Lab 2: Bootstrap methods

PB HLTH 250C

February 8, 2023

Review of concepts (see Carpenter and Bithell (2000) for details)

A confidence interval for parameter θ is a random interval (l, u) that is expected to contain that parameter $(1-\alpha) \times 100\%$ of the time. In other words, we would like

$$\Pr\left(\theta > l \cap \theta < u\right) = 1 - \alpha.$$

Typically, we estimate confidence intervals using knowledge about the sampling distribution of some estimator $\hat{\theta}$ of θ. Most commonly, we assume

$$\hat{\theta} - \theta \sim \text{Normal}\left(0, \text{var}\left(\hat{\theta}\right)\right).$$

Let's take that idea and put it to the side and return to the first equation. For simplicity, let's make our CI one-sided by putting $l \mapsto -\infty$. Then, we have

$$\begin{aligned}
1 - \alpha &= \Pr\left(\theta > l \cap \theta < u\right) \\
&= \Pr\left(\theta > -\infty \cap \theta < u\right) \\
&= \Pr\left(\theta < u\right) \\
&= \Pr\left(\theta + (\hat{\theta} - \theta) < u + (\hat{\theta} - \theta)\right) \\
&= \Pr\left(\hat{\theta} < u + (\hat{\theta} - \theta)\right) \\
&= \Pr\left(\hat{\theta} - \theta > \hat{\theta} - u\right) \\
&= 1 - \Pr\left(\hat{\theta} - \theta \leq \hat{\theta} - u\right) \\
\Rightarrow \Pr\left(\hat{\theta} - \theta \leq \hat{\theta} - u\right) &= \alpha.
\end{aligned}$$

Using the assumption above, we should pick u such that $\hat{\theta} - u$ is the $(\alpha \times 100)^{\text{th}}$ percentile of $\text{Normal}\left(0, \text{var}\left(\hat{\theta}\right)\right)$ i.e. $u = \hat{\theta} - F^{-1}(\alpha)$ where $F^{-1}(\cdot)$ is the inverse cumulative distribution function. When symmetry is satisfied, we can also pick u such that $\hat{\theta} - u$ is the $[(1-\alpha) \times 100]^{\text{th}}$ percentile.

Non-parametric bootstrap for the interaction contrast ratio

The interaction contrast ratio (ICR) is an estimand used to assess the presence of additive interaction when only relative measures are available. Take p_{11}, p_{10}, p_{01}, and p_{00} to be the conditional probabilities of some outcome Y when $X_1 = 1$ and $X_2 = 1$; when $X_1 = 1$ and $X_2 = 0$; and so forth. The additive interaction contrast is the expected difference in the risk differences:

$$p_{11} - p_{00} - ((p_{10} - p_{00}) + (p_{01} - p_{00}))$$

Dividing the expression by the baseline risk p_{00} gives the expression for the ICR:

$$\frac{p_{11}}{p_{00}} - \frac{p_{10}}{p_{00}} - \frac{p_{01}}{p_{00}} + 1$$

Consider the following log-binomial model for stroke:

$$\log\left(\Pr\left(Y = 1 \mid x, \beta\right)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2)$$

where

- $\Pr\left(Y = 1 \mid x, \beta\right)$ is the conditional risk of stroke Y given covariates x and parameters β
- x is a vector of covariates including x_1 diabetes status and x_2 smoking status
- β is a vector containing the coefficients β_0, β_1, and β_2 of covariates 1, x_1 and x_2, respectively
  - β_0 is the log risk of stroke among those with no diabetes and no smoking
  - β_1 is the log risk ratio of stroke comparing those with diabetes at baseline to those without, holding smoking status constant at no smoking
  - β_2 is the log risk ratio of stroke comparing smokers at baseline to non-smokers at baseline, holding diabetes status constant at no diabetes
  - β_3 is left for the reader to interpret

In the context of the model above, the ICR is

$$\exp\left(\beta_1 + \beta_2 + \beta_3\right) - \exp\left(\beta_1\right) - \exp\left(\beta_2\right) + 1$$

Goal: Estimate BS intervals for the ICR by Normal approximation (Wald-type), the percentile method, and the bias corrected and accelerated method (BC_A).

Implementation: Write a function that returns a single estimate of the ICR given data frame `dataset` and a vector `index` of rows to use from `dataset`.

```
library(fastglm)
```

```
Loading required package: bigmemory
```

```
icr.fun <- function(dataset, index) {
                                                                    ①
                                                                    ②
                                                                    ③
  return(icr)
}
```

① Select rows of ‘dataset‘ using ‘index‘.

② Using the data frame with rows given by ‘index‘, fit a log binomial regression to estimate the parameters of the model given above.

- Indicator of stroke status is `stroke`
- Diabetes status is `diabetes`
- Smoking status is `cursmoke`

③ Return the ICR by extracting the coefficient estimates and applying the formula above.

Compute $R = 5000$ BS estimates of the ICR using the `boot()` function. Runtime can be reduced by specifying `parallel = "multicore"` and `ncpus = parallel::detectCores() - 1` (one less than cores available). Compute BS 95% CIs by normal approximation, the percentile method, and BC_A using the `boot.ci()` function.

```
library(boot)
set.seed(1108)
R <- 5000 # Must be greater than `nrow(stroke.data)` for skew adjustment
icr.boot <- boot(
    stroke.data,
    icr.fun,
    R,
    parallel = "multicore",
    ncpus = parallel::detectCores()  - 1)
boot.ci(icr.boot, type = c("norm", "perc", "bca"))
```

Parametric bootstrap for the attributable fraction (Greenland, 2004)

Consider the following expression giving the adjusted attributable fraction:

$$\mathrm{AF_p} = \frac{\mathrm{RR_a} - 1}{\mathrm{RR_a} + 1/\mathrm{O_0}}$$

where $\mathrm{P_0}$ is the exposure prevalence, $\mathrm{O_0} = \mathrm{P_0}/(1 - \mathrm{P_0})$ is the prevalence odds, and $\mathrm{RR_a}$ is the adjusted relative measure of association. We assume that exposure prevalence is independent of the adjusted measure of association.

Suppose we estimated $\mathrm{RR_a}$ and $\mathrm{O_0}$ using maximum likelihood estimation. From the two model results, we have

$$\log\left(\widehat{\mathrm{RR}}_\mathrm{a}\right) = 0.519, \quad \widehat{\mathrm{Var}}\left(\log\left(\widehat{\mathrm{RR}}_\mathrm{a}\right)\right) = 0.159^2$$
$$\log\left(\widehat{\mathrm{O}}_0\right) = -3.041, \quad \widehat{\mathrm{Var}}\left(\log\left(\widehat{\mathrm{O}}_0\right)\right) = 0.153^2.$$

The parametric BS procedure for $i = 1, \ldots, R$ is as follows:

Step 1: Suppose the model results were generated using adequately large data so that the sampling distributions are approximately normal. Draw two R-length vectors $\left\{\log\left(\widehat{\mathrm{RR}}_\mathrm{a}\right)^{(i)}\right\}_{i=1}^{R}$ and $\left\{\log\left(\widehat{\mathrm{O}}_0\right)^{(i)}\right\}_{i=1}^{R}$ from the implied sampling distributions.

```
R <- 5000
set.seed(1140)
```

Step 2: Calculate $\left\{\widehat{\mathrm{AF}}_\mathrm{p}^{(i)}\right\}_{i=1}^{R}$ using the $\left\{\log\left(\widehat{\mathrm{RR}}_\mathrm{a}\right)^{(i)}\right\}_{i=1}^{R}$ and $\left\{\log\left(\widehat{\mathrm{O}}_0\right)^{(i)}\right\}_{i=1}^{N}$ generated from the previous step.

Step 3: Transform $\left\{\widehat{\mathrm{AF}}_\mathrm{p}^{(i)}\right\}_{i=1}^{R}$ to get $\left\{\hat{\mathrm{L}}_\mathrm{p}^{(i)}\right\}_{i=1}^{R}$ for better behavior under normal approximation:

$$\mathrm{L_p} = \log(1 - \mathrm{AF_p})$$

Step 4: Compute the BS 95% CI for $\hat{\mathrm{L}}_\mathrm{p}$ using normal approximation (or some other method). Transform the interval endpoints back to the scale of $\mathrm{AF_p}$ to get the BS 95% CI for $\widehat{\mathrm{AF}}_\mathrm{p}$.

References

Carpenter, James, and John Bithell. 2000. "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians." *Statistics in Medicine*. Wiley Online Library.