Stat 454 Final project:

Bootstrap and Jackknife variance estimation for Ratio and Regression estimators

Table of Contents

Abstract

Bootstrap and Jackknife sampling are commonly used methods to estimate properties of an estimator (such as its mean and standard error). In this project, we compare the Bootstrap sampling and Jackknife sampling methods. For each of those methods, we would like to use two estimators, the ratio estimator, and the regression estimator for estimation. As for the sample selection from the simulated population, we would compare the differences in results by using Probability Proportional Sampling (PPS) and Simple Random Sampling Without Replacement (SRSWOR). These will be done with the help of the commonly used statistical computation program R.

Introduction

Bootstrap sampling is a statistical computer-intensive resampling method for computing an estimator from the original sample. It is proposed by Bradley Efron in 1979. Bootstrap sampling makes use of Monte Carlo sampling to generate estimate of the sampling distribution. There are many different types of Bootstrap sampling. For this report we will be using the most basic resampling Bootstrap. The main idea of Bootstrap resampling is to repeatedly select a sub-sample from a specific sample and use the sub-sample to estimate the parameter of interest.

Jackknife sampling is a resampling method which is more systematic and when repeated on the same data, produces exactly the same result each time, thus is a more popular choice in cases where estimates need to be verified several times. The main idea of Jackknife sampling is that it systematically re-computes the estimate while leaving out one observation at a time from the sample set. We then compute the estimate from each of those replicates.

Bootstrap and Jackknife sampling are used to estimate the variability of a statistic through repeated sampling the samples we have. They are both very easy to understand and require no complex algorithm or formulas. Both methods can also be computer intensive when sample size is large.

The main difference is that Bootstrap sampling selects sub-samples from the same sample, whereas Jackknife sampling selects sub-samples from the original sample with one element removed each time.

Below are some important key concepts used throughout the report.

Sampling methods

1) Simple Random Sampling With Replacement (SRSWR)

SRSWR is a subset of randomly selected units from a larger set. Each unit has an equal probability of being selected from the sample. After each unit has been selected, it is replaced in the larger set, as such; it is possible for an element to appear more than once in the selected set.

2) Simple Random Sampling Without Replacement (SRSWOR)

SRSWOR is a subset of randomly selected units from a larger set. Each unit has an equal probability of being selected from the sample. After each unit has been selected, it is removed in the larger set, as such; each element cannot appear more than once in the selected set.

3) Probability Proportional to Size sampling (PPS)

PPS is a method of selecting a subset such that the probability of each element being picked is set to be proportional to its size measure. However, the probability of selection cannot be more than 1.

Estimators

1) Regression Estimator (REG)

$y_i = \beta_0 + \beta_1 x_i + \in_i$

(y_i is the i-th measure of interest)

(x_i is the i-th auxiliary information)

($\in_i$: random shock epsilon of mean 0 and variance σ^2)

(β_0 and β_1 : constants.)

2) Ratio Estimator (RATIO)

$y_i = \frac{\bar{y}}{\bar{x}} x_i$

(y_i is the i-th measure of interest)

(x_i is the i-th auxiliary information)

($\bar{x} = \sum_{i \in s} x_i$)

($\bar{y} = \sum_{i \in s} y_i$)

Simulation

Bootstrap sampling

Assuming our simulated population is of size N, the procedures for bootstrap resampling are as follows:

1. First we use R to simulate an artificial population of size N=1000.

   These values correspond to X = $(x_1, x_2, \dots, x_{1000})$.
2. Next we use the regression model to calculate Y = $(y_1, y_2, \dots, y_{1000})$. (We assume β_0=7 and β_1=14 for this case)
3. A sample size of size n=100 is selected from the population using either PPS or SRSWOR.
4. After that, we create a Bootstrap sample by selecting n units from the sample (of size n) by SRSWR.
5. We repeat step 4 and create 1000 Bootstrap samples.
6. For each Bootstrap sample, we can estimate the corresponding value of μ_y using either the REG model or RATIO model.
7. We can then calculate the variance and construct the confidence intervals using the set of μ_y.

Jackknife sampling

The main idea of Jackknife sampling is that it systematically re-computes the estimate while leaving out one observation at a time from the sample set. We then compute the estimate from each of those replicates.

Assuming our simulated population is of size N, the steps are as follow:

1. We use the same population from step 1 in Bootstrap resampling.

   X = $(x_1, x_2, \dots, x_{1000})$.
2. Similar to Bootstrap resampling, we use the regression model to calculate Y = $(y_1, y_2, \dots, y_{1000})$. (Assume β_0=7 and β_1=14).
3. We then select a sample size of size n=100 from the population using either PPS or SRSWR.
4. We then remove the i_{th} element from the sample and the remaining 99 units will be the i_{th} Jackknife sample.
5. Repeat step 4 for i = 1,2, … ,n. We now have n=100 Jackknife samples.
6. For each Jackknife sample, we can use the desired model to obtain μ_y.
7. We can then find the variance and construct the confidence intervals using the set of μ_y.

(Note: Since the variance Jackknife variance estimator is $V_J = \frac{n-1}{n} * \sum_{i=1}^{n} (\hat{\mu}_{.i} - \bar{\mu})^2$, we need to multiply the calculated variance in R by $\frac{\mathbf{(nsim-1)^2}}{\mathbf{nsim}}$ **(** $= \frac{nsim-1}{nsim} * nsim$)

Results

The results of two trials of simulations are compiled into the table shown below. Results from R are rounded off to two decimal places for readability.

The actual mean for this first and second simulation is **27.22** and **26.93** , which is obtained from the mean of the Y parameter of the population in step 2 of each procedure mentioned above.

Trial 1) Bootstrap sampling

	Variance	Confidence Interval
SRSWOR sample/ REG	0.0100	(27.07, 27.46)
SRSWOR sample / RATIO	0.1164	(26.56, 27.90)
PPS sample / REG	0.0108	(26.86, 27.27)
PPS sample / RATIO	0.1155	(25.93, 27.27)

Trial 1) Jackknife sampling

	Variance	Confidence Interval
SRSWOR sample/ REG	0.1058	(26.63, 27.91)
SRSWOR sample / RATIO	1.2852	(24.99, 29.43)
PPS sample / REG	0.1103	(26.42, 27.72)
PPS sample / RATIO	1.1117	(24.52, 29.27)

Trial 2) Bootstrap sampling

	Variance	Confidence Interval
SRSWOR sample/ REG	0.0076	(26.95, 27.29)
SRSWOR sample / RATIO	0.1347	(26.58, 28.02)
PPS sample / REG	0.0131	(26.68, 27.12)
PPS sample / RATIO	0.0775	(26.13, 27.22)

Trial 2) Jackknife sampling

	Variance	Confidence Interval
SRSWOR sample/ REG	0.0832	(26.55, 27.68)
SRSWOR sample / RATIO	1.3653	(24.99, 29.57)
PPS sample / REG	0.1314	(26.19, 27.61)
PPS sample / RATIO	0.8259	(23.88, 29.06)

Conclusion

By comparing the two sampling methods, we can see that Bootstrap resampling has a lower variance/standard-error compared to Jackknife estimators, thus resulting in a smaller 95% confidence interval. This shows that Bootstrap is a better choice for this model. Using Regression estimator to find an estimate also seems to be significantly more efficient compared to Ratio estimator in this case, which we can see from the difference in variance comparing both. It is also observed that sample selection using SRSWOR tends to result in a slightly lower variance compared to selection using PPS sampling.

From our simulation, there is also a case where the confidence interval does not contain the actual mean ($\bar{Y}$). It occurred on the Bootstrap sampling on the second trial when we used SRSWOR to select the sample and Regression estimator. The 95% confidence interval (26.95, 27.29) does not contain the true mean of 26.93. This is not at all unexpected because there is still a small chance, although small, that the 95% CI will not capture the true mean.

References

Links from Wikipedia used for the project:

- http://en.wikipedia.org/wiki/Bootstrapping_%28statistics%29
- http://en.wikipedia.org/wiki/Resampling_%28statistics%29