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 asympt 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
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
   
   $X_{\text{bar}} = \frac{1}{n} \sum_{i=1}^{n} X_i$

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) \( \theta_{n,b} = \bar{X}_{n,b} = n \sum_{i=1}^{n} X_i \)

(4) Calculate variance estimate

\[ V_{n,B} = n \sum_{i=1}^{n} (\bar{X}_{n,b} - \bar{X}_n)^2 \]

Notes:

- \( \bar{X}_n \) is the true mean of the \( \theta_{n,b} \) distribution
- Typical people then test

\[ T_{n,B} = \bar{X}_n / (V_{n,B})^{1/2} \]

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) $\theta_{n,b} = T_{n,b} = \bar{X}_{n,b} / (V_{n,b})^{1/2}$  [Note the “little b”]

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

- Estimate the critical value $t_\alpha$ such that $G(t_\alpha) = \Pr[T_{n,b} \leq t_\alpha] = \alpha$
- Do this with $G_\text{hat}$ defined by $G_\text{hat}(t) = \frac{1}{B} \sum_{b=1}^{B} 1[T_{n,b} \leq t]$.
- Estimated critical value is $t_\text{hat}_\alpha$: fraction $\alpha$ of $T_{n,b}$ realizations is less than $t_\text{hat}_\alpha$
Why Resampling Works

- Suppose we want to estimate $V(Y_{\text{bar}})$
- 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:
  - $\hat{\alpha}_{BSSE} - \alpha \to 0$, but $n^r(\hat{\alpha}_{BSSE} - \alpha)$ does not for $r>0$
  - $\hat{\alpha}_{FOAT} - \alpha \to 0$, but $n^r(\hat{\alpha}_{FOAT} - \alpha)$ does not for $r>0$
  - $\hat{\alpha}_{BS-T} - \alpha \to 0$, but $n^r(\hat{\alpha}_{BS-T} - \alpha)$ does for $r<1/2$

- So, when the sample size is small, BS-T is better
Convergence Rate - 1
The secret to getting improved convergence is

- Bootstrap something whose a.d. is known
- 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
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 \( \beta \), e.g., \( H_0: \beta = 0 \)

\[
\begin{align*}
y_{ig} &= x_{ig}' \beta + u_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \\
y_g &= X_g \beta + u_g, \quad g = 1, \ldots, G, \\
y &= X\beta + u,
\end{align*}
\]
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 across 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 $(1/G) \times$ middle matrix

\[
\hat{V}_{CR}[\hat{\beta}] = (X'X)^{-1} \left( \sum_{g=1}^{G} X_g \tilde{u}_g \tilde{u}_g' X_g' \right) (X'X)^{-1}. \tag{3}
\]
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
### 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)

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

<table>
<thead>
<tr>
<th>Estimator #</th>
<th>Method</th>
<th>Number of Groups (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>Assume i.i.d.</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.015)</td>
</tr>
<tr>
<td>3</td>
<td>Cluster-robust</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.013)</td>
</tr>
<tr>
<td>5</td>
<td>Pairs cluster bootstrap-se</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.012)</td>
</tr>
<tr>
<td>10</td>
<td>Pairs cluster bootstrap-t</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.009)</td>
</tr>
<tr>
<td>13</td>
<td>Wild cluster bootstrap-t</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.007)</td>
</tr>
</tbody>
</table>
Re-examining a BDM Design-1

- Notation wrinkles:
  - $g$ denotes cluster
  - $i$ denotes year
  - $n$ denotes individual

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

<table>
<thead>
<tr>
<th>Estimator #</th>
<th>Method</th>
<th>Number of States</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Size</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>CRVE cluster on state-year</td>
<td>0.440</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.031)</td>
</tr>
<tr>
<td>3</td>
<td>CRVE cluster on state</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.023)</td>
</tr>
<tr>
<td>11</td>
<td>Wild bootstrap-t cluster on state</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Size</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>CRVE cluster on state-year</td>
<td>0.444</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.031)</td>
</tr>
<tr>
<td>3</td>
<td>CRVE cluster on state</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.019)</td>
</tr>
<tr>
<td>11</td>
<td>Wild bootstrap-t cluster on state</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.014)</td>
</tr>
</tbody>
</table>

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
  - \( \beta \hat{\text{hat}} \)
  - \( U \hat{\text{hat}} \equiv \{ u \hat{\text{hat}}_g \}, g = 1, 2, \ldots G. \)
- Then re-sample G times from \( U \hat{\text{hat}} \) & create
  \[ y_{g,b} = X_g \beta \hat{\text{hat}} + u \hat{\text{hat}}_{g,b} \]
- Then re-run OLS of \( y_{g,b} \) on \( X_g \) to get \( \beta \hat{\text{hat}}_b \)
- Then do inference using \( \{ \theta \hat{\text{hat}}_b \} \)
The bootstrap can fail:

- **Boundary problems**
  - $X \sim U(a, b)$

- **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 without replacement
  - Use sample sizes m<n; have to choose m
- Asymptotic theory
- Student’s $t$ with df correction