Perturbation Bootstrap

Zachary McCaw

2024-01-09

Perturbation Bootstrap

The perturbation bootstrap (Zhezhen 2001, Das 2019) is a method of approximating the distribution of $\hat{\theta} - \theta$. The general approach is first to obtain an influence function representation:

$$\sqrt{n}\big(\hat{\theta} - \theta\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1)$$

Next, the influence contributions of all subjects $\{\hat{\psi}_i\}_{i=1}^n$ are estimated. For b in $1, \ldots, B$ iterations, mean 1, variance 1 weights $\{\omega_i^{(b)}\}_{i=1}^n$ are sampled from a known distribution, independently of the observed data. These weights are used to perturb the sum:

$$(\theta^*_{(b)} - \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i \omega_i^{(b)}$$

Upon evaluating the right-hand side, each set of weights provides a realization of $(\theta^*_{(b)} - \hat{\theta})$, the deviation of a parameter estimate $\theta^*_{(b)}$ based on the perturbed data around the observed value $\hat{\theta}$. It has been shown that the distribution of $(\theta^*_{(b)} - \hat{\theta})$ can be used to approximate that of $(\hat{\theta} - \theta)$ (Zhezhen 2001, Das 2019). For example, the standard deviation of the B realizations $\{(\theta^*_{(b)} - \hat{\theta})\}_{b=1}^B$ is an estimate for the standard error of $\hat{\theta}$.

Kaplan-Meier Example

Consider survival data of the form $\big\{(U_i, \delta_i)\big\}_{i=1}^n$, where $U_i = \min(T_i, C_i)$ is the minimum of the event time T_i and the censoring time C_i, and $\delta_i = \mathbb{I}(T_i \leq C_i)$ is the status indicator. The Kaplan-Meier estimate has an influence function representation of the form (Andersen 1993):

$$\sqrt{n}\big\{\hat{S}(t) - S(t)\big\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(t) + o_p(1),$$
$$\psi_i(t) = -S(t) \int_0^t \frac{dM_i(u)}{n^{-1} \sum_{i=1}^n Y_i(u)}.$$

Here $Y_i = \mathbb{I}(U_i \geq t)$ is an at risk indicator, $dM_i(t) = dN_i(t) - Y_i(t)dA(t)$ is the increment in the counting process martingale, $N_i(t) = \mathbb{I}(U_i \leq t, \delta_i = 1)$ is the event counting process, and $A(t) = \int_0^t \alpha(u)du$ is the cumulative hazard.

Data subject to censoring are generated according to an exponential distribution with rate $\lambda = 1$. Consider estimation of the survival probability at time $\tau = 1$. The exact calculation is $S(1) = \exp(-1) = 0.368$.

```r
set.seed(101)

# Generate data.
n <- 1e3
data <- SurvUtils::GenData(n = n)
tau <- 1.0

# Estimated survival probability.
prob <- SurvUtils::OneSampleRates(data, tau = tau) %>%
  dplyr::mutate_if(is.numeric, function(x) {round(x, digits = 3)})

show(prob)
```

```
##   tau  rate    se lower upper
## 1   1 0.364 0.017 0.332 0.397
```

The influence contributions at time t can be estimated as:

$$\hat{\psi}_i(t) = -\hat{S}(t) \sum_{u \leq t} \frac{dN_i(u) - Y_i(u) d\hat{A}(u)}{n^{-1} \sum_{j=1}^{n} Y_j(u)},$$
$$d\hat{A}(u) = \frac{\sum_{j=1}^{n} dN_j(u)}{\sum_{j=1}^{n} Y_j(u)}$$

where the sum is taken over unique event times u up to time t, and $d\hat{A}(u)$ is the estimated increment in the cumulative hazard at time u. The influence function contributions can be estimated using `KMInfluence`:

```r
# Estimate influence contributions.
psi <- SurvUtils::KMInfluence(data = data, tau = tau)
head(round(psi, digits = 3))
```

```
## [1]  0.760  0.210 -0.424 -0.405  0.760 -0.436
```

The mean of the influence function contributions will approximate zero:

```r
head(round(mean(psi), digits = 10))
```

```
## [1] 0
```

Writing:

$$\hat{S}(t) - S(t) = \frac{1}{n} \sum_{i=1}^{n} \psi_i(t) + o_p(n^{-1/2})$$

the variance of $\hat{S}(t) - S(t)$ is:

$$\mathbb{V}\{\hat{S}(t) - S(t)\} = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(\psi_i) + o_p(n^{-1})$$

As the influence function contributions have mean zero $\mathbb{E}(\psi_i) = 0$, their variance $\mathbb{V}(\psi_i) = \mathbb{E}(\psi_i^2)$. The sampling variance of $\hat{S}(t)$ can be approximated as:

$$\hat{\mathbb{V}}\{\hat{S}(t) - S(t)\} = \frac{1}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \psi_i^2.$$

Empirically, the estimated standard error of $\hat{S}(t)$ based on $\{\hat{\psi}_i\}_{i=1}^n$ is:

```
# Estimate sampling variance and standard error.
sampling_var <- (1 / n) * mean(psi^2)
se <- sqrt(sampling_var)
show(round(se, digits = 4))
```

```
## [1] 0.0166
```

For comparison, the standard error provided by the `survival` package is:

```
# Reference standard error.
km_fit <- survival::survfit(
  survival::Surv(time, status) ~ 1, data = data)
km_fit <- summary(km_fit)

closest_time <- max(km_fit$time[km_fit$time <= tau])
ref_se <- km_fit$std.err[km_fit$time == closest_time]

show(round(ref_se, digits = 4))
```

```
## [1] 0.0167
```

When estimating standard errors with `survfit`, note that those provided by default (i.e. before applying `summary`) are not properly scaled, see the discussion by Magirr 2022.

To estimate the standard error by perturbation bootstrap, the summand is perturbed with mean 1, variance 1 weights $\{\omega_i^{(b)}\}_{i=1}^n$, generated from a known distribution, such as $\Gamma(1,1)$, independently of the observed data:

$$(\theta_{(b)}^* - \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_i \omega_i^{(b)}$$

Upon taking the weighted sum, each set of weights yields a realization of $(\theta_{(b)}^* - \hat{\theta})$. From the collection of $\{(\theta_{(b)}^* - \hat{\theta})\}_{b=1}^B$, properties of $(\hat{\theta} - \theta)$ may be approximated, notably the standard error:

```
# Perturbation bootstrap.
bootstraps <- 2e3
perturb <- rep(0, bootstraps)
for (b in 1:bootstraps) {
  draw <- rgamma(n = n, shape = 1)
  perturb[b] <- mean(psi * draw)
}
perturb_se <- sd(perturb)
show(round(perturb_se, digits = 4))
```

```
## [1] 0.0166
```

RMST Example

The case of the restricted mean survival time (RMST) is similar. Based on the simulated exponential $\lambda = 1$, the analytical RMST at time $\tau = 1$ is $\int_0^1 e^{-t} dt = 1 - e^{-1} = 0.632$. The estimated RMST is:

```
# Estimated RMST.
tau <- 1
rmst <- SurvUtils::OneSampleRMST(data, tau = tau)
show(round(rmst, digits = 3))
```

```
##   tau   auc    se lower upper
## 1   1 0.635 0.012 0.612 0.658
```

The influence function of the RMST at time t is (Andersen 1993):

$$\sqrt{n}\{\hat{R}(t) - R(t)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(t) + o_p(1),$$

$$\psi_i(t) = -\int_0^t \frac{\mu_t(u) dM_i(u)}{n^{-1} \sum_{i=1}^{n} Y_i(u)},$$

$$\mu_t(u) = \int_u^t S(t) dt.$$

The influence function contributions may be estimated using `RMSTInfluence`:

```
# Estimate influence contributions.
psi <- SurvUtils::RMSTInfluence(data = data, tau = tau)
head(round(psi, digits = 3))
```

```
## [1]  0.408  0.244 -0.260 -0.358  0.408 -0.188
```

The standard error estimated directly from the influence function is:

```
# Estimate sampling variance and standard error.
sampling_var <- (1 / n) * mean(psi^2)
se <- sqrt(sampling_var)
show(round(se, digits = 4))
```

```
## [1] 0.0117
```

The standard error estimated by perturbation bootstrap is:

```
# Perturbation bootstrap.
bootstraps <- 2e3
perturb <- rep(0, bootstraps)
for (b in 1:bootstraps) {
  draw <- rgamma(n = n, shape = 1)
  perturb[b] <- mean(psi * draw)
}
perturb_se <- sd(perturb)
show(round(perturb_se, digits = 4))
```

```
## [1] 0.0121
```