Using Large Samples in Econometrics

James G. MacKinnon (Queen's University)

McGill University Oct. 13, 2023

Introduction

Inference in large samples is often **harder** than in small samples.

Standard results suggest that, if $\hat{\beta}$ is an estimator of a parameter β, then $\hat{\beta}$ **converges** to β quite rapidly as $N \to \infty$.

In many cases, $\text{Var}(\hat{\beta} - \beta)$ is proportional to $1/N$. Thus $\text{se}(\hat{\beta})$, the standard error of $\hat{\beta}$, is proportional to $N^{-1/2}$.

- Can we ignore statistical randomness when N is very large?
- Can we at least be confident that, when $\text{se}(\hat{\beta})$ is small because N is large, we can rely on the usual 95% confidence interval

$$\left[\hat{\beta} - 1.96\text{se}(\hat{\beta}), \quad \hat{\beta} + 1.96\text{se}(\hat{\beta})\right] \; ? \tag{1}$$

Probably not!

The observations in most samples are not truly independent.

Small amounts of dependence probably do not matter for small samples, but they do for large ones.

A Little History

The samples that economists employ tend to be much larger than they once were, and sample sizes seem to be growing rapidly.

- In the early 1970s, datasets rarely had more than a few thousand observations. Often they contained 15-20 years of quarterly data, or 60-80 observations.
- Many datasets were stored on cards. "Large" ones were stored on magnetic tapes.
- Storing data on disks, even removable disks, was expensive.
- Data were often obtained by copying them from official publications onto paper and then onto cards.
- Access to a time-sharing system like TROLL (MIT & NBER, 1966+), which provided on-line access to time-series data and econometric procedures, was enormously expensive.

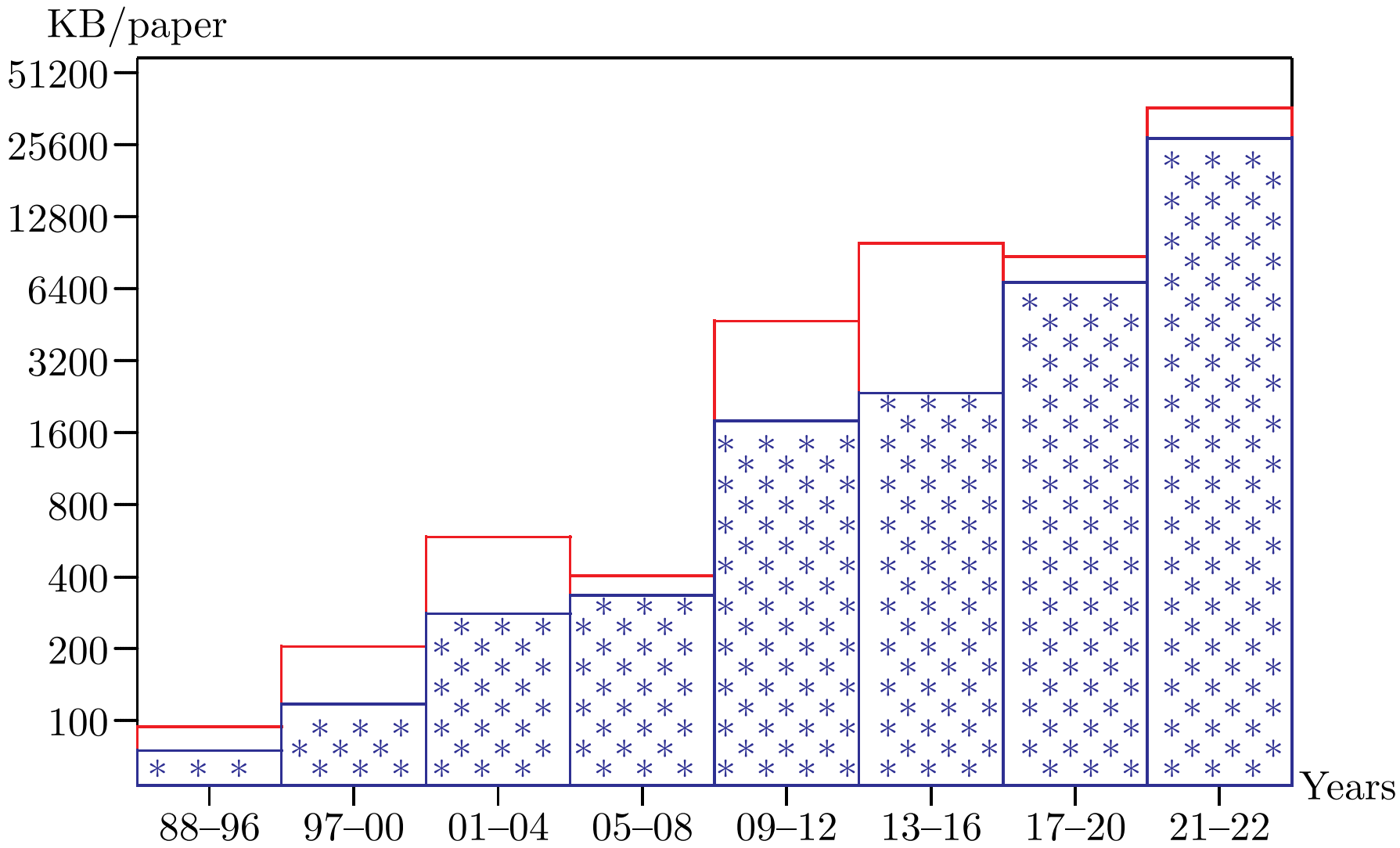
As the founder and maintainer of the **JAE Data Archive** until late 2022, I have observed that datasets are steadily getting larger.

Figure 1 shows total space per paper over eight time periods, most of them four years long.

- The reported numbers are the space for each paper, summed over all papers in each period.
- This includes readme files, zipped data files, zipped program files, and sometimes supplementary PDF files.
- However, it excludes the space that would be needed to store confidential data.
- Among papers published in 1988–1995, just 3% (one paper) used confidential data.
- Among papers since 2021, 36% used confidential data.

The difference between the red and blue bars is the space taken up by the largest single paper in each period.

Figure 1: Average Space per Paper in JAE Data Archive



Bias and Sample Size

For correctly-specified models, any bias typically vanishes as the sample size becomes larger.

However, this is not true when the bias arises from misspecification.

- If we regress y on x, but y actually depends on both x and z, the coefficient on x will usually be biased.
- If we use OLS instead of IV when x is correlated with the disturbances, the coefficient on x will be biased.

If $\hat{\beta}$ estimates β, with true value β_0, then

$$\text{MSE}(\hat{\beta}) = \text{Var}(\hat{\beta}) + \big(\text{E}(\hat{\beta}) - \beta_0\big)^2. \tag{2}$$

The variance term is usually proportional to $1/N$, but the squared bias term does not depend on N, so it eventually dominates the MSE.

Thus the **bias-variance tradeoff** changes with the sample size.

Tests, Confidence Intervals, and Sample Size

It is conventional to test hypotheses at the .05 level and to construct 95% confidence intervals (implicitly by inverting such a test).

Even if these numbers make sense for small samples, they do not make sense for large ones.

There is a tradeoff between the size of a test and its power, and a tradeoff between the coverage of a confidence interval and its length.

These tradeoffs depend on the sample size, in exactly the same way as the bias-variance tradeoff does.

- As $N \uparrow$, tests get more powerful. To balance size and power, we must reduce the level of tests as $N \uparrow$.
- As $N \uparrow$, confidence intervals get shorter. To balance coverage and length, we must increase the coverage of intervals as $N \uparrow$.

Precision and Sample Size

More observations $\longrightarrow$ more precise estimates.

When the sample size N is large enough, can we stop worrying about statistical inference?

According to what we learned in Econometrics 101, we can.

Consider the **simple regression model**

$$y_i = \beta_1 + \beta_2 x_i + u_i, \tag{3}$$

where we assume that the u_i are independent and identically distributed.

- The distribution of $\hat{\beta}_2$ is approximately normal with mean β_2 and variance proportional to $1/N$.
- As $N \to \infty$, this variance tends to 0, so $\hat{\beta}_2 \to \beta_2$.

In order to show the distribution of an estimator like $\hat{\beta}$ graphically, we need to plot something that relates frequency to the value of $\hat{\beta}$.

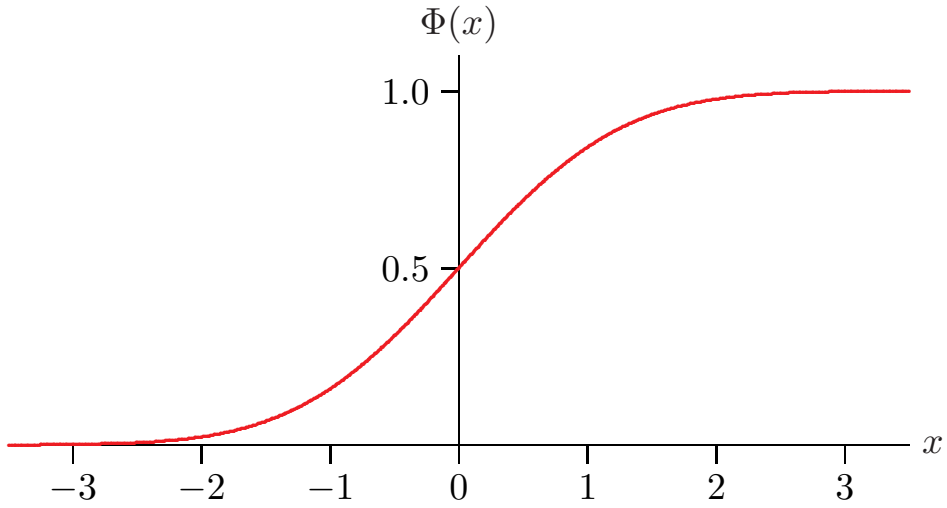
- A **probability density function (PDF)**, or one of its empirical counterparts like a **histogram** or a **kernel density function**, are easy to interpret.
- A **cumulative distribution function (CDF)**, or one of its empirical counterparts like an **empirical distribution function (EDF)** or a **kernel distribution function**, provide the same information.
- However, it is much easier to put CDFs for wildly different sample sizes on the same axes.

Figure 2 shows the PDF and CDF of the standard normal distribution.

- Notice that the largest value of the PDF corresponds to the steepest slope of the CDF.
- When the CDF is flat, as it is for larger absolute values of x, the density is very small.

Figure 2: The Standard Normal PDF and CDF

Standard Normal CDF:



Standard Normal PDF:

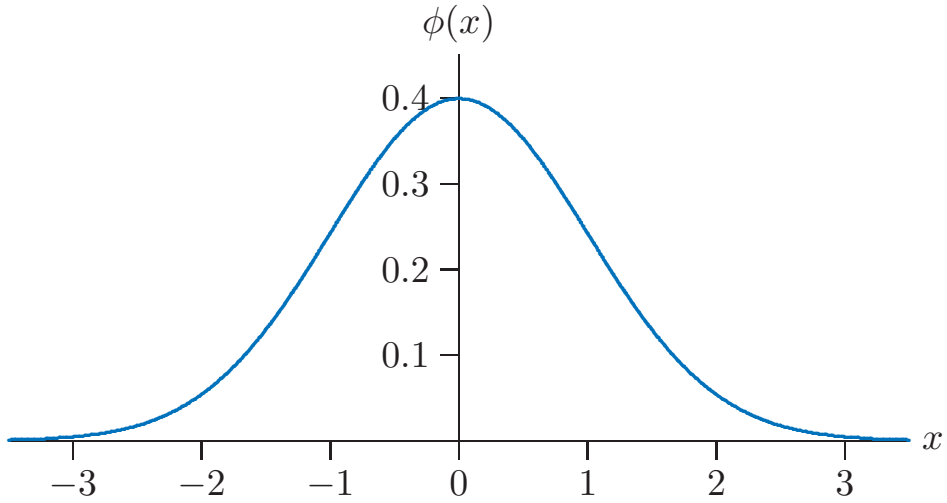


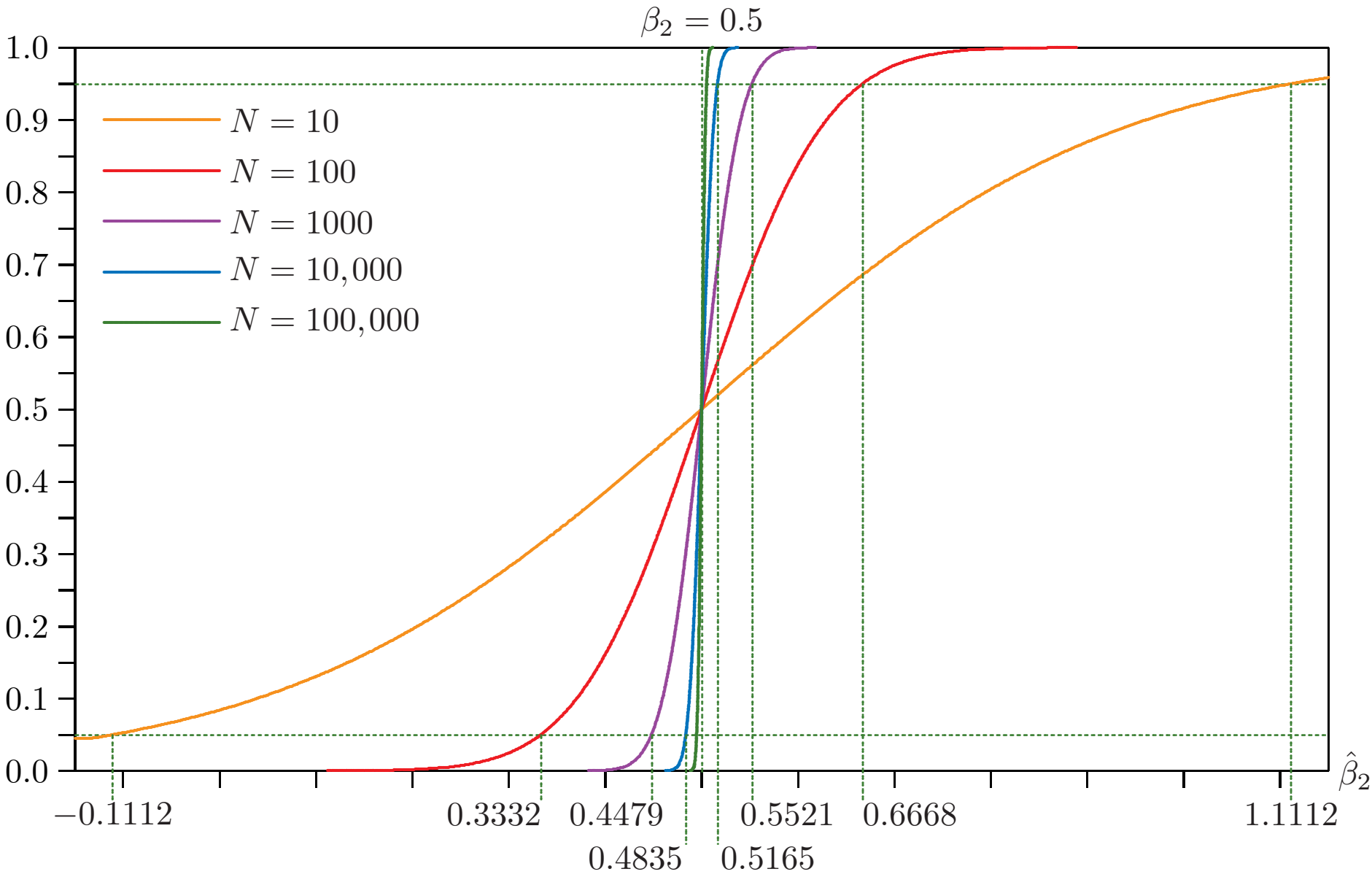
Figure 3 shows (estimates of) the cumulative distribution function (or CDF) of $\hat{\beta}$ for several sample sizes. These are based on either 9,999,999 or 999,999 replications.

In this case, $\beta_1 = 1$, $\beta_2 = 0.5$, and both x and u are distributed as standard normal.

- As N increases, the CDFs become steeper and less spread out.
- The .05 and .95 quantiles are converging to 0.5, which is the true value of β_2 in these experiments.
- For $N = 1{,}000{,}000$ (not shown), the .05 and .95 quantiles are 0.4984 and 0.5016.
- Even the .0001 and .9999 quantiles are 0.4963 and 0.5037.
- Thus $\hat{\beta}_2$ really does converge to $\beta_2 = 0.5$.

In this case, it does seem that, for $N \geq 100{,}000$, we can treat $\hat{\beta}_2$ as if it is the true value.

Figure 3: Distributions of $\hat{\beta}_2$ with Independence



The CDFs in Figure 3 converge for the same reason that a sample mean converges to the population mean as $N \to \infty$.

The usual formula for the variance of a sample mean is

$$\text{Var}(\bar{y}) = \frac{1}{N^2} \sum_{i=1}^{N} \text{Var}(y_i) = \frac{1}{N} \sigma^2, \tag{4}$$

where $\sigma^2 = \text{E}(y_i - \mu_y)^2$ if the y_i all have the same variance.

This result is also true if the $\text{Var}(y_i)$ differ (i.e., there is **heteroskedasticity**), and we define σ^2 as the average variance.

The result (4) says that $\text{Var}(\bar{y} - \mu_y)$ is proportional to $1/N$.

- But (4) requires that the y_i be **uncorrelated**.
- This is a very strong assumption.
- Violations of it are much more consequential for large samples than for small ones.

A general formula for the variance of the sample mean is

$$\text{Var}(\bar{y}) = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}(y_i) + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\text{Cov}(y_i, y_j). \tag{5}$$

If all the $\text{Cov}(y_i, y_j)$ are zero, then only the first term in (5) is non-zero, and we obtain (4) if the average of the $\text{Var}(y_i)$ is σ^2.

But the second term involves two summations. As N becomes large, the second term becomes dominant if the $\text{Cov}(y_i, y_j) \neq 0$.

Depending on what happens to the $\text{Cov}(y_i, y_j)$ as $N \to \infty$, $\bar{y}$ either does not converge at all, or it converges more slowly than $N^{-1/2}$.

- It is usually unrealistic to assume that all the $\text{Cov}(y_i, y_j) = 0$.
- For small samples, minor violations of this assumption do not affect the properties of $\hat{\beta}_2$ very much.
- But for large samples, they can have a very substantial effect.

Why Might Observations be Correlated?

1. **Time series** almost always display serial dependence.

   • With standard ways of modeling serial correlation, correlations between y_t and y_{t+j} diminish as j increases.

   • Thus, if we retain the same frequency while making N larger, $\hat{\beta}_2$ should converge to β_2 at the usual rate.

   • But if we increase N by sampling at greater frequencies, correlations between y_t and y_{t+j} may diminish more slowly with j.

   • Using daily data with $260N$ obs. instead of annual date with N obs. does not reduce the uncertainty in $\hat{\beta}_2$ by a factor of $\sqrt{260}$.

   • This is even more true for financial market data observed at, say, 5-second frequencies (roughly 1.6 million per year).

   • Large time-series samples are not just small time-series samples with more observations!

2. Some of the largest modern samples are **panel data**, which have both a time dimension and a cross-section dimension.
   - Modern panel samples often involve N_t (number of units for t^{th} time period) very large but T quite small.
   - But asymptotic results often depend on both T and the N_t becoming large.
   - Large panels almost always involve complicated patterns of correlation among non-vanishing subsets of the observations.
   - Including time fixed effects, and maybe also cross-section fixed effects, almost certainly reduces these correlations, but it cannot entirely eliminate them.
   - It is impossible to estimate millions of cross-section fixed effects accurately, because each depends on at most T observations.
   - This can make it hard to estimate the parameters of interest reliably, even with many millions of observations.

3. One way to model dependence in possibly large samples is to assume that the data fall into G separate **clusters**.

- Clusters might be states or provinces, municipalities, ethnic groups, villages, industries, firms, families, schools or classrooms, time periods, or many other things.
- Unobserved features of units within each cluster are correlated.
- There can be two or more dimensions of clustering. One dimension might be time, and the other might be space.
- A recent reference is

  James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb, "Cluster-robust inference: A guide to empirical practice," *Journal of Econometrics*, 232, 2023, 272–299.

- When an investigator incorrectly assumes independence, or clusters at too fine a level, there can be serious errors of inference.
- The larger the sample, the worse these errors tend to be.

Consequences of Using Clustered Data

Failing to allow for clustering (in the right way) has increasingly severe consequences as N increases.

Recall the simple regression model (3), where we care about β_2. There are now $G = 10$ clusters, with $N_g = N/G$:

$$y_{gi} = \beta_1 + \beta_2 x_g + u_{gi}, \quad g = 1, \dots G, \quad i = 1, \dots N_g. \tag{6}$$

The x_g vary only at the cluster level. The u_{gi} are independent across clusters, but equi-correlated with correlation ρ within clusters.

- Clustering only causes problems if *both* the regressors and the disturbances are correlated within clusters.

Let $\rho = 0.0$. For $N = 100{,}000$, the .05 and .95 quantiles of $\hat{\beta}_2$ are 0.4939 and 0.5061. For $N = 100$, they are 0.3068 and 0.6931. Ratio of intervals is $(0.6931 - 0.3068)/(0.5061 - 0.4939) = 31.66 \cong \sqrt{1000} = 31.62$.

Figure 4: Distributions of $\hat{\beta}_2$ for $N = 100$ with Clustering

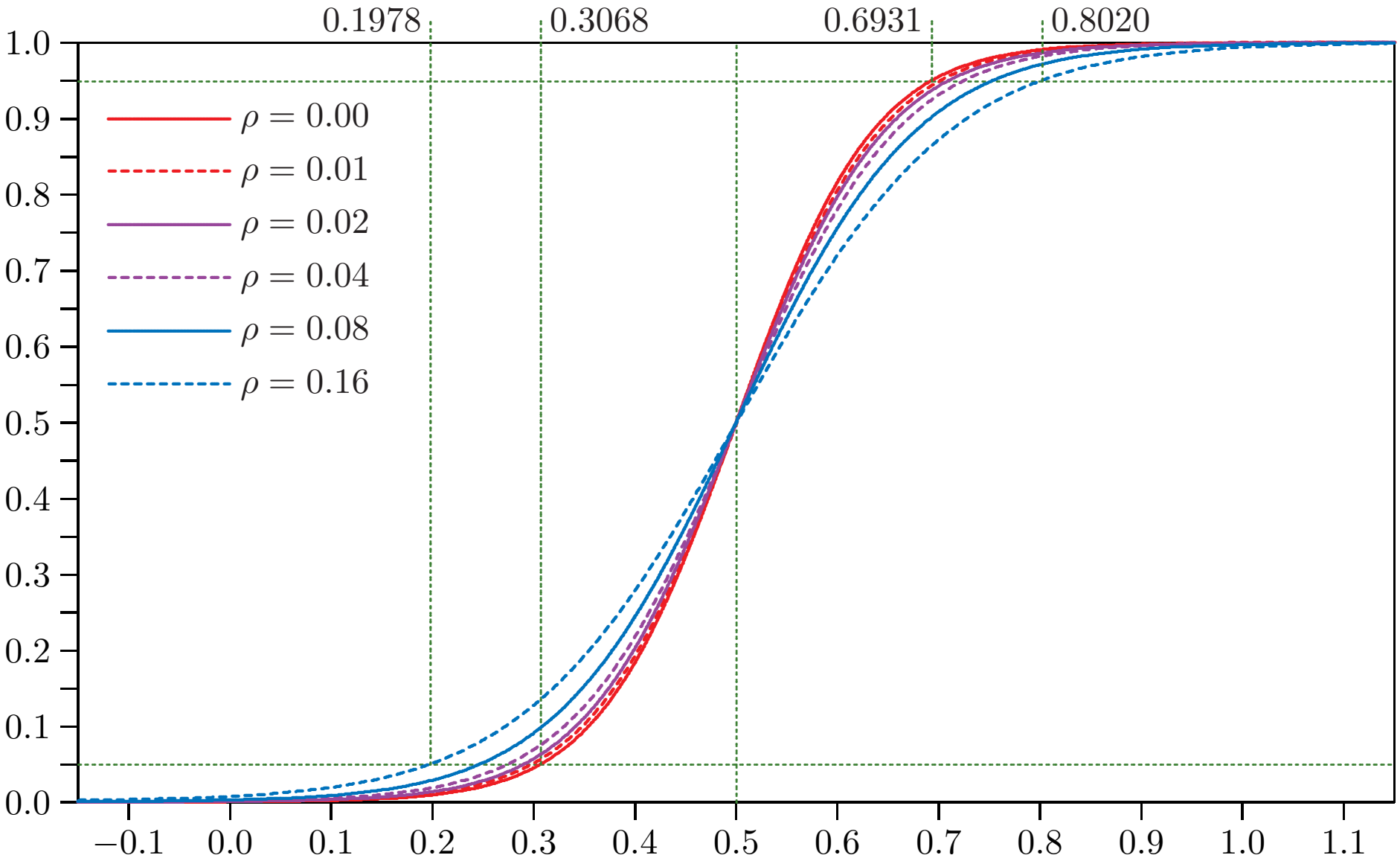
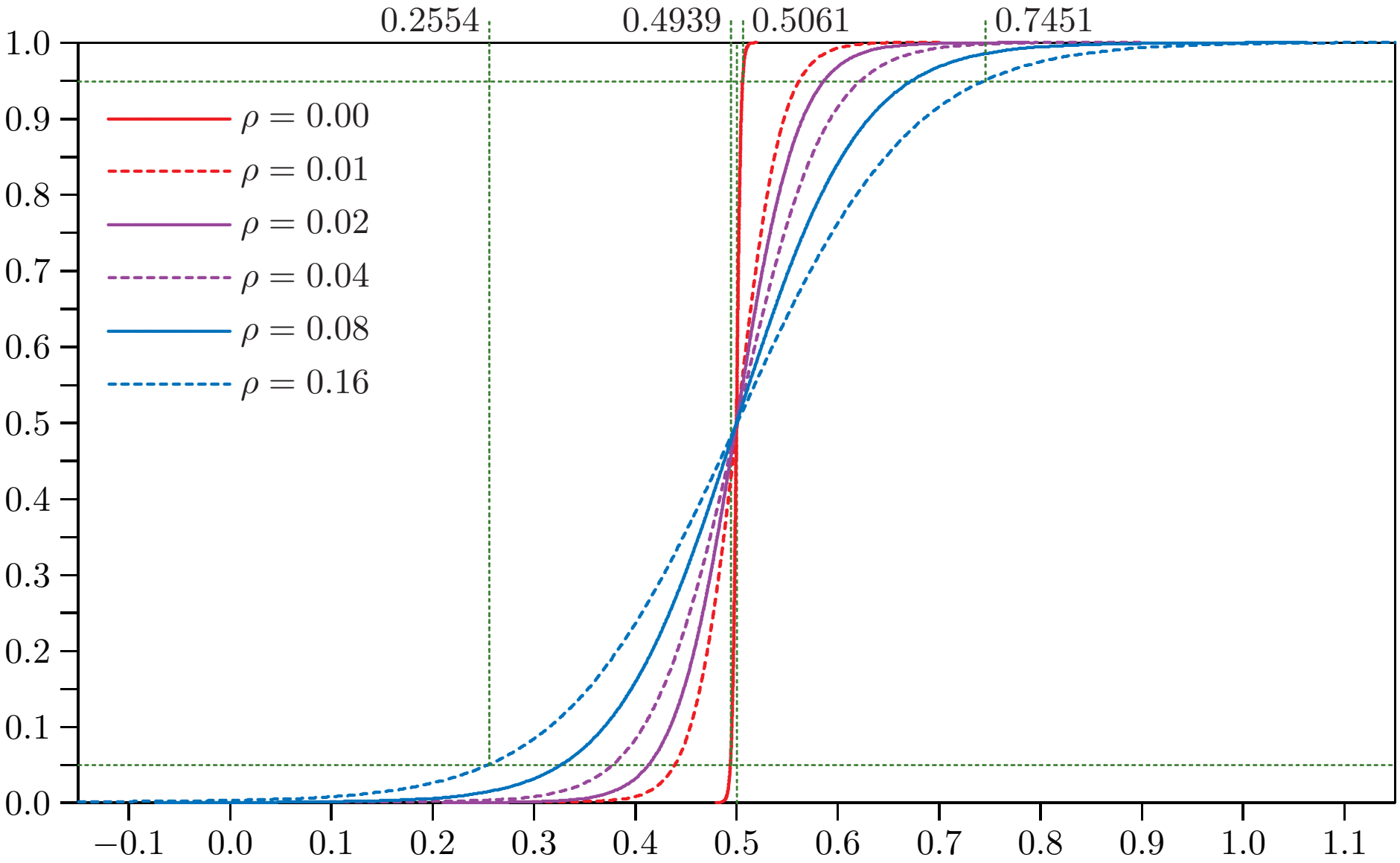


Figure 5: Distributions of $\hat{\beta}_2$ for $N = 100{,}000$ with Clustering



But when $\rho = 0.16$, the same ratio is

$$\frac{0.8020 - 0.1978}{0.7451 - 0.2554} = 1.23 << 31.62.$$

Another interesting number is the **effective sample size**

$$100 \left(\frac{0.6931 - 0.3068}{0.7451 - 0.2554} \right)^2 = 62.2.$$

Numerator is .05–.95 interval for 100 uncorrelated observations, and denominator is interval for 100,000 equi-correlated ones with $\rho = 0.16$.

Intra-cluster correlation has turned a large sample into a tiny one!

Even when $\rho = 0.01$, the effective sample size is just 990 observations for a sample of $N = 100{,}000$.

Effective sample sizes for 1 million observations are 62.5 when $\rho = 0.16$ and 998 when $\rho = 0.01$. The last 900,000 observations add almost no information!

The **information content** of "large" samples may not be anything like as large as N suggests.

However, this result depends on the u_{gi} being equi-correlated. They were generated as

$$u_{gi} = v_g + \epsilon_{gi}, \quad i = 1, \ldots, N_g, \tag{7}$$

where the ϵ_{gi} are independent.

This is the classic **random effects** specification. It implies that

$$\bar{u} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{N_g} u_{gi} = \frac{1}{N} \sum_{g=1}^{G} N_g v_g + \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \epsilon_{gi}. \tag{8}$$

Increasing N and the N_g makes the second term converge to zero.

But the first term does not converge. Even for infinitely large samples, $\bar{u}$ (and hence $\hat{\beta}_1$ and $\hat{\beta}_2$) vary a lot with $G = 10$. Values of N_g/N matter.

Clustering in Real Data

In MacKinnon (2016, 2019), I used data from the U.S. Current Population Survey for 1979–2015 to estimate the **earnings equation**

$$y_{gti} = \beta_1 + \sum_{j=2}^{5} \beta_j \mathrm{ED}_{gti}^{j} + \beta_6 \mathrm{Age}_{gti} + \beta_7 \mathrm{Age}_{gti}^2 + \gamma_t + \eta_g + u_{gti}. \quad (9)$$

Here y_{gti} is the log of weekly earnings for person i in year t in state g.

There are 1,156,597 observations and 93 regressors, of which 50 are state dummies and 36 are year dummies.

(9) is a linear regression model with **two-way fixed effects**.

Do the fixed effects account for all the intra-cluster correlation, as they would in the random effects model (7)?

If so, using heteroskedasticity-robust standard errors is valid.

Placebo Regressions

One way to see whether a procedure yields reliable inferences in practice is to estimate **placebo regressions**, in which we add a completely fake regressor to a regression using real data.

- This idea was used in Bertrand, Duflo, and Mullainathan (2004), with treatment regressors that they called "placebo laws."
- The "placebo regressor" should have no explanatory power, so tests at the .05 level should reject 5% of the time.
- We do not need to make any assumptions about how the data are generated, but we generate the placebo regressor(s) ourselves.
- Whatever within-cluster correlations may be present for any level of clustering simply exist in the data.

For the CPS dataset, I ran 10,000 placebo regressions for various numbers of falsely treated states.

The placebo regressors were constructed as follows:

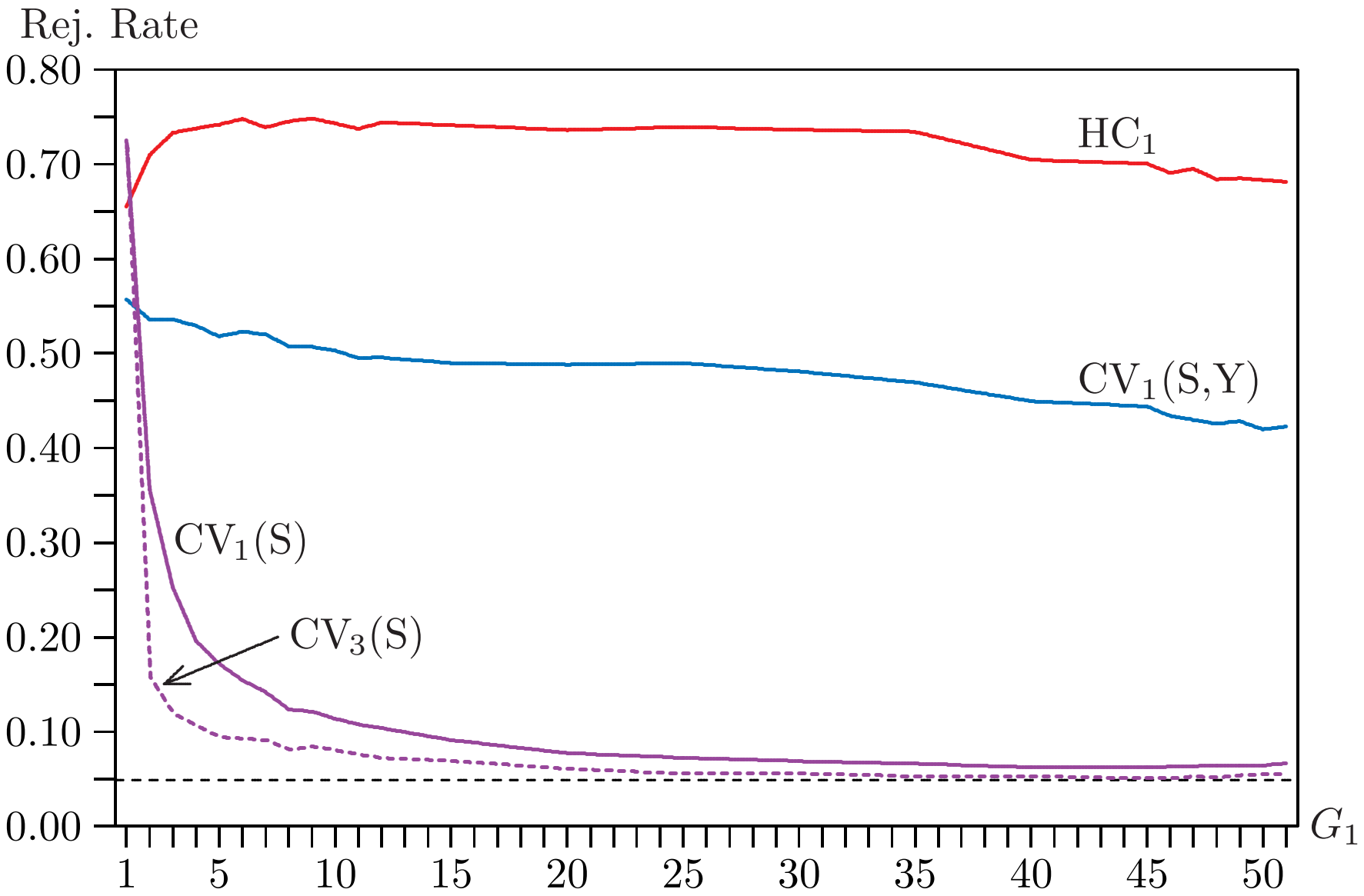
- Given G_1, the number of treated states, the first year of treatment (T_g) was chosen randomly to be between 1983 and 2011.
- For each treated state, all observations of the placebo regressor for each year starting with T_g were set to 1.
- A randomly-generated placebo regressor was added to (9) for each of the 10,000 placebo regressions.
- The same placebo regressors might occur more than once. This is most likely to happen when G_1 is small or large.

Figure 6 shows rejection frequencies, as a function of G_1, for t-tests based on several ways of obtaining standard errors.

HC_1 just allows heteroskedasticity. CV_1(S,Y) allows clustering by state-year intersections. CV_1(S) and CV_3(S) allow clustering by state.

Despite the state and time fixed effects, not clustering, or clustering by state-year intersections, leads to severe over-rejection.

Figure 6: Rejection Frequencies for Placebo Regressions



Larger sample sizes make things worse! Figure 7 shows what happens as we shrink the sample size.

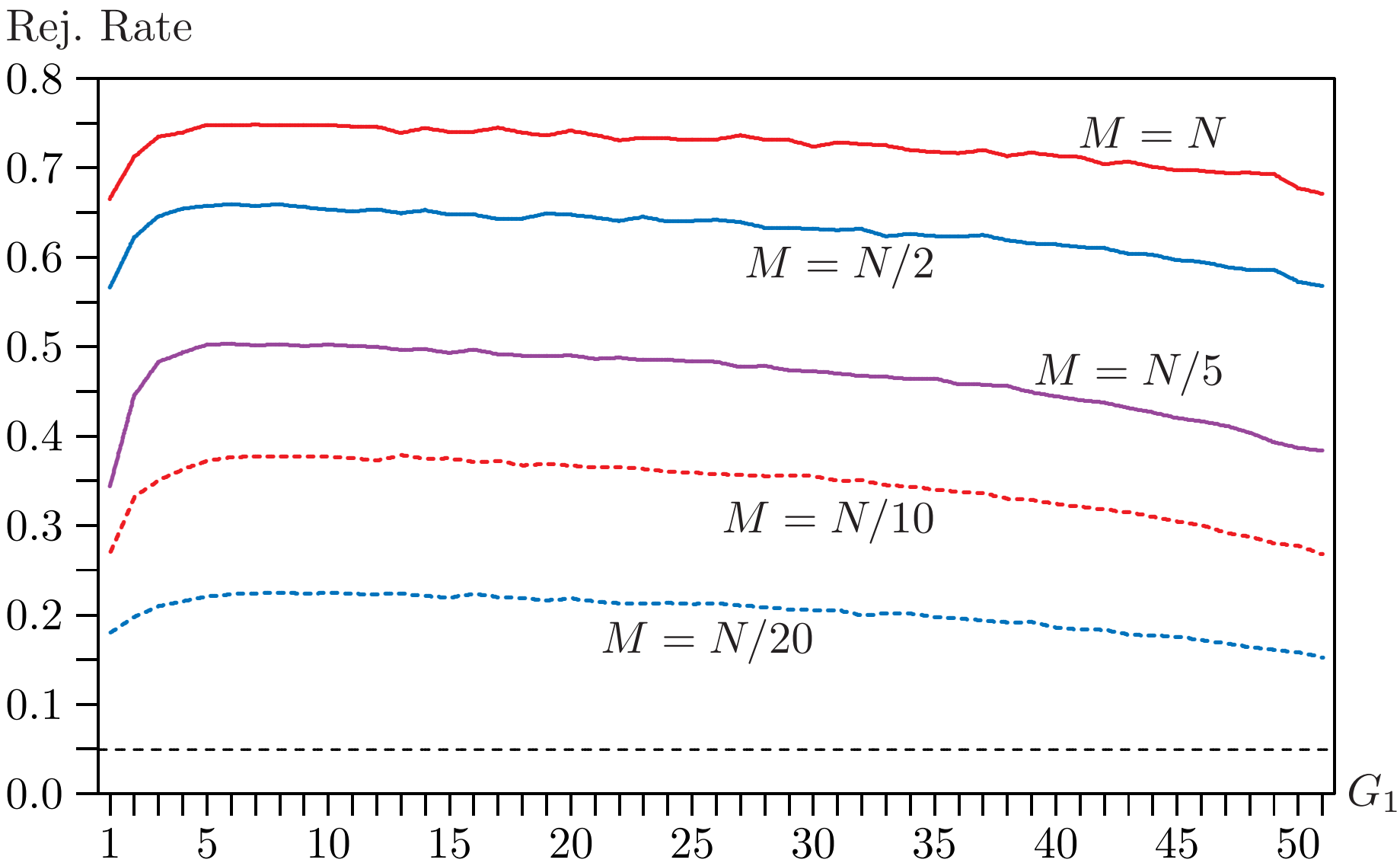
- To obtain smaller samples, I threw away 1/2, then 4/5, then 9/10, and finally 19/20 of the observations.
- I did this 2, 5, 10, or 20 times and averaged the results so as to minimize the random effects of subsampling.
- The smallest samples have $M = N/20 \approx 57,830$ observations, which would have been considered quite large 30 years ago.
- Using heteroskedasticity-robust standard errors would have led to fairly accurate inferences for samples of just a few thousand.

At least with data like these, it is large sample sizes that make it essential to use cluster-robust standard errors.

We only observe serious over-rejection with heteroskedasticity-robust standard errors for large samples.

This is also true of standard errors based on clustering at the level of state-year intersections.

Figure 7: HC_1 Rejection Frequencies, $N = 1{,}156{,}597$



Computational Issues

Prior to the mid-1960s, a computer was a person equipped with a Friden or Monroe calculator.

Econometricians cared deeply about efficient computation!

Computational considerations led to the Frisch-Waugh-Lovell result:

- "Partialing out" regressors not of intrinsic interest does not affect the coefficients or standard errors of the remaining regressors.
- This result is still very valuable for estimating models with large numbers of fixed effects.

Between 1986 and 1996, personal computer processors became hundreds of times faster for least-squares computations.

Datasets got larger, but (usually) not by nearly as much. So efficiency in computation could often be neglected.

But computation seems to be becoming more important again.

Computational costs inevitably increase as N increases.

- Luckily, the cost of OLS (for a given number of regressors) rises at the same rate as N itself.
- We need to avoid computations for which the cost increases more rapidly than N does.
- We especially need to avoid storing $N \times N$ matrices, or any objects with a number of elements proportional to N^2.

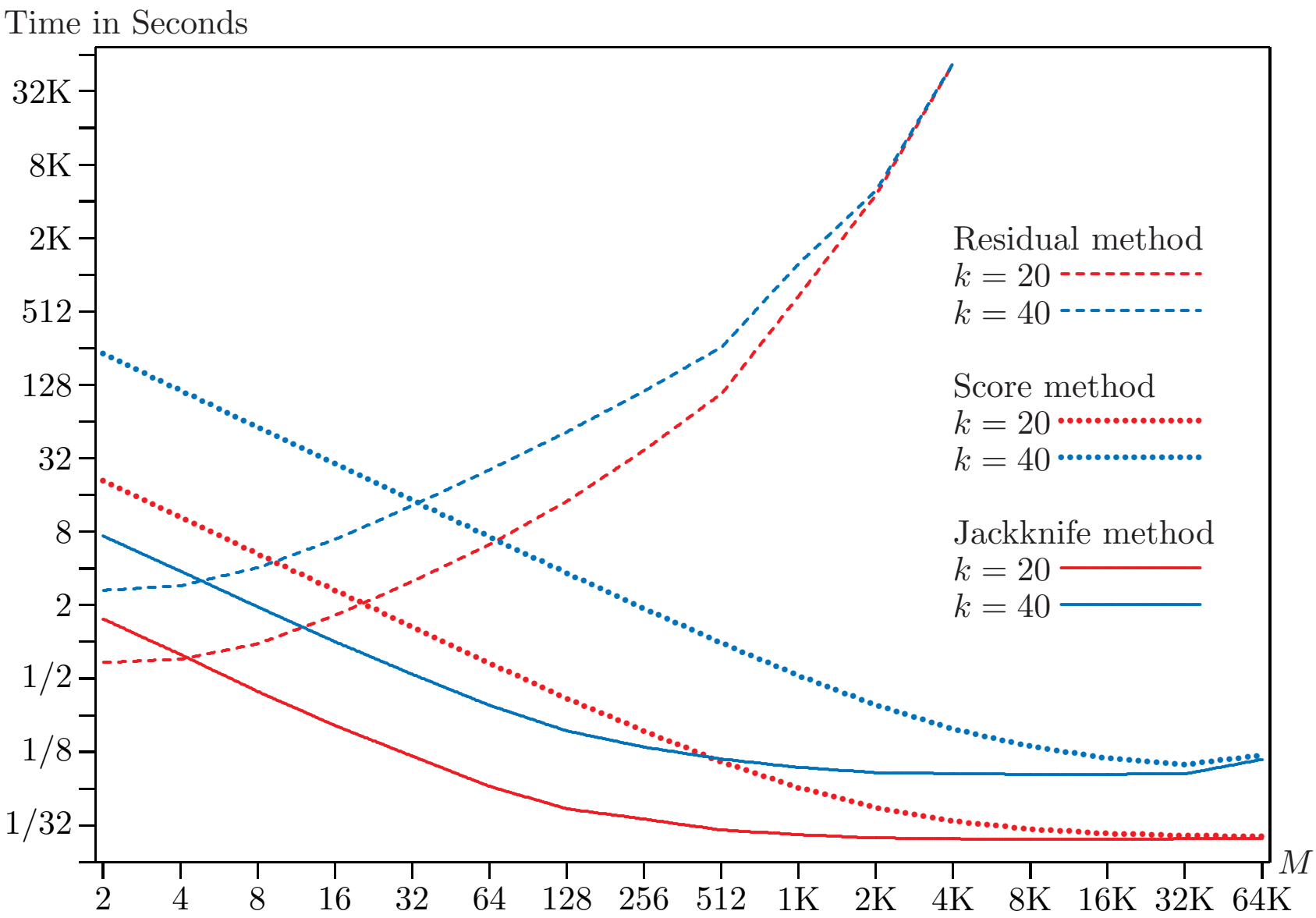
There may be several methods with different computational costs.

Figure 8 shows time in seconds for three methods of computing the CV3, or cluster jackknife, covariance matrix. It is from MacKinnon, Nielsen, and Webb (JAE, 2023).

Here $N = 1{,}048{,}576$. There are G clusters. Each of them contains $M = N/G$ observations.

The most efficient method for very small M becomes by far the most expensive method for moderate and large cluster sizes.

Figure 8: Timings for Three Ways to Compute CV_3



Bootstrap Methods

Inferential methods based on the **bootstrap** often yield more reliable P values and confidence intervals than ones based on asymptotic theory.

- Why not just rely on asymptotic theory when N is large?
- Asymptotic theory may not be reliable when there is dependence, such as clustering, even when N is very large.

Bootstrap methods involve drawing B bootstrap samples and computing estimates and/or test statistics for each of them.

This can be expensive when N is large. It is often $B + 1$ times as expensive as just estimating a model once. **Infeasible**?

The **wild cluster bootstrap** is due to Cameron, Gelbach, and Miller (2008). New variants in MacKinnon, Nielsen, and Webb (JAE, 2023).

Original WCR/WCU-C bootstraps, plus new WCR/WCU-S ones, are available in **boottest** for `Stata` and **fwildclusterboot** for `R`.

The wild cluster bootstrap can be valuable when G is small, or cluster sizes vary a lot, or there are few treated clusters, even if N is large.

Thanks to some neat algebra, the cost of the wild cluster bootstrap does not have to be proportional to $N(B+1)$.

Some computations are proportional to N, and others are proportional to B. Thus **boottest** and **fwildclusterboot** can handle very large samples with B large (say, 99,999) using very little computer time.

The wild cluster bootstrap can be much cheaper than purely analytic methods when the latter involve computations proportional to N^2.

Even the **resampling**, or **pairs** bootstrap, where we generate bootstrap samples by resampling from the rows of the data matrix, can be made reasonably fast for linear regression models with clustering; see MacKinnon (E&S, 2023).

Computational tricks can be used in other cases to make bootstrap methods feasible for large samples. So bootstrap methods should not be rejected out of hand.

More about Computation

When samples are small, we usually do not need to think about computational efficiency. When samples are large, we must.

If we are writing `Stata` or `R` packages, we must think about efficiency, because we do not know how large sample sizes will be!

- When a sample contains only a little more information than subsamples of $N/10$ or $N/20$ observations, perhaps we can do much of the estimation using a subsample.
- For nonlinear models, methods developed for machine learning can reduce the costs of numerical optimization, perhaps at the cost of not getting quite the right answer.
- It is rarely essential to obtain fully accurate estimates for bootstrap samples. One method of saving time in nonlinear models was proposed in Davidson and MacKinnon (IER, 1999).

Final Remarks

If large samples were just like the samples in a first-year statistics course, only larger, then dealing with them would be easy.

Using standard methods, we would obtain highly accurate estimates with tiny standard errors, at least for correctly specified models.

But large samples are generally not as simple as the ones in STATS 101, because they typically involve some dependence across observations.

This can lead to serious errors of inference. When N is large,

1. we need to think very carefully about dependence, because failing to do so can get us into a lot of trouble;
2. we may need to think about computation a lot more carefully than we usually do;
3. we should not assume that bootstrap methods are either not needed or infeasible.