. In Example 4, with $p(\theta) = P_\theta(Y_i = 1 \mid X_i = x)$, we had $h(\epsilon,\theta) = \mathbb{1}\{x'\theta + \epsilon \geq 0\}$, and f is the density of ϵ. A frequency simulator estimates this quantity as

$$\hat{p} = \frac{1}{M}\sum_{m=1}^{M} h(\epsilon^m,\theta), \qquad \operatorname{var}(\hat{p}) = \frac{1}{M}\int (h(\epsilon,\theta) - p(\theta))^2 f(\epsilon)\,\mathrm{d}\epsilon.$$

Suppose that we use importance sampling, and we simulate u^m from density g, yielding the estimate

$$\hat{p} = \frac{1}{M}\sum_{m=1}^{M}\frac{h(u^m,\theta)f(u^m)}{g(u^m)}, \qquad \operatorname{var}(\hat{p}) = \frac{1}{M}\int \left(\frac{h(u,\theta)f(u)}{g(u)} - p(\theta)\right)^2 g(u)\,\mathrm{d}u.$$

Notice that we can make the variance zero if we set $g(u) = g(u,\theta) = h(u,\theta)f(u)/p(\theta)$ (this integrates to one, so it's a density), which, unfortunately, requires knowledge of p. However, the variance expression suggests that we want to make $g(u)$ large if $f(u)h(u,\theta)$ is large, that is, make g large for those u's that contribute a lot to the integral: hence the name "importance sampling". Intuitively, we want to make the integrand $h(u,\theta)f(u)/g(u)$ to be as close to constant as possible. Other observations:

- The importance distribution does not have to be positive everywhere. It is enough to have $g(u) > 0$ whenever $f(u)h(u,\theta) \neq 0$ (we used this previously)
- We do need to prevent $h(u,\theta)f(u)/g(u)$ from getting very large (to keep the variance from exploding), so that the tails of $g(u)$ should be at least as thick as those of $h(u,\theta)f(u)$.

See Sauer and Taber (2021) for an application of a version of this idea to indirect inference.

This is a specific method evaluating the probability that a correlated multivariate normal random variable falls into a rectangular region. Since the choice probabilities in discrete choice models with normal errors (the panel probit model, or the multinomial probit model) take this form, it is commonly used in such models. The Geweke-Hajivassiliou-Keane (GHK) simulator is named after Geweke (1989) and Hajivassiliou and McFadden (1998) and Keane (1994). It reduces the problem of calculating a J-dimensional integral to a sequence of univariate integrals via a conditioning argument. It is a special version of importance sampling, and as such it has the advantage of producing simulated probabilities that are smooth functions of the model parameters.

GENERAL ALGORITHM We first describe the general algorithm. Suppose we want to evaluate the probability that $Z \sim \mathcal{N}_J(0, LL')$, falls into a rectangular region: $A = \{a_j \leq Z_j \leq b_j\}$. Here L is lower-triangular, corresponding to the Cholesky decomposition of the covariance matrix. Write $Z = L\epsilon$, so that $\{Z \in A\}$ is equivalent $\{\epsilon \in B\}$, with $B = \{L^{-1}a \leq \epsilon \leq L^{-1}b\}$.

Now, $\{a_j \leq Z_j \leq b_j\} = \{a_j \leq \sum_{k=1}^{j} L_{jk}\epsilon_j \leq b_j\}$. Therefore, $\epsilon_j \mid \epsilon_1, \ldots, \epsilon_{j-1}, B$, is a univariate standard normal, truncated to

$$\frac{a_j - \mu_j}{L_{jj}} \leq \epsilon_j \leq \frac{b_j - \mu_j}{L_{jj}}, \quad \mu_j = \begin{cases} \sum_{k=1}^{j-1} L_{jk}\epsilon_j & \text{if } j > 1, \\ 0 & \text{if } j = 1. \end{cases}$$

Hence, by the law of total probability, the density of ϵ given B is simply

$$\begin{aligned} g(\epsilon \mid B) &= g(\epsilon_1 \mid B)g(\epsilon_2 \mid \epsilon_1, B) \cdots g(\epsilon_J \mid \epsilon_1, \ldots, \epsilon_{j-1}, B) \\ &= \prod_{j=1}^{J} \frac{\phi(\epsilon_j)}{\Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})} \end{aligned}$$

Since $g(\epsilon \mid B)$ only has support over B, so that $g(\epsilon \mid B) = \mathbb{1}\{\epsilon \in B\}g(\epsilon \mid B)$, it follows that

$$\begin{aligned} P(Z \in A) &= \int \mathbb{1}\{\epsilon \in B\} \prod_j \phi(\epsilon_j) \mathrm{d}\epsilon \\ &= \int \left[\prod_{j=1}^{J} (\Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})) \right] g(\epsilon \mid B) \mathrm{d}\epsilon, \end{aligned}$$

where the second line follows by multiplying and dividing by $\prod_{j=1}^{J} \Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})$.

Since sampling from $g(\epsilon \mid B)$ is simple, we can evaluate $P(Z \in A)$ by sampling from this density, and using importance weights given in the square parentheses in the above expression. This is the GHK algorithm:

1. Draw $\tilde{\epsilon}_1^m, \ldots, \tilde{\epsilon}_{J-1}^m$ from the density $g(\epsilon \mid B)$. In particular, for $j = 1, \ldots, J-1$, draw $\tilde{\epsilon}_j^m$ from standard normal, truncated above at $(b_j - \tilde{\mu}_j^m)/L_{jj}$, and below at $(a_j - \tilde{\mu}_j^m)/L_{jj}$, where $\tilde{\mu}_j^m = \sum_{k=1}^{j-1} L_{jk}\tilde{\epsilon}_j^m$. This is done by drawing a $(J-1)$-vector $\tilde{U}^m$ of independent uniforms, and setting

$$\tilde{\epsilon}_j^m = \Phi^{-1}(\tilde{U}_j^m(\Phi((b_j - \tilde{\mu}_j^m)/L_{jj}) - \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})) + \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})).$$

2. Repeat the previous step M times, and form the estimate

$$\hat{P}(Z \in A) = \frac{1}{M}\sum_{m=1}^{M}\prod_{j=1}^{J}(\Phi((b_j - \tilde{\mu}_j^m)/L_{jj}) - \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})).$$

A detailed description, tailored to the multinomial probit model, can be found in Train (2009, Chapter 4).

MULTINOMIAL PROBIT Let us now show how this simulator can be applied to the multinomial probit model, as in Example 1. Let $Y_{ij}^* = V_{ij} + U_{ij}$ denote the utility from choice $j = 0, \ldots, J$, with $U_i = A\epsilon_i$, and $\epsilon_i \sim \mathcal{N}_{J+1}(0, I)$. Here $V_{ij} = X_{ij}'\beta$ denotes the non-stochastic part of the utility.

- The attractive feature of this model is that it doesn't restrict which choices are close: this allows for rich substitution patterns. In a random effects approach, for example, only choices that are close in terms of observables can be close.
- The problem is that if J is large, there are too many parameters in the covariance matrix to estimate.

VARIANCE NORMALIZATION First, we'll need some normalization on the covariance matrix $\Sigma = AA'$. Let A denote the Cholesky decomposition of Σ, so that A is lower triangular. Let $\tilde{U}_{ik} = U_{i,-k} - U_{ik}$ denote the differences relative to alternative k. Then $\tilde{U}_{ik} \sim \mathcal{N}_J(0, \tilde{\Sigma}_k)$, where $\tilde{\Sigma}_k = M_k\Sigma M_k'$, where M_k is the matrix $I_{J+1} - \iota_{J+1}e_k'$ with the kth row (which is all zeros), removed. Here ι is a $(J+1)$-vector of ones, and e_k is a vector of zeros with 1 in the kth position. For instance, if $J = 2$, then

$$M_0 = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}.$$

It is clear that since only utility differences matter, we could just parametrize the model in terms of $\tilde{\Sigma}_0$, say the utility differences with respect to the outside option. This is still not enough, since we need to normalize the scale of the utility. Typically, this is done by setting $\tilde{\Sigma}_{0,11} = 1$. To ensure $\tilde{\Sigma}_0$ is positive definite, it is convenient to parametrize the model terms of its Cholesky factor, the lower-triangular matrix A_0 with $A_{0,11} = 1$ so that $\tilde{\Sigma}_{0,11} = 1$. The remaining elements of A_0 can vary freely. So while Σ has $(J+1)J/2$ parameters, after normalization, we are left with $J(J-1)/2 - 1$ free parameters. This

means we can normalize $J+1$ parameters in the original matrix. A convenient choice is to set the first row and column to zero, in which case

$$\Sigma(A_0) = \begin{pmatrix} 0 & 0 \\ 0 & A_0 A_0' \end{pmatrix}', \qquad A_0 = \begin{pmatrix} 1 & 0 & \cdots & \\ a_{21} & a_{22} & 0 & \cdots \\ \vdots & & \ddots & \end{pmatrix}.$$

With this normalization, we can derive the matrix $\tilde{\Sigma}_j$ of covariances when we take utility differences with respect to any alternative j.

Example 5. For instance, with $J = 2$, we have 2 free parameters,

$$\Sigma = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & a_{12} \\ 0 & a_{12} & a_{12}^2 + a_{22}^2 \end{pmatrix}, \quad \tilde{\Sigma}_0 = A_0 A_0' = \begin{pmatrix} 1 & a_{12} \\ a_{12} & a_{12}^2 + a_{22}^2. \end{pmatrix}, \quad A_0 = \begin{pmatrix} 1 & 0 \\ a_{12} & a_{22} \end{pmatrix}.$$

and

$$\tilde{\Sigma}_1 = \begin{pmatrix} 1 & 1 - a_{12} \\ 1 - a_{12} & (1 - a_{12})^2 + a_{22}^2 \end{pmatrix}, \quad \tilde{\Sigma}_2 = \begin{pmatrix} a_{12}^2 + a_{22}^2 & -a_{12} + a_{12}^2 + a_{22}^2 \\ -a_{12} + a_{12}^2 + a_{22}^2 & (1 - a_{12})^2 + a_{22}^2 \end{pmatrix}. \quad \boxtimes$$

APPLYING THE GHK SIMULATOR The GHK simulator works with utility differences with respect to the choice the probability of which we are simulating. To simulate $p_j(\theta, X_i) := P_\theta(Y_{ij} = 1 \mid X_i)$, let $\tilde{\Sigma}_j = LL'$ denote the Cholesky decomposition, so that $\tilde{U}_{ij} = L\epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}_J(0, I)$, and $\theta = (\beta, A_0)$, and let $\tilde{V}_{ik} = V_{ik} - V_{ij}$. Given the normalization above, for a given matrix A_0, we can calculate L easily as the Cholesky decomposition of $M_j \Sigma(A_0) M_j'$.

$$p_j(\theta, X_i) = P(\tilde{V}_i + L\epsilon_i \leq 0) = P(L\epsilon_i \leq -\tilde{V}_i).$$

Thus, we apply the GHK simulator with $a = -\infty$, $b = -\tilde{V}_i$, L given by the Cholesky decomposition of $\tilde{\Sigma}_j$, and the uniform draws $\tilde{U}^m$ given by a $(J-1)$ vector of uniforms $\tilde{U}_{ij}^m$, which we keep fixed if we optimize over θ.

REFERENCES

Ackerberg, Daniel A. 2009. "A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation." *Quantitative Marketing and Economics* 7, no. 4 (December): 343–376. https://doi.org/10.1007/s11129-009-9074-z.

Armstrong, Tim, A Ronald Gallant, Han Hong, and Huiyu Li. 2017. "The Asymptotic Distribution of Estimators with Overlapping Simulation Draws." December. https:

//cpb-us-w2.wpmucdn.com/campuspress.yale.edu/dist/2/277/files/2017/12/olsim_dec2017-1fvorn4.pdf.

Berry, Steven T., James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63, no. 4 (July): 841–890. https://doi.org/10.2307/2171802.

Bruins, Marianne, James A. Duffy, Michael P. Keane, and Anthony A. Smith Jr. 2018. "Generalized Indirect Inference for Discrete Choice Models." *Journal of Econometrics* 205, no. 1 (July): 177–203. https://doi.org/10.1016/j.jeconom.2018.03.010.

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.

Gallant, A Ronald, and George Tauchen. 1996. "Which Moments to Match?" *Econometric Theory* 12, no. 4 (October): 657–681. https://doi.org/10.1017/S0266466600006976.

Geweke, John. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57, no. 6 (November): 1317–1339. https://doi.org/10.2307/1913710.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. "Indirect Inference." *Journal of Applied Econometrics* 8, no. S1 (December): S85–S118. https://doi.org/10.1002/jae.3950080507.

Hajivassiliou, Vassilis A., and Daniel L. McFadden. 1998. "The Method of Simulated Scores for the Estimation of LDV Models." *Econometrica* 66, no. 4 (July): 863–896. https://doi.org/10.2307/2999576.

Hong, Han, Huiyu Li, and Jessie Li. 2021. "BLP Estimation Using Laplace Transformation and Overlapping Simulation Draws." *Journal of Econometrics* 222, no. 1A (May): 56–72. https://doi.org/10.1016/j.jeconom.2020.07.026.

Hong, Han, Aprajit Mahajan, and Denis Nekipelov. 2015. "Extremum Estimation and Numerical Derivatives." *Journal of Econometrics* 188, no. 1 (September): 250–263. https://doi.org/10.1016/j.jeconom.2014.05.019.

Huber, Joel, and Kenneth Train. 2001. "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths." *Marketing Letters* 12 (3): 259–269. https://doi.org/10.1023/A:1011120928698.

Keane, Michael P. 1994. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica* 62, no. 1 (January): 95–116. https://doi.org/10.2307/2951477.

Kloek, T., and Herman K. van Dijk. 1978. "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo." *Econometrica* 46, no. 1 (January): 1–19. https://doi.org/10.2307/1913641.

Lerman, Stveven R., and Charles F. Manski. 1981. "On the Use of Simulated Frequencies to Approximate Choice Probabilities." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by Charles F. Manski and Daniel L. McFadden, 305–319. Cambridge: MIT Press. https://eml.berkeley.edu/~mcfadden/discrete.html.

McFadden, Daniel L. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration." *Econometrica* 57, no. 5 (September): 995–1026. https://doi.org/10.2307/1913621.

McFadden, Daniel L., and Kenneth E. Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15, no. 5 (September): 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Chap. 36 in *Handbook of Econometrics*, edited by Robert F. Engle and Daniel L. McFadden, 4:2111–2245. New York, NY: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.

Pakes, Ariel, and David Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57, no. 5 (September): 1027–1057. https://doi.org/10.2307/1913622.

Sauer, Robert M., and Christopher Taber. 2021. "Understanding Women's Wage Growth Using Indirect Inference with Importance Sampling." *Journal of Applied Econometrics* 36, no. 4 (June): 453–473. https://doi.org/10.1002/jae.2818.

Smith, Anthony A. 1993. "Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions." *Journal of Applied Econometrics* 8, no. S1 (December): S63–S84. https://doi.org/10.1002/jae.3950080506.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge, UK: Cambridge University Press. https://eml.berkeley.edu/books/choice2.html.

SYNTHETIC CONTROLS

Michal Kolesár*

April 29, 2024

1. BASICS OF THE SYNTHETIC CONTROL METHOD

To understand the strengths and weaknesses of any causal inference method, it is useful to break it down into two related pieces: (a) what is the precise estimand—the precise counterfactual that the method targets, and (b) how is the counterfactual constructed. Constructing the counterfactual is an extrapolation exercise, and it is useful consider the functional form restrictions, as well as the time periods or units that we are using for the extrapolation. For instance, we may be less concerned with linearity restrictions if we are only extrapolating over a short run.

Consider a simple difference-in-differences (DiD) setup with two time periods $t = 0, 1$ and two groups of units: control units that never get treated, and treated units that start treatment in period 1. We have seen than the DiD estimand corresponds to the average treatment effect for the treated (ATT), so that we are only constructing a counterfactual $Y_{i1}(0)$ for the treated units. Furthermore, the estimator the estimator can be decomposed as $\hat{\beta} = N_1^{-1} \sum_{i:\, D_{i1}=1} \hat{\tau}_i$, with $N_d = \sum_{i=1}^n \mathbb{1}\{D_{i1} = d\}$ denoting the number of units in each group, and

$$\hat{\tau}_i = Y_{i1} - \hat{Y}_{i1}(0), \quad \hat{Y}_{i1}(0) = \underbrace{\left(Y_{i0} - \frac{1}{N_0} \sum_{j:\, D_{j1}=0} Y_{j0} \right)}_{\hat{\mu}_i} + \frac{1}{N_0} \sum_{i:\, D_{j1}=0} Y_{j1}, \tag{1}$$

In other words, the DiD estimator puts equal weight $w_i = 1/N_0$ on each of the control units, and estimates the counterfactual outcome as $\hat{Y}_{it}(0) = \hat{\mu}_i + \sum_i w_i Y_{it}$, where $\hat{\mu}_i$ captures the differential level of outcome between the treated unit i and the control units. Equivalently, we're using the average change in the outcomes $\frac{1}{N_0} \sum_j (Y_{j1} - Y_{j0})$ to estimate $Y_{i1}(0) - Y_{i0}(0)$. If there are multiple post-treatment periods, and we use an event-study approach, the impact of the treatment in the later periods is estimated analogously, obtaining a dynamic treatment effect path.

*Email: mkolesar@princeton.edu.

Table 1: Difference-in-differences estimate on unemployment rates (African-American workers). Adapted from Card (1990).

	Year		
	1979	1981	Difference
Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
Comparison Cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
Difference	−2.0 (1.9)	−3.0 (2.0)	−1.0 (2.8)

Imposing that each control unit has the same weight may be restrictive: after all, in cross-section methods for estimating the average treatment effect (ATE) or ATT under unconfoundedness, regression as well as propensity score weighting methods weight the control units by how similar their covariates are to the covariates of the treated units. Similarly, in regression discontinuity (RD), units farther away from the cutoff receive little or no weight in constructing the counterfactual.

The synthetic control (SC) estimator of Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) provides a data-driven procedure to create a comparison unit in comparative case studies. The method is based on the observation that a weighted average of control units (a "synthetic control") often does a "better job" reproducing the characteristics of a treated unit than a simple average of control units (as DiD estimators do), and a better job than picking a single comparison unit (as nearest neighbor matching estimators do). As such, the ideas behind the SC method have a lot of potential, which is why Athey and Imbens (2017) call it "arguably the most important innovation in the policy evaluation literature in the last 15 years".

To motivate this method, consider the setup in Card (1990). Card is interested in the effect of the Mariel boatlift on the labor market outcomes of native workers in Miami. The Mariel boatlift was a mass emigration of Cubans, who traveled from Cuba's Mariel Harbor to the United States between April and October 1980, after Castro declared on April 20 that Cubans wishing to emigrate to the US were free to do so.

Using a difference-in-differences design, Card (1990) concludes that in spite of increasing the Miami labor force by 7%, the influx of immigrants did not have an effect on wages and unemployment rates of less-skilled native workers. In particular, Card uses four other cities in the south of the United States (Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg) as a control group to approximate the change in native unemployment rates that would have been observed in Miami in the absence of the Mariel Boatlift. The summary of his analysis is in Table 1.

However, one may be concerned that the conclusions are sensitive to how the comparison units were chosen: is there a way of picking the comparison units in a less ad hoc fashion? This is the motivation for the reanalysis of the data in Peri and Yasenov (2019). Indeed, one can think of the SC method as trying to make the control group selection less ad hoc. If we can do that, then we can also hope to formalize the uncertainty about the ability of the control group to reproduce the counterfactual of interest.

1.1. Setup

Suppose there are $N_0 + 1$ units, $i = 0, \ldots, N_0$, with unit 0 exposed to intervention in periods $T_0 + 1, \ldots, T$. Units $1, \ldots, N_0$ are the "donor pool" in the sense that they are potential controls. Our goal is to estimate $\tau_{0t} = Y_{0t}(1) - Y_{0t}(0)$ for $t > T_0$. Let X_i denote the vector of pre-intervention characteristics of unit i. Typically, $X_i = (Z_i, Y_{i1}, \ldots, Y_{iT_0})$, where Z_i is a vector of fixed characteristics.

Similarly to the DiD case, we do not need to have a notion of potential outcomes $Y_{it}(1)$ for the control units. For example, Abadie, Diamond, and Hainmueller (2015) are interested in estimating the economic impact of German reunification on West Germany, using a donor pool of 16 OECD counties. It is hard to conceptualize what it would mean for these countries to "reunify".

The SC estimator constructs a single synthetic control unit as a weighted average of the units from the donor pool. These weights w^* minimize the discrepancy between the characteristics X_0 of the treated unit, and the characteristics of the synthetic control:

$$w^* = \underset{w}{\operatorname{argmin}} \|X_0 - \sum_i w_i X_i\| = \|X_0 - X'w\| \qquad \text{st} \quad \sum_{i=1}^{N_0} w_i = 1, \quad w_i \geq 0. \tag{2}$$

The synthetic control estimator is

$$\hat{\tau}_{0t} = Y_{0t} - \hat{Y}_{0t}(0), \qquad \hat{Y}_{0t}(0) = \sum_i w_i^* Y_{it}. \tag{3}$$

If there is more than one treated unit, we can estimate the treatment effect of each of them in the same way.

Remark 1. The restriction that the weights are positive plays three roles. First, it "precludes extrapolation" in the sense that the predicted counterfactual outcome $Y_{0t}(0)$ is always in the convex hull of the outcomes for the control units. However, allowing for extrapolation may be important if the treated unit is substantively different from the control units: in such cases, extrapolation may be necessary. Second, this restriction also allows the procedure to obtain unique weights even when the number of lagged outcomes is small relative to the number of control units, a common setting in applications, which aids interpretation. Finally, it ensures that, if X_0 doesn't belong to the convex null of the rows of X, then w^* is unique and sparse: thus, the restriction serves to regularize the estimator. In particular, if $w_i \geq 0$, then the synthetic control is the closest point to X_0 from the convex hull of $X_1, \ldots, X_{N_0}$. But points on the surface of a convex hull are convex combinations of a few units: at most $\dim(X_i)$ if $X_1, \ldots, X_{N_0}$ are in a "general position" (A set of points in a k-dimensional Euclidean space is in general position if no m of them lie in a $(m-2)$-dimensional hyperplane for $k = 2, \ldots, p+1$). This sparsity makes the counterfactual outcome estimate very transparent, a big reason for the method's popularity.

Remark 2 (Comparison with DiD). Comparing eq. (1) with eq. (3), we see that there are

two key differences between DiD and SC. First, SC doesn't allow for a non-zero intercept in the counterfactual prediction $\hat{Y}_{it}(0) = \mu + \sum_i w_i Y_{it}$: this is a critical feature of the DiD approach, allowing for permanently different levels for different units. On the other hand, DiD restricts the weights w_i to be equal: this is the key restriction that SC relaxes.

Doudchenko and Imbens (2017) explore this perspective further. Taking a machine learning perspective, they view the problem of constructing the counterfactual outcome $Y_{it}(0)$ as a prediction problem. Instead of imposing that the weights are non-negative, sum to one, and that there is no intercept, as the SC method does, for the case where X_i only consists of pre-treatment outcomes, they propose minimizing an elastic-net objective function $\|X_i - \mu - X'w\|_2^2 + \lambda((1-\alpha)\|w\|_2^2/\alpha + \alpha\|w\|_1)$.

To implement the SC method, one needs to pick the norm $\|\cdot\|$ in eq. (2). Abadie, Diamond, and Hainmueller (2015) use the weighted Euclidean norm $\|a\| = a'Va$, where V is a diagonal matrix with weights chosen by a split-sample method: divide pre-pretreatment period into a training and validation period. Then pick the weights to minimize the prediction error in the validation period. That is, let $t = 1, \ldots, T_0/2$ denote the training periods, and let $t = T_0/2 + 1, \ldots, T_0$ denote the validation period. Compute the mean squared prediction error (MSPE) $\sum_{t=T_0/2+1}^{T_0}(Y_{0t} - \sum_i w_i(V)Y_{it})^2$ over the validation period, using weights computed based on $X_i = (Z_i, Y_{i1}, \ldots, Y_{i,T_0/2})$ and minimizing eq. (2) over the training period. Minimize this over V, yielding V^*. Then use the resulting V^* along with the data $X_i = (Z_i, Y_{i,T_0/2}, \ldots, Y_{i,T_0})$ for the last $T_0/2$ periods before the intervention to estimate the $w^*(V^*)$.

Alternatively, Abadie and Gardeazabal (2003) propose choosing V to minimize the MSPE of the outcome variable for the pre-intervention periods, that is, they $\sum_{t=1}^{T_0}(Y_{0t} - \sum_i w_i^*(V)Y_{it})^2$ over V: if the only covariates in X_i are pre-intervention outcomes, this amounts to just using $V = I_{T_0}$.

Ultimately, the how V should be chosen depends on the relative importance of elements of X_0 as a predictor of the counterfactual post-intervention outcomes. Since the counterfactual outcomes are unobserved, the split-sample approach approximates them using the pre-intervention data.

We illustrate this approach using data from Abadie, Diamond, and Hainmueller (2015), who are interested in estimating the effect of the 1990 German reunification on the GDP of West Germany. The data spans 1960–2003, and the donor pool consists of 16 OECD countries. The covariates also include the inflation rate, trade openness (sum of exports plus imports as a percentage of GDP), industry share of value added (ten-year average), high school completion rate (five-year frequency), and ratio of domestic investment to GDP (five-year average). Years 1971–1980 are used to select the weights V. Of the 16 donors, only 5 receive weights greater than 0.01: USA 0.219, Austria 0.418, Netherlands 0.090, Switzerland 0.111 and Japan 0.155. Remarkably reasonable choices. Figure 1 plots the counterfactual estimates.

A note of caution when interpreting these types of figures: compared to DiD methods, the pre-treatment fit in these figures always looks remarkably good: but of course, this is because we choose the weights precisely to fit the pre-treatment outcome path!

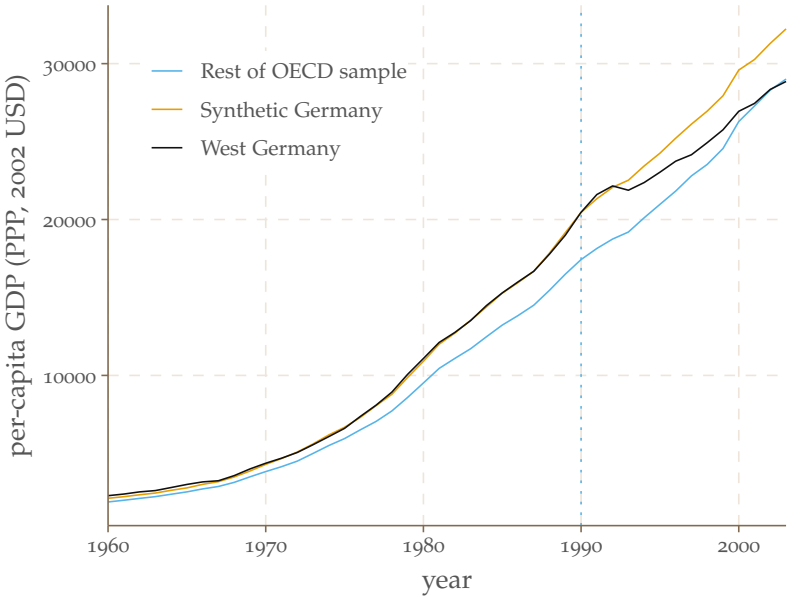


Figure 1: SC method estimates of the effect of German reunification.

This issue is magnified in this example because GDP is not stationary. For comparison, Figure 2 plots GDP growth rates, and we see that the fit is not quite as impressive. If we target GDP growth rates in constructing the synthetic weights, we get quite similar comparison group, with just Switzerland dropping out and replaced by Japan (USA 0.169; Austria 0.448; Netherlands 0.138; Japan 0.245). Figure 3 shows the counterfactual estimates. The fit does not improve, the root MSPE is about 0.5 percentage points for the period 1980–1989, and both ways of estimating the treatment effect yield root MSPE of about 1.5 percentage points post-intervention.

Remark 3. Similar to DiD studies, by saying that $Y_{it} = Y_{it}(0)$ if $t \leq T_0$ or if $i \geq 1$, we are assuming Stable unit treatment value assumption (SUTVA), which rules out spillover effects (a strong assumption for the German reunification example) and anticipation effects.

1.2. *Falsification*

- We can move the treatment date forward in time. For instance, we can suppose the reunification happened in 1980 and estimate the effect on 1981–1990 GDP

While the method has been increasingly popular, part of the limitation is that the assumptions underlying are not easily testable, for two reasons. First is that it's not entirely clear what the key underlying assumption is in the first place. In contrast, the parallel trends assumption in DiD is very clear. Second, there appear to be more researcher degrees of freedom, so it's a bit easier to adjust the estimator so we pass any

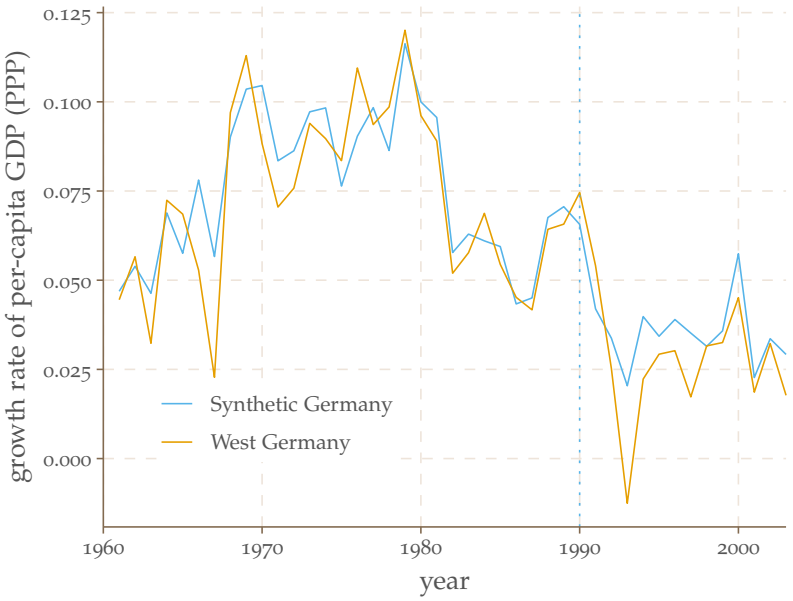


Figure 2: SC method estimates of the effect of German reunification on GDP growth rates. Method fitted on GDP.

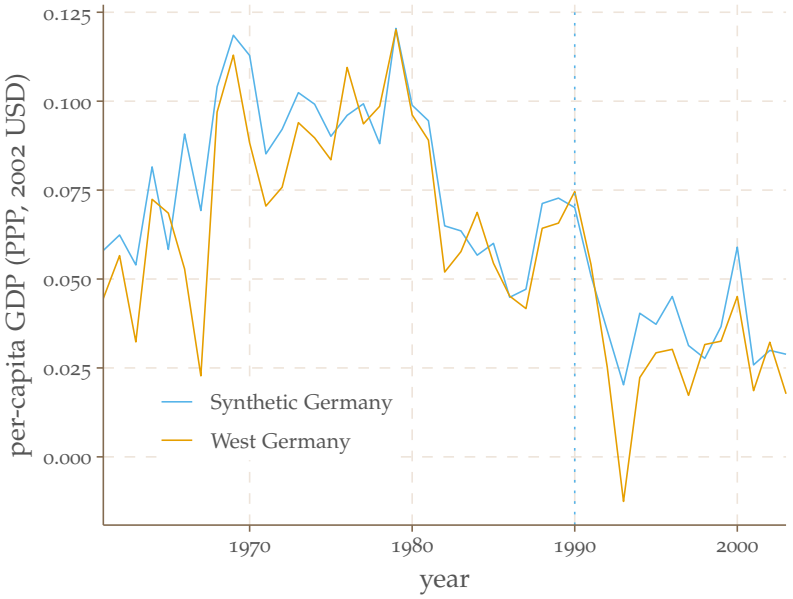


Figure 3: SC method estimates of the effect of German reunification. Method fitted on GDP growth rates.

placebo checks.

1.3. *Formal properties*

There is only one formal result available:

Proposition 4 (Abadie, Diamond, and Hainmueller 2010). Suppose $Y_{it}(0)$ is given by a factor model

$$Y_{it}(0) = \delta_t + Z_i'\theta_t + \lambda_t'\mu_i + \epsilon_{it},$$

where Z_i are observed covariates, λ_t are common factors, and μ_i are factor loadings. Suppose that the smallest eigenvalue of $\sum_{t=1}^{T_0} \lambda_t'\lambda_t / T_0$ is bounded away from zero, that $E|\epsilon_{it}|^p$ is bounded, and that that errors ϵ_{it} are independent of the factors, mean zero, and independent across time. Suppose also that λ_t has bounded support.

Suppose that the weights w satisfy $Z_0 = \sum_i w_i Z_i$ and $Y_{0t} = \sum_i Y_{it}$ for $t = 1, \ldots, T_0$. Then the prediction bias is bounded by

$$|E[Y_{0t}(0) - \sum_i w_i Y_{it}]| \leq K \frac{N_0^{1/p}}{T_0} \max\{(T_0 \overline{m}_p^{1/p}, (T_0 \overline{m}_2)^{1/2}\}.$$

where $K = C(p)^{1/p} \frac{\dim(\lambda_t) \max_{sj} |\lambda_{sj}|}{\min \text{eig}((\lambda^{P\prime}\lambda^P / T_0))}$, $C(p)$ is a constant depending only on p given in Ibragimov and Sharakhmetov (2002), and $\overline{m}_p = \max_i T_0^{-1} \sum_{t=1}^{T_0} E|\epsilon_{it}|^p$.

Proof. Using the factor model, we can write.

$$Y_{0t}(0) - \sum_j w_j Y_{jt}(0) = (Z_0 - \sum_j w_j Z_j)'\theta_t + \lambda_t'(\mu_0 - \sum_j w_j \mu_j) + \sum_j w_j(\epsilon_{0t} - \epsilon_{jt}). \tag{4}$$

Let Y_j^P denote the vector of pre-intervention outcomes, and let λ^P denote the matrix of pre-intervention factors. Then under the assumption that we match Y_0^p and Z_0, we have

$$0 = \lambda^P(\mu_0 - \sum_j w_j \mu_j) + \sum_j w_j(\epsilon_0^P - \epsilon_j^P) = \lambda^P(\mu_0 - \sum_j w_j \mu_j) - \sum_j w_j \epsilon_j^P + \epsilon_0^P \tag{5}$$

Multiplying eq. (5) by $\lambda_t'(\lambda^{P\prime}\lambda^P)^{-1}\lambda^{P\prime}$, and subtracting it from eq. (4), we obtain

$$Y_{0t}(0) - \sum_j w_j Y_{jt}(0) = \sum_j w_j \left[\lambda_t'(\lambda^{P\prime}\lambda^P)^{-1}\lambda^{P\prime}\epsilon_j^P\right] + \sum_j w_j(\epsilon_{0t} - \epsilon_{jt}) - \lambda_t'(\lambda^{P\prime}\lambda^P)^{-1}\lambda^{P\prime}\epsilon_0^P.$$

If $t > T_0$, then the second term has mean zero. Furthermore, the third term always has mean zero since ϵ_{0t}^P is not correlated with the factors. Write

$$\sum_j w_j \left[\lambda_t'(\lambda^{P\prime}\lambda^P)^{-1}\lambda^{P\prime}\epsilon_j^P\right] = \sum_j w_j v_j,$$

where $v_j = \sum_{s=1}^{T_0} \kappa_{js}\epsilon_{js}$, with $\kappa_{ts} = \lambda_t'(\lambda^{P\prime}\lambda^P)^{-1}\lambda_s$. Now, by Hölder's inequality, since $\sum_j w_j^q \leq$

$\sum_j w_j = 1$, we have $|\sum_j w_j v_j| \leq (\sum_j |v_j|^p)^{1/p}$. Therefore, by Jensen's inequality,

$$E|\sum_j w_j v_j| \leq (\sum_j E|v_j|^p)^{1/p}.$$

Furthermore, by Cauchy-Schwarz inequality,

$$|\kappa_{ts}| \leq \max_s \lambda_s (\lambda^{P\prime}\lambda^P)^{-1}\lambda_s \leq \max \text{eig}((\lambda^{P\prime}\lambda^P)^{-1}) \dim(\lambda_t) \max_{sj} |\lambda_{sj}| = \frac{\dim(\lambda_t) \max_{sj}|\lambda_{sj}|}{T_0 \min \text{eig}((\lambda^{P\prime}\lambda^P/T_0))}$$

Now, by Rosenthal's inequality[1], letting $\overline{m}_p = \max_j T_0^{-1} \sum_{s=1}^{T_0} E[|\epsilon_{js}|^p]$

$$E|v_j|^p \leq C(p) \max_{ts} |\kappa|_{ts}^p \max\{T_0 \overline{m}_p, (T_0 \overline{m}_2)^{p/2}\}.$$

Hence,

$$E|\sum_j w_j v_j| \leq N_0^{1/p} C(p)^{1/p} (\max_{ts} |\kappa|_{ts}^p)^{1/p} \max\{T_0^{1/p} \overline{m}_p^{1/p}, (T_0 \overline{m}_2)^{1/2}\},$$

which yields the result. □

Some notes on this result:

1. The factor model is more flexible than assuming constant λ_t as DiD methods do. However, the result only gives a bound on the bias: it doesn't guarantee an unbiased estimate. Further, the constants in this bound are not known: we cannot use it directly for inference.

2. The result requires that we match X_0 exactly. If we do, and T_0 is large enough so that $T_0 \overline{m}_p < (T_0 \overline{m}_2)^{p/2}$, then the bias is of the order $N_0^{1/p} \overline{m}_2^{1/2} / T_0^{1/2}$. In other words, if N_0 is small relative to T_0, and we still manage to match X_0 exactly, then the bias is likely to be small. The intuition is that matching X_0 exactly means that we must be approximately matching the factor loadings μ_0.

   In other words, SC estimates are most credible if we ex ante reduce the donor pool to units comparable to unit 0, and have a sizable number of preintervention periods: otherwise tracking the treated unit's characteristics and outcomes over the pre-treatment period is not indicative of good post-treatment performance.

3. If we really believed this factor model, why not directly estimate it?

1.4. *Inference*

If treatment randomly assigned given X (a strong assumption!), can reassign the treatment to a random unit and recompute the treatment effect over $t = T_0 + 1, \ldots, T$. We can

1. Let Z_i be independent, mean zero, with $E[|Z_i|^t] < \infty$ for some $t > 2$. Then

$$E[|\sum_i Z_i|^t] \leq C(t) \max\{\sum_i E[|Z_i|^t], (\sum_i E[Z_i^2])^{t/2}\}$$

where $C(t)$ is a constant. See Ibragimov and Sharakhmetov (2002).

plot these estimates to get a visual sense of how unusual the actual treated unit is (see Figure 4 in Abadie, Diamond, and Hainmueller (2010)). This method also delivers an exact p-value by comparing say the MSPE for the treated periods. As a test statistic, Abadie, Diamond, and Hainmueller (2010) propose using the ratio between pre-intervention and post-intervention MSPE,

$$r_j = \frac{\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}(Y_{it} - \hat{Y}_{it}(0))^2}{\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_{it} - \hat{Y}_{it}(0))^2}.$$

A related approach, suggested in Doudchenko and Imbens (2017) is to view the treatment unit as exchangeable with the control units in the absence of the treatment. So we can try to estimate $Y_{it}(0)$ for units $i \geq 1$ using the same estimator: drop unit 0 from the sample and let units other than i constitute the donor pool. Make a prediction $\hat{Y}_{it}(0)$. Do this for each unit i. Then we can estimate the variance of $\hat{Y}_{0t}(0)$ as $N_0^{-1}\sum_{i=1}^{N_0}(\hat{Y}_{it}(0) - Y_{it})^2$.

Because in most observational settings assignment to the intervention is not randomly assigned, the interpretation of these approaches in practice is not entirely clear. It is also unclear how one could construct a confidence interval for our predictions.

2. EXTENSIONS

In the SC prediction problem, there are three groups of control outcomes: pre-intervention outcomes for the treated unit, and pre- and post-intervention outcomes for the donor pool. The DiD method combines three groups of control units in a simple fashion, justified by a two-way fixed effects (2WFE) model, $Y_{it}(0) = \alpha_i + \lambda_t + \epsilon_{it}$. Arkhangelsky et al. (2021) augment the objective function to allow unit weights $\hat{w}_i$ and time weights $\hat{\lambda}_t$

$$\sum_{it}(Y_{it} - \mu - \alpha_i - \beta_t - D_{it}\tau)^2\hat{w}_i\hat{\lambda}_t.$$

The unit weights are chosen to align pre-exposure trends of the treated and untreated, $\sum_{i=1}^{N_0}\hat{\omega}_i Y_{it} \approx \frac{1}{N_1}\sum_{i=N_0+1}^{n} Y_{it}$, while the time weights balance pre- and post-exposure time periods $\sum_{i=1}^{N_0}\sum_t \lambda_t Y_{it} \approx \frac{1}{T-T_0}\sum_i\sum_{t=T_0+1}^{T} Y_{it}$. In contrast, one can show that the SC estimator also minimizes the above objective if $\hat{\lambda}_t = 1$, subject to the constraint that $\alpha_i = 0$.

The justification for their procedure is provided by a factor model for $Y_{it}(0)$. They reason that because they are interested in treatment effects, rather than recovering the factor structure, they may achieve consistency for the treatment effects under weaker assumptions than those in Bai (2009).

Research Question. While the assumptions imposed in Bai (2009) are indeed stronger, is it the case that *consistency* of the factor-model estimators for treatment effects estimation does indeed require stronger assumptions? ⊠

- See JASA special issue Volume 116, Issue 536 that contains a number of articles building on the basic SC estimator.

REFERENCES

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105, no. 490 (June): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

———. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, no. 2 (February): 495–510. https://doi.org/10.1111/ajps.12116.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93, no. 1 (March): 113–132. https://doi.org/10.1257/000282803321455188.

Arkhangelsky, Dmitry, Susan Athey, David Hirshberg, Guido Imbens, and Stefan Wager. 2021. "Synthetic Difference In Differences." *American Economic Review* 111, no. 112 (December): 4088–4118. https://doi.org/10.1257/aer.20190159.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31, no. 2 (May): 3–32. https://doi.org/10.1257/jep.31.2.3.

Bai, Jushan. 2009. "Panel Data Models With Interactive Fixed Effects." *Econometrica* 77, no. 4 (July): 1229–1279. https://doi.org/10.3982/ECTA6135.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43, no. 2 (January): 245–257. https://doi.org/10.2307/2523702.

Doudchenko, Nikolay, and Guido W. Imbens. 2017. "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," September. arXiv: 1610.07748.

Ibragimov, R., and Sh. Sharakhmetov. 2002. "The Exact Constant in the Rosenthal Inequality for Random Variables with Mean Zero." *Theory of Probability & Its Applications* 46, no. 1 (January): 127–132. https://doi.org/10.1137/S0040585X97978762.

Peri, Giovanni, and Vasil Yasenov. 2019. "The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift." *Journal of Human Resources* 54 (2): 267–309. https://doi.org/10.3368/jhr.54.2.0217.8561R1.