Bootstrap Methods

Bob Stine
Department of Statistics, Wharton School
University of Pennsylvania
Philadelphia, PA 19104

stine@wharton.upenn.edu

April 14, 2000

Topics

- Foundations and heuristics
- Applications in "regression" problems
- Confidence intervals
- Caveats to casual application

Illustrative Question

Health Status

- What is the average level of osteoporosis in postmenopausal women in the US?

Small Sample

- 20 postmenopausal women
  - sample of "typical" patients
  - collection of clinics
- Osteoporosis measured by hip x-ray
  - converted to a "t-score"
  - "young normal" has mean 0 and SD 1.

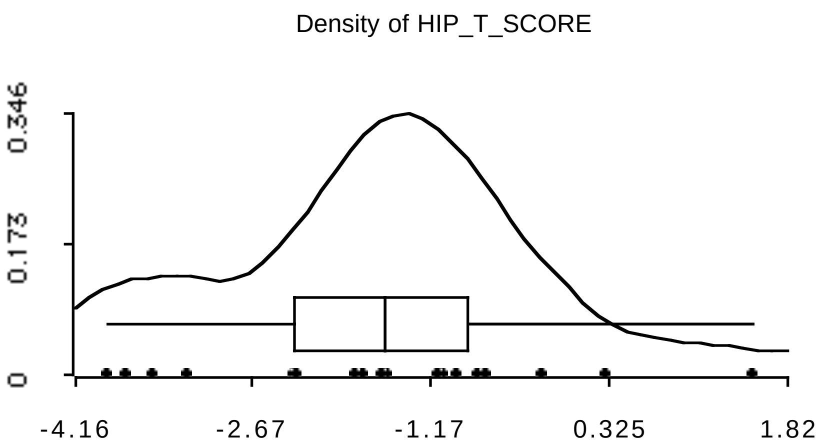
Data Analysis

- Initial summary statistics
  Mean t-score is –1.58 with SD 1.36
- Data "roughly" normally distributed

What can one infer from this data?

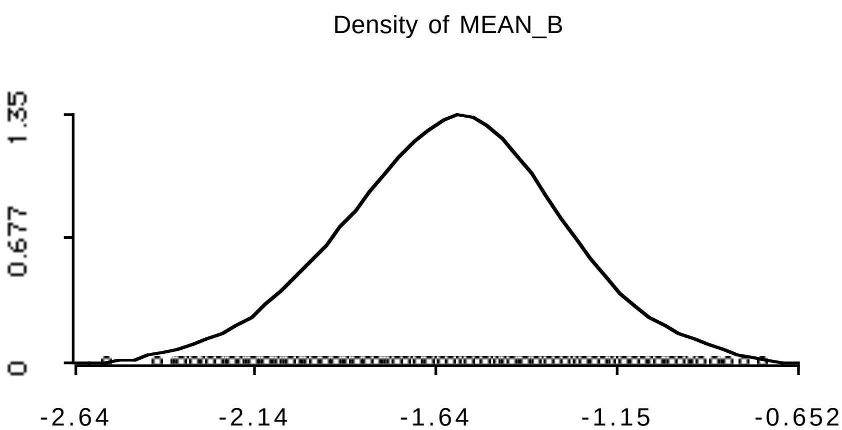
Bootstrap Approach

Observed histogram



Treat this sample as population

- Samples with replacement from this collectio of 20 values, as though it were a population.
- Calculate the average from each sample.



- Estimate SE of mean as 0.286
  95% CI as [–2.2, –1.03]

Classical Approach

Standard Error and Normality

- Estimate standard error using formula as

$$SE(\bar{Y}) = s \big/ \sqrt{n}, \qquad s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

- Form a confidence interval as

$$\bar{Y} \pm t_{\alpha/2}\left(s \big/ \sqrt{n} \right)$$

- Requires
  - Knowledge of t-distribution
  - Normality of sample
  - Expression for standard error

- From data

SE = 1.58/√20 = 0.30

and the associated 95% (two-sided) interval is

[–2.22 , –.945]

Why are these so similar to bootstrapping results?

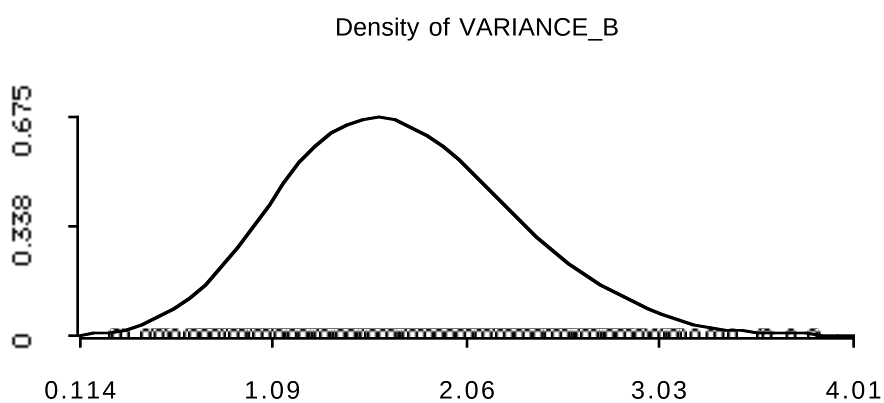
Are they always so similar?

What do standard error and CI mean?

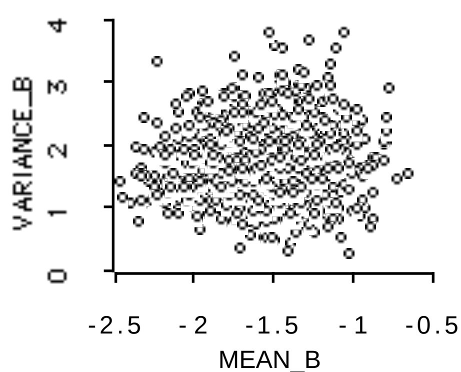
Intervals for the Variance

Simple bootstrap approach

- Treat observed data as population.
- Compute s^2 for each of many samples of size 20 from this “bootstrap population”, sampling with replacement to get different samples.



- Obtain estimated SE(sample var)=0.56 and a 95% interval of [0.8, 3].
- Bonus: Plot resampled variance on mean



Classical Approach for Variance

Assuming Normality

If the data are normal, then

$$s^2 = \sigma^2 \frac{\chi^2_{n-1}}{n-1} \quad \text{Ł} \quad Var(s^2) = \frac{2\sigma^4}{n-1}$$

and the 95% confidence interval is

$$\left[\frac{(n-1)s^2}{\chi^2_{.975,n-1}}, \quad \frac{(n-1)s^2}{\chi^2_{.025,n-1}} \right], \quad P\left(\chi^2_{n-1} \le \chi^2_{\alpha,n-1}\right) :$$

Results for this Sample

The observed $s^2 = 1.36^2 = 1.85$ so that

$SE(s^2) \approx \sqrt{2}(1.85)/\sqrt{19} = 0.60$ (vs 0.56)

and the 95% CI is

$[^{19(1.85)}/_{32.8},\ ^{19(1.85)}/_{8.9}] = [1.07,\ 3.95]$

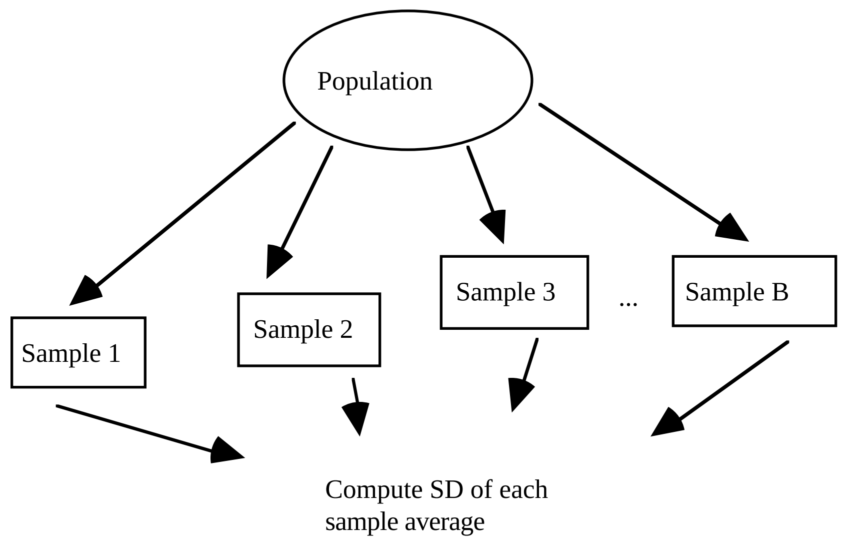
(vs [0.8, 3] for BS)

SE's again close, but not the interval?

Frequentist Confidence Intervals

Models and Assumptions

- Standard methodology (e.g.,t-test) assumes
  - independent observations
  - constant precision (equal variance)
  - normal population
- Idealized sampling picture



- Existential experiment + math implies

$$Y_i \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \bar{Y} \sim N(\mu, \sigma^2 / n)$$

- Mathematical model of sampling variation
  - Describes sample-to-sample variation
  - Derivation of t-quantile $t_{\alpha/2}$

Alternatives to Classical Methods

Simulation

- Make the existential sampling real.
- Pretend the population is, e.g., normal with some mean and variance.

Ranks and permutations

- Exact inference
- Analysis based on order statistics
- Hard to extend to some multivariate methods

Jackknife

- Tukey's 1958 abstract
- Re-compute statistic leaving out one
- Does not generalize well
    - Jackknife samples are too close
    - Fails for the median
- Closely related to bootstrap
    - Type of approximation

Bootstrap

- Simulation with original data as population.
- Compute *observable* sampling distribution.

Bootstrap Resampling

Key idea

- Sample represents all you know about the population, so use it as the "population".
- Assumptions remain
  - independence
  - sampling one population
- "Shape" of the population not assumed.

Key Condition for Statistic

- Depends "smoothly" on underlying populatio
- Mean-like statistics fare well.
- Role of theory is to establish this equivalence

Computing Bootstrap Samples

- Sample with replacement
- Number of replications depends on problem.
- Empirical distribution of sample treated as population with probability 1/n at each obs.
- n^n samples are possible
- Elaborate methods available, but not general.
  - Estimate P(N(0,1)>5) by simulation?

Bootstrap Notation (see references)

Original process

Population $\rightarrow$ (y_1, y_2, ..., y_n) $\rightarrow$ $\overline{Y}$

Resampling process

BS Sample 1: (y_3, y_7, ..., y_2) $\rightarrow$ $\overline{Y}_1^*$

BS Sample 2: (y_8, y_1, ..., y_1) $\rightarrow$ $\overline{Y}_2^*$

….

BS Sample B: (y_4, y_9, ..., y_{11}) $\rightarrow$ $\overline{Y}_B^*$

Resampling analogy re-expressed

$$T(\hat{F}) - T(F) \quad \Leftrightarrow \quad T(\hat{F}^*) - T(\hat{F})$$

where

F = population

$\hat{F}$ is the empirical distribution and

$\hat{F}^*$ is EDF of a bootstrap sample.

Just a Computational Method?

Computer is not really needed

- Bootstrapping is a perspective, not computing
- Computing becoming easier and easier!
- Key analogy is fundamental

  Resampling from the sample resembles
  the process that generated the original data

Bootstrap algebra

- Don't need a computer to find the bootstrap estimate of the standard error of a mean

$$Var^*(\overline{Y}^*) = Var^*\left(Y_1^* + Y_2^* + \cdots + Y_n^*\right)/n^2$$

$$= \left(v^2 + v^2 + \cdots + v^2\right)/n^2 = v^2/n$$

- v^2 is the biased ML estimate of the variance,

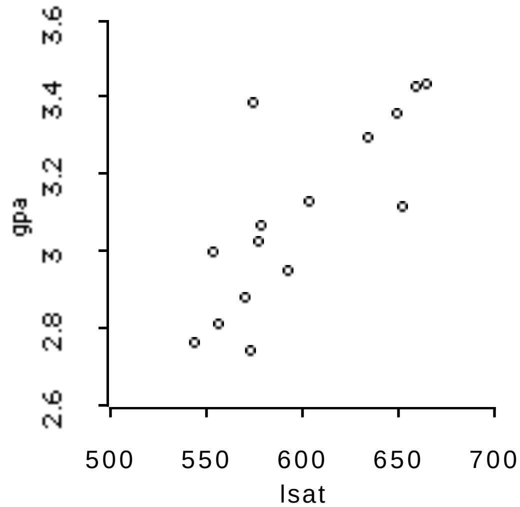
$$v^2 = \sum (Y_i - \overline{Y})^2 / n$$

- The SE for any linear statistic (i.e., a fixed weighted average of the response) can be obtained without computing.

Bootstrapping a Correlation

Classic bootstrap example

- LSAT and GPA values for 15 law schools



- What can one infer about the "population" correlation? The sample correlation is

$$r = 0.776$$

- PS. What is the population anyhow?

Properties of the correlation

- What are the SE/CI for correlation?
  - Both depend on the population ρ.
- Fisher's z transformation
  - Makes SE almost invariant of ρ

Sample results

- Would not make much sense to use an interva of the classic form estimate ± 2 SE(estimate).

- Fisher's transformation gives the 90% confidence interval
  [0.507 , 0.907] = [.776-.269, .776+.131]

- This interval is *not* of the form
  [estimate ± 2 SE of estimate]
  but rather is very asymmetric.

How to bootstrap?

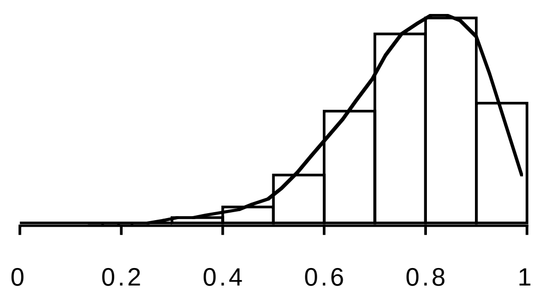
- Keep the data paired – resample observations
  - What happens if sample separately
  - Idea of *bootstrap testing*

- Same basic iteration
  - Draw sample of pairs with replacement from the observed sample.
  - Calculate the correlation for each such bootstrap samples

- Summarizing
  - Use SD of r* as estimate of SE(r)
  - Use percentiles of collection of r* to form a confidence interval

Bootstrap results

- Bootstrap distribution is skewed and clearly not a normal distribution.
- Values accumulate at the upper limit of 1.



• With 3000 replications, the 90% bootstrap interval for the correlation is

[0.520, 0.943] = [.776 - .220, .776+.167]

whereas the Fisher interval is

[0.507 , 0.907] = [.776 - .269, .776+.131]

• Both are skewed and within the range [0,1].

• The bootstrap works without knowing or requiring Fisher's transformation – or the normality it presumes.

• It would not make sense to use the ±2 SE approach since the distribution is not normal and you might easily get a value > 1.

Resampling in Regression

Two Approaches to Resampling

- Random X:
  Resample observations as with correlation example or in one case of t-test.

- Fixed X:
  Resample residuals as follows
  - Fit a model and compute residuals
  - Generate BS data by

$$Y^* = (\text{Fit}) + (\text{BS sample resids})$$

Comparison

	Observations	Residuals
Model-dependent	No	Yes
Preserves X values	No	Yes
Maintains (X,Y) assoc	Yes	No
Conditional inference	No	Yes
Agrees with usual SE	Maybe	Yes
Computing speed	Fast	Faster

Differences are most apparent with outliers.

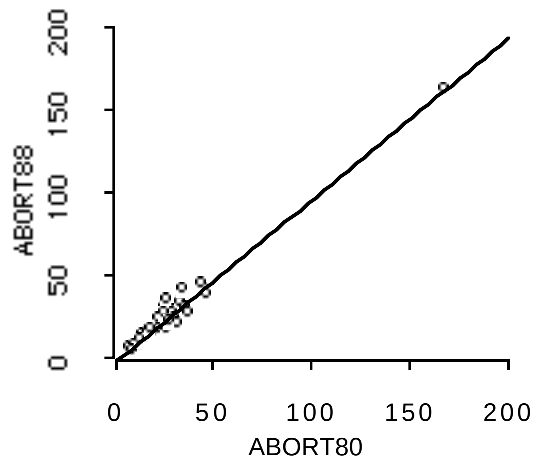
Model Dependence

Suppose that original data are heteroscedastic…

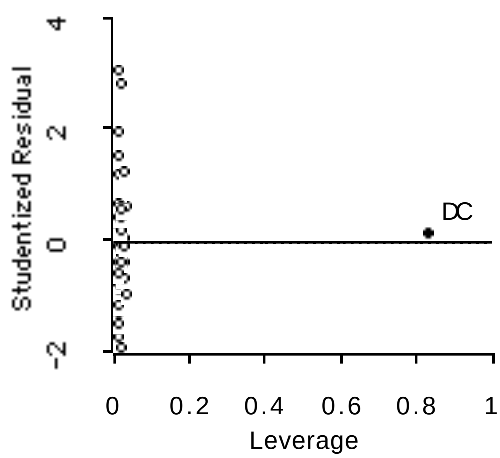
Appearance of bootstrap samples

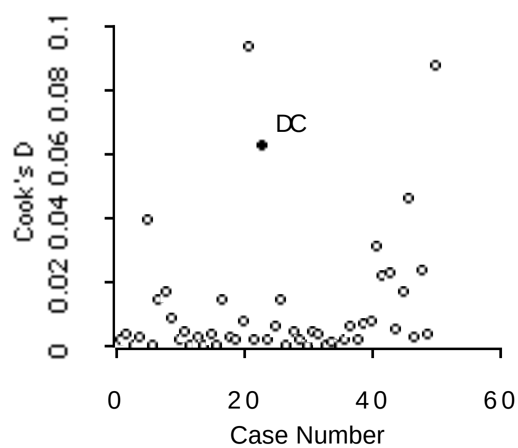
Example: Observation vs Residual BS

Abortion Rates



- DC is leveraged, but not very influential





- Slope standard error

  b = 0.978 SE(b) = 0.0251 (t ≈ 40)

Observation resampling

- Sample “states” as pairs.
- SE*(b) = 0.036 ... bigger than OLS claims

Residual resampling

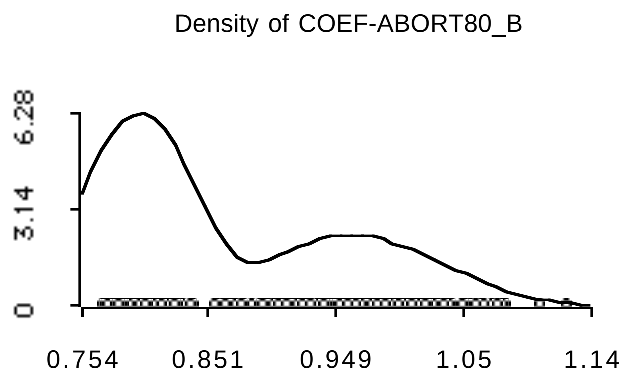
- Sample residuals of fitted model.
- Can compute BS std error without computer.
- SE*(b) = 0.026 ... about same as OLS claim.

Observation SE > Residual SE**

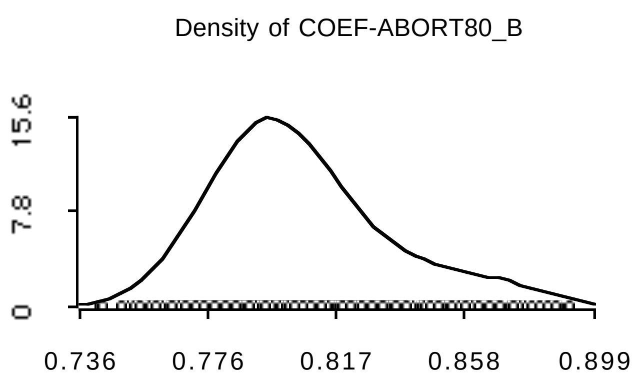
- Is X random or is X fixed?
- Residual resampling estimates Var(b|X)
- Observation resampling estimates Var(b)
- Var(b|X) ≤ Var(b)

Comparison of bootstrap distributions

• Observation resampling binds residual to X location, leading to bimodal distributions.



• Residual resampling “smears” the residual of the outlier, giving a “normal” distribution.



Which Method is Right?

Asymptotically

- Methods converge for large n

Observation Resampling Tradeoffs

+ Does not assume so much of fitted model
  Example with unequal variance.
  Example with curvature.

± Estimates unconditional variation of the slop
rather than the conditional variation.

± Does not always agree with classical SE

– Not appropriate in Anova designs, patterned X's such as time trends
(at least not without special care!)

– Slower to compute (less important these days

What would happen for "another sample"?

- Would you get another outlier for this X?
- Would it again have a negative residual?
- Might expect DC to be an outlier, but not so clear that its error would be negative again.

Locating a Maximum

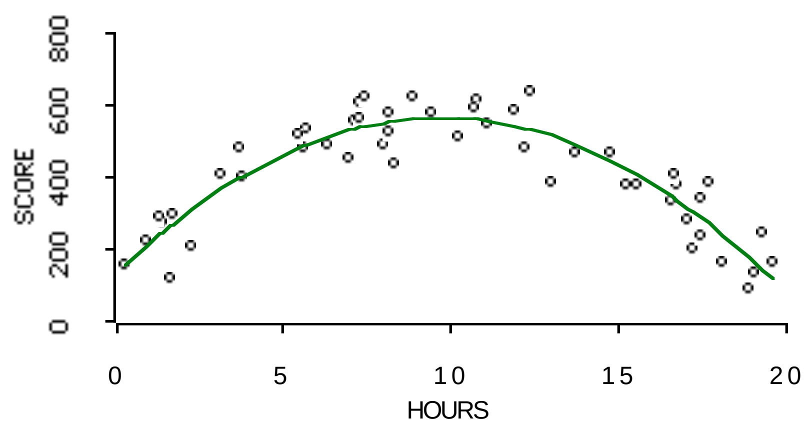
Do you need bootstrapping in regression?

- After all, if fix X, it's a linear estimator…

Where's the maximum

- For what amount of preparation time in hours does maximum test score occur?

Results of fitting a quadratic



- Fitting the model

$$\text{fit} = a + b\,x + c\,x^2$$

via least squares estimates gives

$a = 136$ $\quad$ $b = 89.2$ $\quad$ $c = -4.60$

So, where's the maximum and what's a CI?

- Write the fit as

$$f(x) = a + b\,x + c\,x^2$$

and then take the derivative,

$$f'(x) = b + 2cx$$

The peak occurs where the derivative is zero.

- Solving $f'(x^*) = 0$ for x^* gives

$$x^* = -\frac{b}{2c} \approx -89.2/(2)(-4.6) = 9.7$$

- Questions
- What is the precision (standard error) of x*?
- Can you find a confidence interval?
- Is there any bias in the estimate?

Classical alternative

- "Delta method" computes an approximate standard error by treating this ratio as a linear function of the slope estimates.

Bootstrap results

• Manipulate bootstrap results in “natural way”: simply dividing bootstrap estimates of the linear term by minus twice the quadratic

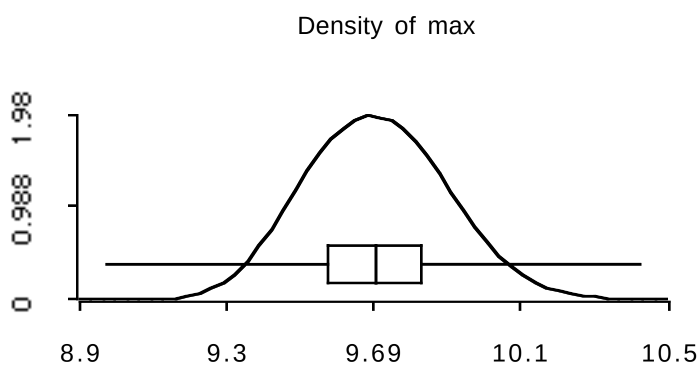
$$max^* = b^* / -2\, c^*$$

• Bootstrap (B=2000) gives usual smaller standard error with fixed resampling…

Observation resampling	SE* = 0.185
Fixed resampling	SE* = 0.174

• Both give a 95% interval of about [9.37, 10]

• The smoothed distribution for the location of the maximum looks pretty normal.



• A quantile plot “heavy tails” as might be expected from a ratio of normals.

Longitudinal Models

Freedman and Peters (1984)

- Regional industrial energy demand
  - 10 DOE regions of the US
- For each region, you observe a short time series, over the 18 years 1961-1978 .

Model

$$Q_{rt} = a_r + b\, C_{rt} + c\, H_{rt} + d\, P_{rt} + e\, Q_{r,t-1} + fV_{rt} + \varepsilon_{rt}$$

where

Q_{rt} = log energy demand in region r, time t
C_{rt}, H_{rt} = log cooling, heating degree days
P_{rt} = log of energy price
V_{rt} = log value added in manufacturing

- Model includes a lagged value of the response as a predictor ("lagged endogenous").

Error assumptions *Block diagonal*

- No remaining autocorrelation
- Arbitrary "geographical" correlation

Generalized Least Squares

Estimators

- Need to know covariance structure in order to get efficient parameter estimates

  $\mathrm{Var}(\varepsilon) = V$ 180x180 block matrix

- Textbook expression

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

- SE for $\hat{\beta}$ comes from

$$\mathrm{VAR}\,\hat{\beta} = (X'V^{-1}X)^{-1}$$

- Problem: Don't know V or its inverse, so you typically estimate it in some fashion from the data itself. However, everyone continues to use the formulas that presume you know the right V.

Results of Simulations

- GLS standard errors that ignore that one has to estimate V are way too small

- BS SE's are larger, but not large enough

Estimation Results

From the paper…

	Est	SE	SE*	SE**
a_1	-0.95	0.31	0.54	0.43
a_2	-1.00	0.31	0.55	0.43
CDD	0.022	0.013	0.025	0.020
HDD	0.10	0.031	0.052	0.043
Price	-0.056	0.019	0.028	0.022
Lag	0.684	0.025	0.042	0.034
Value	0.281	0.021	0.039	0.029

Method of Bootstrap Resampling

- Sample years, since assumed independent over time.
- Use bootstrap to check bootstrap, a so-called bootstrap calibration procedure.
- Values labeled SE** ought to equal SE* (which serve role of true value), but they're less.
- BS is better than nominal, but not enough.

Bootstrap Confidence Intervals

Two basic types

- Percentile intervals that use ordered values of the bootstrapped statistic.

- BS-t intervals have the form of

  estimate ± t-value (SE of estimate)

  Use the bootstrap to find the right multiplier, rather than look up a value in a table.

- I have focused on the percentile intervals
  - I like the pictures of the BS distribution

Alternatives

- Percentile intervals
  - bias-corrected
  - accelerated

- BS-t intervals
  - best if have a SE formula
  - can be very fast to compute

- Double bootstrap methods
  - use the BS to adjust percentiles.
  - another calibration method

- Alternative computing methods
  - importance sampling, “tilting”

Closer Look at Percentile Intervals

Percentile intervals

If g_α denotes the α percentile of the bootstrap distribution of the statistic,

$$P^*(T(X^*) \leq g_\alpha = \alpha,$$

then the 1-α percentile interval is simply

$$[g_{\alpha/2} , g_{1-\alpha/2}]$$

They seem backwards!

Usual confidence interval formed by inverting

$$P\left\{ z_{\alpha/2} \leq \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \leq z_{1-\alpha/2} \right\} = 1 - \alpha$$

with $z_{\alpha/2} = -1.96$ when α=0.05 to

$$P\left\{ \overline{Y} - z_{1-\alpha/2} \sigma \sqrt{n} \leq \mu \leq \overline{Y} - z_{\alpha/2} \sigma \sqrt{n} \right\} = 1 - $$

So, why do they work at all?

Percentile Intervals – Basic Conditions

Utopian conditions

Population parameter	θ
Statistic	$T = T(Y) \sim N(\theta,\nu^2)$
Bootstrap	$T^* = T(Y^*) \sim N(T,\nu^2)$

Ideal 95% confidence interval

$$[T - 1.96\,\nu,\ T + 1.96\,\nu]$$

Percentile interval

Upper endpoint is that value U such that

$$0.975 = P^*(T^* \le U)$$

In other "words", U satisfies

$$\begin{aligned} 0.975 &= P^*((T^*-T)/\nu \le (U-T)/\nu) \\ &= P^*(N(0,1) \le (U-T)/\nu) \end{aligned}$$

so that

$$(U-T)/\nu = 1.96$$

and

$$U = T + 1.96\,\nu$$

Just what we wanted, but those conditions …

Percentile Intervals and Transformation

Unknown transformation

Population parameter θ

Statistic $h(T) \sim N(h(\theta),\nu^2)$

Bootstrap $h(T^*) \sim N(h(T),\nu^2)$

Ideal 95% confidence interval for θ

$$h^{-1}[h(T) - 1.96\,\nu,\ h(T) + 1.96\,\nu]$$

Percentile interval

Upper endpoint is that value U such that

$$0.975 = P^*(T^* \le U) = P^*(h(T^*) \le h(U))$$

or U such that

$$\begin{aligned} 0.975 &= P^*((h(T^*)-h(T))/\,\nu \le (h(U)-h(T))/\,\nu) \\ &= P^*(\,N(0,1) \le (h(U)-h(T))/\,\nu) \end{aligned}$$

so that

$$(h(U)-h(T))/\,\nu = 1.96$$

and

$$h(U) = h(T) + 1.96\,\nu$$

But not all estimators meet these conditions…

Going Further

Generalize further?

- Does not require normality as the common distribution, but this is most likely.
- Can be adjusted to accommodate bias.
  Bias corrected percentile intervals
- Can be further adjusted to accommodate the variance changing with the location
  Accelerated, bias corrected (ABC)

Consequences of generality

- Adjusting for bias, “acceleration” lead to more variation in procedure.
- On average it’s right, but with high variance.
  - Think of trivial 95% interval
- Adjustments can be difficult to accomplish with complex estimators.

Alternative methods

- Bootstrap t intervals
  Make your own t-table, if you can find a standard error to use.
- Double bootstrap methods.

Bootstrapping Variances

Variance for normal sample

$$s^2 \sim \sigma^2\, \chi^2_{n-1}/(n-1)$$

- Both the mean and variance of s^2 depend upo the value of σ^2, unlike a "location" problem.
- No transformation (the "h" used previously) exists for this problem.
- How would you measure the failure of bootstrapping?

Simulation for bootstrap

- Assume that population is normal(0,1).
- Draw samples of size 20.
- For each sample,
  - find the percentile interval
  - see if it covers the truth (σ^2=1)

Simulation results for nominal 95% interval

- Only 409 out of 500 covered, 0.82 (se = .013

Double Bootstrap

Check Percentile Intervals

- Know population, for which $\sigma^2 = 1$.
- Sample population, Y
- Resample Y to obtain the percentile interval
- Compute coverage of nominal 95% interval

Double Bootstrap

- Know the "bootstrap population", with variance v^2.
- Sample the "population", Y*
- Resample Y* to obtain the percentile interval
- Compute the coverage of the interval.

Adjust the Percentiles

- If the nominal 95% percentile interval does not cover, what's the coverage of the nominal 98% interval?
- Tune the *nominal* coverage so that you get the desired level of *actual* coverage.

Double BS Plots

Review for Percentile Intervals

When does it work?

- Suppose BS analogy is perfect.
  - percentile intervals work
- Suppose there is a transformation to perfectio
  - percentile intervals still work
- Suppose there is also some bias.
  - need to re-center
  - bias-corrected intervals
- Allow the variance to change as well
  - need further adjustments
  - accelerated intervals

Example of LSAT data

- Enhanced intervals tend to become more skewed.
- No need to believe that the Gaussian interval is correct ... is this small sample really normal

Second Example for the Correlation

Initial analysis

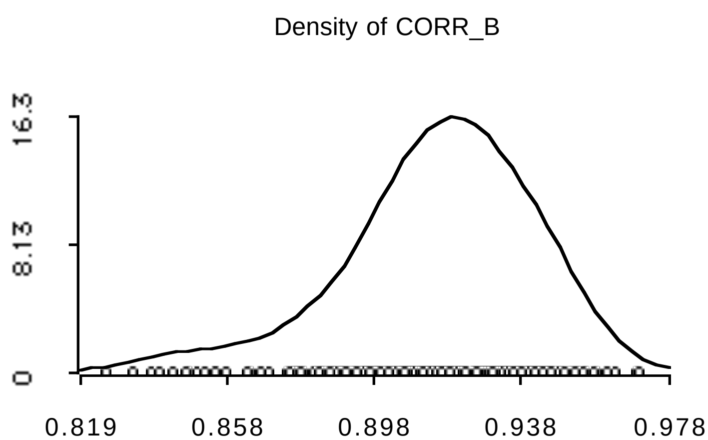
- State abortion rates, with DC removed (50 ob
  - Use filter icon to select those not = D(
- Sample correlation and interval

  corr(88, 80) = 0.915

  90.0% interval = [0.866 0.946]
- Standard interval relies on a transformation which makes it asymmetric.

Bootstrap analysis

• Percentile interval	[0.861, 0.951]
• Bias-corrected percentile	[0.854, 0.946]
• Accelerated percentile	[0.852, 0.946]



Back to Basics - Flaws

Behavior at Extremes

- $M = \text{Max}(X_1, ..., X_n)$
- 95% Percentile is roughly $(x_{(4)}, x_{(1)})$

BUT…

- Expected value of max M is larger than the observed max about 1/2 of the time,

$$\Pr [\ E\ X_{(1)} \geq x_{(1)}\ ] \geq 0.5\ ,$$

so the bootstrap distribution misses a lot of the probability.

Why does the bootstrap fail?

The statistic of interest depends on just the single most extreme observation, regardless of sample size. Getting a larger sample does not improve things.

Regression without a Constant

Leave Out the Constant

- Force the intercept in the fit to be zero.
- Residual average is no longer zero.

Effect on Residual-Based Bootstrap

- If resample residuals, then distribution from which you sample has a non-zero mean value

  BUT

  by assumption the true distribution of the errors has mean zero.

- The lack of a fixed mean of zero in the sampled residuals implies that the bootstrap estimates of variation no longer improve as the sample size increases.

Whose fault is this?

- You need to pay attention when resampling!

Bootstrapping Dependent Data

Sample average

- Example: standard error of mean
- Data: "equal correlation" model

$$\mathrm{Corr}(X_i, X_j) = 1 \quad i=j \qquad \mathrm{Var} = \sigma^2$$

$$\mathrm{Corr}(X_i, X_j) = \rho \quad i \neq j$$

True standard error of average

$$\begin{aligned}\mathrm{Var}(\overline{X}) &= (1/n^2)\ \mathrm{Var}\ (\Sigma X_i) \\ &= (1/n^2)\ (\Sigma\ \mathrm{Var}(X_i) + \Sigma\ \mathrm{Cov}(X_i, X_j)) \\ &= \frac{\sigma^2}{n} + \frac{\rho\sigma^2\ n(n-1)}{n} \\ &= \frac{\sigma^2}{n}(1 + \rho(n-1))\end{aligned}$$

- Does not go to zero with larger sample size!

What happens for bootstrap

- Treats data as independent!
- Adjustments based on blocking data.

Wrapping Up

Bootstrap resampling does

- Produce reliable standard errors and CI's for virtually any estimator.
- Presumes that the resampling parallels the original data generating process.
- Frees time to think about problem, use methods for which CI is hard to come by.
- Shed insights by inspecting the distribution of the bootstrap replications:
    - close to normal, usual methods work
    - far from normal, need to be careful
- Allow one to adaptively select the estimator for a particular data set.
- Can be enhanced, but at a cost.
  Is it worthwhile to make such adjustments?

Bootstrap resampling does not

- Work if resampling done improperly.
- Make good things happen with bad data.
- Fix flaws in your research paradigm.