Cluster-robust inference: A guide to empirical practice

(James G. MacKinnon, Morten Ørregaard Nielsen & Matthew D. Webb)

Journal of Econometrics, 2023

(*Disclaimer: These slides include extra materials and personal understanding beyond the paper's scope. They may contain errors and are meant solely for discussion purposes.*)

Presentor: Zeyu CHEN

School of Economics, Renmin University of China

Last updated： 2023–07–18

Outline

Outline

Preliminary

- Model: $y = X\beta + u$, with $\mathrm{E}[u|X] = 0$, $\mathrm{Var}[u] = \mathrm{E}[uu'] = \Omega$
- OLS estimator:
$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$$
- The variance of $\hat{\beta}$:
$$\mathrm{Var}[\hat{\beta}] = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$
- Non-clustered assumption:
  - Spherical disturbance: $\Omega = \sigma^2 I$, then $\mathrm{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$. The consistent estimator is
$$\widehat{\mathrm{Var}}[\hat{\beta}] = s^2(X'X)^{-1}, \text{ where } s^2 = \hat{u}'\hat{u}/(n-k)$$
  - Heteroskedasticity: Ω is diagonal. The heteroskedasticity-robust estimator is (White, 1980):
$$\widehat{\mathrm{Var}}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{u}_i^2\right)(X'X)^{-1}$$

Preliminary: Clustered structure

- In many cases, we expect that Ω is not diagonal but with a clustered structure. Suppose there are G clusters and the g-th cluster has N_g observations, then Ω can be regarded as a diagonal partitioned matrix:

$$\Omega = \text{Var}[u] = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix} \tag{1}$$

- The model can be re-written as

$$y_g = X_g\beta + u_g,\ g = 1, \cdots, G,$$

with the OLS estimator

$$\hat{\beta} - \beta = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g'u_g\right) = (X'X)^{-1}\sum_{g=1}^{G} s_g,$$

where $s_g \equiv X_g'u_g$.

Preliminary: Clustered structure

- Similarly, the variance of $\hat{\beta}$ can be re-written as:

$$\text{Var}[\hat{\beta}] = (X'X)^{-1}\text{Var}\left[\sum_{g=1}^{G} s_g\right](X'X)^{-1}$$

- Define $\Sigma_g \equiv \text{Var}[s_g] \equiv \text{Var}[X_g' u_g]$. We assume errors within the same cluster are correlated but errors from different clusters are uncorrelated. That is, $\forall g \neq g'$, we have $\text{E}[s_g s_{g'}'] = 0$. Therefore,

$$\text{Var}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G} \Sigma_g\right)(X'X)^{-1} \tag{2}$$

Remarks: Since the clustered structure is a more generalized setting, equation (2) also applies to the spherical and heteroskedastic structures.

The consequence of ignoring the clustered structure

- Re-write s_g as the sum of the sub-samples in cluster g:

$$s_g \equiv X_g' u_g = \sum_{i=1}^{N_g} X_{gi} u_{gi} \equiv \sum_{i=1}^{N_g} s_{gi},$$

then one can prove that

$$s_g s_g' = \left(\sum_{i=1}^{N_g} X_{gi} u_{gi} \right) \left(\sum_{i=1}^{N_g} X_{gi} u_{gi} \right)' = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} s_{gi} s_{gj}'$$

- Therefore,

$$\Sigma_g = \mathrm{E}[s_g s_g'] = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \Sigma_{g,ij},$$

where $\Sigma_{g,ij} \equiv \mathrm{E}[s_{gi} s_{gj}']$. Here, we express Σ_g as the sum of the correlation of each sample-pair in cluster g.

The consequence of ignoring the clustered structure

- Further,

$$\Sigma_g = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\Sigma_{g,ij} = \sum_{i=1}^{N_g}\Sigma_{g,ii} + \sum_{i=1}^{N_g}\sum_{j\neq i}\Sigma_{g,ij} \tag{3}$$

  If we assume a heteroskedastic error structure (i.e., no intra-cluster correlation, $\forall i \neq j$, $\Sigma_{g,ij} = 0$), it is equivalent to setting the last term in (3) to 0.

- Error terms within the same cluster are typically positively correlated because these samples share common (uncontrolled) characteristics. Consequently, the last term in Equation (3) is usually positive. This implies that setting the last term to 0 would result in an underestimation of Σ_g, and consequently, the standard error.

Three cluster-robust standard errors

- We have shown that the variance of the OLS estimator is

$$\text{Var}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G} \text{E}[s_g s_g']\right)(X'X)^{-1}, \text{ where } s_g \equiv X_g' u_g$$

- Only s_g (specifically, u_g) is unknown, so we need to construct a consistent estimator for s_g. We have three ways to replace u_g in s_g:
  - predicted residuals: $\hat{s}_g = X_g' \hat{u}_g$
  - standardized residuals (transformed residuals): $\grave{s}_g = X_g' M_{gg}^{-1/2} \hat{u}_g$
  - leave-one-out residuals (jackknife residuals): $\acute{s}_g = X_g' M_{gg}^{-1} \hat{u}_g$
- Use the predicted residuals to obtain CRVE_1 (Type-1 Cluster-Robust Variance Estimator), which is the default option when using `cluster()` in `Stata`:

$$\text{CRVE}_1\text{: } \frac{G(n-1)}{(G-1)(n-k)}(X'X)^{-1}\left(\sum_{g=1}^{G} \hat{s}_g \hat{s}_g'\right)(X'X)^{-1}$$

Three cluster-robust standard errors

- Use the standardized residuals to obtain CRVE_2:

$$\text{CRVE}_2\text{: } (X'X)^{-1}\left(\sum_{g=1}^{G} \grave{s}_g \grave{s}_g'\right)(X'X)^{-1}$$

- Use the leave-one-out residuals to obtain CRVE_3:

$$\text{CRVE}_3\text{: } \frac{G-1}{G}(X'X)^{-1}\left(\sum_{g=1}^{G} \acute{s}_g \acute{s}_g'\right)(X'X)^{-1}$$

Three cluster-robust standard errors

- It is recommended to use CRVE_3 when sample size is small, since CRVE_3 has better small-sample properties (it is unbiased) and is relatively conservative.
- How to use CRVE_3 in `Stata`?
  1. Add `vce(jackknife, cluster())` in option (available for `areg` but not for `reghdfe`):
     `> areg y x, ... vce(jackknife, cluster(cluster_var))`
  2. Add `jackknife:` before the command and set `cluster()` in option (available for both `reghdfe` and `areg`):
     `> jackknife: reghdfe y x, ... cluster(cluster_var)`
  3. Use `summclust` to estimate CRVE_3 faster.

Outline

Large sample with clustered structure

- "Using large samples in econometrics" (MacKinnon, 2023, *JoE*)
  - Large sample theory tells us that as the sample size increases, the variance of the estimate decreases and eventually converges by probability. For example, if we care about the sample mean of $y_i \sim_{\text{i.i.d.}} (\theta, \sigma^2)$, its variance is

$$\text{Var}[\bar{y}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[y_i] = \frac{\sigma^2}{n}$$

  - However, if samples are correlated, then the variance of sample mean is

$$\text{Var}[\bar{y}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[y_i] + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \text{Cov}(y_i, y_j),$$

    where the first term is $O(1/n)$ but the second term is $O(1)$. Therefore, if the second term is not 0, the speed of convergence will largely decrease (or even divergent).
  - In fact, this is not a new issue; the contribution of this paper is to enhance the computational efficiency within this context (but let's skip it now).

Large sample with clustered structure

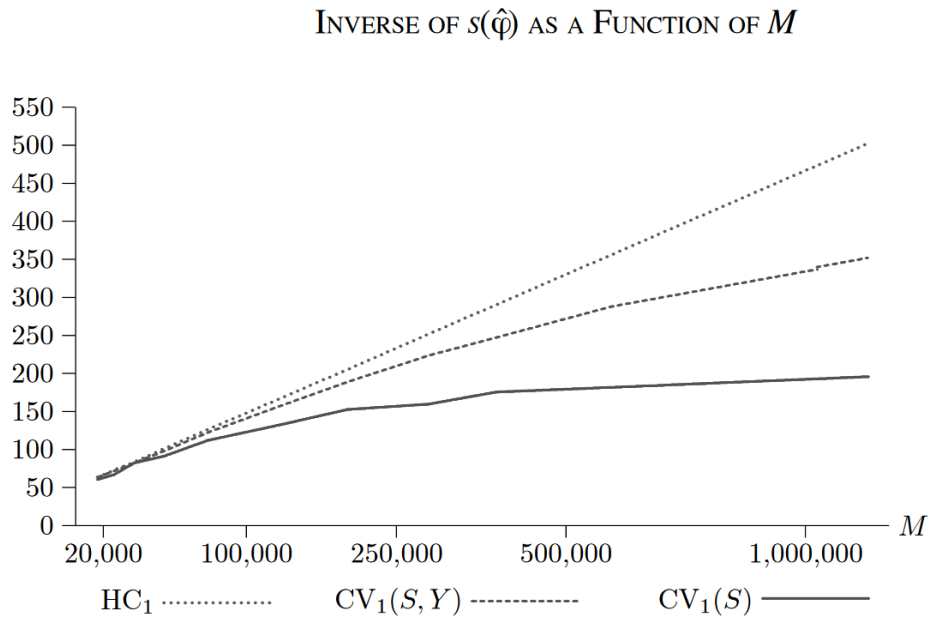
- So, let's have a second look at Equation (3):

$$\Sigma_g = \sum_{i=1}^{N_g} \Sigma_{g,ii} + \sum_{i=1}^{N_g} \sum_{j \neq i} \Sigma_{g,ij}$$

  If there exists a clustered structure, the first term increases in a speed of sample size N, but the second term increases in a speed of N^2.
- Therefore, as the sample size increases, the second term will increasingly become dominant in the variance, and ignoring the clustered structure would result in progressively larger underestimation.

Large sample with clustered structure

- ”Inference with Large Clustered Datasets” (MacKinnon, 2017, *L'Actualité économique*)
  - Heteroskedasticity-robust standard errors HC_1, standard errors clustered at the ”state × year” level $CV_1(S, Y)$, and standard errors clustered at the state level $CV_1(S)$ are estimated in an existing dataset (obs.=1,156,597) by randomly selecting sub-samples of different sample sizes.
  - As the sample size increases, there seems to be a lower bound on the shrinkage of the standard error clustered at the state level, and the underestimation of the standard error due to ignoring the clustered structure becomes increasingly larger.



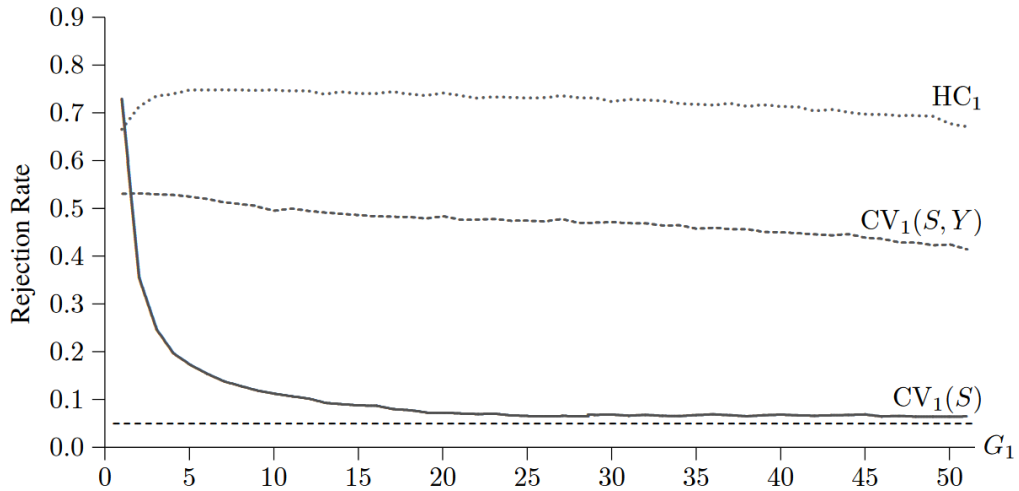
Large sample with clustered structure

- ”How Much Should We Trust Differences-In-Differences Estimates?” (Bertrand et al., 2004, *QJE*)
  - Due to serial correlation, there may be over-rejection of inference in DD estimates.
  - The authors construct many ”placebo samples”, where regressors are generated completely artificial at random (e.g., randomly assigning treatment and control groups), and perform a t-test for each ”placebo sample”.
  - Because a placebo regressor is artificial, we would expect valid significance tests at level α to reject the null close to α% of the time when the experiment is repeated many times.
  - *Results:* Clustering by state performs well, while clustering by ”state × year” performs poorly, and not clustering at all performs even worse.
  - *Conclusion:* When using DD estimates, it is important to cluster at least to the geographic level where the treatment is assigned.

Large sample with clustered structure

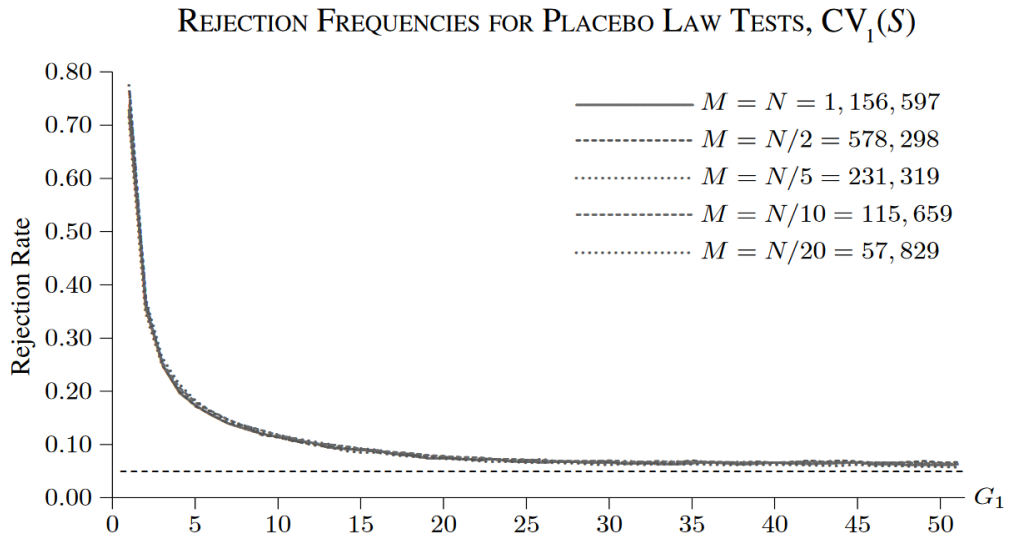
- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - Use almost the same procedure as that in Bertrand et al. (2004) to construct many "placebo samples". G_1 refers to the number of treatment states.
  - When we set 5% significance and perform a *t*-test for each "placebo sample", it is expected that about 5% of them reject the null if the cluster level is correct.
  - When the number of treatment groups are small, all the three standard errors over-reject the null. But when treatment states are more than 20, using standard errors clustered at the state level correctly rejects about 5% "placebo samples".

REJECTION FREQUENCIES FOR PLACEBO LAW TESTS, $N = 1{,}156{,}597$



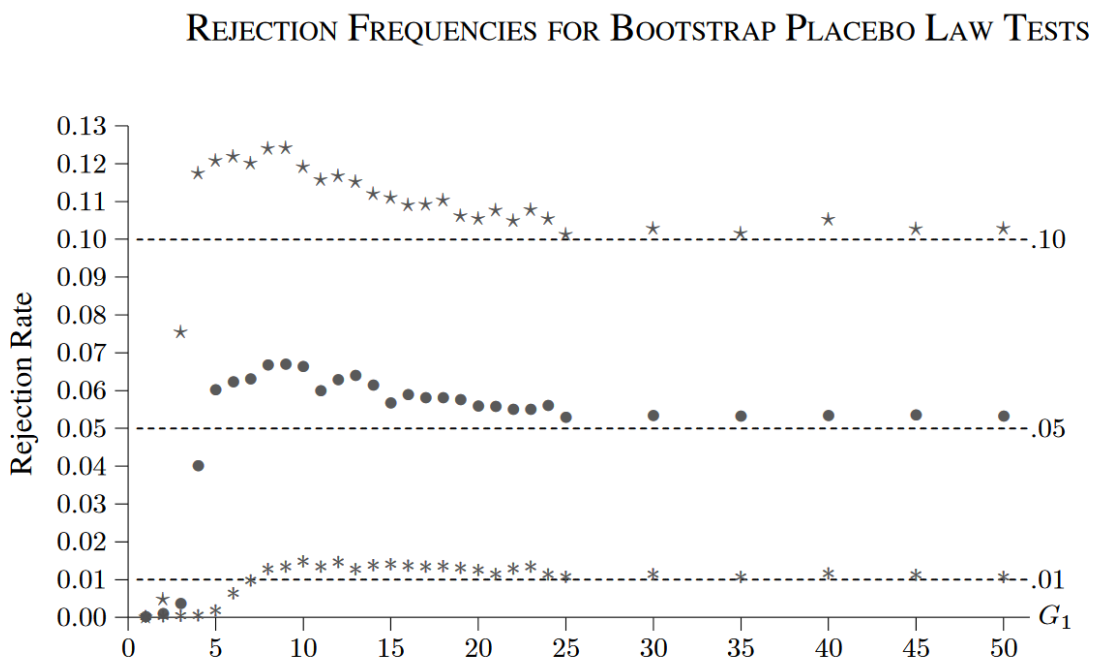
Large sample with clustered structure

- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - Given the number of clusters unchanged (no the 52nd state), increasing the sample size provides limited improvement on inference. (In the following figures, rejection rates calculated with different sample sizes are close.)
  - *Intuition:* As said earlier, adding a new sample to an existed cluster also introduces correlation with all existed samples in the same cluster.
  - To improve inference accuracy, it is more helpful to introduce a new cluster rather than to add some new samples in existed clusters.



Large sample with clustered structure

- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - If the number of treatment clusters (G_1) are small, don't use asymptotic cluster-robust standard errors ($CRVE_1$, $CRVE_2$, and $CRVE_3$) but use wide-cluster bootstrap standard errors (introduced later).
    - Use `boottest` in `Stata` after running `areg`, `ivreg2`, and so on.
  - Wide-cluster bootstrap SEs are less likely to over-reject when G_1 is small.



Outline

Modelling intra-cluster correlation

- So what causes intra-cluster correlation?
- A simple random effect model:
  - Samples in the same cluster share some common characteristics ε_g with the same coefficient vector λ_g:

$$u_{gi} = \lambda_g \varepsilon_g + \varepsilon_{gi}, \text{ where } \varepsilon_{gi} \sim_{\text{i.i.d.}} (0, \omega^2) \text{ and } \varepsilon_g \sim_{\text{i.i.d.}} (0, 1)$$

  Therefore, the covariance of any two errors u_{gi} and u_{gj} in the same cluster g is:

$$\text{Cov}(u_{gi}, u_{gj}) = \lambda^2$$

- Consider a slight more flexible model:
  - Samples in the same cluster share some common characteristics ε_g but allowing different coefficients λ_{gi}:

$$u_{gi} = \lambda_{gi} \varepsilon_g + \varepsilon_{gi}$$

  e.g., y_{gi} is the test score of student i in class g and ε_g refers to the teaching abilities of teachers in class g.
  - When $\lambda_{gi} = \lambda_{gj} = \lambda_g$ $(\forall i, j \in g)$, this model degrades to the random effect model.

Necessary to cluster after controlling for FEs?

- We know that controlling for FEs is equivalent to demeaning each variable by group. So in the flexible model, the demeaned error term is:

$$u^*_{gi} = u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g)\varepsilon_g + (\varepsilon_{gi} - \bar{\varepsilon}_g)$$

  $\forall i \neq j$, we have

$$\text{Cov}(u^*_{gi}, u^*_{gj}) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gj} - \bar{\lambda}_g)$$

  Therefore, only when $\forall i$, $\lambda_{gi} = \bar{\lambda}_g$ (random effect model), the demeaned errors in the same cluster are not correlated.

- *Conclusion:* Assuming that the intra-cluster correlation is solely attributable to the random effect structure, the control for fixed effects would completely absorb such correlation. However, in the vast majority of cases, this assumption is impractical.

Outline

The consequence of clustering at a wrong level

- Suppose there are two possible cluster level: one is *coarse* and the other is *fine*, and the fine level is nested in the coarse level. There are G clusters when using the coarse level, and coarse cluster g comprises M_g coarse clusters.
- According to (2), if the actual level is the coarse one, the variance of the coefficient is $(X'X)^{-1}(\sum_{g=1}^{G}\Sigma_g)(X'X)^{-1}$. Re-write it to yiled:

$$\text{Var}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g}\Sigma_{g,h_1h_2}\right)(X'X)^{-1}$$

  where h_1 and h_2 denote the fine clusters nested in coarse cluster g.
- However, if we wrongly choose the fine cluster, then we assume the variance is:

$$\text{Var}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G}\sum_{h=1}^{M_g}\Sigma_{gh}\right)(X'X)^{-1}$$

The consequence of clustering at a wrong level

- If the actual level is the coarse level but we choose the fine level, the bias is:

$$\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g}\Sigma_{g,h_1h_2}-\sum_{g=1}^{G}\sum_{h=1}^{M_g}\Sigma_{gh}=\sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2\neq h_1}\Sigma_{g,h_1h_2} \tag{4}$$

  This bias comes from the inter-(fine-)cluster correlation in the same coarse cluster. The bias increases as the sample size increases.
- If the actual level is the fine level but we choose the coarse level (i.e., the right side of (4) is zero), there is no bias. However, since we use some information to estimate those inter-(fine-)cluster uncorrelation, it's at the cost of decreased estimation efficiency.

Two rules of thumb to choose the level

- An intuitive rule: to choose the most coarse one among all possible levels (Cameron and Miller, 2015).
- "A Practitioner's Guide to Cluster-Robust Inference" (Cameron and Miller, 2015, *Journal of Human Resources*)
  - "It is possible for cluster-robust errors to actually be smaller than default standard errors."
    - In some rare cases errors may be negatively correlated, most likely when $G = 2$.
    - If clustering has a modest effect so cluster-robust and default standard errors are similar in expectation, then cluster-robust may be smaller due to noise.
  - "In cases where the cluster-robust standard errors are smaller, they are usually not much smaller than the default, whereas in other applications they can be much, much larger."
  - "There is no general solution to this tradeoff, and there is no formal test of the level at which to cluster. The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters."

Two rules of thumb to choose the level

- It's possible that the intra-cluster correlation is negative. In such cases, ignoring the cluster structure could lead to even larger standard errors.
- "The Power of the Street: Evidence from Egypt's Arab Spring" (Acemoglu et al., 2018, *RFS*)
  - "All standard errors we report throughout are robust to heteroscedasticity. In addition, because there might be other factors correlated across connected firms, we report adjusted standard errors and portfolio-based results that account for potential cross-firm correlation of residual returns in the appendix. These robustness checks consistently show that residual returns are negatively correlated with the group of politically connected firms, such that adjusted standard errors tend to be narrower than unadjusted standard errors. To be conservative, we therefore report the wider (robust) standard errors in the main text."
  - "In column 4, we adjust standard errors for the cross-correlation of error terms estimated in 2010 data, with very similar results and somewhat smaller standard errors, reflecting the (aforementioned) fact that the residual correlation between connected firms is negative."

Two rules of thumb to choose the level

Table 2
Mubarak's fall

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				CR[0,8]			*CAR[0,8]*
NDP	−0.086*	−0.131**	−0.142**	−0.131**	−0.142**	−0.200***	−0.145**
	(0.049)	(0.049)	(0.059)	(0.046)	(0.054)	[−0.099,0.101]	(0.056)
Military	0.048*	0.032	0.075**	0.032	0.035	0.053	0.051
	(0.028)	(0.030)	(0.021)	(0.026)	(0.033)	[−0.066,0.082]	(0.035)
Islamic	−0.031	−0.064	−0.058	−0.064	−0.090	−0.159***	−0.125*
	(0.054)	(0.051)	(0.063)	(0.041)	(0.058)	[−0.107,0.130]	(0.066)
β^{World}		0.037**	0.023	0.037	0.050**		0.132**
		(0.016)	(0.023)	(0.023)	(0.013)		(0.046)
β^{Egypt}		−0.028	−0.021	−0.028	−0.093**		
		(0.018)	(0.025)	(0.023)	(0.030)		
β^{Unrest}		2.134*	0.897	2.134	1.812		11.219**
		(1.182)	(1.337)	(2.253)	(2.039)		(4.632)
Size		0.024**	0.022**	0.024**	0.016*		0.014
		(0.007)	(0.007)	(0.007)	(0.009)		(0.009)
Leverage		−0.024	−0.003	−0.024*	−0.028		0.017
		(0.017)	(0.019)	(0.014)	(0.022)		(0.027)
R^2	0.252	0.320	0.138	0.320	0.387		0.451
N	145	143	143	143	136		143
Sector fixed effects	yes	yes	no	yes	yes	no	yes
Adjusted standard errors	no	no	no	yes	no	no	no
Weights	no	no	no	no	yes	no	no
Matching estimator	no	no	no	no	no	yes	no

Two rules of thumb to choose the level

- A conservative rule: to report the largest standard error for the coefficient of interest across all estimated at varying possible cluster levels. (Angrist and Pischke, 2008). This rule leads to the most conservative standard error with the higher risk of a significant loss of estimated efficiency.
- *Mostly harmless econometrics: An empiricist's companion* (Angrist and Pischke, 2008)
  - This viewpoint is proposed in this book when the authors are comparing heteroskedasticity-robust standard errors to conventional standard errors based on the spherical disturbance assumption.
  - "[R]obust standard errors are no panacea. They can be smaller than conventional standard errors for two reasons: the small sample bias we have discussed and the higher sampling variance of these standard errors. We therefore take empirical results where the robust standard errors fall below the conventional standard errors as a red flag. This is very likely due to bias or a chance occurrence that is better discounted."
  - "In this spirit, we like the idea of taking the maximum of the conventional standard error and a robust standard error as your best measure of precision. This rule of thumb helps on two counts: it truncates low values of the robust estimators, reducing bias, and it reduces variability."

Outline

MNW method

- "Testing for the appropriate level of clustering in linear regression models" (MacKinnon et al., 2023, *JoE*)
  - *Idea:* If the true level is the fine one, the standard error estimated at the fine level should be close to that estimated at the coarse level. In other words, the right hand side of (4) is 0.
  - Formally, define $\Sigma_c \equiv \sum_{g=1}^{G} \sum_{h_1=1}^{M_g} \sum_{h_2=1}^{M_g} \Sigma_{g,h_1h_2}$ and $\Sigma_f \equiv \sum_{g=1}^{G} \sum_{h=1}^{M_g} \Sigma_{gh}$, then the null hypothesis is:

$$\mathrm{H}_0\text{: } \lim_{N\to\infty} \Sigma_f \Sigma_c^{-1} = I \text{ and } \mathrm{H}_1\text{: } \lim_{N\to\infty} \Sigma_f \Sigma_c^{-1} \neq I$$

    Accordingly, by constructing four consistent estimators: $\hat{\Sigma}_f$, $\hat{\Sigma}_c$, $\widehat{\mathrm{Var}}[\hat{\Sigma}_f]$, and $\widehat{\mathrm{Var}}[\hat{\Sigma}_c]$, the authors prove that:

$$\text{Under the null, } \tau_\sigma = \hat{\boldsymbol{\theta}} / \sqrt{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})} \to_{\mathrm{d}} N(0,1), \text{ where } \hat{\boldsymbol{\theta}} \equiv \mathrm{vech}\left(\hat{\Sigma}_c - \hat{\Sigma}_f\right)$$

  - Command in `Stata`: `mnwsvt`, available at `http://qed.econ.queensu.ca/pub/faculty/mackinnon/svtest`.

Permutation test

- The formerly discussed ”placebo sample method” (also called ”permutation test”) can also be used to test the cluster level.
- *Idea:* Since the treatment regressor is randomly assigned, it must be uncorrelated with the dependent variable. Therefore, this procedure simulates the distribution of the coefficient of interest when it is actually 0. As we don't make any parametric assumptions about the error structure, this procedure does not suffer from the bias due to wrongly choosing the cluster level.
- Use `permute` to implement the permutation test in `Stata`.

Permutation test

- The permutation test ("placebo test") is commonly used in DD estimation (especially in Chinese literature). We should note that it is used to support the statistic inference rather than the causal identification.
- How to interpret the result of permutation test?
  "Salience and Taxation: Theory and Evidence" (Chetty et al., 2009, *AER*)
  - "A concern in DD analysis is that serial correlation can bias standard errors, leading to over-rejection of the null hypothesis of no effect. To address this concern, we implement a nonparametric permutation test for $\delta = 0$."
  - "Intuitively, if the experiment had a significant effect on demand, we would expect the estimated coefficient to be in the lower tail of estimated placebo effects. Since this test does not make parametric assumptions about the error structure, it does not suffer from the over-rejection bias of the t-test."
  - "Figure 1 illustrates the results of the permutation test by plotting the empirical distribution of placebo effects G for log quantity. The vertical line in the figure denotes the treatment effect reported in Table 4. For log quantity, $G(\delta) = 0.07$. An analogous test for log revenue yields $G(\delta) = 0.04$. Although these p-values are larger than those obtained using the t-tests, they confirm that the intervention led to an unusually low level of demand."

Permutation test

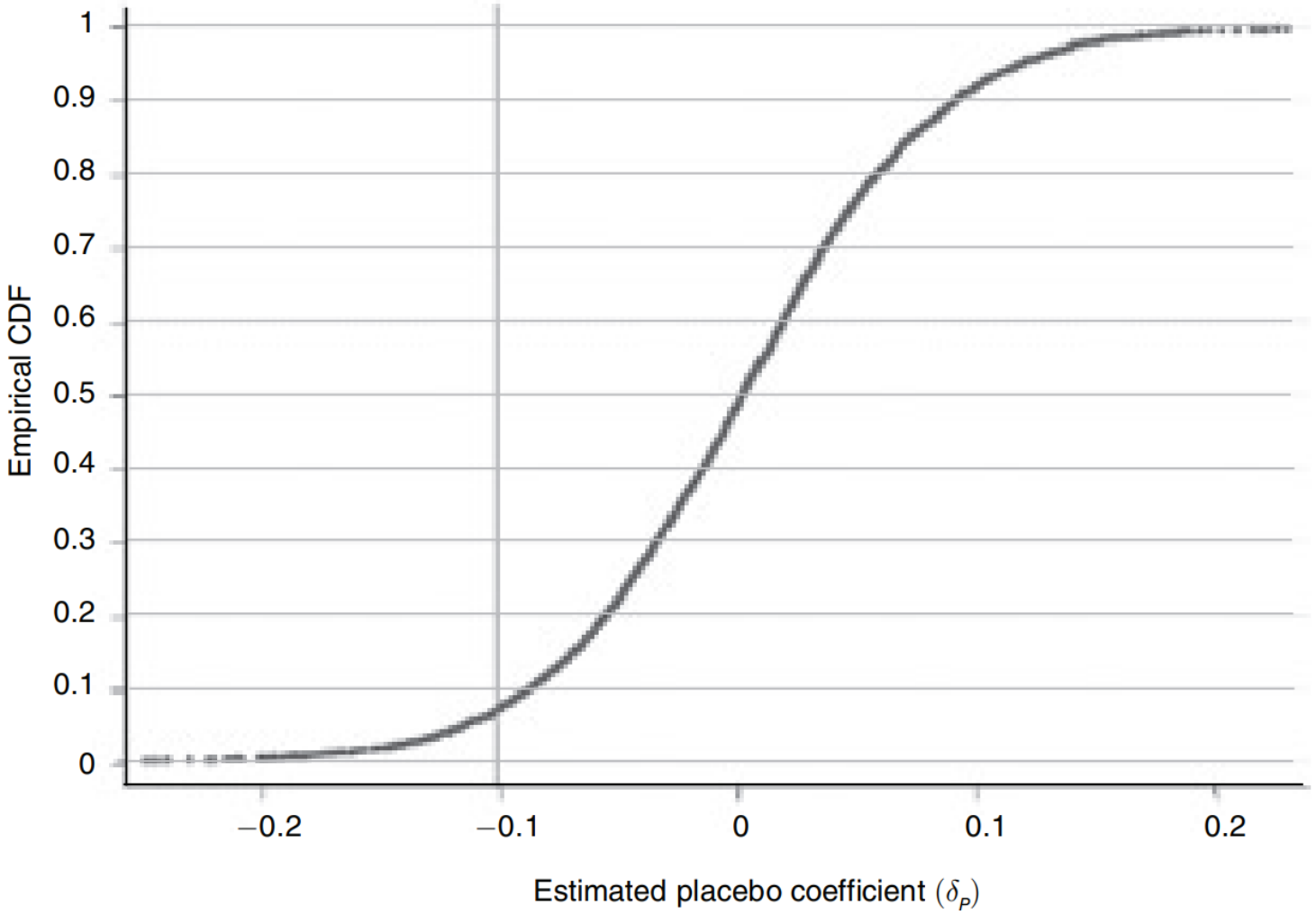


FIGURE 1. DISTRIBUTION OF PLACEBO ESTIMATES: LOG QUANTITY

The issue of pre-tests

- *Intuition:* Pre-tests usually lead to over-rejection, since pre-tests in and of themselves could make mistakes.
- Therefore, joint hypotheses test should be applied when implementing t-tests in regressions. (i.e., those t-tests based on the cluster level resulted from the pre-tests should also take into account the probability of making mistake in pre-tests.)
- In practice, however, such joint hypotheses test is seldomly used.
- *Conclusion:* Although there are effective pre-tests for choosing the cluster level, they are not as conservative as the second rule of thumb. The ideal procedure is to cluster based on the second rule of thumb and use pre-tests to further support the choice.

Outline

Two-way clustering

- In some cases, there are various dimensions of intra-cluster correlation. For example, errors could be correlated because they are in the same geographic region *or* in the same time period.
- Suppose there are two clustered dimension: dimension A with G clusters (e.g., state level) and dimension B with H clusters (e.g., year level), indexed by g and h, respectively. Now the regression model is:

$$y_{gh} = X_{gh}\beta + u_{gh},\ g = 1, \cdots, G,\ h = 1, \cdots, H$$

  where subscript gh denotes the sample is grouped to the g-th cluster in dimension A and the h-th cluster in dimension B.
- *Remarks:* Two-way clustering is different to the "state × year" level we discussed earlier, which assumes that only samples in the same state *and* in the same year are correlated. However, by assuming a two-way clustered structure, we mean that u_{gh} are correlated as long as they share the same g *or* h.

Two-way clustering

- With the two-way clustered structure, the variance of coefficient is

$$\begin{aligned}\text{Var}[\hat{\beta}] &= (X'X)^{-1}\text{Var}\left[\sum_{g,h} s_{gh}\right](X'X)^{-1} \\ &= (X'X)^{-1}\left(\sum_{g,h,g',h'} \text{E}[s_{gh}s'_{g'h'}]\right)(X'X)^{-1},\end{aligned}$$

  where $s_{gh} \equiv X'_{gh}u_{gh}$. Only when $g \neq g'$ *and* $h \neq h'$, $\text{E}[s_{gh}s_{g'h'}] = 0$ always holds.
- One can prove that

$$\sum_{g,h,g',h'} \text{E}[s_{gh}s'_{g'h'}] = \sum_{g=1}^{G} \text{E}[s_g s'_g] + \sum_{h=1}^{H} \text{E}[s_h s'_h] - \sum_{g=1}^{G}\sum_{h=1}^{H} \text{E}[s_{gh}s'_{gh}]$$

Two-way clustering

- Similar to the earlier, by replacing the error term in s_g, s_h, and s_{gh} with the residuals (see Slide #9), we obtain an estimator for the variance:

$$\widehat{\text{Var}}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{g=1}^{G}\hat{s}_g\hat{s}_g' + \sum_{h=1}^{H}\hat{s}_h\hat{s}_h' - \sum_{g=1}^{G}\sum_{h=1}^{H}\hat{s}_{gh}\hat{s}_{gh}'\right)(X'X)^{-1}$$

- Using in `Stata`: add `cluster(cluster_var1, cluster_var2)` in the option (available for `reghdfe` and `ivreg2`).

```
> reghdfe y x, ... cluster(cluster_var1, cluster_var2)
```

Outline

Asymptotic theories

- Now let's focus on specific methods of inference.
- When using asymptotic standard errors, we are concerned about how well the asymptotic theories work.
- For example, we are interested in the null hypothesis H_0: $a'\beta = a'\beta_0$, thus we construct a t-statistic:

$$t_a = \frac{a'(\hat{\beta} - \beta_0)}{\sqrt{a'\hat{V}a}},$$

  where $\hat{V}$ is an estimator for the variance-covariance matrix of the estimated coefficient $\hat{\beta}$ (which can be $CRVE_1$, $CRVE_2$, or $CRVE_3$).
- *The core idea of asymptotic inference*: For statistic inference, we need to know the distribution of the inference estimator. However, as we never know the actual distribution of the error term, we never know the actual distribution of t_a theoretically. Therefore, we consider to use its *asymptotic* distribution to replace its actual distribution.

Key conditions for asymptotic theories

- *Econometrics* (Hansen, 2022, Section 4.22)
  - "In many respects cluster-robust inference should be viewed similarly to heteroskedaticity-robust inference where a 'cluster' in the cluster-robust case is interpreted similarly to an 'observation' in the heteroskedasticity-robust case."
  - "In particular, the effective sample size should be viewed as the number of clusters, not the 'sample size' n. This is because the cluster-robust covariance matrix estimator effectively treats each cluster as a single observation and estimates the covariance matrix based on the variation across cluster means."
- In other words, we can regard the asymptotic inference as two steps:
  - *Step 1:* Calculate the variance-covariance matrix $\hat{\Sigma}_g$ of score vector $s_g \equiv X_g' u_g$ cluster by cluster.
  - *Step 2:* Regard each cluster as a whole (now no correlation between errors) and inference with a heteroskedasticity structure.

Key conditions for asymptotic theories

- Whether asymptotic standard errors work well depends on whether asymptotic theories are satisfied. Based on the "two steps" in the earlier slide, we need to focus on:
  - The law of large number (LLM) should hold to ensure $\sum_{g=1}^{G} \hat{s}_g \hat{s}_g'$ to converge to the actual variance-covariance matrix $\sum_{g=1}^{G} \Sigma_g$ (i.e., enough samples in each cluster are required).
  - The central limit theorem (CLT) should hold to ensure $\sum_{g=1}^{G} s_g$ to converge to a multivariate normal distribution with the variance of $\sum_{g=1}^{G} \Sigma_g$ (i.e., enough clusters are required).
- With a clustered structure, what does $n \to \infty$ mean in asymptotic theories?
  - *Case 1:* Keep the number of samples in existed clusters remained (but still enough for LLN) and add new clusters (i.e., large number of clusters).
  - *Case 2:* Keep the number of clusters remained and increase the number of samples in each cluster (i.e., small number of large clusters).
- Intuitively, Case 1 is more in line with the requirements of the "two-steps" inference procedure.

Key conditions for asymptotic theories

- Again intuitively, given the number of clusters, if there is less heterogeneity among clusters (specifically, among s_g), the CLT could work better.
- In summary, we have gained three insights so far:
  1. With respect to the effectiveness of the CLT, the number of clusters is crucial. Therefore, with a clustered structure, we should not only care about the sample size, but also the number of clusters.
  2. The number of clusters and the heterogeneity among clusters both determine whether asymptotic theories work well. Therefore, there will not be a universal threshold of clusters G^* beyond which we can be assured of effective inference.
  3. We should pay attention to the heterogeneity among clusters (e.g., some clusters with a large proportion of total samples).

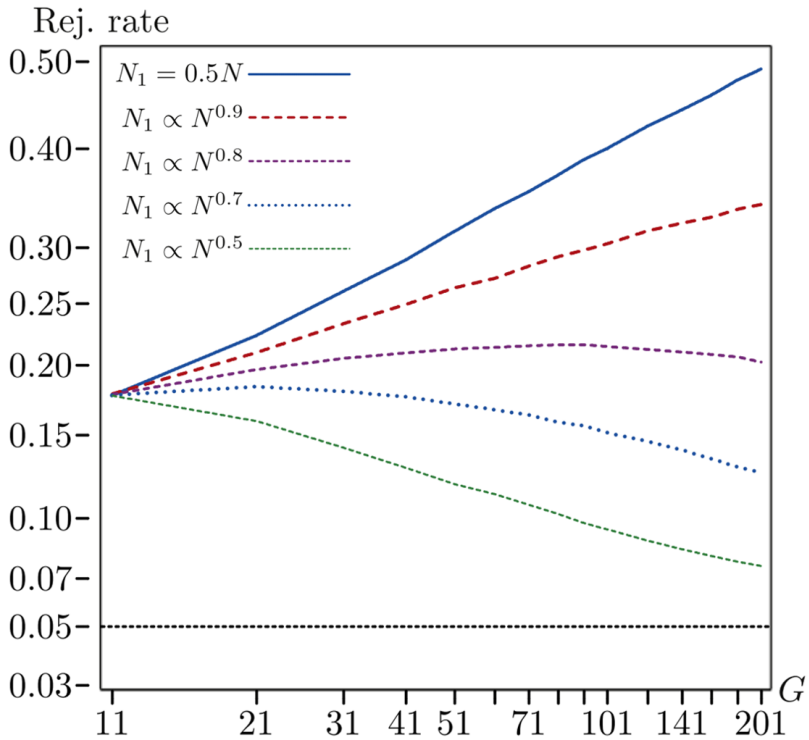
Outline

Large number of clusters

- Suppose that each cluster has M samples, it can be proved that, under the null hypothesis, we have $\sqrt{G}(\hat{\beta} - \beta_0) \xrightarrow{\text{d}} N(0, G\hat{V})$.
- However, the sample size of each cluster is different in practice, which could be a potential threaten to inference. For instance, if the sample size of a specific cluster is much larger than others, this cluster may have a large variance of s_g (i.e., $\Sigma_g \equiv E[s_g s_{g'}]$), making it has a dominant influence when applying the CLT. In such a scenario, the CLT typically couldn't work well.

Large number of clusters

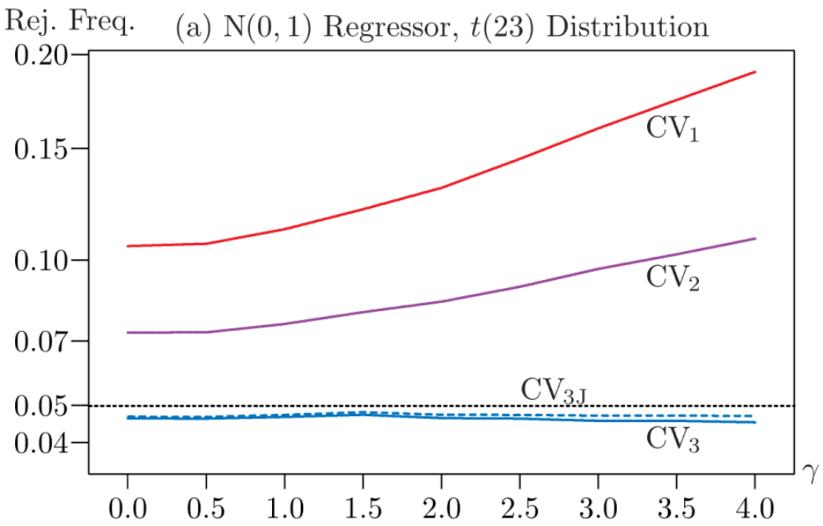
- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - When there is a large cluster (e.g., $N_1 = 0.5N$), increasing the number of clusters inversely aggravates over-rejection.

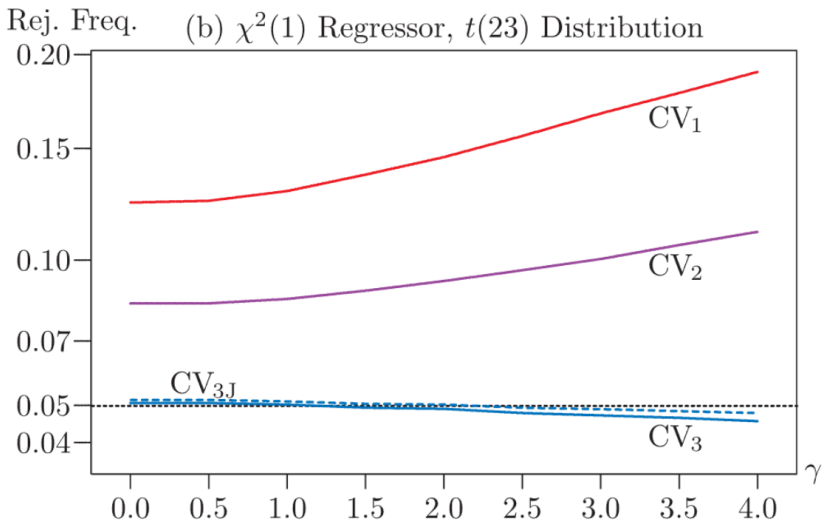


(a) CRVE t-tests

Large number of clusters

- "Fast and reliable jackknife and bootstrap methods for cluster-robust inference" (MacKinnon et al., 2023, *Journal of Applied Econometrics*)
  - Simulation: γ denotes the heterogeneity of sample sizes among clusters. When $\gamma = 0$, samples are distributed uniformly; when $\gamma = 2$, the sample sizes increase from 130 to 899.
  - It's shown that, compared to $CRVE_1$ and $CRVE_2$, $CRVE_3$ works much better under larger heterogeneity. But regrettably, $CRVE_3$ is not a magic bullet!
    "However, as we shall see, there are also many cases in which CV_3 overrejects, and CV_{3J} therefore overrejects slightly more. In practice, it would be perfectly reasonable to report either CV_3 or CV_{3J}."





Large number of clusters

- Accordingly, even though with case 1, there still exists at least two reasons leading to over-rejection:
  - Small number of treatment clusters (please review slide #1715).
  - Some clusters have a large proportion of samples.
- Regarding the second reason, we need some methods to judge the influence of a single cluster and report the results in studies.

Influence and leverage

- Use the leave-one-out coefficient to evaluate the influence of a specific cluster:

$$\hat{\beta}^{(g)} = (X'X - X_g'X_g)^{-1}(X'y - X_g'y_g)$$

- Similarly to the leverage of a single sample, we can use the hat matrix to evaluate the leverage of a single cluster. We know elements in the main diagonal of hat matrix $P_x = X'(X'X)^{-1}X'$ presents the leverage of each sample, thus adding them yields the leverage of a cluster:

$$L_g = \text{tr}(H_g) = \text{tr}[X_g'X_g(X'X)^{-1}]$$

  This expression explains again why a large cluster is influential.

- Compare each cluster's leverage to the average, which is

$$\frac{1}{G}\left(\sum_g L_g\right) = \text{tr}[X(X'X)^{-1}X'] = \text{tr}[X'X(X'X)^{-1}] = \frac{k}{G}$$

Influence and leverage

- In practice, we usually solely care about the coefficient of a specific explanatory variable, thus we can use the FWL theorem to calculate the partial leverage for each cluster:

$$L_{gj} = \frac{\acute{x}'_{gj}\acute{x}_{gj}}{\acute{x}'_{j}\acute{x}_{j}}$$

  where $\acute{x}_j$ is the estimated residual from the regression with the explanatory variable of interest x_j as the dependent variable.
- Use `summclust` to calculate influence, leverage, partial leverage, and in `Stata`. Of course, we can calculate them on our own since these estimators are not that complicated.

Outline

Small number of large clusters

- Firstly, all threats in case 1 exist in case 2!
- Additionally, the small number of clusters leads to extra concerns to the satisfaction of the CLT.
- "Inference with dependent data using cluster covariance estimators" (Bester et al., 2011, *JoE*)
  - This paper proves that the CLT holds in case 2 with some strict assumptions:
    - "... all the clusters are assumed to be the same size M."
    - "... it limits the amount of dependence within each cluster and requires it to diminish quite rapidly as $M \to \infty$." (This is usually unrealistic in panel data settings.)

Summary: inference with asymptotic standard errors

- *Threat:* We use the asymptotic distribution to replace the actual distribution of inference estimator. But the asymptotic distribution can sometimes quite differ from the actual one.
  - Small number of clusters G, large heterogeneity among clusters, ..., which could weaken the effectiveness of the large sample theories.
  - Unfortunately, it is challenging (if not impossible) to establish a universal criterion for determining the number of clusters and the level of heterogeneity that can ensure the satisfaction of the CLT.
- Can we directly estimate the true distribution of the inference estimator instead of relying on its asymptotic distribution?
- "Instead of basing inference on an asymptotic approximation to the distribution of a statistic of interest, it is often more reliable to base it on a bootstrap approximation. [...] We therefore recommend that at least one variant of the WCR bootstrap be used almost all the time."

Outline

The idea of bootstrap

- *The core idea of bootstrap inference*: For statistic inference, we need to know the distribution of the inference estimator. However, as we never know the actual distribution of the error term, we never know the actual distribution of t_a theoretically. Therefore, we consider to estimate its *actual* distribution and use this *empirical distribution* to replace its actual distribution.
- Suppose that we are inferring with an estimator τ, we can resample the original samples to yield B bootstrap samples (e.g., randomly-sampled sub-samples) and then calculate B inference estimators τ_b^* $(b = 1, \cdots, B)$, which provides the empirical distribution of τ. Theoretically, the empirically distribution can be a good approximation of the actual distribution.
- Considering the trade-off between accuracy and computational time, we usually choose that $B = 9{,}999$ or $99{,}999$.
- Based on the empirical distribution, we can estimate the p-value and the confidential interval of the inference estimator.

Outline

Wild-cluster bootstrap

- Wild-cluster bootstrap is a residual-based resampling method.
- *Restricted wild-cluster bootstrap (WCR):* Define the estimated coefficient resulting from the Constrained Least Squares (CLS) with a restriction $a'\beta = a'\beta_0$ as $\tilde{\beta}$. The restricted residuals for the g-th cluster is then $\tilde{u}_g = y_g - X_g\tilde{\beta}$. Therefore, we can construct the bootstrap sample b as follows:

$$y_g^{*b} = X_g\tilde{\beta} + u_g^{*b}, \text{ where } u_g^{*b} = v_g^{*b}\tilde{u}_g \text{ and vector } v_g^{*b} \sim_{i.i.d.} (0,1)$$

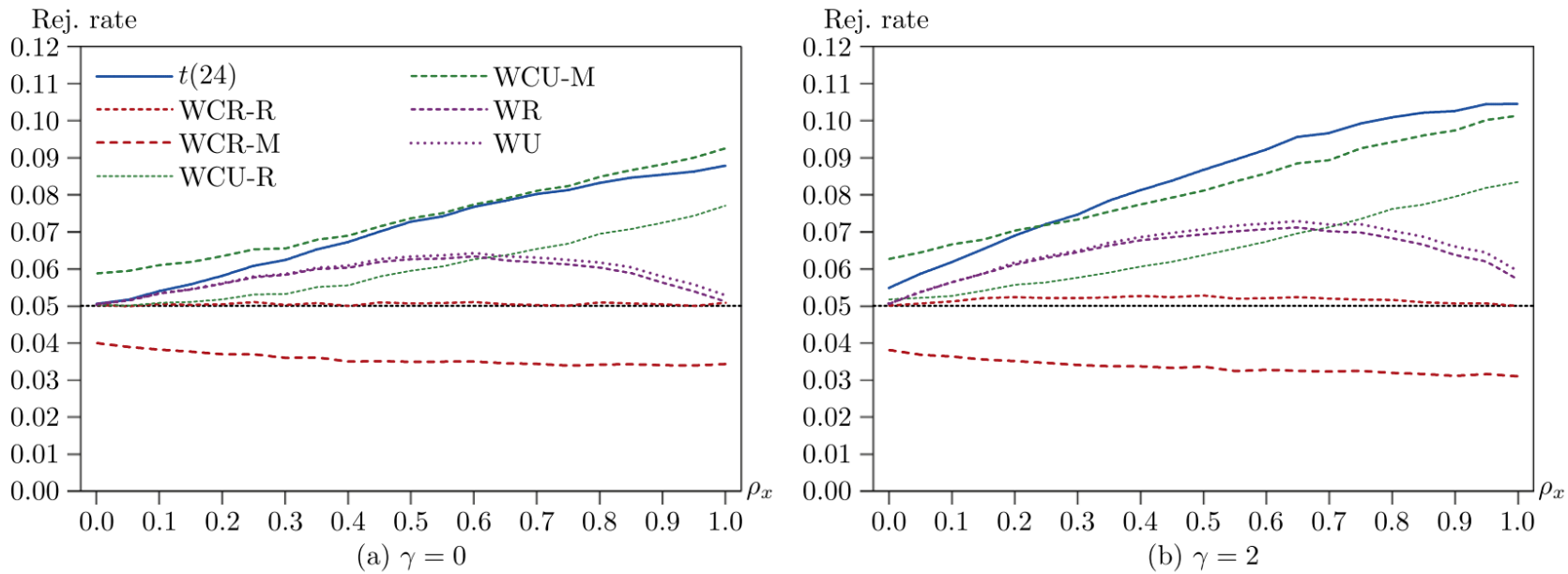
  We usually use Rademacher distribution (variable taking -1 or 1 with probability of 0.5, respectively) when generating v_g^{*b}.
- *Unrestricted wild-cluster bootstrap (WCU)*: Similar to WCR, the only difference is to use OLS instead of CLS.
- *Intuition*: Since v_g^{*b} is independent among clusters, the DGP of bootstrap samples ensures that there is no inter-cluster correlation.

Wild-cluster bootstrap

- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - In finite samples, WCR usually performs better than WCU.
  - In addition, we can generate the new error terms at the sample level rather than at the cluster level (i.e., generating i.i.d. v_i^{*b} for each sample rather than generating i.i.d. v_g^{*b} for each cluster). The corresponding method is known as *restricted wild bootstrap (WR)* and *unrestricted wild bootstrap (WU)*. However, in many cases, WCR performs better than WR; additionally, WR requires much more computational time when sample size is large.
- Use `boottest` in `Stata` to apply WCR.
  - It is actually an easy and convenient command in `Stata`, please refer to the help file for guidance on its usage.
  - *A crash course:* Click *here* to download the replication package for Section 8. Then read the do-file and text in Section 8 carefully, where the authors show how to use `boottest` to apply WCR.

Wild-cluster bootstrap

- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - Based on simulation, with ρ_X (denotes the level of intra-cluster correlation) increases, asymptotic standard errors ($t(24)$ in the figure) tend to over-reject. WCR-R (WCR with using Rademacher distribution to generate those new error terms) performs the best among all variants of wild bootstrap.



Remarks on using WCR

- "We recommend using at least one variant of the WCR bootstrap (preferably with at least $B = 9,999$) almost all the time."
- WCR is also not a magic bullet. When the number of treatment groups are small, WCR would become extremely conservative (see slide #19). Instead, we can consider to use WR in such scenario.
- If our sample is similar to the case 2 discussed ealier, WCR-R can sometimes yield accurate inferences, but this is not always the case. ("The wild bootstrap with a 'small' number of 'large' clusters", Canay et al., 2021, *REStat*)
- If G is small, we should avoid using two-point distributions like Rademacher distribution (choosing other distributions in the option of `boottest`).

Outline

What should applied researchers report?

1. In addition to the default CRVE_1, report at least 1 or 2 other types of standard errors, e.g., CRVE_3 and a variant of WCR.
2. As discussed earlier, the CLT relies on enough clusters rather than solely on large sample size. Therefore, it is important to report the number of clusters.
3. The heterogeneity among clusters is also important for inference, thus researchers should report the median, minimum, and maximum of sample sizes among all clusters.
4. Additionally, report the leverage L_g, partial leverage L_{gj}, and leave-one-out coefficient $\beta_j^{(g)}$ of each cluster.
5. When using two-way clustering (or clustering with more than 2 dimensions), researchers should report the above information of each cluster and each of their intersections.

Outline

Overall recommendations

1. List all plausible clustering dimensions and levels for the data at hand and make an informed decision regarding the clustering structure. The decision may depend on what is to be estimated and why. A conservative approach is simply to choose the structure with the largest standard error(s) for the coefficient(s) of interest, subject to the number of clusters not being so small that inference risks being unreliable (It's more common to report SEs clustered in different levels). Some pre-tests (MNW test and placebo test) can be helpful in making this decision.
2. If using DD specification, we should at least cluster at the treatment level.
3. After choosing the cluster level, it is suggested to report at least the minimum, maximum, mean, and median of the sample sizes among clusters.
4. For the key regression specification(s) considered, report information about leverage, partial leverage, and influence. This may be particularly informative for DD and other treatment models. Inferences may not be reliable when a few clusters are highly influential or have high (partial) leverage.

Overall recommendations

5. In addition to, or instead of, the usual CRVE_1, employ the CRVE_3 and at least one variant of the restricted wild cluster (WCR) bootstrap as a matter of course for both tests and confidence intervals. In many cases, especially when G is reasonably large and the clusters are fairly homogeneous, these methods will yield very similar inferences that can likely be relied upon. However, in the event that they differ, it would be wise to try other methods as well, including additional variants of the WCR bootstrap and some of the alternative methods.
6. For models with treatment at the cluster level, where either the treated clusters or the controls are few in number and/or atypical, cluster-robust inference can be quite unreliable, even when it is based on CRVE_3 or the WCR bootstrap. In such cases, it is important to verify that the results are (or perhaps are not) robust. This can often be done by using methods based on placebo test (permutation test).