# Bootstrap Basics

Jonah B. Gelbach

CELS Presentation

November 4, 2011

# What *Is* the Bootstrap?

- Method for inference for parameters that
  - Depend on an unknown probability distribution *F*
  - Example:
    *F* as distribution of $\varepsilon_i$ in regression model
- Sometimes defined by Monte Carlo re-sampling
  - Re-sampling, yes.
  - MC not strictly necessary (though virtually always done)
- BS can be used for bias correction, too
  - I won't discuss this—focus on inference

# Why Use the Bootstrap

Three potential reasons

1. Standard theory correct but difficult to implement
2. Multiple testing problems
3. Small  n makes asymp approximation problematic

# Asymptotic v. BS Methods

- Conventional asymptotic methods
  - Estimate parameter of interest
  - Use asymptotic theory to estimate parameter's a.d.
  - Base inference on estimated asymptotic distribution
- BS methods
  - Estimate parameter of interest
  - Find a consistent estimate of *F*; call it *F-hat*
  - Use *F-hat*'s properties to calculate either
    - Test statistic or
    - Critical values of test statistic's distribution in size-n sample

# Asymptotic Inference

- Typical method
  - Construct statistic $T_n$
  - Use asymptotic theory to find critical values
  - Example: A.d. of t-statistics is N(0,1)
  - Reject if $|T_n| > z_{1-a}$ or $z_{1-a/2}$, as appropriate

# The Basic BS Routine

1) Take observed sample $S_n = \{X_1, X_2, \ldots, X_n\}$
2) Calculate statistic of interest

   $$Xbar = n^{-1}\Sigma_{i=1 \text{ to } n} X_i$$

3) Do the following routine B times
   i. Draw n times w/replacement to create

      $$S_{n,b} = \{X_{1,b}, X_{2,b}, \ldots, X_{n,b}\}$$

   ii. Calculate some statistic $\theta_{n,b}$
4) Do something with $\Theta_{n,B} = \{\theta_{n,1}, \theta_{n,2}, \ldots \theta_{n,B}\}$

# Example 1: BS Standard Error

(3) (ii) $\theta_{n,b}$ = $Xbar_{n,b} = n\ ^{1}\Sigma_{i=1 \text{ to } n}X_i$

(4) Calculate variance estimate

$$V_{n,B} = n\ ^{1}\Sigma_{i=1 \text{ to } n}(Xbar_{n,b} - Xbar_n)^2$$

Notes:

- $Xbar_n$ is the true mean of the $\theta_{n,b}$ distribution
- Typical people then test

$$T_{n,B} = Xbar_n/(V_{n,B})½$$

against the critical values of the SN distribution

- This makes sense only if it's tough to estimate V

# Example 2: BS T

(3)(ii) $\theta_{n,b} = T_{n,b} = Xbar_{n,b}/(V_{n,b})^{1/2}$ [Note the "little b"]

(4) Let $G(t) = \Pr[T_{n,b} \leq t]$

- Estimate the critical value $t_\alpha$ such that

  $$G(t_\alpha) \equiv \Pr[T_{n,b} \leq t_\alpha] = \alpha$$

- Do this with Ghat defined by

  $$\hat{G}(t) \equiv B^{-1}\Sigma_{b=1 \text{ to } B} 1[T_{n,b} \leq t].$$

- Estimated critical value is t-hat $_\alpha$:

  fraction $\alpha$ of $T_{n,b}$ realizations is less than t-hat $_\alpha$

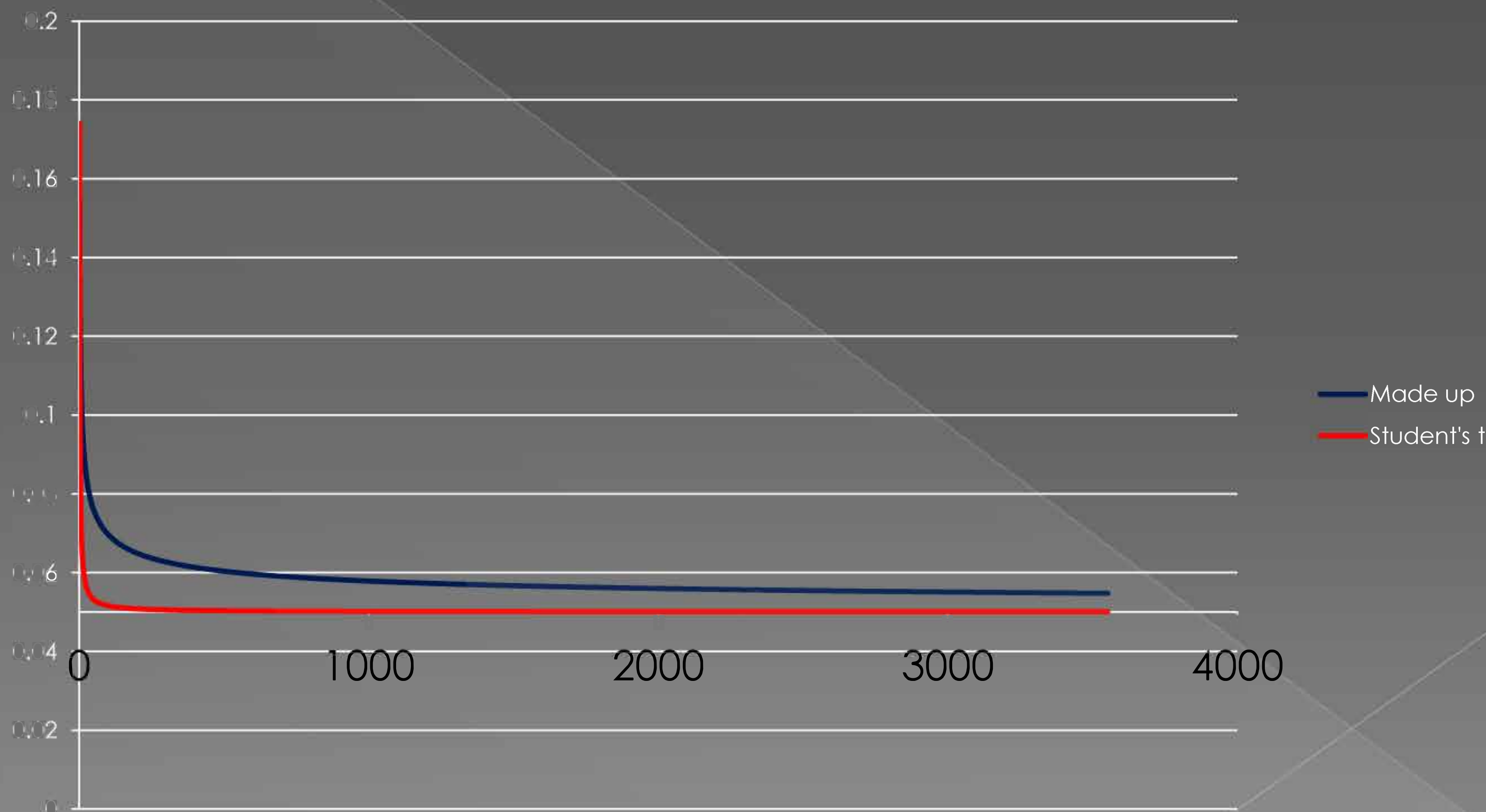
# Why Resampling Works

- Suppose we want to estimate V(Ybar)
- Typically there's an analytical estimator.
- But for the sake of argument:
  - Population resampling would be kosher (if infeasible)
  - A very large, representative subpop would work, too
    - Why? Subpop EDF is consistent for population distribution
    - Have to do the resampling with replacement
  - Sample EDF itself also consistent for pop distribution!
    - So we can just as well resample from the original sample
- Census wages example: Paper 1, Table 1

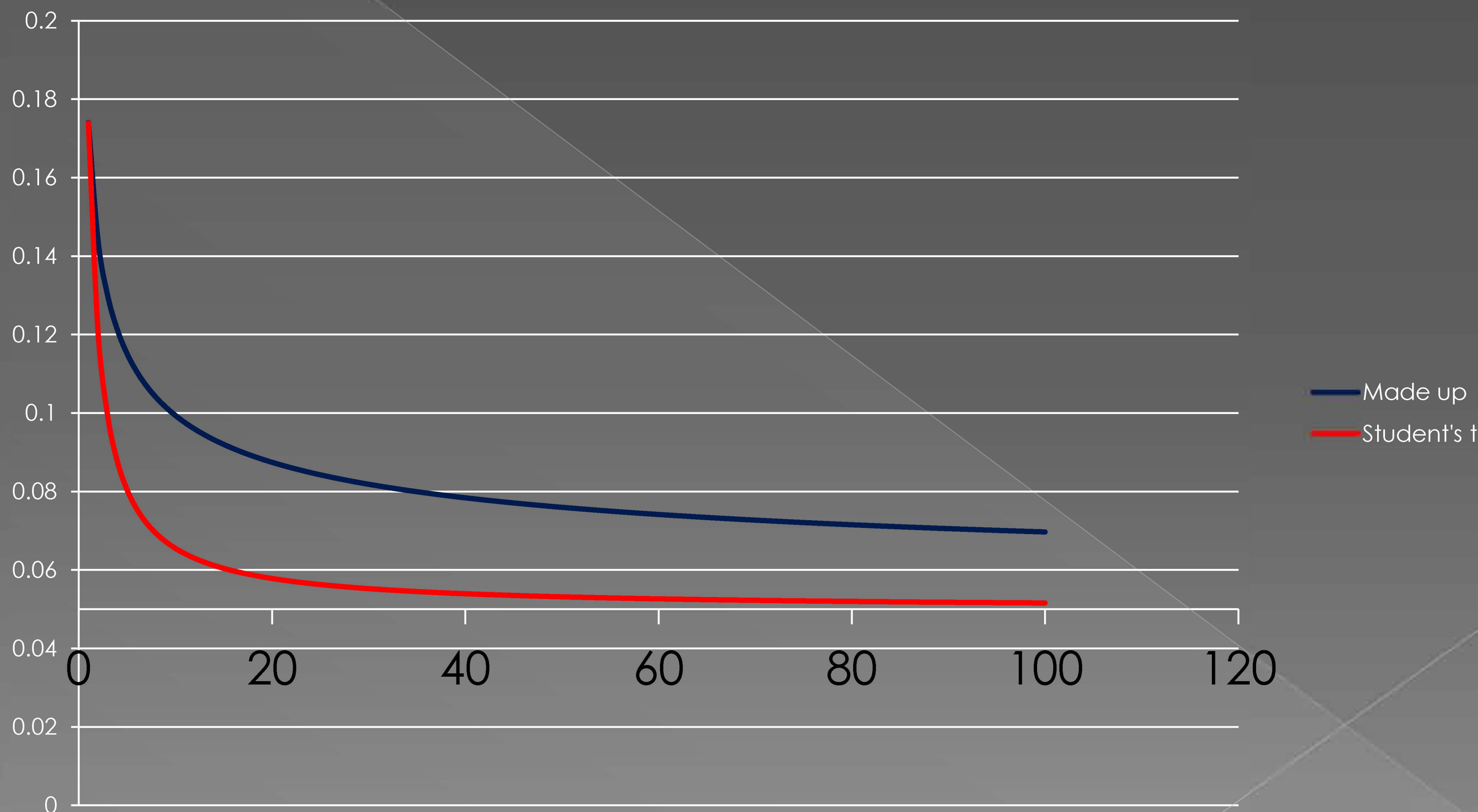
# The Bootstrap-T is Better

- BSing the s.e.
  - First-order equivalent to using asymptotic theory
- BSing the critical value of the t-statistic
  - Increases convergence rate by order of magnitude
- Error rates:
  - $\alpha hat_{BSSE} - \alpha \rightarrow 0$, but $n^r(\alpha hat_{BSSE} - \alpha)$ does not for $r>0$
  - $\alpha hat_{FOAT} - \alpha \rightarrow 0$, but $n^r(\alpha hat_{FOAT} - \alpha)$ does not for $r>0$
  - $\alpha hat_{BS\ T} - \alpha \rightarrow 0$, but $n^r(\alpha hat_{BS\ T} - \alpha)$ <u>does</u> for $r<1/2$
- So, when the sample size is small, BS-T is better

# Convergence Rate - 1



# Convergence Rate - 2



# Asymptotic Refinement

- The secret to getting improved convergence is
  - Bootstrap something whose a.d. is <u>known</u>
  - The t-statistic has a.d. N(0,1)
    - It is "asymptotically pivotal"
  - The s.e. has unknown a.d.
    - So t-statistic based on BSSE is not asymptotically pivotal
- Can generalize idea to many a.p. statistics
- Reasons are deep and pretty technical

# Application: Regression With Clustering

The Basic Model

- There are G groups (clusters). Ex: Federal circuits.
- $N_g$ observations in each group that can vary
  - Cross-sectionally
  - Over time
- Want to do inference on β, e.g., $H_0$: β=0

$$y_{ig} = \mathbf{x}'_{ig}\boldsymbol{\beta} + u_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G,$$

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g, \quad g = 1, \ldots, G, \tag{1}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

# Application: Regression With Clustering

- People usually follow famous BDM (2004) paper
  - Use “cluster-robust” covariance estimator
  - E.g., cluster on federal circuit
  - This works under two assumptions:
    1. No correlation in residuals <u>across</u> circuits
    2. Enough G for “middle matrix” to be good approximation to its expectation.
  - (Problem: C=12 not big!)

# Application: Regression With Clustering

- The middle matrix is based on
  - Huber-White-Eicker etc idea
  - No cross-cluster conditional dependence in u
  - Arbitrary within-cluster dependence
    - Within-g info not used to prove convergence
    - But see C. Hansen, J. Econometrics 2007
  - As. properties depend on (1/G)*middle matrix

$$\hat{V}_{\mathrm{CR}}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G} \mathbf{X}_g \tilde{\mathbf{u}}_g \tilde{\mathbf{u}}_g' \mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}. \quad (3)$$

# A BS Solution to Small-G

- Cameron, Gelbach & Miller (2008, *ReStat*)
  - Treat clusters like individual observations
  - Re-sample whole clusters at a time (BDM did this)
  - But: use BS-t to take advantage of convergence rate
- Simulations in CGM show BS-t does very well
  - Useful to impose null hypothesis via "wild bootstrap"
  - Difficult to do in nonlinear models (e.g., probit)
    - But Pat Kline & Andres Santos show alternative

# Some Results From CGM

TABLE 3.—1,000 SIMULATIONS FROM DGP WITH GROUP-LEVEL RANDOM ERRORS AND HETEROSKEDASTICITY
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

| Estimator # | Method | Number of Groups ($G$) 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|
| 1 | Assume i.i.d. | 0.302<br>(0.015) | 0.288<br>(0.014) | 0.307<br>(0.015) | 0.295<br>(0.014) | 0.287<br>(0.014) | 0.297<br>(0.014) |
| 3 | Cluster-robust | 0.208<br>(0.013) | 0.118<br>(0.010) | 0.110<br>(0.010) | 0.081<br>(0.009) | 0.072<br>(0.008) | 0.068<br>(0.008) |
| 5 | Pairs cluster bootstrap-se | 0.174<br>(0.012) | 0.111<br>(0.010) | 0.109<br>(0.010) | 0.085<br>(0.009) | 0.074<br>(0.008) | 0.070<br>(0.008) |
| 10 | Pairs cluster bootstrap-t | 0.079<br>(0.009) | 0.067<br>(0.008) | 0.074<br>(0.008) | 0.058<br>(0.007) | 0.054<br>(0.007) | 0.053<br>(0.007) |
| 13 | Wild cluster bootstrap-t | 0.053<br>(0.007) | 0.056<br>(0.007) | 0.058<br>(0.007) | 0.048<br>(0.007) | 0.041<br>(0.006) | 0.044<br>(0.006) |

# Focusing On Smaller G

| Estimator # | Method | Number of Groups ($G$) 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| 1 | Assume i.i.d. | 0.302 (0.015) | 0.288 (0.014) | 0.307 (0.015) | 0.295 (0.014) |
| 3 | Cluster-robust | 0.208 (0.013) | 0.118 (0.010) | 0.110 (0.010) | 0.081 (0.009) |
| 5 | Pairs cluster bootstrap-se | 0.174 (0.012) | 0.111 (0.010) | 0.109 (0.010) | 0.085 (0.009) |
| 10 | Pairs cluster bootstrap-t | 0.079 (0.009) | 0.067 (0.008) | 0.074 (0.008) | 0.058 (0.007) |
| 13 | Wild cluster bootstrap-t | 0.053 (0.007) | 0.056 (0.007) | 0.058 (0.007) | 0.048 (0.007) |

# Re-examining a BDM Design-1

- Notation wrinkles:
  - g denotes cluster
  - i denotes <u>year</u>
  - n denotes individual

$$y_{nig} = \alpha_g + \gamma_i + \mathbf{x}'_{nig}\boldsymbol{\delta} + \beta_1 I_{ig} + u_{nig},$$

# Re-examining a BDM Design-2

TABLE 6.—250 SIMULATIONS FROM BDM (2004) DESIGN USING MICRODATA
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

| Estimator # | Method | Number of States ($G$) 6 Size | 10 Size |
|---|---|---|---|
| 1 | CRVE cluster on state-year | 0.440 (0.031) | 0.444 (0.031) |
| 3 | CRVE cluster on state | 0.148 (0.023) | 0.100 (0.019) |
| 11 | Wild bootstrap-t cluster on state | 0.080 (0.017) | 0.048 (0.014) |

Note: Micro regressions control for a quartic in age, three education dummies, and state and year fixed effects. Number of Monte Carlo replications $R = 250$. Number of bootstrap replications $B = 199$.

# Resampling Methods

- There are different types of bootstraps
  - Resampling: parametric or nonparametric?
    - Nonparametric: “pairs” resampling
    - Parametric
      - *A priori* known distribution
      - Based partly on empirical distribution
      - For example, draw from fitted residuals
  - Is resampling iid or cluster-based?
  - Is the null hypothesis imposed?

# Whether $H_0$ is Imposed

- Helpful for power
  - Also related to size based on Monte Carlos
- Impossible with pairs resampling
- Possible with parametric forms
  - Fully parametric bootstrap
  - Residual bootstrap
  - Wild bootstrap

# How the Wild BS Works

- Recall

$$y_g = X_g\beta + u_g$$

- First estimate (say, via OLS), to get
  - βhat
  - Uhat ≡ {$uhat_g$}, g=1,2,…G.
- Then re-sample G times from Uhat & create

$$y_{g,b} = X_g\beta hat + uhat_{g,b}$$

- Then re-run OLS of $y_{g,b}$ on $X_g$ to get $\beta hat_b$
- Then do inference using {$\theta hat_b$}

# When Bad Things Happen to Good Re-samples

The bootstrap can fail:

- Boundary problems
  - X ~ U(a,b)
- Non-smooth problems
  - But note that quantiles/qreg OK
- Mass points
- Nearest-neighbor matching

# Other Methods of Inference

BS isn't only "non-standard" approach

- Randomization inference
  - Requires known null hypothesis
- Sub-sampling
  - Re-sample <u>without</u> replacement
  - Use sample sizes $m<n$; have to choose m
- Asymptotic theory
- Student's *t* with df correction