# The Impact of Sampling Procedure on Statistical Inference with Clustered Data

Shuai Jin
University of Iowa

## Overview

- OLS underestimates the variance of parameters with clustered data.
- Researchers develop and compare multiple methods of handling clustering.
- However, the sampling procedure is not considered in the comparison.
- The sampling procedure determines the structure of clustering in the data, and therefore affects the performance of the different methods.
- I manipulate three sampling procedures in Monte Carlo experiments, and compare eight methods, focusing on bootstrap and jackknife techniques.
- The sampling method affects the variance estimates of both individual-level and group-level factors. Simple random sampling produces stable results.
- Jackknife cluster standard errors perform well across different circumstances.

## Data Generating Process in Monte Carlo simulations

$$y_{ic} = 0 + 5 \times x_{ic} + 3 \times z_c + u_c + e_{ic}$$

$$\rho = \frac{var(u_c)}{var(u_c) + var(e_{ic})}$$

- $x_{ic} \sim N(3.4, 16.36)$, $z_c \sim N(2, 9)$, $corr(x_{ic}, z_c) = 0.2$
- Population: Number of clusters $C = 50$, within each cluster $n = 200$, $N = 10,000$.
- population 1: $\rho = 0.5$, $u_c \sim N(0, 1)$, $e_{ic} \sim N(0, 1)$
- population 2: $\rho = 0.1$, $u_c \sim N(0, 1)$, $e_{ic} \sim N(0, 9)$

## Three Ways of Sampling

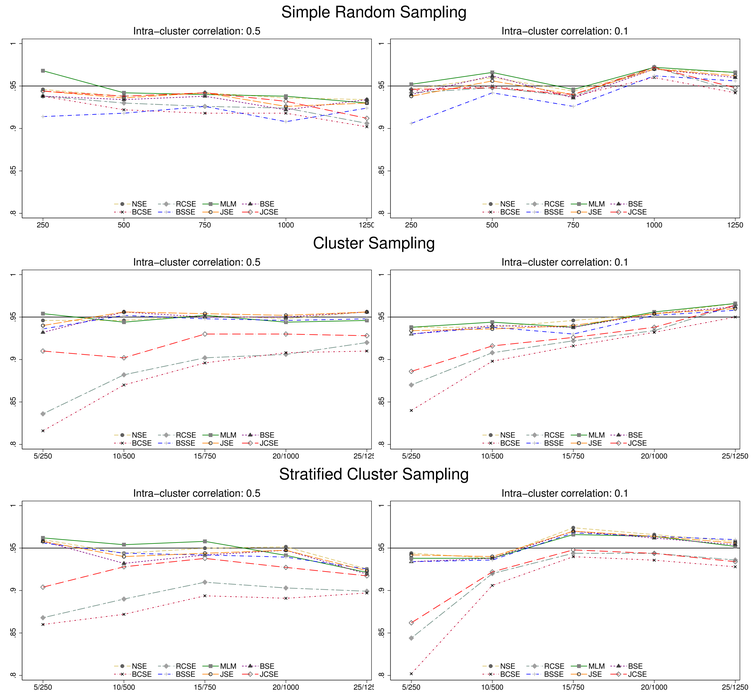
| | |
|---|---|
| Simple random sampling w/o replacement (SRS) | 5 samples |
| No. of clusters | no control |
| Size of clusters | no control |
| Sample size | 250, 500, 750, 1000, 1250 |
| Cluster sampling (CS) | 5 samples |
| No. of clusters | 5, 10, 15, 20, 25 |
| Size of clusters | 50 |
| Sample size | 250, 500, 750, 1000, 1250 |
| Stratified cluster sampling (SCS) | 5 samples |
| No. of strata | 5 (10 clusters in each) |
| No. of clusters from each stratum | 1, 2, 3, 4, 5 |
| No. of clusters | 5, 10, 15, 20, 25 |
| Size of clusters | 50 |
| Sample size | 250, 500, 750, 1000, 1250 |

$500$ replications for each sample $\times 15$ samples $\times 2$ populations

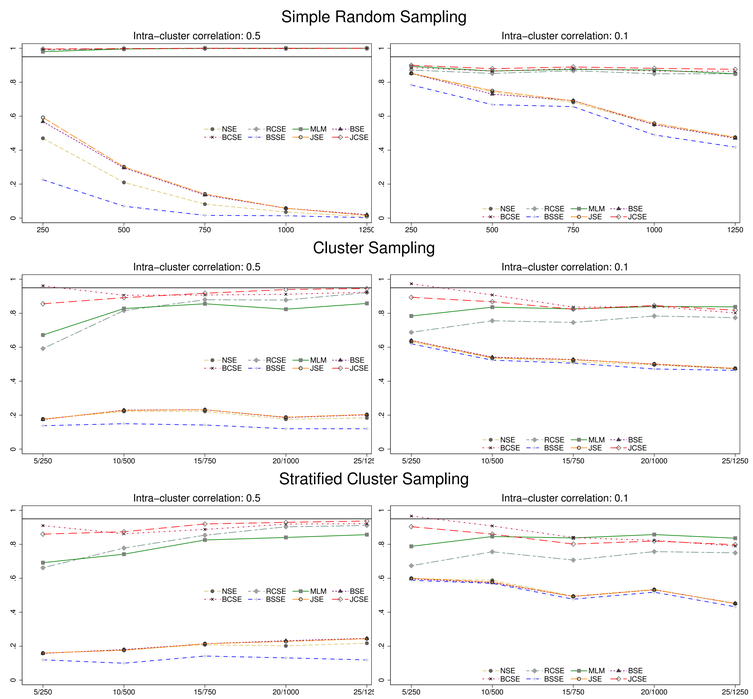
## Eight Methods of Handling Clustering

- OLS Normal standard error (NSE)
- Robust clustered standard error (RCSE)
- Multilevel model (MLM)
- Bootstrap standard error (BSE): resampling 250 times
- Bootstrap cluster standard error (BCSE)
- Bootstrap strata standard error (BSSE)
- Jackknife standard error (JSE)
- Jackknife cluster standard error (JCSE)

## Monte Carlo Results for Individual-Level Factor: $x$ (coverage probability)




## Monte Carlo Results for Group-Level Factor: $z$ (coverage probability)




## Monte Carlo Summaries

- The sampling procedure affects the variance estimates and the performance of the different methods. SRS produces relatively stable results.
- With SRS, all methods perform well for $x$, but this is not the case for the other two sampling methods.
- With CS and SCS, almost all methods underestimate the standard error for $z$, but this is not the case for SRS.
- NSE, BSE, BSSE, and JSE perform well for $x$, but very poorly for $z$.
- BCSE and RCSE are often the worst for $x$.
- MLM perform well for $x$, but not for $z$, especially when $\rho$ is high.
- JCSE is the best method of handling clustering.

## Application to Chinese National Survey Data

- 2005 China General Social Survey data
- Representative national sample: $C = 24, N = 8,042$
- Multi-step stratified cluster sampling, unbalanced clusters
- Linear model with a continuous dependent variable: tolerance of inequality
  - Ratio between the perceived boundaries of being rich and poor
- MLM estimation of $\rho$: 0.01
- Results from real data are consistent with the results from simulations

## Application Results: $\beta$s and $SE$s of selected covariates

| | Education | Internet Access | Gini Index | Marketization Index |
|---|---|---|---|---|
| MLM | 0.030 | 0.483 | 2.054 | 1.417 |
| | (0.061) | (0.180) | (13.222) | (0.451) |
| NSE | -0.011 | 0.496 | 5.502 | 1.698 |
| | (0.061) | (0.181) | (5.623) | (0.201) |
| RCSE | (0.067) | (0.161) | (8.102) | (0.358) |
| BSE | (0.056) | (0.176) | (4.688) | (0.201) |
| BCSE | (0.062) | (0.159) | (12.152) | (0.571) |
| BSSE | (0.056) | (0.155) | (4.998) | (0.240) |
| JSE | (0.053) | (0.163) | (5.270) | (0.217) |
| JCSE | (0.073) | (0.164) | (11.364) | (0.505) |

- All methods except MLM have the same $\beta$s.
- With SCS, C= 24, and a small $\rho$, the eight methods have similar $SE$s for individual-level factors.
- For group-level factors, JCSE, BCSE and MLM have similar $SE$s, which are bigger than RCSE, which in turn are bigger than the rest four methods.

## Discussion and Conclusion

- When analyzing clustered data, researchers are well aware of the downward bias of the OLS variance estimates. Among the alternative methods, studies show BCSE often outperform other methods.
- However, researchers overlook the impact of sampling procedure on the structure of clustering in the data.
- This study shows sampling procedure affects the performance of the different methods, for both individual-level and group-level covariates.
- Among the three sampling procedures, SRS is recommended.
- When SRS is not possible, this study suggests using JCSE to analyze the data.