STA6171: Statistical Computing for DS 1

Bootstrap

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

2020.11.04

1. The Bootstrap Principal
2. Basic Methods
3. Permutation Tests

Introduction

- Bootstrapping is a computational intensive method that allows researchers to simulate the distribution of a statistic.
- The idea is to repeatedly resample the observed data, each time producing an empirical distribution function from the resampled data.
- For each resampled data set—or equivalently each empirical distribution function—a new value of the statistic can be computed, and the collection of these values provides an estimate of the sampling distribution of the statistic of interest.
- In this manner, the method allows you to “pull yourself up by your bootstraps” (an old idiom, popularized in America, that means to improve your situation without outside help).
- Bootstrapping is nonparametric by nature, and there is a certain appeal to letting the data speak so freely.

The Bootstrap Principal

- Let $\theta = T(F)$ be an interesting feature of a distribution function, F, expressed as a functional of F.
- For example, $T(F) = \int z dF(z)$ is the mean of the distribution.
  - Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be data observed as a realization of the random variables $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim F$.
  - We use $\mathbf{X} \sim F$ to denote that $\mathbf{X}$ is distributed with density function f having corresponding cumulative distribution function F.
  - Let $\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ denote the entire dataset.

The Bootstrap Principal

- If F is the empirical distribution function of the observed data, then an estimate of θ is $\hat{\theta} = T(\hat{F})$. For example, when θ is a univariate population mean, the estimator is the sample mean,

$$\hat{\theta} = \int z d\hat{F}(z) = \sum_{i=1}^{n} X_i / n$$

- Statistical inference questions are usually posed in terms of $T(\hat{F})$ or some $R(\mathcal{X}, F)$, a statistical function of the data and their unknown distribution function F. For example, a general test statistic might be

$$R(\mathcal{X}, F) = \frac{T(\hat{F}) - T(F)}{S(\hat{F})},$$

where S is a functional that estimates the standard deviation of $T(\hat{F})$.

The Bootstrap Principal

- The distribution of the random variable $R(\mathcal{X}, F)$ may be intractable or altogether unknown. This distribution also may depend on the unknown distribution F.
- The bootstrap provides an approximation to the distribution of $R(\mathcal{X}, F)$ derived from the empirical distribution function of the observed data.
- Let $\mathcal{X}^*$ denote a bootstrap sample of *pseudo-data*, which we will call a *pseudo-dataset*. The elements of are iid random variables with distribution $\hat{F}$.
- The bootstrap strategy is to examine the distribution of $R(\mathcal{X}^*, \hat{F})$, that is, the random variable formed by applying R to $\mathcal{X}^*$.

Simple Illustration

- Suppose $n = 3$ univariate data points, namely $\{x_1, x_2, x_3\} = \{1, 2, 6\}$, are observed as an i.i.d. sample from a distribution F that has mean θ.
- At each observed data value, $\hat{F}$ places mass $\frac{1}{3}$.
- Suppose the estimator to be bootstrapped is the sample mean $\hat{\theta}$, which we may write as $T(\hat{F})$ or $R(\mathcal{X}, F)$, where R does not depend on F in this case.

Simple Illustration

- Let $\mathcal{X}^* = \{X_1^*, X_2^*, X_3^*\}$ consist of elements drawn iid from $\hat{F}$.
- There are $3^3 = 27$ possible outcomes for $\mathcal{X}^*$.
- Let $\hat{F}^*$ denote the empirical distribution function of such a sample, with corresponding estimate $\hat{\theta}^* = T(\hat{F}^*)$.
- Since $\hat{\theta}^*$ does not depend on the ordering of the data, it has only 10 distinct possible outcomes.

Simple Illustration

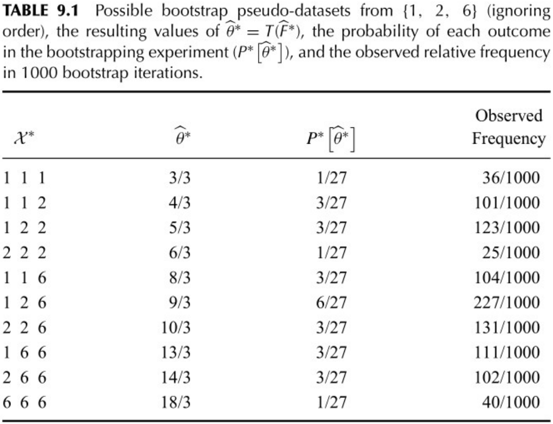

TABLE 9.1 Possible bootstrap pseudo-datasets from {1, 2, 6} (ignoring order), the resulting values of $\widehat{\theta}^* = T(\widehat{F}^*)$, the probability of each outcome in the bootstrapping experiment ($P^*\left[\widehat{\theta}^*\right]$), and the observed relative frequency in 1000 bootstrap iterations.

$\mathcal{X}^*$	$\widehat{\theta}^*$	$P^*\left[\widehat{\theta}^*\right]$	Observed Frequency
1 1 1	3/3	1/27	36/1000
1 1 2	4/3	3/27	101/1000
1 2 2	5/3	3/27	123/1000
2 2 2	6/3	1/27	25/1000
1 1 6	8/3	3/27	104/1000
1 2 6	9/3	6/27	227/1000
2 2 6	10/3	3/27	131/1000
1 6 6	13/3	3/27	111/1000
2 6 6	14/3	3/27	102/1000
6 6 6	18/3	1/27	40/1000

Nonparametric Bootstrap

- For realistic sample sizes the number of potential bootstrap pseudo-datasets is very large, so complete enumeration of the possibilities is not practical.
- Instead, B independent random bootstrap pseudo-datasets are drawn from the empirical distribution function of the observed data, namely F.
- Denote these $\mathcal{X}_i^* = \{\mathbf{X}_{i1}^*, \cdots, \mathbf{X}_{in}^*\}$ for $i = 1, \cdots, B$.
- The empirical distribution of the $R(\mathcal{X}_i^*, \hat{F})$ for $i = 1, \cdots, B$ is used to approximate the distribution of $R(\mathcal{X}, F)$, allowing inference.

Nonparametric Bootstrap

- The simulation error introduced by avoiding complete enumeration of all possible pseudo-datasets can be made arbitrarily small by increasing B.
- Using the bootstrap frees the analyst from making parametric assumptions to carry out inference, provides answers to problems for which analytic solutions are impossible, and can yield more accurate answers than given by routine application of standard parametric theory.
- A fundamental requirement of bootstrapping is that the data to be resampled must have originated as an iid sample. If the sample is not iid, the distributional approximation of $R(\mathcal{X}, F)$ by $R(\mathcal{X}^*, \hat{F})$ will not hold.

Parametric Bootstrap

- The ordinary nonparametric bootstrap described above generates each pseudo-dataset $\mathcal{X}^*$ by drawing $\mathbf{X}_1^*, \cdots, \mathbf{X}_n^*$ iid from $\hat{F}$.
- When the data are modeled to originate from a parametric distribution, so $\mathbf{X}_1, \cdots, \mathbf{X}_n \sim$ iid $F(\mathbf{x}, \boldsymbol{\theta})$, another estimate of F may be employed.
- Suppose that the observed data are used to estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$.
- Then each parametric bootstrap pseudo-dataset $\mathcal{X}^*$ can be generated by drawing $\mathbf{X}_1^*, \cdots, \mathbf{X}_n^*$. When the model is known or believed to be a good representation of reality, the parametric bootstrap can be a powerful tool, allowing inference in otherwise intractable situations and producing confidence intervals that are much more accurate than those produced by standard asymptotic theory.

Bootstrapping Regression

- Consider the ordinary multiple regression model, $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ where the $i = 1, \cdots, n$ are assumed to be i.i.d. mean zero random variables with constant variance.
- $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are p-vectors of predictors and parameters, respectively.
- A naive bootstrapping mistake would be to resample from the collection of response values a new pseudo-response, say Y_i^*, for each observed $\mathbf{x}_i$, thereby generating a new regression dataset.

Bootstrapping Regression

- Then a bootstrap parameter vector estimate, $\hat{\beta}^*$, would be calculated from these pseudo-data.
- After repeating the sampling and estimation steps many times, the empirical distribution of $\hat{\beta}^*$ would be used for inference about β.
- The mistake is that the $Y_i \mid \mathbf{x}_i$ are not iid - they have different conditional means. Therefore, it is not appropriate to generate bootstrap regression datasets in the manner described.

Bootstrapping Regression

We must ask what variables are iid in order to determine a correct bootstrapping approach. The ϵ_i are iid given the model. Thus a more appropriate strategy would be to bootstrap the residuals as follows.

- Start by fitting the regression model to the observed data and obtaining the fitted responses $\hat{y}_i$ and residuals $\hat{\epsilon}_i$.
- Sample a bootstrap set of residuals, $\{\hat{\epsilon}_1^*, \cdots, \hat{\epsilon}_n^*\}$, from the set of fitted residuals, completely at random with replacement.
- Create a bootstrap set of pseudo-responses, $Y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$, for $i = 1, \cdots, n$
- Regress Y^* on $\mathbf{x}$ to obtain a bootstrap parameter estimate $\hat{\boldsymbol{\beta}}^*$.
- Repeat this process many times to build an empirical distribution for $\hat{\boldsymbol{\beta}}^*$ that can be used for inference.

Bootstrapping Regression

- This approach is most appropriate for designed experiments or other data where the $\mathbf{x}_i$ values are fixed in advance.
- The strategy of bootstrapping residuals is at the core of simple bootstrapping methods for other models such as autoregressive models, nonparametric regression, and generalized linear models.
- Bootstrapping the residuals is reliant on the chosen model providing an appropriate fit to the observed data, and on the assumption that the residuals have constant variance.
- Without confidence that these conditions hold, a different bootstrapping method is probably more appropriate.

Bootstrapping Regression

- Suppose that the data arose from an observational study, where both response and predictors are measured from a collection of individuals selected at random.
- In this case, the data pairs $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ can be viewed as values observed for iid random variables $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ drawn from a joint response-predictor distribution.
- To bootstrap, sample $\mathbf{Z}_1^*, \cdots, \mathbf{Z}_n^*$ completely at random with replacement from the set of observed data pairs, $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$.
- Apply the regression model to the resulting pseudodataset to obtain a bootstrap parameter estimate $\hat{\beta}^*$.
- Repeat these steps many times, then proceed to inference as in the first approach. This approach of bootstrapping the cases is sometimes called the paired bootstrap.

Bootstrapping Regression

- If you have doubts about the adequacy of the regression model, the constancy of the residual variance, or other regression assumptions, the paired bootstrap will be less sensitive to violations in the assumptions than will bootstrapping the residuals.
- The paired bootstrap sampling more directly mirrors the original data generation mechanism in cases where the predictors are not considered fixed.

Bootstrapping Regression Example

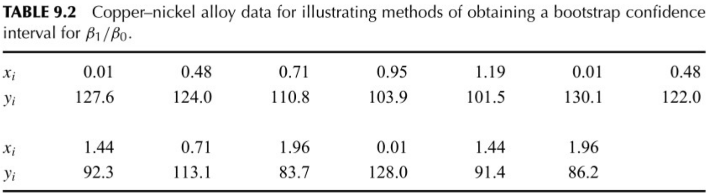

TABLE 9.2 Copper–nickel alloy data for illustrating methods of obtaining a bootstrap confidence interval for β_1/β_0.

x_i	0.01	0.48	0.71	0.95	1.19	0.01	0.48	1.44	0.71	1.96	0.01	1.44	1.96
y_i	127.6	124.0	110.8	103.9	101.5	130.1	122.0	92.3	113.1	83.7	128.0	91.4	86.2

Bootstrapping Regression Example

- 13 measurements of corrosion loss (y_i) in copper-nickel alloys, each with a specific iron content (x_i).
- Of interest is the change in corrosion loss in the alloys as the iron content increases, relative to the corrosion loss when there is no iron. Thus, consider the estimation of $\theta = \beta_1 / \beta_0$ in a simple linear regression

Bootstrapping Regression Example

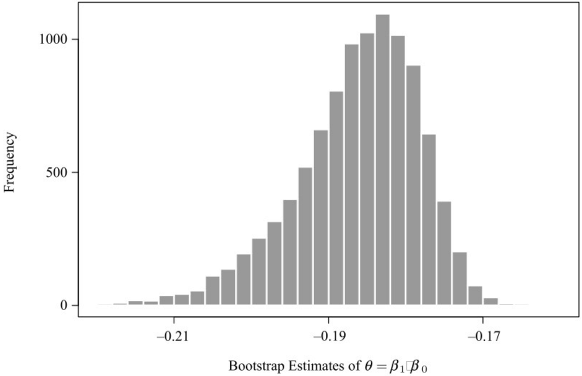

FIGURE 9.1 Histogram of 10,000 bootstrap estimates of β_1/β_0 from the nonparametric paired bootstrap analysis with the copper–nickel alloy data.

Bootstrap Bias Correction

- A particular interesting choice for bootstrap analysis when $T(F) = \theta$ is the quantity $R(\mathcal{X}, F) = T(\hat{F}) - T(F)$. This represents the bias of $T(\hat{F}) = \hat{\theta}$, and it has mean equal to $E\{\hat{\theta}\} - \theta$. The bootstrap estimate of the bias is
$$\sum_{i=1}^{B} \frac{\left(\hat{\theta}_i^* - \hat{\theta}\right)}{B} = \bar{\theta}^* - \hat{\theta}.$$
- An improved bias estimate requires only a little additional effort.
- Let $\hat{F}_j^*$ denote the empirical distribution of the j-th bootstrap pseudo-dataset, and define
$$\bar{F}^*(\mathbf{x}) = \sum_{j=1}^{B} \hat{F}_j^*(\mathbf{x})/B.$$

Then, $\bar{\theta}^* - T(\bar{F}^*$ is a better estimate of bias.

Permutation Tests Examples

Consider a medical experiment where rats are randomly assigned to treatment and control groups.

- The outcome X_i is then measured for the i-th rat.
- Under the null hypothesis, the outcome does not depend on whether a rat was labeled as treatment or control.
- Under the alternative hypothesis, outcomes tend to be larger for rats labeled as treatment.
- A test statistic T measures the difference in outcomes observed for the two groups. For example, T might be the difference between group mean outcomes, having value t_1 for the observed dataset.

Permutation Tests Examples

- Under the null hypothesis, the individual labels “treatment” and “control” are meaningless because they have no influence on the outcome.
- Since they are meaningless, the labels could be randomly shuffled among rats without changing the joint null distribution of the data.
- Shuffling the labels creates a new dataset: Although one instance of each original outcome is still seen, the outcomes appear to have arisen from a different assignment of treatment and control.
- Each of these permuted datasets is as likely to have been observed as the actual dataset, since the experiment relied on random assignment.

Permutation Tests Examples

- Let t_2 be the value of the test statistic computed from the dataset with this first permutation of labels.
- Suppose all M possible permutations (or a large number of randomly chosen permutations) of the labels are examined, thereby obtaining t_2, $\cdots$, t_M.
- Under the null hypothesis, t_2, $\cdots$, t_M were generated from the same distribution that yielded t_1.
- Therefore, t_1 can be compared to the empirical quantiles of t_1, $\cdots$, t_M to test a hypothesis or construct confidence limits.

Permutation Tests

- To pose this strategy more formally, suppose that we observe a value t for a test statistic T having density f under the null hypothesis.
- Suppose large values of T indicate that the null hypothesis is false.
- Monte Carlo hypothesis testing proceeds by generating a random sample of $M - 1$ values of T drawn from f.
- If the observed value t is the k-th largest among all M values, then the null hypothesis is rejected at a significance level of k/M.
- If the distribution of the test statistic is highly discrete, then ties found when ranking t can be dealt with naturally by reporting a range of p-values.

Permutation Tests

- There are a variety of approaches for sampling from the null distribution of the test statistic.
- The permutation approach works because “treatment” and “control” are meaningless labels assigned completely at random and independent of outcome, under the null hypothesis.
- This simple permutation approach can be broadened for application to a variety of more complicated situations.
- In all cases, the permutation test relies heavily on the condition of exchangeability.
- The data are exchangeable if the probability of any particular joint outcome is the same regardless of the order in which the observations are considered.

Permutation Tests

There are two advantages to the permutation test over the bootstrap.

- If the basis for permuting the data is random assignment, then the resulting p-value is exact (if all possible permutations are considered).
    - For such experiments, the approach is usually called a randomization test.
    - In contrast, standard parametric approaches and the bootstrap are founded on asymptotic theory that is relevant for large sample sizes.

Permutation Tests

There are two advantages to the permutation test over the bootstrap.

- Permutation tests are often more powerful than their bootstrap counterparts.
    - However, the permutation test is a specialized tool for making a comparison between distributions, whereas a bootstrap tests hypotheses about parameters, thereby requiring less stringent assumptions and providing greater flexibility.
    - The bootstrap can also provide a reliable confidence interval and standard error, beyond the mere p-value given by the permutation test.
    - The standard deviation observed in the permutation distribution is not a reliable standard error estimate.