Bootstrap

Max Turgeon

STAT 3150–Statistical Computing

Lecture Objectives

- Use bootstrap to estimate the bias and variance of an estimator.
- Understand how the empirical CDF is related to resampling techniques.

Motivation

- As with jackknife, the main motivation is to study the sampling distribution of an estimator.
- Jackknife can be used to estimate bias and standard error.
    - But it doesn't always work (e.g. sample median)
- Bootstrap is another resampling method that takes a more direct approach to estimating the sampling distribution.

Bootstrap estimate of the standard error i

- Let $X_1, \dots, X_n$ be a random sample from a distribution F.
- Suppose we use this sample to compute an estimate $\hat{\theta}$ of a population parameter θ.
- Imagine a situation where we can generate B additional samples of size n from the same distribution F.
- For each sample, we could compute an estimate $\hat{\theta}^{(b)}$, where $b = 1, \dots, B$.
- We could then estimate the *standard error* of $\hat{\theta}$ by taking the *sample standard deviation* of the additional estimates $\hat{\theta}^{(b)}$.
- Of course, we can't really generate these additional samples...

Bootstrap estimate of the standard error ii

- **Bootstrap** mimics this situation by sampling **with replacement** from the original sample $X_1, \ldots, X_n$.
    - Generate a sample $X_1^{(b)}, \ldots, X_n^{(b)}$ of size n by sampling with replacement from the original sample.
    - Compute $\hat{\theta}^{(b)}$ using that bootstrap sample.

Example i

- Let's look at the sample median with both jackknife and bootstrap

```r
# "Population" is all integers between 1 and 100
population <- seq(1, 100)
median(population)
```

```
## [1] 50.5
```

Example ii

```r
# Generate B samples from sampling distribution
B <- 5000
n <- 10
results <- replicate(B, {
    some_sample <- sample(population,
                          size = n)
    median(some_sample)
})
sd(results) # Unbiased estimate
```

```
## [1] 13.04957
```

Example iii

```r
# Take a single sample from population
one_sample <- sample(population, size = n)
median(one_sample)
```

```
## [1] 28.5
```

Example iv

```r
# Jackknife----
theta_hat <- median(one_sample)
theta_i <- numeric(n)
for (i in 1:n) {
    theta_i[i] <- median(one_sample[-i])
}
# Too small...
sqrt((n-1)*mean((theta_i - mean(theta_i))^2))
```

```
## [1] 1.5
```

Example v

```r
# Bootstrap----
# How do we sample with replacement?
sample(n, n, replace = TRUE)
```

```
##  [1]  4 10  1  8  4  4  4  7  8  2
```

Example vi

```r
# Bootstrap estimate of SE
boot_theta <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  median(one_sample[indices])
})
# Closer to true value
sd(boot_theta)
```

```
## [1] 8.14544
```

Example i

- We will revisit the `law` dataset in the `bootstrap` package, which contains information on average `LSAT` and `GPA` scores for 15 law schools.
- We are interested in the correlation ρ between these two variables

```r
library(bootstrap)
# Estimate of rho
(rho_hat <- cor(law$LSAT, law$GPA))
```

```
## [1] 0.7763745
```

Example ii

```
# Bootstrap estimate of SE
n <- nrow(law)
boot_rho <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  # We're sampling pairs of observations
  # to keep correlation structure
  cor(law$LSAT[indices], law$GPA[indices])
})
```

```
sd(boot_rho)
```

Example iii

```
## [1] 0.1360481
```

Empirical CDF i

- The **empirical CDF** of a sample $X_1, \ldots, X_n$, denoted $\hat{F}_n$, is the CDF of a *discrete* distribution whose support is the data points $\{X_1, \ldots, X_n\}$, and where each point has mass $1/n$.
- Mathematically, we have

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

- **Why do we care?** We already argued that we can't easily generate more samples from F. Instead, bootstrap generates more samples from the distribution $\hat{F}_n$.
  - Sampling with replacement is the same as sampling from the empirical CDF!

Empirical CDF ii

- Since $\hat{F}_n \to F$, we can often translate this convergence in terms of the bootstrap estimates.

$$\begin{array}{rccccc} \text{Real world:} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & \hat{\theta} = g(X_1, \ldots, X_n) \\ \text{Bootstrap world:} & \hat{F}_n & \Rightarrow & X_1^{(b)}, \ldots, X_n^{(b)} & \Rightarrow & \hat{\theta}^{(b)} = g(X_1^{(b)}, \ldots, X_n^{(b)}) \end{array}$$

Bootstrap estimate of bias

- Just as with jackknife, we can use bootstrap to estimate the bias of $\hat{\theta}$.
- Let $\hat{\theta}^{(b)}$ be the estimates computed using the bootstrap samples, and let $\bar{\theta} = n^{-1} \sum_{b=1}^{B} \hat{\theta}^{(b)}$ be their sample mean.
- The **bootstrap estimate of bias** is given by

$$\widehat{bias}(\hat{\theta}) = \bar{\theta} - \hat{\theta}.$$

Example i

```r
# law dataset
rho_hat <- cor(law$LSAT, law$GPA)

# Bootstrap estimate of bias
B <- 5000
n <- nrow(law)
```

Example ii

```
boot_rho <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  # We're sampling pairs of observations
  # to keep correlation structure
  cor(law$LSAT[indices], law$GPA[indices])
})

(bias <- mean(boot_rho) - rho_hat)
```

```
## [1] -0.004382551
```

Example iii

```r
# Debiased estimate
rho_hat - bias
```

```
## [1] 0.780757
```

Final remarks

- So when should we use jackknife vs bootstrap?
- In some way, the jackknife is an *approximation* of the bootstrap, and as a consequence, the bootstrap almost always outperforms the jackknife.
- However, for small sample sizes, the jackknife will be more computationally efficient:
    - Jackknife requires $n + 1$ computations of the estimate.
    - Bootstrap requires $B + 1$ computations of the estimate, where B is usually at least 1000.
- Bootstrap performs better when the sampling distribution is skewed (see next lecture).
- Jackknife does **not** work with some estimators, e.g. sample median and sample quantiles.