Cluster-Robust Variance Estimators for Instrumental Variables

Theory and Application in Assessing the Impact of Gender Representation on Carbon Emissions

Sam Lee

04/15/2024

Background

- Bertrand et. al (2004) → cluster standard errors in DiD
    - Standard errors otherwise inflated
- MacKinnon et. al (2023)
    - Guide for cluster-robust variance estimators
    - Three “feasible” CRVEs
        - Derived from OLS Asymptotic Variance
    - “For such models [with clustered data estimated by instrumental variables (IV)], neither the current state of econometric theory nor the available simulation evidence allows us to make recommendations with any confidence”

Three Feasible CRVEs

Assuming data are generated at $\beta = \beta_0$,

$$\sqrt{n}(\hat{b} - \beta_0) \xrightarrow{d} N\left(0, \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[x_i x_i' u_i^2] \mathbb{E}[X_i X_i']^{-1}\right)$$

Assuming data can be divided into G disjoint groups, it follows that a reasonable way to "cluster" standard errors by group (g) will depend on the residual errors from that group:

$$\mathbb{E}[x_i x_i' u_i^2] \approx \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}\, \mathrm{x_i'} \hat{u}_i^2 \rightarrow \frac{1}{n} \sum_{g=1}^{G} X_g{}' \hat{u}_g \hat{u}_g' X_g$$

Thus, the first CRVE follows as (and with a degree of freedom correction applied),

$$CV_1 : n \frac{G(N-1)}{(G-1)(N-k)} (X'X)^{-1} \sum_{g=1}^{G} \hat{s}_g \hat{s}_g{}' \ (X'X)^{-1}; \hat{s}_g = X_g' \hat{u}_g$$

Three Feasible CRVEs

However, sometimes $X_g'\hat{u}_g$ are not the best estimators for $\hat{s}_g$ (Bell & McCaffery, 2002).

Two Alternative CRVEs:

$$CV_2\colon n(X'X)^{-1}\sum_{g=1}^{G}\grave{s}_g\grave{s}_g{}'\ (X'X)^{-1};\ \grave{s}_g = X_g'M_{gg}^{-\frac{1}{2}}\hat{u}_g;M_{gg} = I_{Ng} - X_g(X'X)^{-1}X_g{}'$$

$$CV_3\colon n\frac{G-1}{G}(X'X)^{-1}\sum_{g=1}^{G}\acute{s}_g\acute{s}_g{}'\ (X'X)^{-1};\acute{s}_g = X_g'M_{gg}^{-1}\hat{u}_g$$

A GMM Framework for Just-Identified IV

Assumptions for Identification in GMM:

$$\mathbb{E}\left[m_i(\beta) = Z_i\left(Y_i - X_i'\beta\right)\right] = 0; \quad G(\beta) \coloneqq \left[\frac{\partial m_i(\beta)}{\partial \beta'}\right] = -\mathbb{E}\left[Z_i X_i'\right] \quad \text{(has rank k)}$$

$$\sqrt{n}\left(\hat{b} - \beta_0\right) \xrightarrow[d]{} N\left(0, \left[G(\beta)\mathbb{E}[m_i(\beta)m_i(\beta)']^{-1}G(\beta)'\right]^{-1}\right)$$

$$\rightarrow \sqrt{n}\left(\hat{b} - \beta_0\right) \xrightarrow[d]{} N\left(\left[\mathbb{E}\left[Z_i X_i'\right]\mathbb{E}\left[Z_i Z_i' u_i^2\right]\mathbb{E}\left[Z_i X_i'\right]'\right]^{-1}\right)$$

2SLS Variance Estimator with Homoskedasticity:

$$\hat{V} = \left[\left(\frac{1}{n}X'Z\right)\left(\frac{1}{n}Z'Z\hat{u}_2\right)^{-1}\left(\frac{1}{n}X'Z\right)'\right]^{-1}$$

Proposed CRVEs for GMM

For G disjoint clusters,

$$\hat{V}_g = \left[\left(\frac{1}{n} X'Z \right) \left(\frac{1}{n} \sum_{g=1}^{G} Z_g' \hat{u}_g \, \hat{u}_g' Z_g \right)^{-1} \left(\frac{1}{n} X'Z \right)' \right]^{-1}$$

Thus, following MacKinnon et. al, all three CRVEs can be rewritten as follows:

$$\underset{CV_1}{iv} : n \frac{G(N-1)}{(G-1)(N-k)} \left[(X'Z) \left(\sum_{g=1}^{G} \hat{\zeta}_g \hat{\zeta}_g' \right)^{-1} (X'Z)' \right]^{-1} \quad ; \quad \hat{\zeta}_g = Z_g' \widehat{u_g}$$

Proposed CRVEs for GMM

$$\underset{CV_2}{iv} : n\left[(X'Z)\left(\sum_{g=1}^{G}\grave{\zeta}_g\grave{\zeta}_g{}'\right)^{-1}(X'Z)'\right]^{-1} \quad ; \quad \grave{\zeta}_g = Z_g'M_{gg}^{-\frac{1}{2}}\hat{u}_g; M_{gg} = I_{Ng} - Z_g(Z'Z)^{-1}Z_g{}'$$

$$\underset{CV_3}{iv} : n\frac{G-1}{G}\left[(X'Z)\left(\sum_{g=1}^{G}\acute{\zeta}_g\acute{\zeta}_g{}'\right)^{-1}(X'Z)'\right]^{-1} \quad ; \quad \acute{\zeta}_g = Z_g'M_{gg}^{-1}\hat{u}_g$$

An Applied Example

What is the effect that electing higher proportions of women into national legislatures in African and Arab nations has on yearly per capita CO_2 emissions?

- Women are disproportionately affected by negative externalities of climate change

- One recent paper suggests a causal link between higher proportions of women in parliament and stricter climate policies (Mavisakalyan & Tarverdi, 2019)

    - Uses cross-sectional data-set

- Intertemporal link?

Identification Strategy

Two-stage difference-in-differences:

$$(1) \quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Psi_p + \sum_{k=\text{Angola}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

$$(2) \quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Omega_p + \sum_{k=\text{Angola;}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Identification Strategy

Two-stage difference-in-differences:

$$(1)\quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Psi_p + \sum_{k=\text{Angola}\ k\neq\text{Qatar, Iraq}}^{Zimbabwe} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

$$(2)\quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Omega_p + \sum_{k=\text{Angola; }k\neq\text{Qatar, Iraq}}^{Zimbabwe} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Endogenous Treatment

Identification Strategy

Two-stage difference-in-differences:

Instrumental Variable – Years Since Country was Granted Suffrage

$$(1) \quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Psi_p +$$

$$\sum_{k=\text{Angola}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

$$(2) \quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Omega_p +$$

$$\sum_{k=\text{Angola;}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Identification Strategy

Two-stage difference-in-differences:

$$(1) \quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Psi_p + \sum_{k=\text{Angola}}^{Zimbabwe}{}_{k\neq\text{Qatar, Iraq}} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

Response Variable – Per Capita Carbon Emissions for a country *c* in year *t*

$$(2) \quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Omega_p + \sum_{k=\text{Angola;}}^{Zimbabwe}{}_{k\neq\text{Qatar, Iraq}} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Identification Strategy

Two-stage difference-in-differences:

$$(1) \quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p} \Psi_p +$$

$$\sum_{k=\text{Angola}}^{Zimbabwe}{}_{k \neq \text{Qatar, Iraq}} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

A sequence of autoregressive lags on economic covariates

$$(2) \quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p} \Omega_p +$$

$$\sum_{k=\text{Angola;}}^{Zimbabwe}{}_{k \neq \text{Qatar, Iraq}} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Identification Strategy

Two-stage difference-in-differences:

$$(1) \quad W_{ct-1} = \pi_0 + \lambda Z_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Psi_p + \sum_{k=\text{Angola}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \mu_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \eta_j \mathbb{1}(\text{Year}_t = j) + v_{ct}$$

$$(2) \quad Y_{ct} = \beta_0 + \delta W_{ct-1} + \sum_{p=2}^{P} X'_{ct-p}\Omega_p + \sum_{k=\text{Angola;}}^{Zimbabwe}{}_{k\neq \text{Qatar, Iraq}} \beta_k \mathbb{1}(\text{Country}_c = k) + \sum_{j=1999}^{2022} \gamma_j \mathbb{1}(\text{Year}_t = j) + \epsilon_{ct}$$

Country and year fixed effects (difference-in-differences)

Regression Results

Do we trust these results?

How do we know which CRVE to use?

Variance Estimator ($\widehat{\boldsymbol{\delta}} = -\mathbf{107.8425}$)	2.5% Lower C.I.	97.5% Upper C.I.
2SLS (Homoskedastic)	-133.37	-82.32
iv CV_1	-164.88	-50.81
iv CV_2	-165.66	-50.02
iv CV_3	-166.86	-48.83

Monte Carlo Analysis

- Assess empirical performance of CRVEs via simulation.
- Randomly generate a new (placebo) vector for Y_{ct}. (Leave $\mathbb{X}_{ct}$ matrix the same)
- Calculate regression coefficients and respective 95% confidence intervals for each generation.
- Calculate empirical coverage for each estimator
- Evaluate whether inference using each of the CRVEs (if any) are reliable

Monte Carlo Analysis - DGP

The data-generating process for Y_{ct} will follow a bootstrap sampling process.

- Replace the empirical sample Y_{ct} with a new bootstrap sample of Y_{ct} each generation, defined as $\dot{Y}_{ct}$.

Let $\dot{Y}_{ct}$ be defined as

$$\dot{Y}_{ct} = \{\dot{y}_{ct}\}_{c=\text{Angola},\ t=1999}^{c=\text{Zimbabwe},\ t=2022} = \{\dot{y}_{Angola,1999}, \dot{y}_{Angola,2000}, \cdots, \dot{y}_{Zimbabwe,2022}\}$$

$$\Pr(\dot{y}_{ct} = y_{ct}) = \frac{1}{n}, \forall c, t$$

Monte Carlo Analysis - DGP

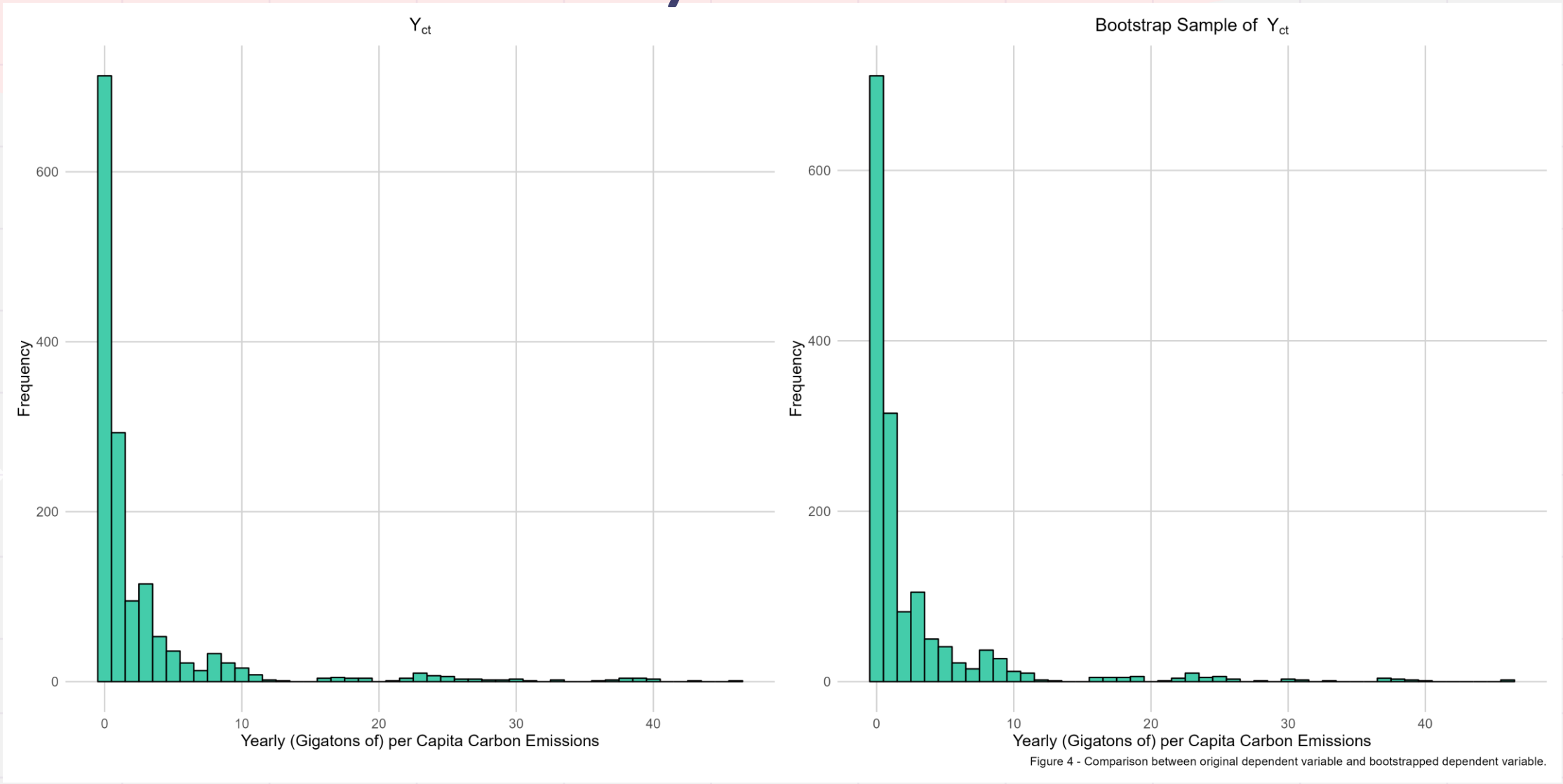

Figure 4 - Comparison between original dependent variable and bootstrapped dependent variable.

Monte Carlo Analysis – Simulation (n=10000)

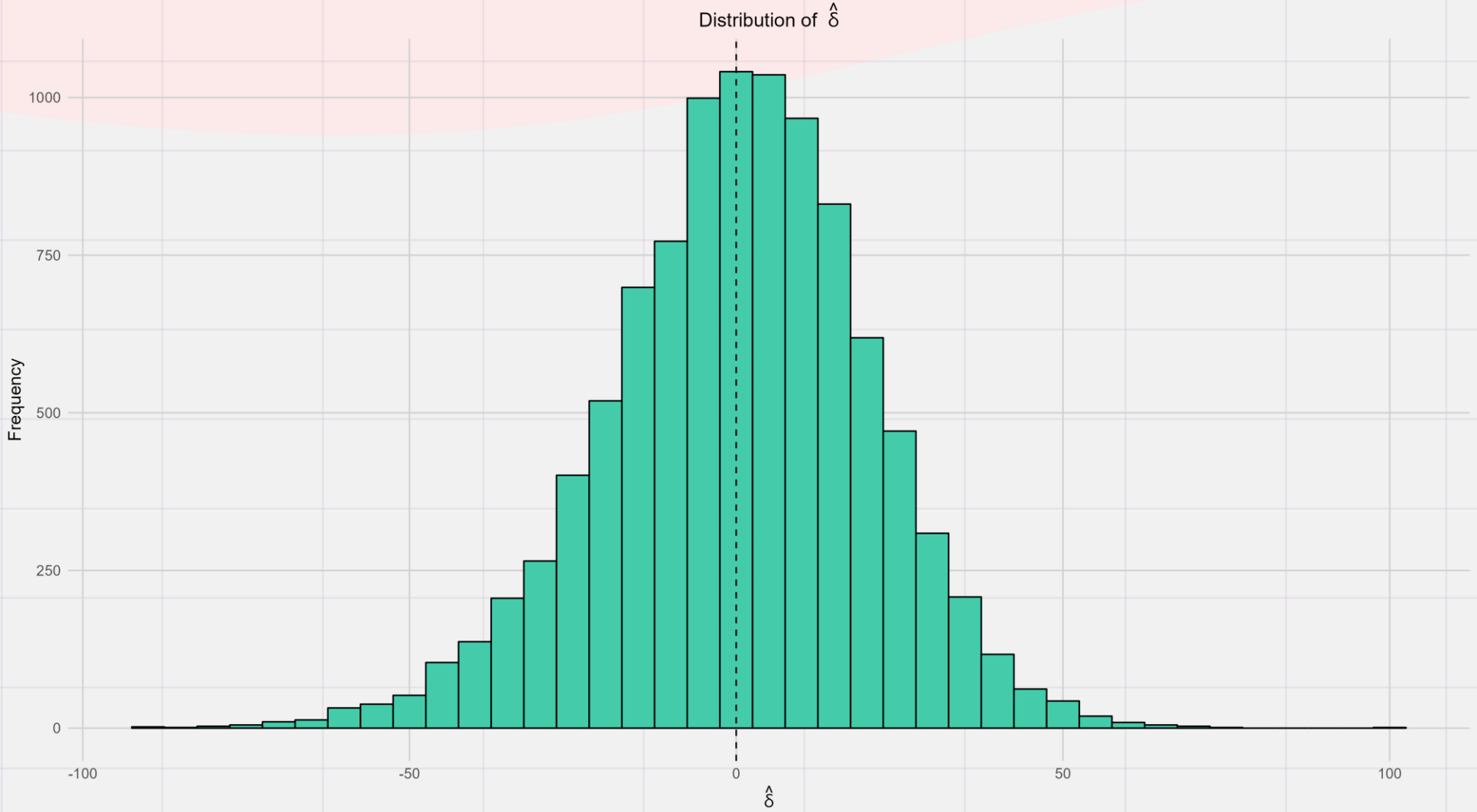

Figure 5 - Sample distribution of the simulated estimators for the local average treatment effect

Monte Carlo Analysis – Simulation (n=10000)

	iv CV_1		iv CV_2		iv CV_3		2SLS	
	(2.5%, 97.%)		(2.5%, 97.%)		(2.5%, 97.%)		(2.5%, 97.%)	
1	-20.24	13.52	-25.83	19.12	-41.9	35.18	-24.45	17.73
2	-20.5	25.35	-22.09	26.94	-24.25	29.1	-16.1	20.95
3	-23.2	9.02	-24.25	10.07	-25.28	11.1	-26.95	12.77
4	-11.07	20.3	-14.11	23.33	-22.51	31.74	-15.92	25.14
5	-10.46	18.29	-12.15	19.97	-19.37	27.19	-12.61	20.43
6	-16.58	31.21	-15.08	29.71	-12.2	26.82	-11.04	25.67
7	-17.97	8.7	-18.8	9.53	-21.25	11.98	-24.55	15.28
8	-16.23	19.49	-20	23.26	-31.98	35.24	-22.71	25.96
9	-20.32	13.15	-20.91	13.74	-23.88	16.71	-21.51	14.35
10	-19.66	12.13	-27.07	19.54	-46.94	39.4	-27.31	19.78
...	...	...	...	...	...	...	...	...
9996	-10.38	22.17	-10.49	22.28	-11.08	22.87	-14.57	26.36
9997	-33.48	15.01	-33.73	15.26	-34.04	15.57	-33.51	15
9998	-13.5	14.85	-15.36	16.71	-19.84	21.19	-18.88	20.23
9999	-43.52	5.72	-47.46	9.66	-57.51	19.71	-39.13	1.33
10000	-13.76	11.52	-13.63	11.39	-14.23	11.99	-20.46	18.22
Empirical Rejection		0.0553		0.0297		0.0118		0.0465

Conclusions

- $\overset{iv}{CV_2}$ *and* $\overset{iv}{CV_3}$ tend to routinely under reject

- $\overset{iv}{CV_3}$ may be the best the CRVE for IV right now

    - Stata's CRVE for IV seems to over reject without DF correction

- If clustering is unnecessary or unfeasible, inference using the GMM asymptotic variance seems to perform just as well

- The test for δ seems appropriately sized

- Hence the interpretation on p-values remains as if higher proportions of women in legislature truly had no effect on yearly carbon emissions within a country (given H_0), the standard errors from $\overset{iv}{C1}$ imply that there would have been about a 0.021% (p-value < 0.001) chance we observed the same effect or larger by random.

- With high confidence, it seems that when more women enter national legislatures, correspondingly, yearly carbon emissions decrease