The Business Case for Bootstrapping

“When you’re stuck with incomplete data, here’s how you make it work!”

2010 IPSA / SCEA Joint Annual Conference

Brett Gelso, PhD
Glenn Grossman, CCE/A, PMP, EVP
Eric Druker, CCE/A

This document is confidential and is intended solely for the use and information of the client to whom it is addressed.

Table Of Contents

- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

A typical client's dilemma involves needing to answer a question using insufficient data

How popular is the tequila?

- Your client owns a college bar and the local fraternity consumes massive amounts of tequila. She recently started selling a new brand of tequila.
- You want to know, on average, how much of the new tequila is consumed so you can provide an estimate for her next purchase order.
- The bar is very busy, and you don't have time to determine how popular the tequila is and you don't have access to the previous purchase records.
- Plus the fraternity and sororities don't want to be interrupted because they are drinking your fantastic tequila

Our client's dilemma applies because she has a high quality sample, but a small sample size

How popular is the tequila?

- Clearly you don't have time to interview everyone at the bar, so you approach random people around the bar:
- The first person consumed 10 shots of tequila
- The second person consumed 5 shots
- The third person consumed 0 shots
- The fourth person consumed 1 shot and
- The fifth person consumed 0 shots
- As such, the average amount of tequila consumed from your "sample" is 3.2 shots. You wonder if the average of everyone in the bar is 3.2 shots?
- In other words, does your "sample" of five people represent everyone in the bar? How can you use your sample information to determine your next purchase order?

Do you sometimes gather high quality client data, but in very limited amounts?

- Have you ever had a small sample and didn't know whether it was representative of the population?
- What if you have several samples from different unknown populations?
- How can you make a scientifically valid comparison of the means?

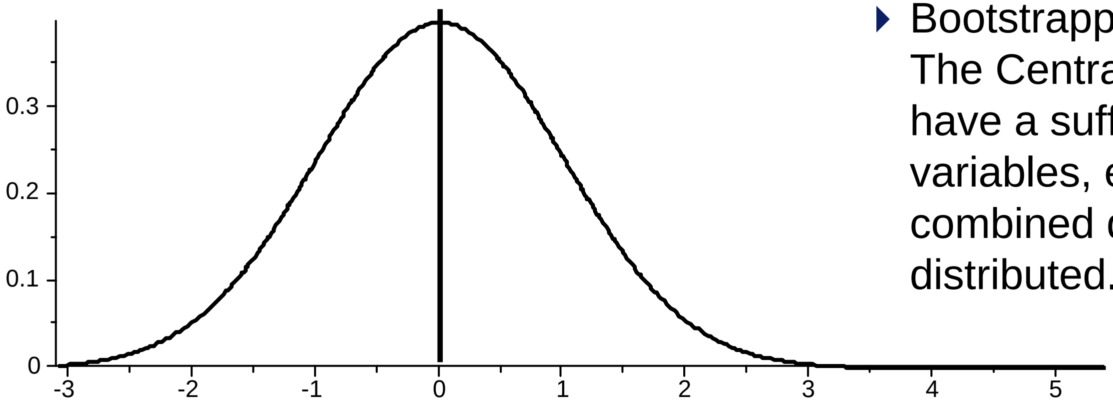
Table Of Contents

- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

Bootstrapping allows analysts to strengthen their conclusions by “resampling” their existing data rather than collecting more

- Called Bootstrapping because if you are stuck in the mud, the only way to get out of the mud is to pull yourself up by your bootstraps. If you are stuck with bad data, you have to figure out something to do with the data you have
- Efron (1982) showed that data that has already been collected can be “resampled” without requiring collecting more data. Data collection is very resource intensive, so avoiding it when possible, is a good goal.
- In short, Bootstrapping enhances our ability to make inferences about existing data. It does this by using a range to capture the population mean with a predefined level of precision
- Usually parametric estimates are relied on for studies conducted in our field. But if the parametric assumptions are in doubt, Bootstrapping is sometimes used as an alternative approach to making inferences.
- The “Tequila” example demonstrates this concept.
- Resampling entails randomly drawing new subsets of samples from the primary sample, each independently drawn, to form a distribution
- The Lottery Ball example demonstrates this concept.

Conceptually, bootstrapping allows us to make new conclusions because it is based on the Central Limit Theorem, a unifying theorem of all inferential statistics



- Also called “Out of Sample” estimation, because this is what your data would look like 9 times out of 10, if you had all the data in the world

- Bootstrapping is based on the Central Limit Theorem. The Central Limit Theorem indicates that when you have a sufficient number of independent, random variables, each with a finite mean and variance, their combined distribution will be approximately normally distributed.

- The Central Limit Theorem states that the limiting value of your sample mean is *always* your population mean

The Central Limit Theorem is robust, allowing the Bootstrapping technique to apply across a diversity of population distributions

- According to Schmuller (2009, page 158):

  "…the Central Limit Theorem says nothing about the population. All it says is that if the sample is large enough (N=30), the sampling distribution of the mean is a normal distribution, with the indicated parameters…", and, "…the population that supplies the samples doesn't have to be a normal distribution for the Central Limit Theorem to hold".

- But what if the population is normal? He notes that

  "…In that case the sampling distribution of the mean is a normal distribution regardless of sample size".

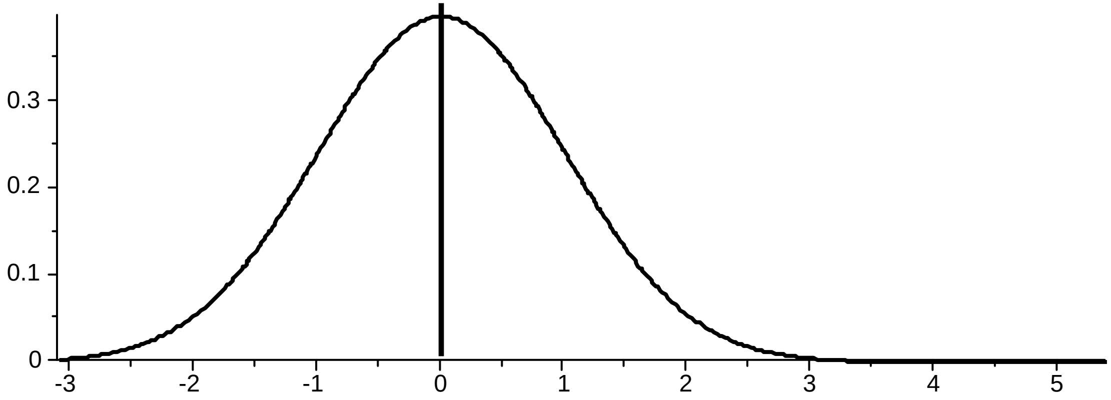


Table Of Contents

- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

The benefits of using Bootstrapping are enormous, adding real value through actionable study plans, increased client confidence, and more compelling research results

Benefits to your Clients:

- **Avoids additional investment and time spent collecting more data if it isn’t necessary**
- **Leverages studies which had failed to deliver actionable results**
- **Increases efficacy of study estimates**
- **Provides our internal & external clients more rigorous and compelling results**
- **Supports management & understanding of risks to the study plan and study variables**

Table Of Contents

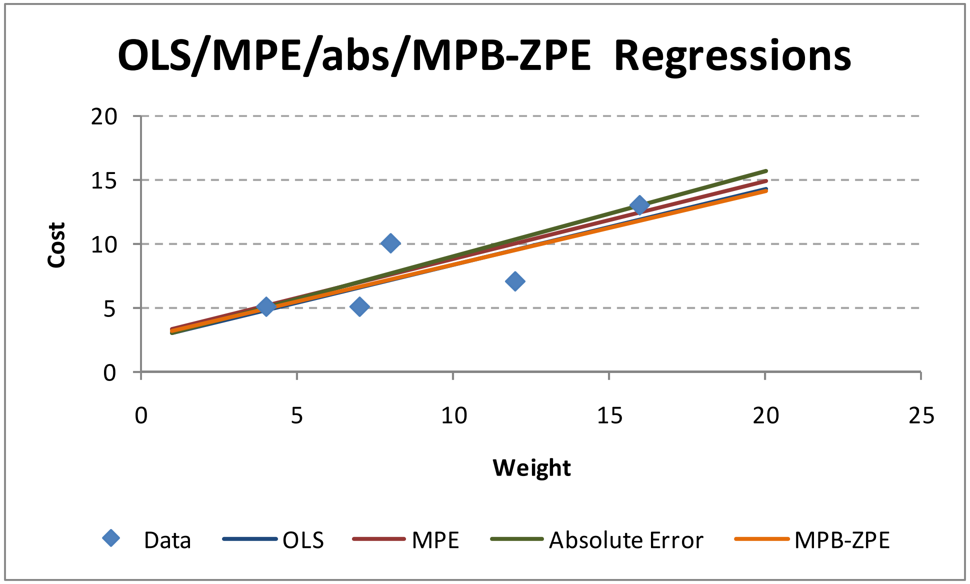
- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

Bootstrapping increases the efficiency of developing Cost Estimating Relationships (CERs) when some assumptions are not met

- Cost estimators typically use Ordinary Least Squares (OLS) regression when developing Cost Estimating Relationships (CERs)
- There are situations, however, where OLS is not an appropriate methodology
  - Independent variables exhibit correlation
  - Residuals are heteroskedastic
- Outside of OLS, there are many regression techniques, some common in cost estimating
  - MPE, MUPE, IRLS, WLS, absolute error, ZMPE
- Techniques differ in that they are minimizing something other than the sum of squared errors
  - Example: MPE = Minimum Percent Error; absolute error: absolute value of errors
- The downside with non-OLS regression methods is that, for the most part, there are no closed form equations for prediction intervals (which are then used in the risk analysis)

Bootstrapping Can Be Used to Develop Prediction Intervals for Non-OLS Regression Techniques

Alternate regression techniques may deliver varying results, without adequately representing the underlying risk distribution



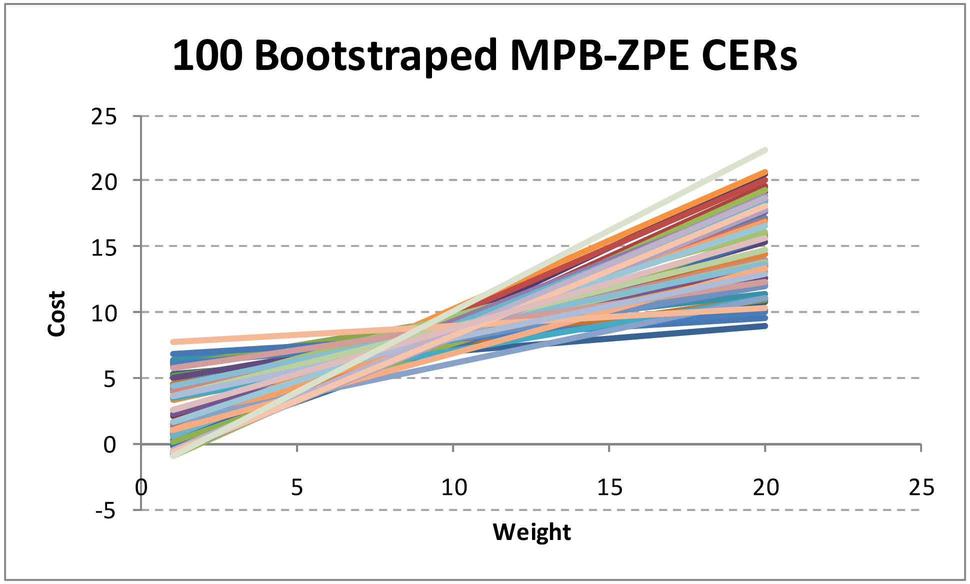
- Above is an example of various regression techniques applied to the sample data set from the Cost Estimating Body of Knowledge's (CEBoK) Regression Module
  - Ordinary Least Squares, Minimum Percent Error, Absolute Error and Minimum Percent Error-Zero Percent Bias

The six step Bootstrapping process enables the analyst to describe the CER risk distribution when some assumptions are not met

6 Step Process

Repeat 5,000 times in Monte Carlo simulation (steps 2–3)

1. Find regression equation based on desired technique (MUPE, abs, etc.) and record residuals
2. Re-sample residuals and apply to data points
3. Re-run regression
4. Sort results
5. Multiply results by the standard error to transform from Confidence Interval to Prediction Interval
6. Convert Prediction interval into risk distribution



Bootstrapping Can Be Used to Develop Prediction Intervals for Non-OLS Regression Techniques

Spreadsheet Example

- Show how to input data
- Show confidence intervals
- Illustrate histogram

Care must be taken with Bootstrapping to ensure that inherent limitations are addressed

- Samples MUST be drawn independently
- Tequila example of non-random sample
  - You only ask people close to you about tequila consumption
- Obtain wide confidence intervals
- Weird name and hard to believe how powerful the technique is

Table Of Contents

- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

Bootstrapping was successfully used to answer questions for the Centers for Medicare and Medicaid Services (CMS), using limited data

- CMS sought to compare success rates for audits with very small samples from different parent organizations, with unknown populations
- Sample sizes were quite small:
  - 99% success rate for Audited Kaiser Permanente claims based on 10 data points and
  - 92% for Blue Cross and Blue Shield with 12 data points.
- The actual number of claims processed was unknown and was probably different across Kaiser and Blue Cross
- Bootstrapping supported more rigorous conclusions, by making an “apples to apples” comparison

Bootstrapping enabled CMS to develop more rigorous and compelling results without the need for additional investment or time spent collecting data

- Bootstrapping enabled CMS to increase the efficacy of their study estimates and develop actionable results, while using only a small amount of existing data
- CMS was able to avoid additional investment in collecting new data and was able to reach sufficient conclusions on a much quicker timeline
- CMS was very satisfied with the results and enthusiastically agreed that bootstrapping was a viable alternative analytic approach in cases such as theirs
- In conclusion, when sample sizes are small and/or CER assumptions cannot be met, bootstrapping analyses may still deliver useful results

Questions?

Table Of Contents

- The Client’s Dilemma: Weak Data
- The Analytic Solution: Bootstrapping
- The Bootstrapping Value Proposition
- How Does Bootstrapping work?
- Successful Client Example of Bootstrapping
- Appendix

Other Types of Resampling

- Jackknife: Resample any statistics and find whether it still hold out of sample, i.e., parameter estimates from regression

- Extremely powerful: If you had all the data in the world, would you regression still give you the same results

Other Types of Resampling

- Cross Validation: Does your data predict itself well?

- Take the difference between your predicted values and actual values in regression to obtain residual and solve for Root Mean Square Error (RMSE); delete the ith predicted value, allow the rest of the data to predict the missing value, and do that for n observations, and solve for Out of Sample RMSE

For more information, please contact:

- **Brett Gelso, PhD**
  Economics & Business Analysis
  703.377.1883 office
  202.412.7580 mobile
  gelso_brett@bah.com

- **Glenn Grossman CCE/A, PMP, EVP**
  Economics & Business Analysis
  301.838.3754 office
  347.744.9288 mobile
  grossman_glenn@bah.com

- **Eric Druker CCE/A**
  Economics & Business Analysis
  314.368.5850 phone
  druker_eric@bah.com