Queen's Economics Department Working Paper No. 1456

Using Large Samples in Econometrics

James G. MacKinnon
Queen's University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

2-2022
5-2022 (minor revisions)

Using Large Samples in Econometrics*

James G. MacKinnon†
Queen's University
mackinno@queensu.ca

May 12, 2022

Abstract

As I demonstrate using evidence from a journal data repository that I manage, the datasets used in empirical work are getting larger. When we use very large datasets, it can be dangerous to rely on standard methods for statistical inference. In addition, we need to worry about computational issues. We must be careful in our choice of statistical methods and the algorithms used to implement them.

Keywords: bootstrap, clustered data, jackknife, statistical computation, statistical inference

JEL Codes: C10, C12, C13, C55

*This paper was prepared as a short talk at the ASSA Meeting on January 9, 2022 in a session organized by the *Journal of Econometrics*. I am grateful to the Social Sciences and Humanities Research Council of Canada (grant 435-2021-0396) for financial support.

†Corresponding author. Address: Department of Economics, 94 University Avenue, Queen's University, Kingston, Ontario K7L 3N6, Canada. Email: mackinno@queensu.ca. Tel. 613-533-2293. Fax 613-533-6668.

1 Introduction

As is documented in Section 2, the datasets used in applied econometrics are getting larger. In many ways, this is a very positive development. However, it has important implications for both inference (Section 3) and computation (Section 4). Methods that work well for samples with a few hundred or a few thousand observations may not be suitable for samples with millions of observations. They may lead to misleading inferences, or they may be computationally infeasible.

2 Datasets Are Getting Larger

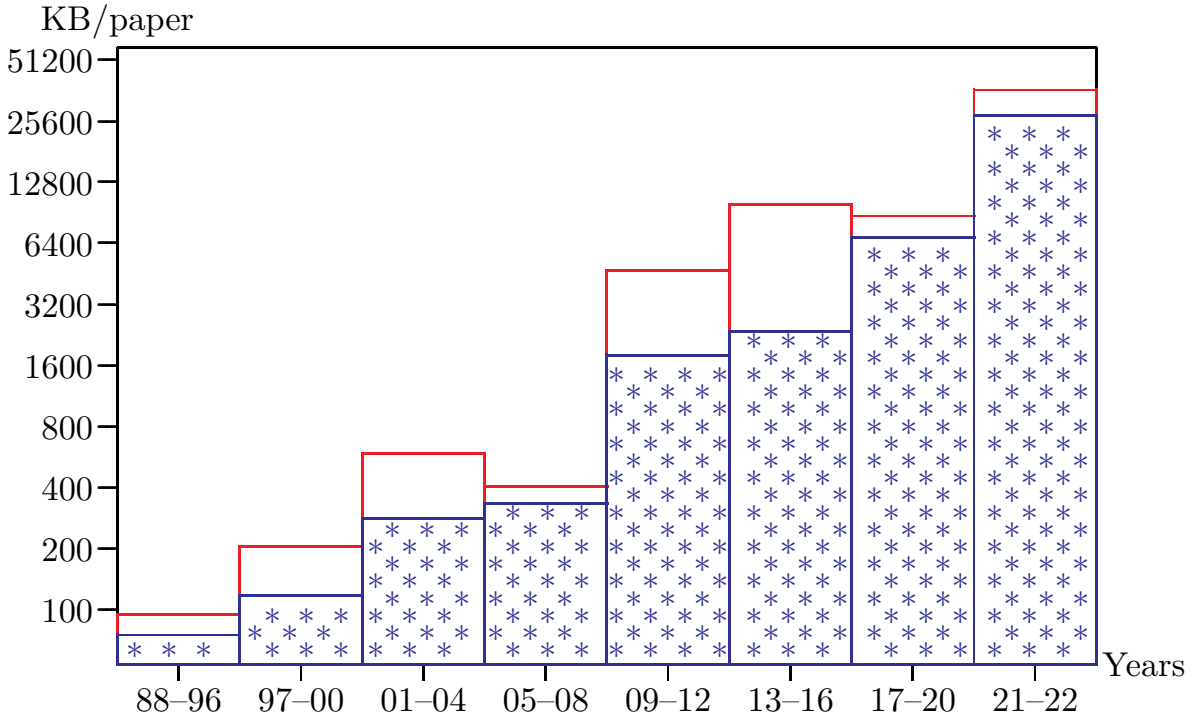
Fifty years ago, almost all of the datasets used in applied econometric work were very small. Many of them contained annual, quarterly, or monthly time-series data with at most a few hundred observations. Even for cross-section and panel data, there were rarely more than a few thousand observations. As the decades have passed, however, datasets with much larger sample sizes have become available. They include high-frequency time-series data, large cross-sections such as individual-level census data, panel data that may involve a modest number of time periods but perhaps millions of individuals, and data for various types of on-line transactions or other activities. Of course, this list is far from exhaustive.

Since 1994, I have managed the Journal of Applied Econometrics Data Archive. The authors of all papers published in the journal since that year have been required to deposit their data in the Data Archive, unless the data are confidential. By observing how the number of bytes of disk storage per paper has grown over time, we can get a very rough idea of how much economic datasets have grown.

Each published paper has a unique directory in the JAE Data Archive. I use the total number of kilobytes (KB) per directory as a measure of how large the dataset(s) associated with each paper are. This measure is, of course, extremely crude. In addition to data, which are almost always compressed, a paper's directory always contains a readme file, and it may also contain program files and supplementary material of various types. This suggests that the space per directory may provide an over-estimate. Especially for recent years, however, these numbers almost certainly understate the average size of actual datasets. The reason is that many papers use at least some data that cannot legally be stored in the Data Archive. Administrative datasets, which tend to be quite large, can very rarely be stored there. Neither can proprietary datasets from commercial providers. In recent years, about one-third of all papers have used confidential data of one sort or another.

Figure 1 shows the number of KB of disk storage per paper, on a logarithmic scale, in

Figure 1: Average Disk Space per Paper in JAE Data Archive



the form of two overlapping bars for each of eight time periods. The most recent period includes all forthcoming papers as of early January, 2022. The taller bars include the disk space used by all papers, and the shorter bars omit the space taken by the single largest paper for each time period. When these outliers are removed, we observe that the shorter bars increase monotonically in height. Thus there is absolutely no indication that the rate of increase is slowing down. On average, a paper from 2021-22 uses about 363 times as much disk space as a paper from 1988-96. This number would be larger if we could correct for missing confidential datasets. However, sample sizes have probably not grown as fast as disk storage, because larger samples often involve larger numbers of variables.

3 Are Larger Samples Always Better?

Since many estimators are $\sqrt{N}$-consistent, where N is the sample size, it is tempting to think that problems of inference become negligible when the sample size becomes sufficiently large. This is not always true, however. In fact, I suspect that it is rarely true. The fundamental problem is that the disturbances in the models we estimate are not totally uncorrelated. Assuming that they are uncorrelated may be a very good approximation when $N = 100$, but not when $N = 10,000,000$.

To illustrate the problem, consider the sample mean $\bar{y} = N^{-1} \sum_{i=1}^{n} y_i$. The usual formula for its variance is

$$\mathrm{Var}(\bar{y}) = \frac{1}{N}\sigma^2, \tag{1}$$

where σ^2 denotes either $\mathrm{Var}(y_i)$, when there is homoskedasticity, or the limit of the average of the $\mathrm{Var}(y_i)$, when there may be heteroskedasticity. Under typical regularity conditions, (1) implies that $\bar{y} - \mu_y = O_p(N^{-1/2})$, where μ_y is the population mean. But this equation is true only if the y_i are completely uncorrelated. With even a very small amount of correlation, it is too optimistic.

A more general formula for the variance of the sample mean is

$$\mathrm{Var}(\bar{y}) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(y_i) + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathrm{Cov}(y_i, y_j). \tag{2}$$

If all the $\mathrm{Cov}(y_i, y_j)$ are zero, then only the first term in (2) is non-zero, and it equals (1). This term is $O(1/N)$. But if the $\mathrm{Cov}(y_i, y_j)$ are on average positive, then the second term is $O(1)$, because it involves two summations over N, as well as a factor of $1/N^2$. Thus, as N becomes large, the second term becomes dominant. Depending on what happens to the $\mathrm{Cov}(y_i, y_j)$ as $N \to \infty$, $\bar{y}$ either does not converge at all, or it converges more slowly than $N^{-1/2}$. Andrews (2005) studies an important case in which it does not converge.

For many types of sample data, it seems unrealistic to assume that all correlations among the observations vanish asymptotically. In fact, non-vanishing correlations can arise in a great many ways. For example, when there is spatial correlation, the observations generally do not become further apart in space as N increases, so that adding additional observations cannot be expected to cause the average correlation among them to diminish. Similarly, with panel data, there are usually both correlations among cross-section units for each time period, and correlations across time periods for the same cross-section units. Asymptotic results often depend on both T, the number of time periods, and N_t, the number of units for the t^{th} time period, becoming large. Although samples with very large values of the N_t are becoming common, ones that also have T very large are exceedingly rare. Thus, even very large panels will almost always involve complicated patterns of correlation among non-vanishing subsets of the observations. Although including time fixed effects, and maybe also cross-section fixed effects, almost certainly reduces these correlations, it seems unlikely that it entirely eliminates them.

One increasingly popular way to deal with correlated observations is to assume that every observation belongs to one of G disjoint clusters. These might correspond to schools or school districts; villages, towns, or cities; counties or states; firms or industries; and so on. The clusters may be small or large. Inference is then based on the assumption that the products

of the regressors and the disturbances (the scores) are correlated within clusters, with an unspecified pattern of variances and covariances, but uncorrelated across clusters. Note that the required lack of correlation across clusters can arise either because the disturbances are uncorrelated across them or because the regressors are. For example, a treatment regressor might be uncorrelated across clusters if treatment were randomly assigned at the cluster level.

Assigning observations to disjoint clusters is often used as a way to approximate the effects of spatial correlation. As an approximation, it is surely imperfect, since there seems very likely to be correlations among observations in neighboring clusters. If they exist, these correlations must cause errors of inference for sufficiently large samples. Nevertheless, clustering by geographic units often seems to work well.

In an influential paper, Bertrand, Duflo and Mullainathan (2004) provide evidence that, in linear regression models with both time and state fixed effects, clustering by geographic unit (U.S. states) performs quite well, while clustering by the intersection of geographic unit and time period performs poorly, and not clustering at all performs even worse. This conclusion is based on "placebo-law" experiments, which use actual data from the Current Population Survey for the regressand and all but one of the regressors. The latter is a treatment dummy variable, similar to the ones in difference-in-differences regressions, which is generated randomly for each replication. Because the treatment dummy is a completely artificial variable, the hypothesis that its coefficient is zero should be rejected about 5% of the time by tests at the .05 level. In the experiments, this is close to being true only for state-level clustering.

MacKinnon (2016) performs a more extensive set of placebo-law experiments based on similar data. The results vary dramatically with the sample size. When the full sample of about 1.16 million observations is used, failing to allow for any level of clustering leads to rejection percentages of up to 75% for tests at the .05 level, depending on how many states are "treated." But as the sample size is reduced, the extent of over-rejection declines. With subsamples of about $58,000$ observations (1/20 of the original sample), the rejection percentages without clustering never exceed 22.4%. A similar, but somewhat less extreme, pattern of over-rejection that increases with the sample size is observed when the standard errors are clustered at the state-year level. However, no such pattern is observed for state-level clustering, which works quite well for samples of any size. Thus, at least for these CPS data, there seems to be strong evidence that clustering at anything less than the state level leads to errors of inference which become more severe as the sample size increases.

Dividing the samples into clusters and using cluster-robust inference can work very well when the number of clusters G is large and the clusters are well balanced, but it can work dreadfully either when G is small or when the clusters are unbalanced. In particular, inference

can be extremely problematical when there are few treated clusters, even if both G and N are large; see Djogbenou, MacKinnon and Nielsen (2019) and MacKinnon and Webb (2017, 2018). In such cases, increasing the sample size while holding G fixed does not make inference more reliable. In general, even having an extremely large sample size does not ensure reliable inference when the data are clustered.

Instead of one-way clustering, it is possible to employ two-way (or three-way, or even more-way) clustering, as proposed in Cameron, Gelbach and Miller (2011). Multi-way clustering can approximate a much wider range of correlation structures than one-way clustering can. However, it involves both theoretical and practical complications; see Davezies, D'Haultfœuille and Guyonvarch (2021), Menzel (2021), and MacKinnon, Nielsen and Webb (2021). Even when multi-way clustering provides a good approximation, obtaining reliable inferences can be challenging.

4 Computational Issues

In the early days of econometrics, a computer was a person equipped with a Friden or Monroe calculator. Needless to say, econometricians cared deeply about efficient computation. The famous result of Frisch and Waugh (1933), that what we now call "partialing out" the regressors which are not of intrinsic interest does not affect the coefficients or standard errors of the remaining regressors, was entirely motivated by computational considerations.

Between the early 1960s and the mid to late 1980s, econometricians primarily used mainframes or perhaps (from the mid 1970s) mini-computers. These had very limited amounts of memory by modern standards, and users often had to pay for CPU time. Thus econometricians still cared about efficient computation.

However, once it became feasible to perform most econometric computations on personal computers, roughly in the late 1980s, econometricians started to think of computing time as free. Figure 1 suggests that the size of the average dataset increased by a factor of perhaps 25 between 1990 and 2010. But the speed of personal computers increased by far more than that during the same period. Thus the importance of efficient computation diminished.

In recent years, however, many interesting datasets seem to be becoming larger more quickly than computers are becoming faster. Moreover, much of the increase in speed over the past decade has come from increasing the number of cores per processor. For large samples, doubling the number of cores often does not translate into doubling the speed of computation, because of cache congestion and constraints on how fast main memory can be accessed. Thus, at least for the very large datasets that econometricians are increasingly using, efficient computation has become important again.

Writing numerical programs so that the calculations are performed efficiently for large samples is often not hard. The key thing is to avoid any computations which are $O(N^\delta)$ for $\delta >> 1$. In particular, any algorithm that involves computations which are $O(N^2)$ will become impractical for sufficiently large N, and any algorithm that requires storing matrices which are $O(N^2)$ will fail for sufficiently large N. In the latter case, N does not have to be all that large by modern standards. Storing an $N \times N$ matrix in 64-bit precision requires 8×10^{12} bytes, or about 7450 gigabytes, when N equals one million. But few personal computers currently have more than 64 gigabytes of main memory.

As an example, consider an OLS regression with N observations and k regressors. There are two common ways to run such a regression. One is to use a QR decomposition, and the other is to form the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix, call a procedure for inverting positive definite matrices, such as a Cholesky decomposition, and then multiply $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ by $\boldsymbol{X}^\top \boldsymbol{y}$. In either case, the principal calculations are $O(k^2 N)$, so that either algorithm is feasible for very large samples, provided that k is fixed or at least is growing much more slowly than N.

It is wise to avoid computational methods that are $O(N^2)$ when N may be large, and it is essential to avoid storing $N \times N$ matrices. However, econometricians do not always pay sufficient attention to these issues. Consider the projection matrix $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ and its orthogonal complement $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{P_X}$. These are $N \times N$ matrices. Although $\boldsymbol{P_X}$, $\boldsymbol{M_X}$, and other projection matrices are enormously useful in econometric theory, they should not be used for computation except in very rare cases. For example, although it is convenient to write the sum of squared residuals as $\boldsymbol{y}^\top \boldsymbol{M_X} \boldsymbol{y}$, it is enormously faster to compute it as $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ after the least-squares estimates $\hat{\boldsymbol{\beta}}$ have been obtained. Computing the SSR in this way requires additional computations that are $O((k+1)N)$.

It may seem that bootstrap methods are infeasible whenever N is large, because the computational cost of bootstrapping is often $O(NB)$, where B is the number of bootstrap samples. However, there are important cases where bootstrap methods can be surprisingly inexpensive. For linear regression models with clustered disturbances, the cost of the wild cluster bootstrap (Cameron, Gelbach and Miller 2008) can be either $O(k^2 N) + O(BG^2)$ or $O(k^2 N) + O(k^2 BG)$, depending on what algorithm is used. The trick is to perform all the computations that are $O(N)$ before the bootstrap loop begins, and then to generate bootstrap samples at the cluster level instead of the individual level. For details, see Roodman, MacKinnon, Nielsen and Webb (2019) and MacKinnon (2022).

Similarly, while methods that apply the jackknife at the observation level generally involve computations that are $O(N^2)$, tricks like the one employed in MacKinnon and White (1985) to obtain a jackknife covariance matrix estimator for linear regression models can sometimes reduce this to $O(N)$. Moreover, when the jackknife is applied to a linear regression model at

the cluster level, MacKinnon, Nielsen and Webb (2022) shows how to make the computations just $O(k^2N)+O(k^2G)$. The first term here is the same as for OLS estimation, and the second term is typically much smaller than the first unless G/N is large. Thus, for large values of N and small to moderate values of G, the additional cost of jackknifing at the cluster level is negligible.

5 Concluding Remarks

In Section 2, I provided some evidence to support the unsurprising stylized fact that the samples used in applied econometric work are becoming larger. Figure 1 suggests that this trend has been under way for more than a quarter of a century and that it shows no signs of slowing down. I then pointed out that dealing with very large samples raises two important, and often neglected, issues. In Section 3, I argued that obtaining reliable inferences in very large samples may actually be harder than obtaining them in samples of moderate size. In Section 4, I briefly discussed econometric computation, which often receives insufficient attention. For small samples, it may not matter much whether the computational demands of a procedure are $O(N)$ or $O(N^2)$. But for large samples, it can make the difference between calculations that are easy and calculations that are infeasible.

References

Andrews, D.W.K., 2005. Cross-section regression with common shocks. Econometrica 73, 1551–1585.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Quarterly Journal of Economics 119, 249–275.

Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics 90, 414–427.

Cameron, A.C., Gelbach, J.B., Miller, D.L., 2011. Robust inference with multiway clustering. Journal of Business & Economic Statistics 29, 238–249.

Davezies, L., D'Haultfœuille, X., Guyonvarch, Y., 2021. Empirical process results for exchangeable arrays. Annals of Statistics 49, 845–862.

Djogbenou, A.A., MacKinnon, J.G., Nielsen, M.Ø., 2019. Asymptotic theory and wild bootstrap inference with clustered errors. Journal of Econometrics 212, 393–412.

Frisch, R., Waugh, F.V., 1933. Partial time regressions as compared with individual trends. Econometrica 1, 387–401.

MacKinnon, J.G., 2016. Inference with large clustered datasets. L'Actualité économique 92, 649–665.

MacKinnon, J.G., 2022. Fast cluster bootstrap methods for linear regression models. Econometrics and Statistics 21, to appear.

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2021. Wild bootstrap and asymptotic inference with multiway clustering. Journal of Business & Economic Statistics 39, 509–519.

MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2022. Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust. QED Working Paper 1483. Queen's University.

MacKinnon, J.G., Webb, M.D., 2017. Wild bootstrap inference for wildly different cluster sizes. Journal of Applied Econometrics 32, 233–254.

MacKinnon, J.G., Webb, M.D., 2018. The wild bootstrap for few (treated) clusters. Econometrics Journal 21, 114–135.

MacKinnon, J.G., White, H., 1985. Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics 29, 305–325.

Menzel, K., 2021. Bootstrap with cluster-dependence in two or more dimensions. Econometrica 89, 2143–2188.

Roodman, D., MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2019. Fast and wild: Bootstrap inference in Stata using boottest. Stata Journal 19, 4–60.