ENHANCING INFERENCE WITH CLUSTERED STANDARD ERRORS THROUGH BOOTSTRAP METHODS

BOWEN MA (12960780) (AUTHOR)
Faculty of Economics and Business, University of Amsterdam

DR. RUTGER POLDERMANS(THESIS SUPERVISOR)
Faculty of Economics and Business, University of Amsterdam

This thesis investigates the potential advantages of the bootstrap method in dealing with clustered standard errors. Due to its simplicity and the progress in computational power, the bootstrap methods have become an dispensable tool in statistical analysis. Therefore, an overview of bootstrap techniques, the theoretical underpinnings of clustered standard errors, and their implications for linear regression models are included. Using simulation studies, the enhanced performance of bootstrap methods over traditional approaches, particularly in finite sample scenarios, is demonstrated.

KEYWORDS: Pairs Cluster Bootstrap, Wild Bootstrap, CRVE, Clustered errors.

BOWEN MA (12960780) (AUTHOR): bowen.ma@student.uva.nl
Dr. RUTGER POLDERMANS(THESIS SUPERVISOR): r.w.poldermans@uva.nl

Statement of Originality

This document is written by Student Bowen Ma who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

1. INTRODUCTION

The bootstrap method has been widely used in the field of econometrics due to its simplicity and the advancement of computing power. This technique allows researchers to make inferences without relying on strict parametric assumptions, making it particularly valuable in empirical research, MacKinnon (2009). Clustering is common when dealing with empirical data, see e.g. Abadie et al. (2023), as it often arises from the presence of a common unobserved random shock at the group level (Also see Hansen, 2007). These shocks can lead to intra-group correlation, which complicates traditional inferential methods that assume independence. When doing inference on clustered data without taking a grouped structure into consideration, it leads to invalid results.

One of the key assumptions in this thesis, which aligns with the proposal by Colin Cameron and Miller (2015), is that errors are correlated for individuals within the same group but uncorrelated across different clusters. Ignoring the clustered errors can lead to biased and inconsistent conclusion. Some researchers use cluster-robust standard errors for the correction which can be easily implemented in software, while it is only useful conditional on strong statistical assumption, Cameron et al. (2007).

One of the assumptions is that the sample size has tends to infinity. In practical applications, this assumption is often unrealistic. In such cases, bootstrapping offers an alternative by providing a way to refine asymptotic theory without requiring large sample sizes, see e.g. Jeong and Maddala (1993). In this thesis, two different bootstrap methods are implemented to investigate whether bootstrapping can improve the inference for Ordinary Least Squares (OLS) estimation with clusters, particularly in the case of few clusters. By exploring these methods, the aim is to determine their effectiveness in enhancing the reliability of statistical conclusions when traditional methods fall short due to limited data availability.

The remainder of this thesis is organized as follows. Section 2 presents the theoretical framework and a review of the relevant literature. Section 3 describes the methodology, including the technical details of data simulation and the application of both the pairs cluster bootstrap and the wild bootstrap. Section 4 details the results of the analysis. Sections 5 and 6 discuss the limitations of this thesis and provide concluding remarks.

2. THEORETICAL FRAMEWORK AND LITERATURE REVIEW

In this section, comprehensive overviews of bootstrap methods and clustered standard errors are provided, with a focus on the application of bootstrap techniques to clustered standard errors.

2.1. *Bootstrap Method*

Bootstrap is a random sampling method with replacement that replicates the observed information according to the given original sample, and makes statistical inferences about the distribution characteristics of the population. Introduced by Efron (1979), the bootstrap principle is considered more widely applicable and more reliable compared with the jackknife method, although bootstrapping was inspired by the method initially. For instance, the bootstrap can correctly estimate the variance of the sample median, where jackknife is known to fail. Also it is outperforming cross-validation when dealing with the error rates in a linear discrimination problem. The idea of the jackknife theory is to leave out a single observation each time to recompute the parameter vector of interest, such that the total number of estimated parameter vectors is equal to the sample size. On the other hand, the bootstrap method generates subsamples, which are of the same size, out of the original sample. Estimation of a parameter is done in each bootstrap set, such that an approximate distribution of the estimate is obtained, further inference can be conducted. The biggest difference between the jackknife method and bootstrapping is that the former removes one observation at a time, while bootstrapping randomly samples observations with replacement.

The bootstrap method gained prominence in econometrics about three decades ago thanks to improvements in numerical computation (see e.g. MacKinnon (2006), 2006). It has proven effective in situations where conventional methods fall short. The resample method is useful when the distribution of an estimator is unknown or when reliable estimators for its standard error are unavailable, as indicated by Jeong and Maddala (1993) for example. The problems to be solved by using resampling are to assess the stability of test statistics and to provide alternative estimators for the variance of estimated parameters. In practice, while traditional estimators rely on asymptotic theory for approximations, large-sample theory may not always provide clear guidance as pointed out by Roodman et al. (2019). This is where the bootstrap method comes into play. It provides an alternative to refine approximations when asymptotic theories become intractable or inaccurate.

There are several types of bootstraps, such as the simple, the double and the wild bootstrap, which are applied according to special problems encountered in econometrics. In Roodman et al. (2019) for example, the wild bootstrap is particularly useful when dealing with clustered error terms, offering a specialized approach that can provide reliable inference in situations where conventional methods may fail. Bootstrapping has several advantages when it comes to resampling. Firstly, it is a non-parametric method, which makes it more flexible and robust. Secondly, thanks to modern computing power, it becomes more feasible for large datasets and allows for quick inference for the parameters, as further illustrated in the main text MacKinnon (2006). Despite the advantageous properties of bootstrapping, it is important to recognize that applying the bootstrap method to a defective model is meaningless, as pointed out by Jeong and Maddala (1993).

2.2. *Clustered Standard Errors*

A pillar of traditional cross-sectional inference is the assumption that the data are independent. However, this assumption is unrealistic for some sampling models, as mentioned by Angrist and Pischke (2008). Econometricians frequently use micro-level data from populations with grouped structures, such as industry or occupation, to fit regression models. In such cases, the regression errors are often correlated within groups which may be caused by random unobserved shocks, see Moulton (1986), and Abadie et al. (2023), for example.

When fitting a regression model while ignoring the existence of clustered standard errors, it leads to incorrect estimation. Independent and identically distributed (i.i.d.) standard errors implicitly assume no serial correlation and no heteroskedasticity. Heteroskedasticity-robust standard errors assume no autocorrelation. In practice, both approaches tend to underestimate the true standard error when clustering problem appears. Angrist and Lavy (2002) further demonstrated this in their experiment conducted in Israel, in 2002. This underestimation leads to larger absolute values of t-statistics and the over-rejection of the null hypothesis, potentially resulting in the incorrect conclusion that a true zero effect is a nonzero causal effect.

Angrist and Pischke (2008) pointed out that many researchers aware of the clustering problem understand that simply using the ’cluster’ option in Stata may not be sufficient. The asymptotic approximation relevant to clustered standard errors relies on a large number of clusters, but it is often the case that many clusters cannot be obtained. Bootstrapping can be utilized to overcome these issues.

2.3. *Application of Bootstrap Method in Clustered Standard Errors*

Bootstrapping provides a way to obtain more accurate inferences in finite samples compared to using cluster-robust t-statistics. Cameron et al. (2007) focused on the bootstrap-t procedure, which provides asymptotic refinement. They found that this procedure can lead to considerable improvements in inference. An alternative to the wild cluster bootstrap is the pairs cluster bootstrap, where bootstrap samples are constructed by resampling regressors and dependent variable pairs. However, Cameron et al. (2007) found that the pairs cluster bootstrap produced less reliable inferences than the wild cluster bootstrap, attributing this to the instability of the pairs bootstrap in regression models with heteroskedastic errors and as it is not conditional on regressors.

Although Bertrand et al. (2002) found that when dealing with a moderate number of clusters, say ten or twenty, bootstrap based models performed poorly, Cameron et al. (2007) had more optimistic conclusions when conducting the same exercise, but with the wild bootstrap method. Even when dealing with as few as six clusters, they did not observe notable loss of power.

3. METHODOLOGY

This section starts with the notational framework within which all the following work is contained. Then, two data generating processes inspired by Cameron et al. (2007) are outlined. The section concludes with the implementation of pairs cluster bootstrap and variants of wild bootstrap methods.

3.1. *Notational Framework*

Similar to Cameron et al. (2007), a regression model for G disjoint clusters can be written as,

$$y_{ig} = x'_{ig}\beta + \varepsilon_{ig}, \quad i = 1, \cdots, n_g; \quad g = 1, \cdots, G \tag{1}$$

where $\boldsymbol{\beta}$: $k \times 1$, $x_{ig} : k \times 1$ and $\sum_{g=1}^{G} n_g = n$. Alternatively, the model in matrix is expressed as

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} = X\beta + \boldsymbol{\varepsilon} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_G \end{bmatrix}$$

The matrix $X_g : n_g \times k$ contains all the observations on the regressors within a cluster. Similarly, y_g and ε_g are both $n_g \times 1$.

Under correct model specification

$$\hat{\beta} - \beta_0 = (\sum_{g=1}^{G} X'_g X_g)^{-1} \sum_{g=1}^{G} X'_g \varepsilon_g$$

with $\hat{\beta}$ the OLS estimator of β_0.

It is assumed that for all $g = 1, \cdots, G$

$$E[\varepsilon_g] = 0 \quad and \quad E[\varepsilon_g \varepsilon'_g] = \Sigma_g$$

Furthermore, $E[\varepsilon_g \varepsilon_h] = 0$, where $h \neq g \in [1, \cdots, G]$.

Then

$$\sqrt{n}(\hat{\beta} - \beta_0) \overset{a}{\sim} N(0, V[\hat{\beta}])$$

with

$$V[\hat{\beta}] = (X'X)^{-1}(\sum_{g=1}^{G} X_g \Sigma_g X'_g)(X'X)^{-1} \tag{2}$$

The $V[\hat{\beta}]$ is usually larger compared with $V[\hat{\beta}] = \sigma^2(X'X)^{-1}$, which is based on the false assumption of i.i.d errors.

Let β_1 be a scalar component of the $k \times 1$ vector β, then testing the null hypothesis $H_0 : \beta_1 = \beta_{0,1}$ against $H_a : \beta_1 \neq \beta_{0,1}$ can be done by considering the well-known statistic

$$t = \frac{\hat{\beta}_1 - \beta}{s.e.(\hat{\beta}_1)} \tag{3}$$

Here, $s.e.(\hat{\beta}_1)$ is the square root of the diagonal entry of a feasible version of $V[\hat{\beta}]$, corresponding with $\hat{\beta}_1$.

3.2. *Monte Carlo Simulation for Clustered Data*

3.2.1. *Simulations with Homoskedastic Clustered Errors*

Inspired by Cameron et al. (2007), a linear model with an intercept and a single regressor

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} = \beta_0 + \beta_1(z_g + z_{ig}) + (\varepsilon_g + \varepsilon_{ig})$$

is of interest, where z_g, z_{ig}, ε_g and ε_{ig} are independent and drawn from standard normal distribution, $\beta_0 = 0$ and $\beta_1 = 1$. Note that the components z_g and ε_g are common to individuals within group and induce a within group correlation of both regressors and errors. The data simulation in this thesis follows the methodology of Cameron et al. (2007). Specifically, 1,000 simulations were conducted for each sample, with the number of observations per cluster (n_g) set at 15, 30, and 60, and the number of clusters (G) set at 5, 10, 15, 20, 25, and 30.

3.3. *Cluster-Robust Variance Estimator (CRVE)*

Moulton (1986) found that neglecting the cluster structure can lead to severely biased standard errors and significant distortions in size. When dealing with error terms correlated within clusters, it is common to use a cluster-robust variance estimator (CRVE) to calculate t-statistics and Wald statistics Djogbenou et al. (2019).

The model in (1) is estimated using OLS, giving $\hat{\varepsilon}_g = y_g - X_g\hat{\beta}$, and the standard errors are computed based on the cluster-robust variance estimator (CRVE)

$$\frac{G}{(G-1)}(X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\hat{\varepsilon}_g\hat{\varepsilon}_g' X_g\right)(X'X)^{-1} \tag{4}$$

Note that the estimator $\left(\hat{\varepsilon}_g'\hat{\varepsilon}_g/(n-k)\right)(X'X)^{-1}$ based on the assumption of i.i.d errors yields false inference in case of a clustered error structure.

3.4. *Bootstrap Methods*

3.4.1. *Pairs Cluster Bootstrap*

For the pairs clusters bootstrap, samples are constructed by resampling the regressors and dependent variable within each cluster. Then, for each sample, the model is estimated by OLS and the test statistics

$$t_b^* = \frac{\hat{\beta}_{1,b}^* - \hat{\beta}}{s.\hat{e}._{\hat{\beta}_{1,b}^*}}$$

are computed, where $\hat{\beta}_1$ is the OLS estimator of the original sample and $\hat{\beta}^*_{1,b}$ denotes the OLS estimator in a bootstrap sample. The $s.\hat{e}._{\hat{\beta}^*_{1,b}}$ is based on the assumption of clustering and computed as in (4). The bootstrap distribution of the test statistics is used to obtain critical values. The null hypothesis $H_0 : \beta_1 = \beta_{0,1}$ is rejected at level α if $t > t^*_{[1-\alpha/2]}$ or $t < t^*_{[\alpha/2]}$, where t is the test statistic of the original sample.

As indicated by Djogbenou et al. (2019) a primary issue with the pairs cluster bootstrap method is its failure to condition on regressors, unlike the wild bootstrap (WB) and the wild cluster bootstrap (WCB). This lack of conditioning presents two significant drawbacks. Firstly, when cluster sizes differ across clusters, the sample size will vary across bootstrap samples. Secondly, when any of the regressors is a dummy variable that varies at the cluster level, the number of treated clusters and treated observations will fluctuate across bootstrap samples.The wild bootstrap method does not have the same drawbacks.

3.4.2. *Wild Cluster Bootstrap (WCB)*

The defining characteristic of the wild cluster bootstrap DGP lies in the generation of bootstrap error terms. First, in the original sample the null hypothesis $H_0 : \beta_1 = \beta_{0,1}$ is imposed and the remaining parameters are estimated by OLS giving the restricted vector $\hat{\varepsilon}^R_g$. In each bootstrap DGP, the error vector is created by multiplying the restricted residual vector $\hat{\varepsilon}^R_g$ with v^*_g, where $v^*_1, \cdots, v^*_G$ denotes the i.i.d realization of an auxiliary random variable with zero mean and unit variance. Similar to Cameron et al. (2007), the Rademacher weights are used where assigning $\hat{\varepsilon}^*_g = \hat{\varepsilon}^R_g$ or $\hat{\varepsilon}^*_g = -\hat{\varepsilon}^R_g$ with equal probability. After generation of each bootstrap sample a test statistic is calculated in the same way as in the pairs cluster bootstrap mentioned above. This also allow to determine the critical values of the test statistics.

4. RESULTS

The results for clusters with 15, 30, and 60 observations are presented in Tables 1, 2, and 3, respectively. As clearly can be seen in the first row of the table below, falsely assuming i.i.d errors leads to severe over-rejection under the null hypothesis, as expected. If the cluster-robust covariance estimator is utilized the tendency of rejection still exists especially when the cluster size is small. That is the reason why the bootstrap methods are implemented.

TABLE I

ACTUAL REJECTION FREQUENCIES WHEN $n_g = 30$ WITH α=0.05

	Number of clusters					
Method	5	10	15	20	25	30
IID	0.462	0.494	0.509	0.472	0.493	0.474
CRVE	0.202	0.153	0.115	0.092	0.108	0.072
Pairs Cluster Bootstrap	0.087	0.102	0.084	0.071	0.084	0.055
WCB	0.030	0.081	0.065	0.045	0.066	0.043

TABLE II

ACTUAL REJECTION FREQUENCIES WHEN $n_g = 15$ WITH α=0.05

	Number of clusters					
Method	5	10	15	20	25	30
IID	0.352	0.347	0.034	0.358	0.341	0.347
CRVE	0.220	0.119	0.090	0.108	0.081	0.085
Pairs Cluster Bootstrap	0.085	0.081	0.065	0.084	0.067	0.073
WCB	0.045	0.055	0.044	0.057	0.052	0.058

TABLE III

ACTUAL REJECTION FREQUENCIES WHEN $n_g = 60$ WITH α=0.05

	Number of clusters					
Method	5	10	15	20	25	30
IID	0.594	0.650	0.596	0.628	0.622	0.593
CRVE	0.218	0.135	0.104	0.093	0.083	0.073
Pairs Cluster Bootstrap	0.114	0.096	0.077	0.081	0.064	0.060
WCB	0.041	0.066	0.055	0.049	0.047	0.043

The pairs cluster bootstrap and the wild cluster bootstrap methods demonstrate significant improvement over the previously mentioned techniques, indicating asymptotic refinement. When applying the wild bootstrap, performance improves when the cluster size exceeds 10, which is not obvious when using the pairs cluster bootstrap. Given a fixed number of clusters, increasing the number of observations per cluster does not result in obvious improvement. In fact, fewer observations per cluster can yield results closer to the significance level.

5. CONCLUSION

The analysis conducted in this thesis focuses on the application of the bootstrap hypothesis test and the asymptotic test for clustered standard errors. While the results are promising, several aspects warrant further discussion and consideration. The models analyzed in this study primarily consider one regressor. However, in practical applications, models often include multiple regressors to better capture the complexity of the data and to control for various confounding factors. The simulations and analyses in this thesis are based on synthetic data generated to meet the grouped data structure for the demonstration of bootstrap methods. While this approach ensures controlled conditions and clear demonstrations of the statistical properties, applying these methods to empirical data would be highly beneficial. In the data simulation process, all clusters were assumed to have the same sizes. This assumption simplifies the analysis but may not reflect the reality of many empirical datasets, where cluster sizes can vary significantly.

The simulation results show that when using the CRVE method and assuming i.i.d to deal with clustered standard errors, the rejection rates are always much higher than they should be. Hence, they are worse-performing methods compared to the bootstrap methods. The wild cluster bootstrap method has outperformed the other three methods discussed, though its rejection rates are not exactly the same with the significance level. It is noteworthy that bootstrap methods do not consistently over-reject or under-reject.

REFERENCES

ABADIE, ALBERTO, SUSAN ATHEY, GUIDO W IMBENS, AND JEFFREY M WOOLDRIDGE (2023): "When Should You Adjust Standard Errors for Clustering?*," *The Quarterly Journal of Economics*, 138 (1), 1–35. [3, 5]

ANGRIST, JOSHUA D. AND VICTOR LAVY (2002): "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," . [5]

ANGRIST, JOSHUA D. AND JÖRN-STEFFEN PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. [4, 5]

BERTRAND, MARIANNE, ESTHER DUFLO, AND SENDHIL MULLAINATHAN (2002): "How Much Should We Trust Differences-in-Differences Estimates?" . [5]

CAMERON, A. COLIN, JONAH GELBACH, AND DOUGLAS MILLER (2007): "Bootstrap-Based Improvements for Inference with Clustered Errors," Tech. Rep. t0344, National Bureau of Economic Research, Cambridge, MA. [3, 5, 6, 7, 8]

COLIN CAMERON, A. AND DOUGLAS L. MILLER (2015): "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 50 (2), 317–372. [3]

DJOGBENOU, A.A., J.G. MACKINNON, AND M.Ø. NIELSEN (2019): "Asymptotic theory and wild bootstrap inference with clustered errors," *Journal of Econometrics*, 212 (2), 393–412. [7, 8]

EFRON, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7 (1), 1–26, publisher: Institute of Mathematical Statistics. [4]

HANSEN, CHRISTIAN B. (2007): "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, 140 (2), 670–694. [3]

JEONG, JINOOK AND G.S. MADDALA (1993): "21 A perspective on application of bootstrap methods in econometrics," in *Handbook of Statistics*, Elsevier, vol. 11, 573–610. [3, 4]

MACKINNON, JAMES G. (2006): "Bootstrap Methods in Econometrics*," *Economic Record*, 82 (s1). [4]

——— (2009): "Bootstrap Hypothesis Testing," in *Handbook of Computational Econometrics*, ed. by David A. Belsley and Erricos John Kontoghiorghes, Wiley, 183–213, 1 ed. [3]

MOULTON, BRENT R. (1986): "Random group effects and the precision of regression estimates," *Journal of Econometrics*, 32 (3), 385–397. [5, 7]

ROODMAN, DAVID, MORTEN ØRREGAARD NIELSEN, JAMES G. MACKINNON, AND MATTHEW D. WEBB (2019): "Fast and wild: Bootstrap inference in Stata using boottest," *The Stata Journal: Promoting communications on statistics and Stata*, 19 (1), 4–60. [4]