More Regression, More Confidence Intervals More Everything!

Review with some extensions

Questions from Lecture 4
- Robust regression and the handling of outliers

Animated graphics

Lisp-Stat
- Alternative free software package
- Excels at *interactive* graphics
- Written in language lisp

Axis software interface

Comparison of resampling methods

	Observations	Residuals
Equation-dependent	No	Yes
Assumption-dependent	Some	More
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

New things for today…

Longitudinal data
- Longitudinal (panel) data
- Generalized least squares

Logistic regression (a.k.a., max likelihood)
- Estimating the "error rate" of a model

Path analysis, structural equations

Missing data
- A bootstrap version of imputation

Some theory and chinks in the bootstrap
- Dependence
- Special types of statistics (sample max)

Confidence intervals for the BS
- Justification and improvements

Yes. More t-shirts too!

Robust Multiple Regression

Motivation

Exploratory methods need exploratory tools

Classical tools + data editing = problems

Robust regression automatically weights
Analogy to insurance policy

Fitted model using least squares

Duncan occupation data, 45 occupations

Slopes not significantly different

Variable	Slope	SE	t	p-value
Constant	-6.06	4.27	-1.4	0.16
INCOME	0.60	0.12	5.0	0.00
EDUC	0.55	0.10	5.6	0.00

$R^2 = 0.828$ $s = 13.369$

Reformulated to give the difference as estimate.
- Diagnostic plots show outlier effects
- Difference signif on trimmed data (see R script)
- Effect is not significant on full data set (below)

Variable	Slope	SE	t	p-value
Constant	-6.06	4.27	-1.4	0.16
INCOME	0.053	0.20	0.3	0.80
INC+ED	0.55	0.10	5.6	0.00

$R^2 = 0.828$ $s = 13.369$

Robust Fits for Duncan Model

Biweight fit, with explicit difference

Output suggests a significant difference
- Shows asymptotic SE for estimates
- Agrees with our “drop three” analysis.

Robust Estimates (BIWEIGHT, c=4.685):

Variable	Slope	Std Err	t-Ratio	p-value
Constant	-7.42	2.97	-2.497	0.02
INCOME	**0.34**	0.14	2.404	0.02
INC+ED	0.43	0.068	6.327	0.00

More robust fit suggests more significant
- Robust “tuning constant” set to 2
- Note: resulting iterations need not converge

Robust Estimates (BIWEIGHT, c=2):

Variable	Slope	Std Err	t-Ratio	p-value
Constant	-8.44	2.41	-3.496	0.00
INCOME	**0.40**	0.11	3.464	0.00
INC+ED	0.43	0.056	7.663	0.00

Check the weights for this last regression.

What happens with bootstrap resampling?
- Observation resampling
- Residual resampling

Bootstrap Resampling Robust Regression

Random resampling (biweight, c=2)

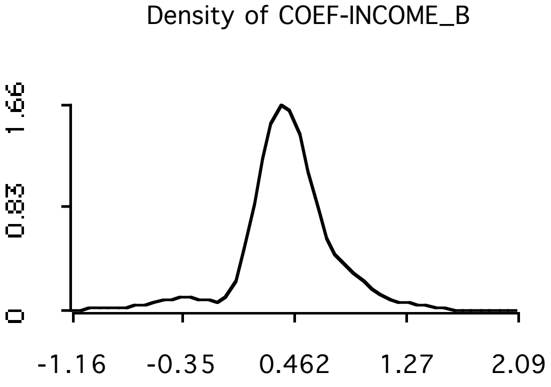
Summary of bootstrap estimates of difference
- Difference in slope of income and education

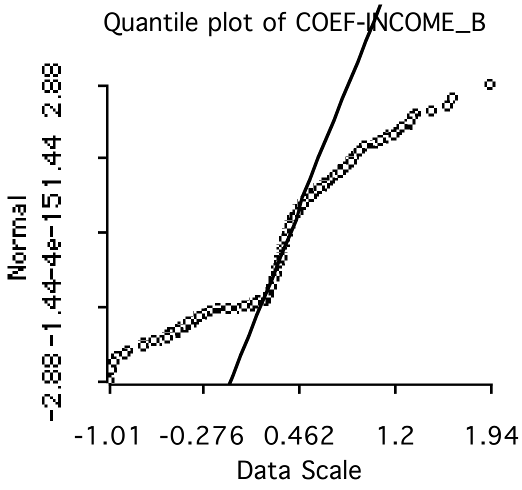
Mean = 0.410 , SD = 0.358 B=500

2.5%	5%	50%	95%	97.5%
-0.497	-0.315	0.380	0.950	1.18

Random resampling
Gives *much* larger estimate of variation (.358 vs .11) and indicates the difference is not significant.

Very non-normal…
- Is the standard deviation meaningful here?





Residual resampling gives…

Numerical summary
- much, much *smaller* SE value
- smaller than original asymptotic value (0.11).

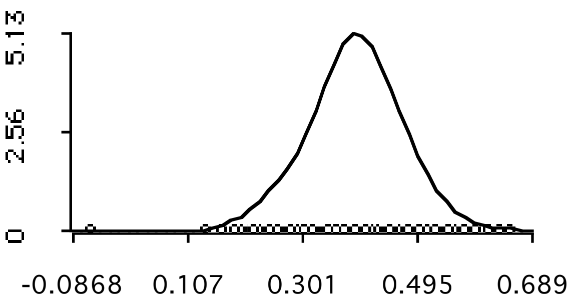
Mean = 0.392 SD = .081 n=500

2.5%	5%	50%	95%	97.5%
0.225	0.254	0.391	0.519	0.551

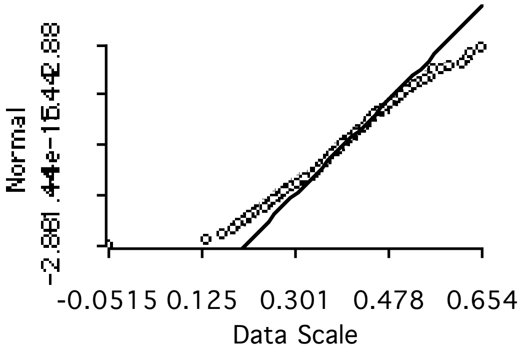
Consequently, it finds a very significant effect.

Bootstrap distribution more normal.

Density of COEF-INCOME_B



Quantile plot of COEF-INCOME_B



What to make of it?

Different conclusions
- Manual deletion gives significant effect
- Resampling with BS does not (random resample)

Bootstrapping a Longitudinal Model

Freedman and Peters (1984)

Full citation in bibliography

Regional industrial energy demand
10 DOE regions of the US

Short time series for each region
18 years 1961-1978 .

Model

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

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 (a.k.a. “lagged endogenous variable”).

Error assumptions *Block diagonal*

No remaining autocorrelation (can’t allow this)
Arbitrary “spatial” correlation

Generalized Least Squares

Estimators

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

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

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 estimate it from the data itself.
- However, most would continue to use the formula that presumes you knew the right V.

Results of F&P's 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

Simulation Results

From the paper of Freedman and Peters…

	Estimate	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
- Assumed independent over time.

Bootstrap calibration
Use bootstrap to check bootstrap (double BS)

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.

Prediction Accuracy

How well will my model predict new data?

Develop and fit model to observed data.

How well will the model predict new data?

Optimistic assessment
If test the model on the data used to construct it, you get an "optimistic" view of its accuracy.

Cross-validation (a.k.a. hold-back sample)
Investigate predictive accuracy on separate data.

Bootstrap approach

Build a bootstrap replication of your fitted model, say M*, based on a bootstrap sample from the original data.

Use the M* to predict the bootstrap population, i.e. use M* to predict the observations Y in the original sample.

Use the error in predicting Y from M* to estimate the accuracy of this model.

Efron and Tibshirani discuss other resampling methods that improve upon this basic idea.

Example of Prediction Error

Duncan regression model

Least squares fit to the sample data
- Estimate □^{2} to be $s^2 = 13.37^2 = 178.8$.

Theory
Prediction error will be a higher than this estimate, by about (1 + k/n), where k denotes the number of predictors. Revises our estimate up to 186.7.

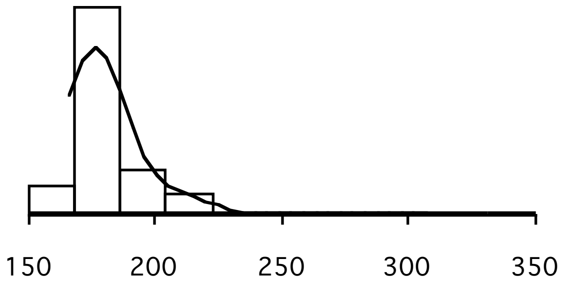
Theory makes big assumption
Presumes that you have fit the “right model”.

Bootstrap results

Indicates that the model predicts about as well as we might have hoped, given the adjustment of 1+2/45.

Mean = 182. SD = 16.1 B=203

2.5%	5%	50%	95%	97.5%
168.	168.	177.	212.	216.



Logistic Regression

Categorical response

Predict choice variable (0/1)

Calculation is iterative least squares algorithm
- same method used in robust regression.

Efron and Gong (1983) discuss logistic regression as well as the problem of model selection.

Classification error
Efron (1986) considers validity of observed error rates and uses bootstrap to estimate "optimism".

Bootstrapping logistic regression

Procedurally similar to least squares
- bootstrap gives distribution for coefficients
- interpretation of coefficients is different

Coefficient standard errors
Output shows asymptotic expressions

Prediction: Is the model as accurate as *it* claims?

Importance of Prediction Error

How do you pick a model?

Interpretation

Prediction

"Natural criterion" since you don't have to make pronouncements of true models.

Pick the model that you think predicts the best… That is, pick the model (or set of predictors) which has the *smallest estimated prediction error*.

Selection bias

When we pick the model that has smallest error, we get an inflated impression of how good it is.

Random variation, not real structure

Such "selection bias" is very severe when we compare more and more models

Happens in context of stepwise regression

Example

stepwise regression and financial data.

Moral

Honest estimates of prediction error are essential in a data-rich environment.

Structural Equation Models

Path analysis

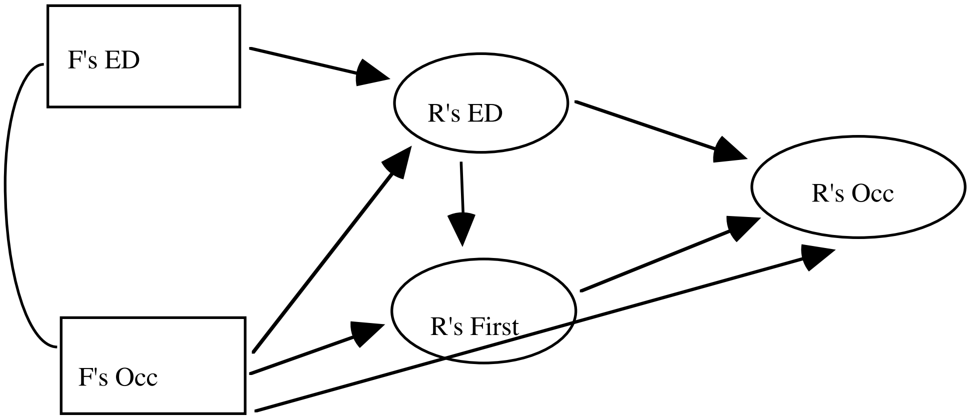
Generalized in Lisrel

Collection of related regression equations

Blau and Duncan recursive model

Comparison of direct and indirect effects

Observation resampling



Computing example

Simulated sample from the Blau&Duncan model.

Recursive

Questions compare direct versus indirect effects.

Multivariate methods

Uncertainty in structural equation models.

General reference
Beran and Srivastava (1985), *Annals Stat*.

Goodness-of-fit in structural equations
Bollen and Stine (1990, 1992). *Soc. Meth*.

Theory for Bootstrap

Sometimes don't need a computer

Simple statistics which are weighted averages
- Sample average
- Regression slope with fixed X.

Bootstrap SE almost usual SE in these cases
- Under fixed resampling in regression

Key analogy revisited

Notation

F is population distribution
F_n is distribution of sample
F_n* is distribution of bootstrap sample
□ is parameter, s is statistic's value

Think in terms of distributions:

$\square = S(F)$ vs. $s = S(F_n)$

Error of the statistical estimate is

$s - \square = S(F_n) - S(F)$

In bootstrap world,

$s = S(F_n)$ vs. $s^* = S(F_n{}^*)$

Error of the statistical estimate is

$s^* - s = S(F_n{}^*) - S(F_n)$

A Flaw - Bootstrapping the Maximum

Behavior at extremes

$M = \mathrm{Max}(X_1, ..., X_n)$

95% Percentile is roughly $(x_{(4)}, x_{(1)})$

BUT...

Expected value of sample 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?

Not a “smooth” statistic
max depends on “small” feature of F_n.

Sampling variation of real statistic

$$S(F_n) - S(F)$$

is not reproduced by the bootstrap version

$$S(F_n{}^*) - S(F_n)$$

Illustration

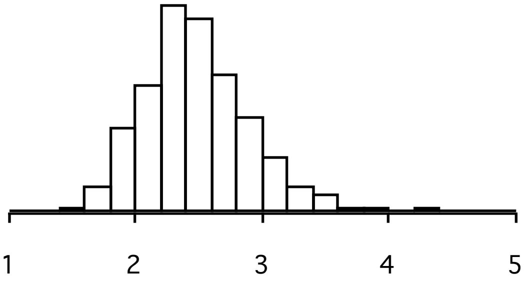
Simulation

Simulate the max of samples of 100 from normal population, using the "bootstrap" command menu item,

Estimator	max
Sampling rule	normal-rand 100
Number trials	1000

Bootstrap distribution

Use AXIS to simulate what the distribution of the sample maximum looks like



Bootstrap results for a random sample

Normal sample

Define sample "norm" of 100 normals, using

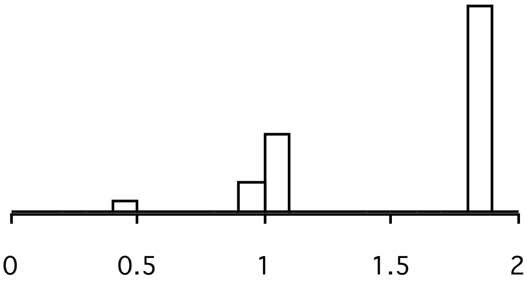
"normal-rand 100"

and be sure to convert to values!

Bootstrap

Resampling from this fixed sample,

Estimator	max
Sampling rule	resample norm
Number trials	1000



The observed max of the data is the max of a bootstrap sample with probability

$$1 - (1 - \frac{1}{n})^n \approx 1 - \frac{1}{e} = 0.63$$

Discussion

Sample alone does not convey adequate information in order to bootstrap maximum.

Have to add further information about “tails” of the population (parametric bootstrap)

Regression without a Constant

Leave out the constant

Force the intercept in the fitted model to be zero.

Average residual
Residual average is no longer zero. Mean of residuals must be zero when have a constant term in the regression model.

Effect on residual-based bootstrap

Resample residuals
The distribution of "bootstrap errors" from which you sample has a non-zero mean value
BUT
by assumption the true distribution of the errors has mean zero.

Consequence: the bootstrap fails.
Bootstrap estimates of variation contain spurious source of variation

Whose fault is this?

Residual resampling requires model validity.

Bootstrapping Dependent Data

Sample average

Example: standard error of mean

Data: “equal correlation” model

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

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

True standard error of average

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

Does not go to zero with larger sample size!

Bootstrap estimate of standard error

Sample with replacement as we have.

Bootstrap estimate is $\frac{s^2}{n}$

Bootstrap does not “automatically” recognize the presence of dependence and gets the SE wrong.

What should be done?

Find a way to remove the dependence.

Preserve dependence
Resample to retain the dependence (variations on random resampling), as in the Freedman and Peters illustration.

Model
Find a model for the dependence and use this model to “build in” dependence into bootstrap.

Generic tools
Recent methods such as block-based resampling and sub-sampling offer hope for model-free methods.

Missing Data and the Bootstrap

Places to read more

Efron (1994) “Missing data and the bootstrap”

Davison and Hinkley (1997)
Bootstrap Methods and their Application

Two approaches to missing data

Key assumption: Missing at random

(1) Use estimator that accommodates missing
e.g., EM algorithm

(2) “Impute” missing and analyze complete data.

Imputation

Multiple imputation is currently “popular”

Refined version of hot deck

Propensity scores

Bootstrap approach to imputation

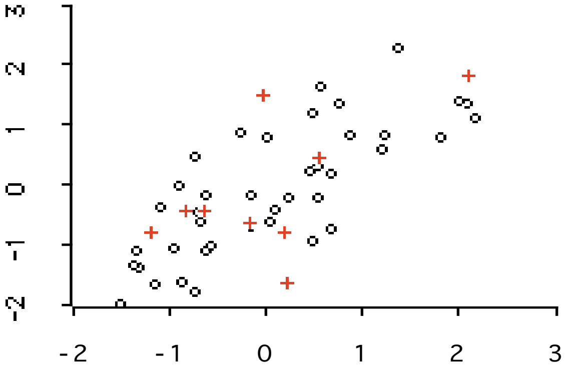
Bootstrap version
- Fill in the missing values preserving variation
- Fit to complete data

Use associated bootstrap estimate of variation

Correlation Example

Setup

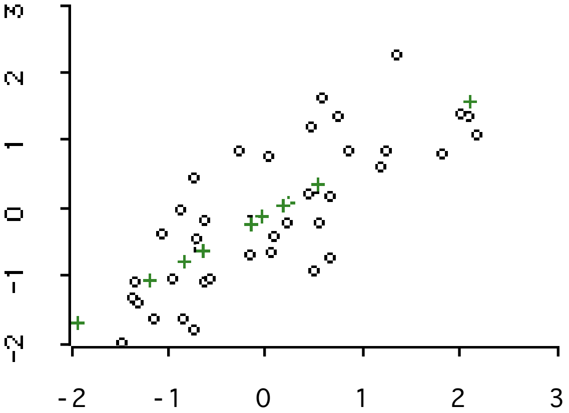
Two variables (X and Y), with missing on Y



Assume linear association (lots of assumptions)
Can predict/fit Y from X

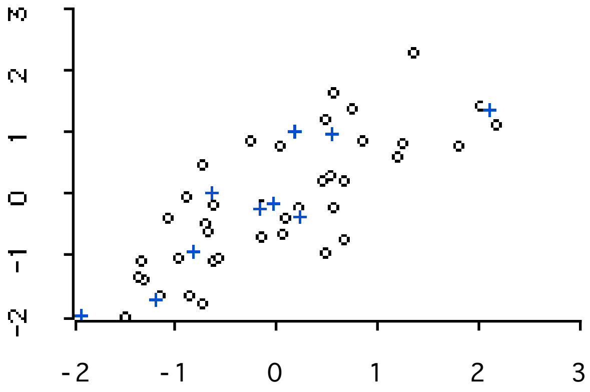
How to generate the bootstrap samples

Cannot just fill in missing Y with predictions
Would understate variation.



Alternative

Fill in Y using the method resembling fixed X resampling from regression



Results

Sensitivity

Example estimates of “sensitivity of analysis” to presence of missing data.

Missing imputation adds variation.

Similar to goal of multiple imputation.

Bootstrap Confidence Intervals

Two basic types

Percentile intervals
Use ordered values of the bootstrapped statistic.

t-type intervals
BS t-intervals have the form of
estimate ± t-value (SE of estimate),
Use the bootstrap to find the right "t-value", rather than looking up in a table.

We have focused on the percentile intervals
- go with the graphs!

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.
- bootstrap the bootstrap.

Standard Percentile Interval

Procedure

Start with large number (B ≈ 2000) reps

Sort the replications and trim off the edges

BS interval is the interval holding remaining

Example with Efron LSAT data

Correlation

Stability in the extremes requires much more data than to compute standard error.

SE is more easy to obtain
Compare SE's based on B=200 to CI based on same replications.

Some Theory for Percentile Intervals

When does it work?

Suppose BS analogy is perfect.
- percentile intervals work

Suppose there is a transformation to perfection.
- percentile intervals still work
- example of Fisher's z-transform for corr.

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 obs)

- Use filter icon to select those not = DC

Sample correlation and interval

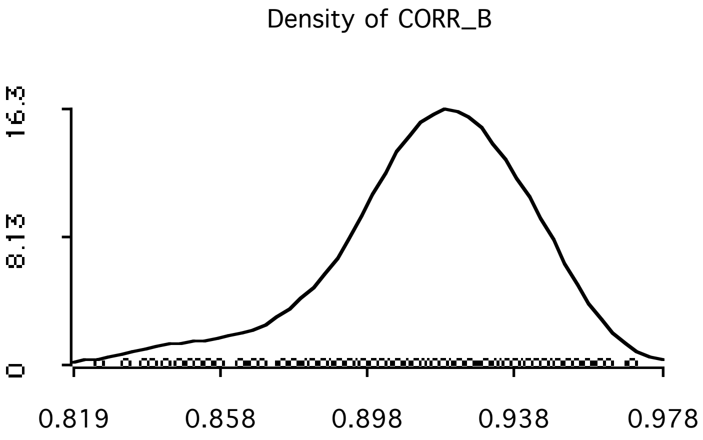
corr(88, 80) = 0.915

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]



Enhancements

Double bootstrap

Use the BS to improve the BS

Review logic of a confidence interval.

Bootstrap the bootstrap
- Similar to idea in Freedman and Peters
- Second layer of BS resampling determines properties of top layer.

Special computing tricks

No longer get histogram/kernel of BS dist.

Balanced resampling
Computational device to get better simulation estimates at the cost of complicating how you can use the BS replications of the statistic.

Importance sampling to learn about extremes.

Things to Take Away

Resampling with longitudinal data

Done to preserve correlation of process.
Requires some assumptions for time series.

Bootstrap for generalized least squares

BS standard error larger than nominal.

Actual SE appears to be larger still.

Resampling in a structural equation

Select observations and fit model to full data set, not one equation at a time.

Many terms in this models are nonlinear combinations of regression coefficients, much like the location of the max for a polynomial.

Percentile intervals

Percentile intervals are easy to obtain.

Enhancements are needed to improve the coverage when the sampling distribution is skewed.

Bootstrap-t designed to be fast and more accurate in *certain* problems, particularly those where you have a standard error formula.

Review Questions

If your data consist of short time series, how should you resample?

Bootstrap resampling should parallel the original data generating process. You should sample the short series! The paper of Freedman and Peters takes this approach.

What feature of generalized least squares does the bootstrap capture, but most procedures ignore?

The BS recognizes the variation in our estimate of the covariance among the observations, and gives estimates that reflect this uncertainty.

Why does the bootstrap fail to correct for dependence without taking special steps?

Sampling with replacement generates a collection of independent observations, regardless of the true structure. For example, residuals in regression are correlated. However, when we sample them as in fixed X resampling, the resulting errors are “conditionally” independent.

What happens when you bootstrap, but the model does not have a constant term?

For residual resampling, the average residual is not forced to be zero and so the average bootstrap error term does not have mean zero, leading to problems.

What important assumptions underlie bootstrap percentile intervals?

These assumptions embody the basic bootstrap analogy: the sampling distribution of the bootstrap statistic has to resemble, up to a transformation, the distribution of the actual statistic.

How do the bias-corrected and accelerated intervals weaken these assumptions? At what cost?

At the cost of more calculation, these allow for bias as well as skewness.

How do BS t-intervals differ from percentile intervals?

BS t-intervals resemble the usual type of interval, with an estimate divided by its standard error.

When is it easy (or hard) to compute the BS t-intervals?

BS t-intervals require a standard error estimate. If you've got one, they work well. If not, you've got a more complex computing problem.

What's the point in iterating a bootstrap procedure? What's a double bootstrap?

The bootstrap is a procedure one can use to estimate a standard error. So, you can use it to check itself. It takes quite a bit more calculation, but is a powerful idea.

How can you use the bootstrap to check for the presence of bias?

Compare the mean of the bootstrap replications (or maybe better, the median) to the orginal statistic . If the two differ by much (relative to the SE of the statistic), then there's evidence of bias.

What feature of GLS does the BS capture that is missed by standard methods?

The formula for the variance of the GLS estimator of the regression slopes assumes that the error covariance matrix is known. That's pretty rare; usually, it's estimated. The usual formula ignores this estimation. The BS does not.

How do structural equation models differ from standard OLS models?

These models have a collection related equations, often joined to form a "causal model".

What is a direct effect (indirect effect) in a structural model?

A direct effect is typically like a regression coefficient. An indirect effect is usually like the sum of products of regression coefs.

What goes wrong if you BS equations separately in structural equation models?

That would be like estimating the different equations using different samples. That's not what is done when you fit these models.

What important assumptions underlie the basic bootstrap percentile intervals?

That the BS estimator and the original estimator have analogous distributions (do not have to be normal, and can have a transformation).

Why do the percentile intervals require so many more bootstrap samples than the SE estimate?*

To accurately estimate the "tail percentiles" requires a very large sample. Variances are easier to estimate.