Bootstrapping and Simulating Quantities of Interest

Advanced Quantitative Methods 2019

David Randahl

Department of Peace & Conflict Research
Uppsala University

April 7, 2019

Bootstrapping - What it is

- Bootstrapping is a resampling method

Bootstrapping - What it is

- Bootstrapping is a resampling method
- Creates new datasets of size n by drawing observations from the original data with replacement

Bootstrapping - What it is

- Bootstrapping is a resampling method
- Creates new datasets of size n by drawing observations from the original data with replacement
- Mimics the underlying distribution of the variables

Bootstrapping - What it is

- Bootstrapping is a resampling method
- Creates new datasets of size n by drawing observations from the original data with replacement
- Mimics the underlying distribution of the variables
- Non-parametric method

Bootstrapping - How its done

1. Index all observations in the dataset of size n

Bootstrapping - How its done

1. Index all observations in the dataset of size n
2. Make a random draw of n numbers from 1 to n with replacement. Each observation with a set probability (usually equal, but can be generalized)

Bootstrapping - How its done

1. Index all observations in the dataset of size n
2. Make a random draw of n numbers from 1 to n with replacement. Each observation with a set probability (usually equal, but can be generalized)
3. Take all the observations which have been drawn in the sequence above into a new dataset, note that some observations will not be selected, and some will be selected multiple times

Bootstrapping - How its done

1. Index all observations in the dataset of size n
2. Make a random draw of n numbers from 1 to n with replacement. Each observation with a set probability (usually equal, but can be generalized)
3. Take all the observations which have been drawn in the sequence above into a new dataset, note that some observations will not be selected, and some will be selected multiple times
4. Conduct the analysis on the new, bootstrapped, dataset, and calculate all quantities of interest

Bootstrapping - How its done

1. Index all observations in the dataset of size n
2. Make a random draw of n numbers from 1 to n with replacement. Each observation with a set probability (usually equal, but can be generalized)
3. Take all the observations which have been drawn in the sequence above into a new dataset, note that some observations will not be selected, and some will be selected multiple times
4. Conduct the analysis on the new, bootstrapped, dataset, and calculate all quantities of interest
5. Repeat 2-4 J times

UPPSALA UNIVERSITET

Bootstrapping - How its done

1. Index all observations in the dataset of size n
2. Make a random draw of n numbers from 1 to n with replacement. Each observation with a set probability (usually equal, but can be generalized)
3. Take all the observations which have been drawn in the sequence above into a new dataset, note that some observations will not be selected, and some will be selected multiple times
4. Conduct the analysis on the new, bootstrapped, dataset, and calculate all quantities of interest
5. Repeat 2-4 J times
6. Use the J bootstrapped estimates of your quantities of interest to calculate the estimated values and confidence bounds for your quantities of interest

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal
  - There is heteroskedasticity in the data

UPPSALA UNIVERSITET

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal
  - There is heteroskedasticity in the data
  - Allows for an assumption-free estimation of the covariance matrix between coefficients

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal
  - There is heteroskedasticity in the data
  - Allows for an assumption-free estimation of the covariance matrix between coefficients
- Allows for a 'brute-force' approach to the distribution of quantities of interest

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal
  - There is heteroskedasticity in the data
  - Allows for an assumption-free estimation of the covariance matrix between coefficients
- Allows for a 'brute-force' approach to the distribution of quantities of interest
- Bootstrap aggregation (bagging) is useful in creating better predictions

Bootstrapping - Why its useful

- Bootstrapping is non-parametric and can therefore be used when other assumptions do not hold
  - Sample size is insufficient and the errors are non-normal
  - There is heteroskedasticity in the data
  - Allows for an assumption-free estimation of the covariance matrix between coefficients
- Allows for a 'brute-force' approach to the distribution of quantities of interest
- Bootstrap aggregation (bagging) is useful in creating better predictions
- Can take multiple forms, including asymmetric bootstrapping, boosting, etc (useful when trying to maximize predictive power on certain classes)

Bootstrapping - Assumption

The sample is large enough to be a representation of the underlying distribution of values

Bootstrapping - R-Time

Example in R

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - What it is

- Simulation of QI uses Mote Carlo techniques to make results more easy to understand

Simulating Quantities of Interest (QI) - What it is

- Simulation of QI uses Mote Carlo techniques to make results more easy to understand
- IMPORTANT: Simulating QI does NOT change the results

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - What it is

- Simulation of QI uses Mote Carlo techniques to make results more easy to understand
- IMPORTANT: Simulating QI does NOT change the results
- Useful when dealing with complex analysis such as non-linear models or interaction effects

Simulating Quantities of Interest (QI) - What it is

- Simulation of QI uses Mote Carlo techniques to make results more easy to understand
- IMPORTANT: Simulating QI does NOT change the results
- Useful when dealing with complex analysis such as non-linear models or interaction effects
- Aims to mimic 'repeated sampling' by making draws from the sampling distribution (typically a multivariate normal distribution)

Simulating Quantities of Interest (QI) - What it is

- Simulation of QI uses Mote Carlo techniques to make results more easy to understand
- IMPORTANT: Simulating QI does NOT change the results
- Useful when dealing with complex analysis such as non-linear models or interaction effects
- Aims to mimic 'repeated sampling' by making draws from the sampling distribution (typically a multivariate normal distribution)
- Similar to bootstrapping, but based on parametric assumptions

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix
  - The assumption that the parameters follow a multivariate normal distribution

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix
  - The assumption that the parameters follow a multivariate normal distribution
- Using the full variance/covariance matrix instead of only the standard errors takes the covariance between parameters into account

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix
  - The assumption that the parameters follow a multivariate normal distribution
- Using the full variance/covariance matrix instead of only the standard errors takes the covariance between parameters into account
- Expected vs. predicted value (Carsey & Harden):

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix
  - The assumption that the parameters follow a multivariate normal distribution
- Using the full variance/covariance matrix instead of only the standard errors takes the covariance between parameters into account
- Expected vs. predicted value (Carsey & Harden):
  - Expected value: 1,000 draws from the sampling distribution for the estimates, calculate $\hat{Y}$, present the distribution of $\hat{Y}$

Simulating Quantities of Interest (QI) - Assumptions and properties

- The sampling distribution is defined by
  - The estimated parameters/quantities of interest
  - Their estimated (variance-)covariance matrix
  - The assumption that the parameters follow a multivariate normal distribution
- Using the full variance/covariance matrix instead of only the standard errors takes the covariance between parameters into account
- Expected vs. predicted value (Carsey & Harden):
  - Expected value: 1,000 draws from the sampling distribution for the estimates, calculate $\hat{Y}$, present the distribution of $\hat{Y}$
  - Predicted value: 1,000 draws from the sampling distribution for the estimates, calculate $\hat{Y}$, draw from the residuals of the model, present the distribution of $\hat{Y} + \varepsilon$

Simulating Quantities of Interest (QI) - Some examples of applications

- Simulating QIs allows for a visualization (or numerical representation) of the marginal effects of certain variables in certain models

Simulating Quantities of Interest (QI) - Some examples of applications

- Simulating QIs allows for a visualization (or numerical representation) of the marginal effects of certain variables in certain models
  - Example 1: Calculating the difference in expected value between a 'high' and 'low' value on a certain variable

Simulating Quantities of Interest (QI) - Some examples of applications

- Simulating QIs allows for a visualization (or numerical representation) of the marginal effects of certain variables in certain models
  - Example 1: Calculating the difference in expected value between a 'high' and 'low' value on a certain variable
  - Example 2: Estimating the different effects of a variable dependent on the value of another variable (interaction)

Simulating Quantities of Interest (QI) - Some examples of applications

- Simulating QIs allows for a visualization (or numerical representation) of the marginal effects of certain variables in certain models
  - Example 1: Calculating the difference in expected value between a 'high' and 'low' value on a certain variable
  - Example 2: Estimating the different effects of a variable dependent on the value of another variable (interaction)
  - Exampel 3: Visualizing the marginal effect of a variable in a non-linear model

Simulating Quantities of Interest (QI) - How its done

- Average case solution

Simulating Quantities of Interest (QI) - How its done

- Average case solution
    1. Estimate the parameters and the covariance matrix of the parameters using the full data

Simulating Quantities of Interest (QI) - How its done

- Average case solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest

Simulating Quantities of Interest (QI) - How its done

- Average case solution
    1. Estimate the parameters and the covariance matrix of the parameters using the full data
    2. Define a variable (or variables) of interest
    3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)

Simulating Quantities of Interest (QI) - How its done

- Average case solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Set all other values to their mean, median, or mode depending on variable type

Simulating Quantities of Interest (QI) - How its done

- Average case solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Set all other values to their mean, median, or mode depending on variable type
  5. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix

Simulating Quantities of Interest (QI) - How its done

- Average case solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Set all other values to their mean, median, or mode depending on variable type
  5. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix
  6. For each of the values in the range in 3, calculate the i, expected values produced by the i parameter sets from 5

Simulating Quantities of Interest (QI) - How its done

- Average case solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Set all other values to their mean, median, or mode depending on variable type
  5. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix
  6. For each of the values in the range in 3, calculate the i, expected values produced by the i parameter sets from 5
  7. Take the mean of 6 as your point estimate, and use appropriate quantiles for the confidence bounds

Simulating Quantities of Interest (QI) - How its done

- Observed value solution

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
    1. Estimate the parameters and the covariance matrix of the parameters using the full data
    2. Define a variable (or variables) of interest
    3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
    4. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix
  5. For each of the values in the range in 3, set all observations' value of the variable to that value and calculate the $n \cdot i$, expected values produced by the i parameter sets from 5

UPPSALA UNIVERSITET

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix
  5. For each of the values in the range in 3, set all observations' value of the variable to that value and calculate the $n \cdot i$, expected values produced by the i parameter sets from 5
  6. For each set of random parameters, calculate the mean of the expected value to produce a total of i expected values per value in the range in 3

Simulating Quantities of Interest (QI) - How its done

- Observed value solution
  1. Estimate the parameters and the covariance matrix of the parameters using the full data
  2. Define a variable (or variables) of interest
  3. Set a range of values to investigate the effect over (common choices are 75th and 25th percentiles, min/max, range of count variables, and 0/1 for dummy variables)
  4. Draw an appropriate number, i, of random parameter sets using a multivariate normal distribution with the estimated parameters as the means and the covariance matrix as the covariance matrix
  5. For each of the values in the range in 3, set all observations' value of the variable to that value and calculate the $n \cdot i$, expected values produced by the i parameter sets from 5
  6. For each set of random parameters, calculate the mean of the expected value to produce a total of i expected values per value in the range in 3
  7. Take the mean of 7 as your point estimate, and use appropriate quantiles for the confidence bounds

UPPSALA UNIVERSITET

Translating a table using simulated QI

Table I. Multinomial Logit: Civil War Terminations, 1944–97

	Outcomes			
Variable	*Government*	*Rebel*	*Truce*	*Treaty*
Bureaucracy	–.38	–7.09**	1.36	–.40
	.84	2.81	2.21	.65
Democracy	–.21	.13	–.17	.08
	.18	.27	.34	.13
Army	.26**	.31**	.57***	.29***
	.11	.14	.139	.10
Duration	–.12*	–.03	.06	–.14**
	.06	.11	.09	.06
Duration^2	.0002	–.0003	–.00008	.0004**
	.0002	.0006	.00028	.0002
Exports	53.75*	80.59	427.00***	59.31**
	31.04	35.90	118.65	30.33
Gini	.12	.25	–2.12***	.06
	.20	.22	.66	.18
Borders	1.17**	3.20***	3.86***	.93*
	.53	1.03	1.11	.48
Ethnicity	[illegible]	[illegible]	[illegible]	[illegible]

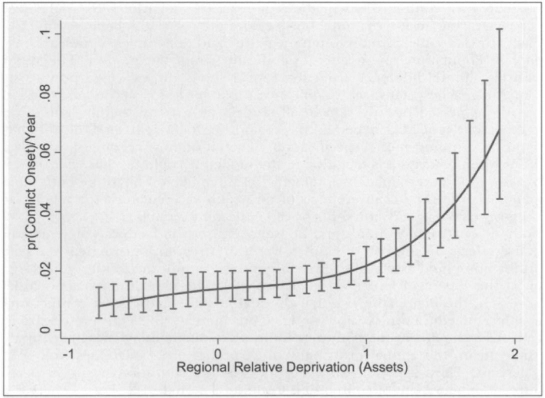
UPPSALA UNIVERSITET

Translating a table using simulated QI

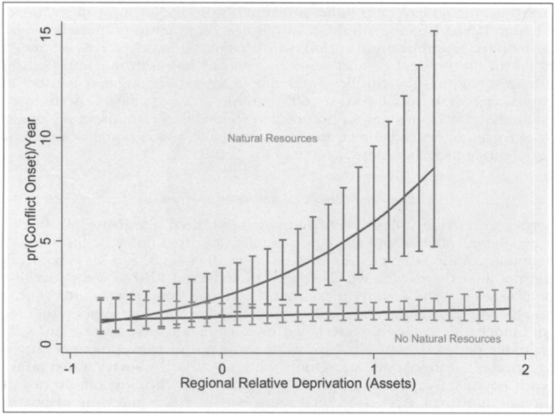
Table II. Probabilities and Expected Durations Based on Reduced Model

Effect	*Ongoing*	*Government*	*Rebel*	*Truce*	*Treaty*	*Duratio*
Baseline	.003(.00,.03)	.64(.30,.89)	.007(.00,.06)	.05(.00,.30)	.31(.07,.65)	65
Bureaucracy 25th	.003(.00,.02)	.69(.27,.92)	.04(0,.31)	.03(.00,.29)	.24(.05,.60)	56
Bureaucracy 75th	.004(.00,.04)	.56(.21,.86)	.003(0,.02)	.08(.00,.36)	.35(.07,.74)	73
Army 25th	.02(.00,.15)	.63(.29,.88)	.007(.00,.04)	.04(.00,.28)	.30(.07,.65)	63
Army 75th	.002(.00,.02)	.64(.29,.89)	.007(.00,.07)	.05(.00,.31)	.30(.07,.64)	65
Dur./dur.^{2} 25th	.002(.00,.01)	.64(.27,.89)	.01(.00,.09)	.04(.00,.38)	.31(.07,.67)	–
Dur./dur.^{2} 75th	.02(.00,.08)	.50(.16,.80)	.01(.00,.06)	.06(.00,43)	.42(.14,.73)	–
Exports 25th	.015(.00,.10)	.68(.30,.92)	.01(.00,.05)	.02(.00,.16)	.28(.05,.66)	71
Exports 75th	.003(.00,.03)	.61(.27,.88)	.007(.00,.06)	.06(.00,.37)	.31(.08,.65)	64
Borders 25th	.008(.00,.06)	.52(.16,.85)	.006(.00,.04)	.05(.00,32)	.41(.08,.82)	77
Borders 75th	.002(.00,.02)	.70(.33,.92)	.01(.00,.10)	.04(.00,.37)	.25(.06,.57)	61
Ethnicity 25th	.003(.00,.02)	.65(.25,.92)	.01(.00,.12)	.10(.00,.45)	.23(.02,.64)	59
Ethnicity 75th	.004(.00,.03)	.60(.26,.86)	.01(.00,.04)	.03(.00,.32)	.36(.11,.68)	69

Visualizing Marginal Effects and Interactions using simulated QI



Visualizing Marginal Effects and Interactions using simulated QI



Bootstrapping - R-Time

Example in R

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis
- Bootstrapping allows for relaxing of assumptions on distributions etc

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis
- Bootstrapping allows for relaxing of assumptions on distributions etc
- QI simulation makes your results clearer and easier to interpret

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis
- Bootstrapping allows for relaxing of assumptions on distributions etc
- QI simulation makes your results clearer and easier to interpret
- Bootstrapping is useful in prediction and can be combined with powerful other tools to maximize predictive power

UPPSALA UNIVERSITET

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis
- Bootstrapping allows for relaxing of assumptions on distributions etc
- QI simulation makes your results clearer and easier to interpret
- Bootstrapping is useful in prediction and can be combined with powerful other tools to maximize predictive power
- Simulating QIs - dependent on parametric assumption

Take aways

- Bootstrapping and simulating QIs are ways of using Monte Carlo simulation in your analysis
- Bootstrapping allows for relaxing of assumptions on distributions etc
- QI simulation makes your results clearer and easier to interpret
- Bootstrapping is useful in prediction and can be combined with powerful other tools to maximize predictive power
- Simulating QIs - dependent on parametric assumption
- Bootstrapping - a non-parametric solution