Bootstrap

俞 声

清华大学统计学研究中心

Bootstrap

- In simple cases, we can derive the distribution of a statistic and its characteristics (such as the standard error) mathematically. For example, the mean of normal samples is also normal. But in most real cases the distribution is too complicated to obtain.

- Bootstrap methods are a class of *nonparametric* Monte Carlo methods that estimate the distributional characteristics of a statistic by *resampling*.

Parametric bootstrap

- There is in fact a parametric version of bootstrap:

If we are using a model, our best guess at P_{X,θ_0} is $P_{X,\hat{\theta}}$, with our best estimate $\hat{\theta}$ of the parameters

THE PARAMETRIC BOOTSTRAP

1. Get data X, estimate $\hat{\theta}$ from X
2. Repeat b times:
   1. Simulate $\tilde{X}$ from $P_{X,\hat{\theta}}$ (simulate data of same size/"shape" as real data)
   2. Calculate $\tilde{T} = \tau(\tilde{X})$ (treat simulated data the same as real data)
3. Use empirical distribution of $\tilde{T}$ as P_{T,θ_0}

Parametric bootstrap

- However, knowing the distribution $P_{X,\theta}$ is a strong assumption in itself and is usually not applicable in real cases.

- Therefore, the most common form of bootstrap is nonparametric bootstrap.

Bootstrap resampling

- Suppose that $x = (x_1, \ldots, x_n)$ is an observed random sample from a distribution with cdf $F(x)$. If X^* is selected at random from x, then

$$P(X^* = x_i) = \frac{1}{n}, i = 1, \ldots, n.$$

- Resampling generates a random sample $X_1^*, \ldots, X_n^*$ by *sampling with replacement* from x. The random variables X_i^* are iid, uniformly distributed on the set $\{x_1, \ldots, x_n\}$.

The idea behind bootstrap

- The empirical distribution function (ecdf) $F_n(x)$ is an estimator of $F(x)$.
- Moreover, $F_n(x)$ is itself the distribution function of a random variable; namely the random variable that is uniformly distributed on the set $\{x_1, \ldots, x_n\}$. Hence the empirical cdf F_n is the cdf of X^*.
- When $F_n(x)$ approximates $F(x)$ closely (typically when n is large), it is reasonable to assume that X^* has the same distribution with X, and that a statistic $T(X)$ has the same distribution with $T(X^*)$.

Nonparametric bootstrap

Suppose θ is the parameter of interest (θ could be a vector), and $\hat{\theta}$ is an estimator of θ. Then the bootstrap estimate of the distribution of $\hat{\theta}$ is obtained as follows.

1. For each bootstrap replicate, indexed $b = 1, \ldots, B$:
   (a) Generate sample $x^{*(b)} = x_1^*, \ldots, x_n^*$ by sampling with replacement from the observed sample $x_1, \ldots, x_n$.
   (b) Compute the b^{th} replicate $\hat{\theta}^{(b)}$ from the b^{th} bootstrap sample.
2. The bootstrap estimate of $F_{\hat{\theta}}(\cdot)$ is the empirical distribution of the replicates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

Uses of the bootstrap samples

- In the following, we will use the bootstrap samples to estimate:
  - Standard error of an estimator
  - Bias of an estimator
  - Confidence interval of the target parameter

Bootstrap estimation of standard error

The bootstrap estimate of standard error of an estimator $\hat{\theta}$ is the sample standard deviation of the bootstrap replicates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$.

$$\widehat{se}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \overline{\hat{\theta}^*})^2}, \tag{7.1}$$

where $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$ [84, (6.6)].

According to Efron and Tibshirani [84, p. 52], the number of replicates needed for good estimates of standard error is not large; $B = 50$ is usually large enough, and rarely is $B > 200$ necessary. (Much larger B will be needed for confidence interval estimation.)

Exercise

The law school data set `law` in the `bootstrap` package contains average LSAT and average GPA for 15 law schools. This data set is a random sample from the universe of 82 law schools in `law82`.

1. Estimate the correlation between LSAT and GPA scores, and compute the bootstrap estimate of the standard error of the sample correlation.

2. Use the `boot` function from package `boot`.

The `boot` function

- The boot function uses 3 arguments: the first argument is the data, the second is a statistic of interest, and the third is the number of repetitions.

```
boot(data, statisitc = fun, B)
```

- The statistic is a function with 2 arguments: the first one represents the data, and the second one represents the indices from resampling.

```
fun = function(dat, ids) {
  ...
  return(...)
}
```

Bootstrap estimation of bias

- The bias of the estimator $\hat{\theta}$ for θ is
$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

- If $bias(\hat{\theta}) = 0$, $\hat{\theta}$ is said to be unbiased. An example of a biased estimator is the MLE estimator of variance.

- The bootstrap estimation of bias is
$$\widehat{bias}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)} - \hat{\theta}.$$

Exercise

- Compute the bootstrap estimation of bias for the law school sample correlation problem.

- The `patch (bootstrap)` data contains measurements of a certain hormone in the bloodstream of eight subjects after wearing a medical patch. The parameter of interest is

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}.$$

- If $|\theta| \leq 0.2$, this indicates bioequivalence of the old and new patches. The statistic is $\bar{Y}/\bar{Z}$. Compute a bootstrap estimate of bias in the bioequivalence ratio statistic.

The normal bootstrap confidence interval

- You can use this method if you believe your estimator $\hat{\theta}$ is unbiased and approximately normal

$$\hat{\theta} \sim N\left(\theta, se\left(\hat{\theta}\right)\right)$$

E.g.

  - Sample mean (by the Central Limit Theorem)
  - $\hat{\beta}$ from linear regression with iid normal noise.

- The approximate $100(1-\alpha)\%$ normal bootstrap confidence interval is

$$\hat{\theta} \pm z_{1-\alpha/2} se\left(\hat{\theta}\right).$$

The basic bootstrap confidence interval

- Suppose that $\hat{\theta}$ is an estimator of θ and a_α is the α quantile of $\hat{\theta} - \theta$. Then

$$P(\hat{\theta} - \theta > a_\alpha) = 1 - \alpha \Rightarrow P(\theta < \hat{\theta} - a_\alpha) = 1 - \alpha$$

Thus, a $100(1-\alpha)\%$ confidence interval is given by $(\hat{\theta} - a_{1-\alpha/2}, \hat{\theta} - a_{\alpha/2})$.

- Since we don't know the distribution of $\hat{\theta} - \theta$, we replace θ with $\hat{\theta}$, and denote the α quantile of $\hat{\theta}^*$ with $\hat{\theta}^*_\alpha$. Then a_α is replaced by $\hat{\theta}^*_\alpha - \hat{\theta}$.

- The bootstrap confidence interval is therefore

$$(2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2}).$$

The percentile bootstrap confidence interval

- The percentile bootstrap confidence interval is simply
$$\left(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}\right).$$

- You can choose to adjust all the introduced confidence intervals by $\widehat{bias}(\hat{\theta})$.

Exercise

- Compute the confidence intervals for the `patch` ratio statistic.
- Compute the confidence intervals using the `boot.ci` function.

BCa confidence intervals

- The usual percentile confidence interval corrected by bias and adjusted for acceleration (skewness) is called the BCa confidence interval.

- The $100(1-\alpha)\%$ BCa interval is $\left(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2}\right)$, where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right),$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right),$$

and $z_\alpha = \Phi^{-1}(\alpha)$. See next pages for the definitions of $\hat{z}_0$ and $\hat{a}$.

BCa confidence intervals

- $\hat{z}_0$ measures the median bias:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}^{(b)} < \hat{\theta})\right),$$

where $I(\cdot)$ is the indicator function.

- $\hat{a}$ is the acceleration factor (skewness) estimated from *jackknife* replicates:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\overline{\theta_{(\cdot)}} - \theta_{(i)}\right)^3}{6\left(\sum_{i=1}^{n}\left(\overline{\theta_{(\cdot)}} - \theta_{(i)}\right)^2\right)^{3/2}}$$

Jackknife replicates

- Jackknife is another resampling technique.
- $\theta_{(i)}$ is the "leave-one-out" estimation of θ without using the i-th sample x_i.
- $\overline{\theta_{(\cdot)}} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}$ is the mean of $\theta_{(i)}$.

BCa confidence intervals

The BCa confidence intervals are transformation respecting and BCa intervals have second order accuracy,

- *Transformation respecting* means that if $(\hat{\theta}^*_{\alpha_1}, \hat{\theta}^*_{\alpha_2})$ is a confidence interval for θ, and $t(\theta)$ is a transformation of the parameter θ, then the corresponding interval for $t(\theta)$ is $(t(\hat{\theta}^*_{\alpha_1}), t(\hat{\theta}^*_{\alpha_2}))$.

- A confidence interval is *first order accurate* if the error tends to zero at rate $\sqrt{n}$ for sample size n, and *second order accurate* if the error tends to zero at rate $1/n$.

Exercise

library(MASS); data(cats)

Exercise

Use linear regression model to predict cat heart weights with cat body weights. Compute the confidence intervals of the regression coefficients with bootstrap resampling.