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.