Cluster-robust inference: A guide to empirical practice

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

Journal of Econometrics, 2023

聚类稳健推断：经验研究指南

CHEN Zeyu

School of Economics, Renmin University of China

Last updated：2023–07–18

目录

目录

模型设定

- 假定线性模型为：$y = X\beta + u$，并有 $\mathrm{E}[u|X] = 0$ 且 $\mathrm{Var}[u] = \mathrm{E}[uu'] = \Omega$。
- 在 OLS 估计下，系数 β 的点估计值为：

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$$

- 于是，$\hat{\beta}$ 的方差为：

$$\mathrm{Var}[\hat{\beta}] = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

- 如果误差项的方差满足经典假设，即 $\Omega = \sigma^2 I$，则 $\mathrm{Var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$，于是可以得到方差的估计量为：

$$\widehat{\mathrm{Var}}[\hat{\beta}] = s^2(X'X)^{-1}, \text{其中} s^2 = \hat{u}'\hat{u}/(n-k)$$

- 如果误差项的方差矩阵是一个对角阵，我们可以使用异方差稳健标准误进行一致推断 (White, 1980)：

$$\widehat{\mathrm{Var}}[\hat{\beta}] = (X'X)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{u}_i^2\right)(X'X)^{-1}$$

聚类结构

- 许多情境中，我们能预先判断 Ω 不是一个对角阵，而存在聚类结构。假设共有 G 个聚类，第 g 个聚类共有 N_g 个样本，则 Ω 是一个分块的对角阵：

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

- 在聚类结构下，回归方程可以改写为：$y_g = X_g\beta + u_g$，$g = 1, \cdots, G$。此时系数点估计值也可以改写为：

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

其中定义了 $s_g \equiv X_g' u_g$。

聚类结构

- 此时，$\hat{\beta}$ 的方差为：

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

- 定义 $\Sigma_g \equiv \mathrm{Var}[s_g] \equiv \mathrm{Var}[X'_g u_g]$，假设 $\forall g \neq g'$ 满足 $\mathrm{E}[s_g s'_{g'}] = 0$，则有：

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

- 由于聚类设定相比起经典假设或异方差假设是更加一般化的，因此式 (2) 对于经典假设和异方差假设同样适用。

忽略聚类结构的后果

- 下面讨论忽略聚类结构会导致什么样的后果。
- 为此，将 s_g 进一步写为 $s_g \equiv X_g' u_g = \sum_{i=1}^{N_g} X_{gi} u_{gi} \equiv \sum_{i=1}^{N_g} s_{gi}$，可以证明：

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

- 于是，有：

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

其中定义了 $\Sigma_{g,ij} \equiv \mathrm{E}[s_{gi} s_{gj}']$。

忽略聚类结构的后果

- 前面的式子可以进一步写为：

$$\Sigma_g = \mathrm{E}[s_g s_g'] = \sum_{i=1}^{N_g} \Sigma_{g,ii} + \sum_{i=1}^{N_g} \sum_{j \neq i} \Sigma_{g,ij} \tag{3}$$

- 在异方差假定下，不存在组内相关，因此 $\forall i \neq j$，有 $\Sigma_{g,ij} = 0$，相当于直接令式 (3) 中最后一项为 0。
- 通常情况下，组内样本间误差项是正相关的 (比如组内样本的误差项中具有某些共同的变量)，因此在聚类结构下，式 (3) 最后一项往往是正的。如果忽略了聚类结构，将导致 Σ_g 被大大低估，从而导致系数标准误被大大低估。

三种聚类稳健标准误

- 根据前面，估计系数的方差为：

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

- 方差中只有 s_g 是未知的，因此关键在于构造对误差项 u_g 的一致估计。根据不同的构造方法，可以得到不同类型的聚类稳健标准误。
  - predicted residuals：$\hat{s}_g = X_g'\hat{u}_g$
  - standardized residuals (transformed residuals)：$\dot{s}_g = X_g' M_{gg}^{-1/2}\hat{u}_g$
  - leave-one-out residuals (jackknife residuals)：$\acute{s}_g = X_g' M_{gg}^{-1}\hat{u}_g$
- 构造 $\hat{s}_g = X_g'\hat{u}_g$，得到 CRVE_1(`STATA` 中 `cluster()` 默认使用)：

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

三种聚类稳健标准误

- 构造 $\grave{s}_g = X_g' M_{gg}^{-1/2} \hat{u}_g$，得到 CRVE$_2$：

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

- 构造 $\acute{s}_g = X_g' M_{gg}^{-1} \hat{u}_g$，得到 CRVE$_3$(STATA 命令：summclust；或在 option 中添加 vce(jackknife, cluster())，该方法在 reghdfe 中不可用而在 areg 中可用；也可以在回归命令前加上 jackknife:，该方法对 reghdfe 和 areg 均适用)：

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

- 样本较小的情况下，更加建议使用 CRVE$_2$ 或 CRVE$_3$，因为两者对 u_g 方差的估计不仅是一致的，而且是无偏的。CRVE$_3$ 是更加推荐使用的，因为相对保守一些。

目录

聚类结构与大样本理论

- "Using large samples in econometrics" (MacKinnon, 2023, *JoE*)
  - 大样本理论告诉我们，随着样本量的增加，估计量的方差会逐渐减少，最终依概率收敛。比如说我们关心 $y_i \overset{\text{i.i.d.}}{\sim} (\theta, \sigma^2)$ 的样本均值 $\bar{y} = \left(\sum_{i=1}^n y_i\right)/n$，则其方差满足：
  $$\mathrm{Var}[\bar{y}] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[y_i] = \frac{\sigma^2}{n}$$
  - 然而，如果变量存在相关性 (不独立)，则实际上样本均值的方差是：
  $$\mathrm{Var}[\bar{y}] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[y_i] + \frac{1}{n^2}\sum_{i=1}^n \sum_{j \neq i} \mathrm{Cov}(y_i, y_j)$$
  其中，前一项是 $O(1/n)$(以 $1/n$ 的速度在变化)，然而后一项却是 $O(1)$。变量间通常是正相关的，因此当存在相关性时，方差收敛的速度会大大下降，甚至不收敛。忽略聚类结构将明显低估估计量的方差。
  - 上述问题早已提出。这篇文章贡献在于指出了大样本下计算机计算效率的问题。显然，如果需要估计的方差-协方差矩阵以 $O(n^2)$ 的速度在变化，那么大样本将对计算机的内存提出巨大的挑战，尤其是使用 Bootstrap 的时候，因此需要对一些算法进行优化。

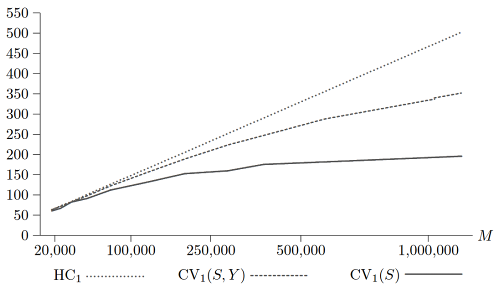
聚类结构与大样本理论

- "How Much Should We Trust Differences-In-Differences Estimates?" (Bertrand et al., 2004, *QJE*)
  - 由于序列相关的存在，DID 方法的推断可能存在过度拒绝的现象。
  - 作者构造了许多安慰剂样本，考察不同类型的标准误对原假设的拒绝率。设定 5% 的显著性水平时，一个合理的标准误应该大约拒绝 5% 的安慰剂样本。
  - 按州聚类的表现很好，而“州 $\times$ 时间”聚类的表现很差，完全不聚类的表现更差。
  - <u>结论：使用 DID 方法的时候，至少要聚类到接受处理的地理层级。</u>

聚类结构与大样本理论

- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - 在一个现有数据集 (obs.=1,156,597) 中随机抽取不同样本量的子样本，考察不同子样本下的异方差稳健标准误 HC_1、聚类到“州 $\times$ 时间”的标准误 $\mathrm{CV}_1(S,Y)$ 和聚类到州的标准误 $\mathrm{CV}_1(S)$。
  - <u>随着样本量的增大，聚类稳健标准误的缩小存在下界，忽略聚类结构导致的标准误低估会越来越严重。</u>

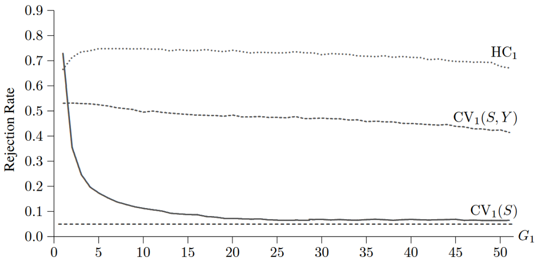
Inverse of $s(\hat{\varphi})$ as a Function of M



聚类结构与大样本理论

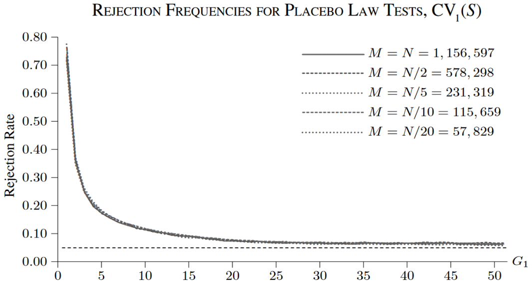
- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - 和 Bertrand et al. (2004) 的做法一样，构造安慰剂样本，G_1 代表构造处理组州的数量，从而考察处理组聚类数量对推断的影响。
  - 在处理组聚类数量很少的时候，三类标准误都有明显的过度拒绝现象，甚至聚类到州的标准误过度拒绝更加严重。<u>但当处理组聚类数量大于 20 以后，聚类到州的稳健标准误基本上就收敛到 5% 了。</u>

Rejection Frequencies for Placebo Law Tests, $N = 1{,}156{,}597$



聚类结构与大样本理论

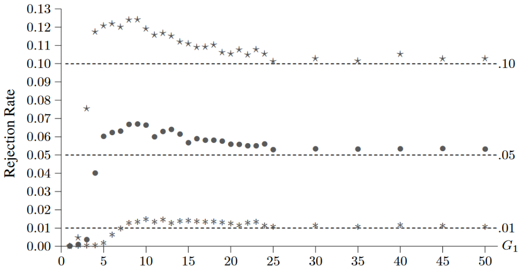
- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - 在聚类结构给定的情况下，增加每个聚类中的样本数量，对于更准确地推断基本没有帮助。
  - 直觉理解：增加聚类内的样本增加了我们对误差项方差的认识，但同时也引入了其与聚类内其他样本的更多的相关性，因此其对于我们更准确地估计误差项的帮助很小，远远小于增加一个新聚类能够带来的帮助。



聚类结构与大样本理论

- "Inference with Large Clustered Datasets" (MacKinnon, 2017, *L'Actualité économique*)
  - 当处理组聚类数量很少的时候，不要直接使用聚类稳健标准误，而是使用 wide-cluster bootstrap 标准误。
  - 这种标准误在聚类数量较少时过度拒绝的程度相对更小一些。
  - STATA 命令：`boottest`，在跑完回归 (`areg`、`ivreg2`……) 后使用。

REJECTION FREQUENCIES FOR BOOTSTRAP PLACEBO LAW TESTS



目录

刻画组内相关的产生

- 使用聚类稳健标准误需要对标准误的聚类结构进行预判，而这种预判往往会先有一个理论上的直觉。因此我们研究一开始就需要考虑的问题是，数据生成过程 (DGP) 可能是怎样的？这一 DGP 如何导致了组内相关的出现？
- 最简单的随机效应模型：
  - 组内相关性来源于组内共同拥有的一个变量的同质影响：$u_{gi} = \lambda\varepsilon_g + \varepsilon_{gi}$，其中 $\varepsilon_{gi} \overset{\text{i.i.d.}}{\sim} (0, \omega^2)$ 且 $\varepsilon_g \overset{\text{i.i.d.}}{\sim} (0, 1)$。
  - 因此，聚类 g 内任意两个样本的误差项协方差为：$\text{Cov}(u_{gi}, u_{gj}) = \lambda^2$。
- 更加一般化的数据生成过程：
  - 组内相关性来源于组内共同拥有的一个变量的异质影响：$u_{gi} = \lambda_{gi}\varepsilon_g + \varepsilon_{gi}$。
  - 例如，y_i 是学生的成绩，聚类 g 代表不同的班级，ε_{gi} 是学生的个人特征，ε_g 是班级 g 的教师质量，教师质量的好坏对于优生和差生成绩的影响显然是不同的。

控制了固定效应还需要聚类吗?

- 如果根据聚类控制固定效应，那么相当于在聚类内进行去平均。在一般化的数据生成过程中，误差项将变为:

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

  于是，$\forall i\neq j$，有
$$\mathrm{Cov}(u_{gi}^{*},u_{gj}^{*})=(\lambda_{gi}-\bar{\lambda}_{g})(\lambda_{gj}-\bar{\lambda}_{g})$$
  因此，当且仅当 $\forall i$，有 $\lambda_{gi}=\bar{\lambda}_{g}$ 时，新的误差项不再具有组内相关性，这便对应随机效应模型。
- 所以，如果有十足的把握认为，误差项来源于随机效应的生成过程，那么控制固定效应后不使用聚类标准误，可以得到更有效的推断；但是如果没有把握，那么控制固定效应后仍需使用聚类标准误。

目录

错误地选择聚类层级的后果

- 假设存在两个聚类层级，其中一个较大 (coarse)，另一个较小 (fine)，较大的层级包含较小的层级。假设较大的层级共有 G 个聚类，而每一个大聚类 g 中包含 M_g 个小聚类。
- 由式 (2)，在大聚类下，系数方差是 $(X'X)^{-1}(\sum_{g=1}^{G}\Sigma_g)(X'X)^{-1}$，将 Σ_g 写成 M_g 个小聚类的方差协方差矩阵之和，可以得到：

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

- 假如选择了小聚类，则系数方差是：

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

错误地选择聚类层级的后果

- 如果真实的聚类结构是大聚类，而操作中选择了小聚类，对上述方差夹心估计量的“心”的低估为：

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

  而且随着样本量增大，错误选择聚类层级导致的标准误低估也会越大。
- 如果真实的聚类结构就是小聚类，相当于式 (4) 中等号右边为 0，聚类到大聚类的方差估计量依然是一致的，但会因为估计过多不必要的数而降低估计效率，从而高估方差。但不意味着聚类到较大的层级永远是较为保守的，前面提到了，如果聚类数量很少，也可能产生过度拒绝。

两种选择聚类层级的经验法则

- 最简单直接的方法:选择所有可行的聚类层级中最大的那一级 (Cameron and Miller, 2015)。如果最大的聚类数量依然较多的话,这种做法是合适的;但如果聚类数量很少或者聚类之间差异性很大的话,这种做法并不好。
- "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."

两种选择聚类层级的经验法则

- 确实在一些例子中，聚类内部的样本之间的误差项是负相关的，此时聚类标准误要比异方差稳健标准误更小，于是可以考虑不聚类。
- "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."

两种选择聚类层级的经验法则

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

两种选择聚类层级的经验法则

- 另一种相对保守的方法：对于感兴趣的系数，汇报所有可行的聚类层级中最大的标准误 (Angrist and Pischke, 2008)。这种做法通常会得到和前一种方法一样的结果，但也不一定，因为当聚类数量很少时，反而可能估得一个很小的标准误。因此这种方法比前一种更加保守，但同时推断效率也会下降。
- *Mostly harmless econometrics: An empiricist's companion* (Angrist and Pischke, 2008)
    - 作者是在比较异方差稳健标准误和经典标准误的时候提到这个经验法则的。
    - "稳健标准误并非灵丹妙药，在两个原因的作用下，它可能比普通标准误更小：我们已经讨论过的小样本偏误以及过大的抽样方差。因此，当稳健标准误小于传统标准误时，我们认为它是经验研究结论可能存在问题的危险信号。很可能是因为存在某种偏误或者没有考虑到的情况，才会出现稳健标准误小于传统标准误。"
    - "在这个看法下，取传统标准误和稳健标准误的最大值可能可以更好地度量估计的精确性。这种经验法则在两个方面有所帮助：它剔除了稳健标准误估计中可能得到的较低值，降低了偏误并且降低了波动。"

目录

一种检验聚类层级的方法

- "Testing for the appropriate level of clustering in linear regression models" (MacKinnon et al., 2023, *JoE*)
  - 核心思想：如果基于小聚类的聚类标准误是有效的，那么小聚类标准误和大聚类标准误应该差不多，即式 (4) 等号右边等于 0。
  - 定义 $\Sigma_c = \sum_{g=1}^{G}\sum_{h_1=1}^{M_g}\sum_{h_2=1}^{M_g}\Sigma_{g,h_1h_2}$ 和 $\Sigma_f = \sum_{g=1}^{G}\sum_{h=1}^{M_g}\Sigma_{gh}$，则原假设为：

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

构造估计量：$\hat{\Sigma}_f$、$\hat{\Sigma}_c$、$\widehat{\mathrm{Var}}[\hat{\Sigma}_f]$ 和 $\widehat{\mathrm{Var}}[\hat{\Sigma}_c]$，然后构造推断统计量：

$$\text{在原假设下,}\ \tau_\sigma = \hat{\boldsymbol{\theta}}/\sqrt{\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})} \xrightarrow{\mathrm{d}} N(0,1),\ \text{其中}\hat{\boldsymbol{\theta}} \equiv \mathrm{vech}(\hat{\Sigma}_c - \hat{\Sigma}_f)$$

  - STATA 命令：`mnwsvt`
    下载网址：`http://qed.econ.queensu.ca/pub/faculty/mackinnon/svtest`，将 `mnwsvt.ado` 文件粘贴到 STATA 安装目录的`./ado/plus/m` 文件夹中，然后把 `mnwsvt-readme.txt` 文件改名成 `mnwsvt.hlp` 也放到该文件夹中 (生成 help 文档)。

一种检验聚类层级的方法

- "Testing for the appropriate level of clustering in linear regression models" (MacKinnon et al., 2023, *JoE*)
  - 需要说明的是，这种检验方法实际上是在做预检验，本身可能导致过度拒绝。比如说，如果真实的聚类结构是大聚类，那么大聚类得到的标准误将比小聚类更大。然而在上述检验中，可能出现没能拒绝原假设而依然使用小聚类的情况。
  - 根本原因在于，任何预检验本身就存在犯两类错误的可能性，因此基于预检验的任何后续推断都应该进行联合假设检验。如果我们没有这么做，就可能导致过度拒绝。
  - 所以，这种预检验方法实际上是不如第二条经验法则那么保守的。操作上，我们根据第二条经验法则选择聚类层级，如果同时又在检验中看到确实应该选择这一层级，那么将使得聚类稳健推断更有说服力。
  - 注：貌似 `mnwsvt` 这个命令不是特别好用，我自己试的时候经常报错。根据原文的理论推导，可能该命令要求检验的小聚类必须嵌套 (nested) 在检验的大聚类中，如果不是这样的聚类设定，命令就会报错。

安慰剂检验

- 前面提到的安慰剂检验也是一种检验手段，即人为重复随机生成安慰剂样本，然后在某一显著性水平下对所有安慰样本进行估计和假设检验，考察所有安慰剂样本中对原假设的拒绝率，并于设定的显著性水平进行比较。
- 安慰剂检验的核心思想在于：由于处理变量是人为随机生成的，因此必然与被解释变量是无关的，如果我们为每个安慰剂样本的假设检验设定 5% 的显著性水平，那么在大量安慰剂样本下，正确的标准误将大约拒绝 5% 的安慰剂样本。由于这一检验并没有对误差的误差项的方差结构做任何参数假设，因此该检验流程并不受到组内相关的影响。
- 貌似没有专门的 `STATA` 命令，但实际上自己写代码实现模拟过程并不困难，参考现有研究中进行 DID 安慰剂检验的代码即可。
  如：`http://ciejournal.ajcass.org/Magazine/show/?id=69658`，选择“下载附件”，参考 do-file 中 287-322 行。
- 和上一页提到的检验方法一样，安慰剂检验也是一种预检验，在没有进行联合假设检验时不如第二条经验法则保守，因此最好还是基于第二条经验法则选择聚类层级，然后用上一页的检验方法和安慰剂检验作为补充证据。

安慰剂检验

- 请一定注意，DID 中随机构造处理组的安慰剂检验方法是用来论证统计推断的有效性的，并不能说明因果识别的好坏。千万别写“通过安慰剂检验表明 DID 估计没有内生性”之类的表述了。
- 如何解释 DID 中安慰剂检验的结果？
- "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."

安慰剂检验

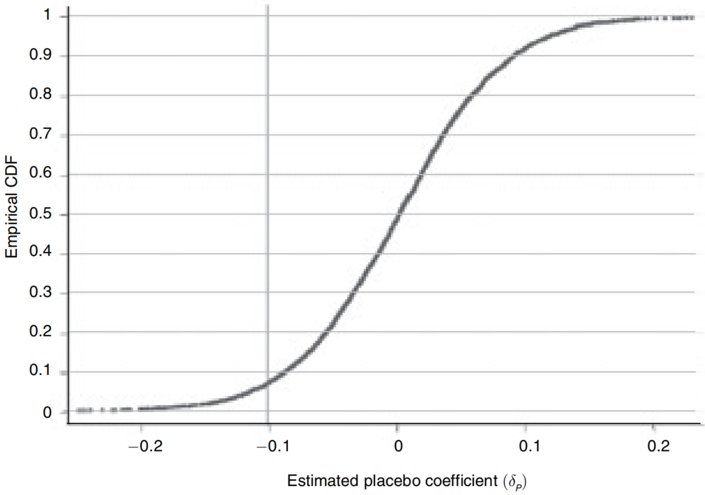


FIGURE 1. DISTRIBUTION OF PLACEBO ESTIMATES: LOG QUANTITY

目录

双向聚类结构

- 有的时候，组内相关并不仅仅存在一种分组的划分中，在多个不同维度的分组中可能同时存在组内相关。例如面板数据中，个体可能在地理层面存在组内相关，同时又在时间层面存在组内相关。
- 假设数据存在两种维度的聚类结构：聚类 A 和聚类 B，分别有 G 和 H 个聚类，用下标 g 和 h 来表示，则估计方程可以写成：

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

  其中下标 gh 代表同时属于第 g 个聚类 A 和第 h 个聚类 B 的样本。
- 请注意，双向聚类与前面提到的“州 $\times$ 时间”的聚类结构不是一回事。“州 $\times$ 时间”的聚类结构假设了仅同一时间且同一州的样本才存在误差项的相关性 (即位于不同 X_{gh} 的样本误差项是不相关的)；然而，双向聚类中同一时间的样本均存在误差项的相关性，同一州的样本也均存在误差项的相关性 (因此位于不同 X_{gh} 的样本也可能存在误差项的相关性)。

双向聚类结构

- 在双向聚类结构下，系数估计值的方差可以写成：

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

其中，$s_{gh} \equiv X'_{gh}u_{gh}$，只有当 $g \neq g'$ 且 $h \neq h'$ 时，才有 $\mathrm{E}[s_{gh}s_{g'h'}] = 0$。

- 与前面类似，定义 $\Sigma_g \equiv \mathrm{E}[s_g s'_g]$, $\Sigma_h \equiv \mathrm{E}[s_h s'_h]$, $\Sigma_{gh} \equiv \mathrm{E}[s_{gh}s'_{gh}]$，可以证明：

$$\sum_{g,h,g',h'} \mathrm{E}[s_{gh}s'_{g'h'}] = \sum_{g=1}^{G}\Sigma_g + \sum_{h=1}^{H}\Sigma_h - \sum_{g=1}^{G}\sum_{h=1}^{H}\Sigma_{gh}$$

双向聚类稳健标准误

- 用残差替换 s_g、s_h 和 s_{gh} 中的误差项，并进行自由度的调整，可得到其样本估计量 $\hat{s}_g$、$\hat{s}_h$ 和 $\hat{s}_{gh}$，进而可以得到系数估计值方差的估计量：

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

- `STATA` 命令：在 option 中添加 `cluster(聚类 A, 聚类 B)`，这一做法对于 `reghdfe` 和 `ivreg2` 均适用。

目录

渐进理论

- 前面的理论部分基本都是在总体上考察，还没有过多地涉及样本对总体的近似。下面我们转向对于渐进理论的考察，即讨论在什么情况下渐进推断能够更好地生效。
- 检验单一假设时，若原假设为 H_0: $a'\beta = a'\beta_0$ 构造的 t 统计量为：

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

其中 $\hat{V}$ 就是系数方差协方差矩阵的估计量，可以是前面提到的 CRVE_1、CRVE_2 和 CRVE_3。

渐进理论成立的条件

- *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."
- 我们可以将聚类稳健推断理解为两步:
  - 第一步,分别在每个聚类内部估计出 score vector $s_g \equiv X_g' u_g$ 的方差协方差矩阵 $\hat{\Sigma}_g$。
  - 第二步,将各个聚类视为样本,此时聚类的数量就是样本的数量,扰动项方差的结构则是异方差结构,如式 (1) 所示。

渐进理论成立的条件

- 渐进推断能够起作用，依赖于大数定律与中心极限定理能够发挥效果。从上一页中提到的“两步”出发，则要求：
  - 首先，需要大数定律成立，使得 $\sum_{g=1}^{G} \hat{s}_g \hat{s}_g'$ 能够依概率收敛到方差协方差矩阵 $\sum_{g=1}^{G} \Sigma_g$(因此需要样本量足够大)。
  - 其次，需要中心极限定理成立，使得向量 $\sum_{g=1}^{G} s_g$ 能够依分布收敛于一个多元正态分布，该分布方差为 $\sum_{g=1}^{G} \Sigma_g$(因此需要聚类足够多)。
- 在聚类结构下，当渐进理论中说样本量 $n \to \infty$ 时，是什么含义呢？
  - 含义 1：保持聚类中样本量不变，让聚类数量 G 趋于无穷大 (large number of clusters)。
  - 含义 2：保持聚类数量不变，让每个聚类中的样本数量 N_g 趋于无穷大 (small number of large clusters)。
- 直觉上，当样本之间的异质性更小的时候，中心极限定理和大数定律在有限样本中能更好地发挥作用。显然，增加一个新聚类引入的误差项相关性明显小于保持聚类不变而增加相应的样本量。在含义 1 下，渐进理论能更好地成立；而在含义 2 下，可能由于聚类数量太少而导致错误推断。

渐进理论成立的条件

- 同样直觉上能判断的是，给定相同的聚类数量，中心极限定理在聚类间差异程度更小的有限样本中 (更具体地说，s_g 之间差异更小的样本中) 能够更好的发挥作用。
- 综上，可以给我们三个方面的启示：
  - 第一，聚类的数量对于中心极限定理的效果至关重要，因此在聚类结构下，我们不能只关心样本量，还应该关心聚类的数量。
  - 第二，在有限样本下，聚类的数量不是渐进理论效果好坏的唯一因素，聚类间的差异性也决定了渐进理论的效果。因此，不会有一个普适的聚类数量 G^* 能够保证，当我们数据中聚类数量 G 大于 G^* 时就可以高枕无忧。
  - 第三，在使用聚类稳健推断的时候，还应该关注聚类的差异性 (比如样本量的差异性)，要警惕一些影响力 (杠杆值) 很大的聚类。

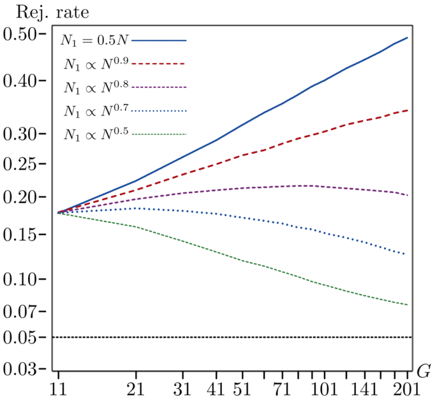
目录

大量聚类的情形

- 假定各聚类都有 M 个样本，易证原假设下有：$\sqrt{G}(\hat{\beta}-\beta_0)\xrightarrow{\mathrm{d}}N(0,G\hat{V})$。
- 然而应用中，各个聚类的样本大小不同，样本量的差异程度将影响 s_g 之间的差异程度，从而影响中心极限定理在有限样本中的效果。比如，如果有一个聚类中的样本比其他聚类都多得多，那么它可能在整个推断中影响很大，但是这样一来中心极限定理就不容易发挥效果。

大量聚类的情形

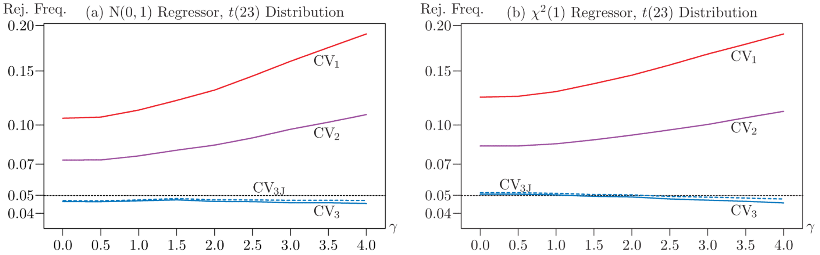
- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - 当一半的样本在一个聚类中的时候，聚类数量的增加反而加剧了过度拒绝。



(a) CRVE t-tests

大量聚类的情形

- "Fast and reliable jackknife and bootstrap methods for cluster-robust inference" (MacKinnon et al., 2023, *Journal of Applied Econometrics*)
  - 构造模拟样本，γ 衡量了聚类间样本量的差异。当 $\gamma = 0$ 时，样本在各聚类均匀分布；当 $\gamma = 2$ 时，各聚类的样本量从 130 递增至 899。
  - 相比起 CRVE_1 和 CRVE_2，CRVE_3 在聚类间样本差异很大的时候表现更好，但并非使用 CRVE_3 就能完美避免聚类间差异性导致的偏误。
  - "However, as we shall see, there are also many cases in which CV_3 overrejects, and CV_3 therefore overrejects slightly more. In practice, it would be perfectly reasonable to report either CV_3 or $\text{CV}_{3\text{J}}$."



大量聚类的情形

- 综上，即便有大量的聚类，依然至少有两种情况会导致聚类的渐进标准误出现严重的过度拒绝：
  - 处理组的聚类很少(这个在 slides 的第15页已经看到了)；
  - 少数聚类的样本量巨大 (高杠杆聚类)，导致其对估计结果的潜在影响力很大。
- 因此，我们需要有一些方法来判断聚类的影响力，并在研究中汇报是否具有高杠杆值的聚类。

聚类的影响力与杠杆值

- 参考刻画单一样本的影响力的思路，如果一个聚类具有很强的印象里，那么将其从样本中删去后，将会对系数的点估计值产生很大的影响。将聚类 g 从样本中去掉后新的系数点估计值：

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

- 首先我们考察聚类的杠杆值。我们知道，帽子矩阵 $P_x = X'(X'X)^{-1}X'$ 的主对角线元素刻画了每个样本的杠杆值，因此将其中属于某一聚类 g 的主对角线元素相加，即可得到聚类 g 的杠杆值，这等价于求 $H_g = X_g(X'X)^{-1}X_g'$ 的迹。因此刻画聚类杠杆值的表达式为：

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

  由此也可以看出为何大聚类的杠杆值会很大，以及为何一个异质性强的大聚类将对结果产生很强的影响力。
- 判断聚类是否为高杠杆聚类，可以比较其杠杆值与平均值。易证：

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

聚类的影响力与杠杆值

- 在经验研究中，我们通常只关心某一个系数，因此可以使用 FWL 定理计算只针对变量 j 的聚类偏杠杆值：

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

其中 $\acute{x}'_j$ 代表以变量 j 作为被解释变量的回归残差。

- 貌似没有现成的 STATA 可以直接计算聚类杠杆值和偏杠杆值，但因为上述计算方法很简单，我们用 Matlab、Python 或 STATA Mata 等自己算就行了。

目录

固定数量的大聚类的情形

- 前面含义 1 中提到的由杠杆聚类导致的偏误在含义 2 中同样存在。
- 同时，在含义 2 中，聚类数量少将导致渐进理论的效果往往不如含义 1，因此还必须施加额外的假设，这些额外的假设本身就体现了含义 2 在推断中将遭遇的问题。
- "Inference with dependent data using cluster covariance estimators" (Bester et al., 2011, *JoE*)
  - 证明了含义 2 下渐进理论也能成立，但必须建立在很强的假设上：
    - "... 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 $M \to \infty$."(这一条假设尤其不现实。)

使用渐进标准误进行聚类稳健推断: 总结

- 这一节讨论的是使用渐进的聚类标准误时，渐进理论有时难以成立的问题。
- 概括起来，问题大概是：我们根据渐进理论推测出系数估计值 $\hat{\beta}$ 在大样本下的渐进分布，然后使用样本去估计这一渐进分布的方差，最后进行推断；然而，在一些情况下，渐进理论在有限样本中的效果并不好，因此系数估计值的真实分布与其渐进分布并不接近，此时我们使用其渐进分布来进行推断就会不准确。
    - 聚类数量 G 很少、聚类间差异性很大……这些都会影响渐进理论在有限样本中的效果。
    - 然而，我们很难提出一个标准去判断到底多大的 G、多小的聚类异质性就能够使渐进理论成立。即使使用相对保守的 CRVE_3，也依然可能过度拒绝。
- 如果有一种方法是对系数估计值的真实分布直接进行估计的 (而不是其渐进分布)，那么将具有更好的有限样本性质。
- "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."

目录

自助法的直觉理解

- 假设我们想要使用某检验统计量 τ 进行推断，我们可以从原始样本中抽样，使用 B 个抽样样本计算出 B 个 bootstrap 统计量 $\tau_b^*(b=1,\cdots,B)$，从而得到 τ 的经验分布，该经验分布往往是 τ 的真实分布的良好近似。
- 显然，B 取越大的值越有利于推断的准确性，但计算时间通常与 $B\times N$ 成正比。如果不担心计算时间太长，令 B 取 9,999 或 99,999 即可。
- 利用该经验分布，我们可以构造出 bootstrap 的 p 值以及系数的置信区间。

聚类自助法

- 当数据具有聚类结构的时候，构造抽样样本的时候必须考虑聚类结构，因此聚类结构下的自助法称为 cluster bootstrap、block bootstrap 或 resampling by cluster。
- 一些使用时的问题：
  - 如果存在一些影响力很大的聚类，那么自助法得到的经验分布可能存在不同的形状，因此即使使用自助法，我们依然需要关心聚类的影响力和杠杆值，同时我们需要留意自助法得到的经验分布的形状。
  - 自助法在大样本下的计算量很大，导致计算时间很长，可以通过一些优化算法来减少计算时间，在 `STATA` 中可以使用 `boottest` 命令来实现更快的计算。
  - 实际上，线性模型中并不推荐使用上述的聚类自助法，有更好的方法可以用，即原始聚类自助法 (wild-cluster bootstrap)。

目录

原始聚类自助法

- wild-cluster bootstrap 是一种基于残差的，resampling 方法，并非在原始样本中抽取子样本，而是通过随机调整原始样本中的残差来构造出大量新样本。
- 有约束的原始聚类自助法 (restricted wild-cluster bootstrap, WCR)：假设在约束条件 $a'\beta = a'\beta_0$ 下，Constrained Least Squares (CLS) 得到的系数估计值是 $\tilde{\beta}$，则有约束的残差是 $\tilde{u}_g = y_g - X_g\tilde{\beta}$，于是，bootstrap 样本 b 的数据生成过程为：

$$y_g^{*b} = X_g\tilde{\beta} + u_g^{*b}, \text{ 其中} u_g^{*b} = v_g^{*b}\tilde{u}_g \text{且标量} v_g^{*b} \overset{\text{i.i.d.}}{\sim} (0,1)$$

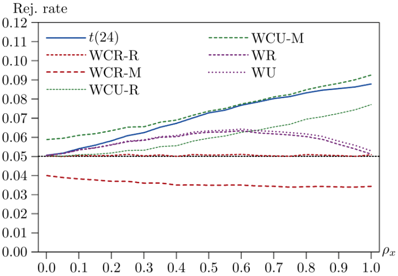
  通常设定 v_g^{*b} 分布是 Rademacher 分布，也有用 Mammen 分布等其他的。
- 无约束的原始聚类自助法 (unrestricted wild-cluster bootstrap, WCU)：和 WCR 思路类似，仅仅是用 OLS 替代 CLS，或者说用无约束的残差 $\hat{u}$ 替代有约束的残差 $\tilde{u}$。
- 直觉：由于 v_g^{*b} 是独立同分布的，因此这一新样本的 DGP 保证了组间不具有相关性。

原始聚类自助法

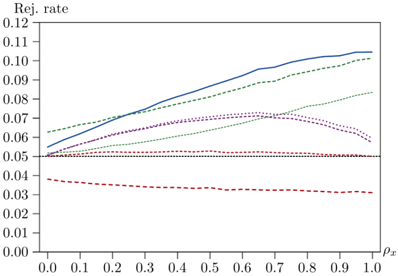
- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - 在有限样本中，WCR 比 WCU 的表现更好。
  - 实际上，我们可以不在聚类的层级上构造样本，而在样本的层级上构建新样本，此时要生成 n 个随机变量 u_i^{*b} 并与每个样本的 (有约束/无约束) 残差相乘。这种做法不仅保证了组间不相关，而且保证了组内也不相关。如果使用有约束残差，则称为有约束的原始自助法 (restricted wild bootstrap, WR)；如果使用无约束残差，则称为无约束的原始自助法 (unrestricted wild bootstrap, WU)。
  - 大多数情况下，WR 的表现不如 WCR；而且在大样本下，WR 的计算时间会很长。
- STATA 命令：`boottest`(上述的所有原始自助法 (WCR、WCU、WR 和 WU) 都可以实现)。

原始聚类自助法

- "Asymptotic theory and wild bootstrap inference with clustered errors" (Djogbenou et al., 2019, *JoE*)
  - 随着 ρ_x 越来越大 (代表聚类内样本误差项相关性越来越大)，聚类稳健标准误 (图中 $t(24)$) 的过度拒绝越来越严重，而在各种原始自助法中，WCR 的表现最好。
  - 通过模拟来考察聚类样本量大小的影响，左图中 $\gamma = 0$ 代表 25 个聚类中样本量都是 100，右图中 $\gamma = 2$ 代表 25 个聚类的样本量从 32 递增至 234。比较左右两张图，则说明在样本分布不均匀的情况下，过度拒绝会更加严重。



(a) $\gamma = 0$



(b) $\gamma = 2$

使用 WCR 的注意事项

- "We recommend using at least one variant of the WCR bootstrap (preferably with at least $B = 9{,}999$) almost all the time."
- WCR 也不是万能的，当处理组的聚类数量很少的时候 (甚至只有一个处理组时)，WCR 往往会变得非常保守 (见 slides 第17页，拒绝率达到了 0)。如果我们发现 WCR 得到的 p 值大的十分可疑，可以尝试使用 WR 方法。
- 当样本的结构是前面提到的含义 2(small number of large clusters) 的时候，WCR-R 方法有时也能得到准确的推断，但并不绝对！("The wild bootstrap with a 'small' number of 'large' clusters", Canay et al., 2021, *REStat*)
- 如果 G 很小，那么对于 Rademacher 分布这种两点分布，其取值的所有可能性只有 2^G 中，可能比 resample 的数量 B 还要小，此时就不再具有抽样的随机性了。因此如果 G 很小，可以尝试使用其他分布 (`boottest` 命令中可以选择不同的分布)。

目录

经验研究者应该汇报什么？

1. 除了汇报大家通常使用的 CRVE_1，至少还应该汇报 1-2 种其他类型的标准误：比如汇报 CRVE_3 以及某一种 WCR 自助法。
2. 正如前面理论部分提到的，在聚类结构下，推断的有效单位是聚类，而不是样本，所以除了汇报样本量 n，更重要的是汇报聚类数量 G。
3. 样本数量在聚类间的差异性是决定中心极限定理效果的因素之一。因此，除了汇报 n 和 G，还应该要汇报各聚类样本数量的中位数、最小值和最大值；如果可以的话，最好把样本量 n_g 的分布图画出来。
4. 为了更好地刻画聚类的异质性，还应该汇报各个聚类的杠杆值 L_g、感兴趣的系数的偏杠杆值 L_{gj} 以及 $\beta_j^{(g)}$。如果聚类数量 G 比较小，那么可以全部汇报出来；如果 G 比较大的话，可以画分布图。
5. 如果使用的是双向聚类 (甚至更高的维度)，那么应该汇报每个聚类及其交集的杠杆值、偏杠杆值与 leave-one-out 系数，所以每个结果至少有三组分布。

目录

实操指南

1. 开始经验研究之前，先把手头数据所有可行的聚类维度列好，然后想想聚类到哪个维度是最合适的。这一判断首先基于理论，即我们认为数据的 DGP 是什么样的，组内相关是怎么产生的。实操中可以使用一种保守的方法，即对感兴趣的系数取所有聚类层级中标准误最大的那个，同时辅以一些正式检验予以支持 (比如 MNW 方法和安慰剂检验，但别忘记在没有进行联合假设检验的情况下是会导致过度拒绝的)。对于 DID 方法，至少要聚类到接受处理的层级。
2. 在选定聚类层级以后，在研究中汇报聚类的数量 G 以及各聚类样本量 n_g 的中位数、最小值和最大值；可以的话，直接呈现 n_g 的分布。
3. 对于研究者最偏好的模型设定，汇报各聚类的杠杆值、偏杠杆值和影响力的信息，这一点对于 DID 方法或者其他估计处理效应的模型来说都非常重要。

实操指南

4. 除了使用 CRVE_1，还应该使用 CRVE_3 和某一种 WCR 自助法。在聚类足够多的情况下，它们得到的结果往往比较类似；但是如果聚类太少或者聚类间差异太大而导致不同聚类标准误的结果差异很大，那么除了汇报上述三者，最好再多多尝试其他类型的标准误。
5. 对于在聚类层面接受处理的模型来说，如果处理组或对照组的数量很少或不典型，即使用 CRVE_3 或者 WCR 也可能得不到可靠的推断，可以考虑使用随机化推断 (randomization inference, RI) 方法 (即 slides 第31-33页讲的安慰剂检验)。