Computer Intensive Statistics
Group HWA2: Bootstrap and Simulation-Based Methods

1. (1p) Consider the **Penguin** data set on Studium. We have measured the bill length (in millimeter), bill depth (in millimeter), flipper length (in millimeter) and the body mass (in gram) of 342 penguins. We are interested in the ratio of expected bill length and expected bill depth

$$\theta = \frac{\mathrm{E}\left[\text{bill length}\right]}{\mathrm{E}\left[\text{bill depth}\right]}.$$

   Write your own code to perform nonparametric bootstrap to construct the percentile bootstrap interval, the basic bootstrap interval, the normal bootstrap interval, and the studentized bootstrap interval.

2. (1p) Consider again the Penguin data set. Suppose that we want to test the hypothesis that the expected body mass is 4200 grams. The t-test with the test statistic

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

   requires that our data follow a normal distribution, where $\bar{X}$ is the sample mean, μ_0 is the hypothesized value, S is the standard deviation of the sample, and n is the sample size. For this data set, the normal distribution assumption may not be plausible. Write your own code to perform a bootstrap test to test such hypothesis. Be explicit on how the bootstrap samples are obtained.

3. (1p) Consider an iid sample from the exponential distribution with density function

$$p_\theta(x) = \exp\left[-(x-\theta)\right], \quad x > \theta.$$

   Its cumulative distribution function is

$$\mathrm{P}\left(X \leq x\right) = 1 - \exp\left[-(x-\theta)\right], \quad x > \theta.$$

   We can estimate θ by $\hat{\theta} = X_{[1]} = \min\left\{X_1, ..., X_n\right\}$. The distribution function of $X_{[1]}$ is

$$\mathrm{P}\left(X_{[1]} \leq z\right) = 1 - \exp\left[-n\left(z-\theta\right)\right].$$

   Consider $T = n\left(X_{[1]} - \theta\right)/\theta$. Is bootstrap valid to approximate the distribution of T? Be clear which bootstrap you consider in this task.

4. (1p) Suppose that we have observed (Y_i, X_i), $i = 1, ..., n$ and that the observations are independent of each other. We fit a parametric logistic model as

$$\mathrm{P}\left(Y_i = 1 \mid X_i = x_i\right) \quad = \quad \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)}.$$

Explain how you can obtain parametric bootstrap samples for such logistic regression.

5. (2p) Consider the model $Z_i \sim \text{Bernoulli}\left(p\right)$ and

$$\begin{aligned} Y_i &\sim \text{Poisson}\left(\lambda_1\right), \text{ if } Z_i = 1, \\ Y_i &\sim \text{Poisson}\left(\lambda_0\right), \text{ if } Z_i = 0. \end{aligned}$$

Suppose that we have observed a random sample $(Y_1, ..., Y_n)$ of size n. Derive the EM algorithm for the MLE of $(p, \lambda_1, \lambda_0)$.

6. (2p) Suppose that we have a data set of size np sampled from the following model: $Z_i \sim N\left(0, 1\right)$ and

$$Y_{ij} \mid Z_i \quad \overset{iid}{\sim} \quad \text{Bernoulli}\left(p_i\right), \quad i = 1, ..., n, j = 1, ..., p,$$

where

$$p_i \quad = \quad \frac{\exp\left(\beta + Z_i\right)}{1 + \exp\left(\beta + Z_i\right)}.$$

But we only observe Y_{ij}, $i = 1..., n$, $j = 1, ..., p$. Develop an MCEM algorithm to estimate the parameter β.

7. (2p) Consider again the penguin data set. We want to fit the linear regression model

$$\text{body mass} \quad = \quad \beta_0 + \beta_1 \text{bill\_length} + \beta_2 \text{bill\_depth} + \beta_3 \text{flipper\_length}.$$

One possible estimator is

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \end{bmatrix} \quad = \quad \begin{bmatrix} n & 1_n^T X \\ X^T 1_n & X^T X + \lambda I_n \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T y \\ X^T y \end{bmatrix},$$

where n is the sample size,1_n is an $n \times 1$ vector of ones, I_n is an $n \times n$ identity matrix, X is a $n \times 3$ matrix with column corresponding to bill length, bill depth, and flipper length, and y is an $n \times 1$ vector for body mass. Here λ is a tuning parameter that we need to choose ourselves. Consider a grid of λ such that 100 values of $\log \lambda$ are equally spaced between 0 and 10. Write your own code to perform leave-one-out cross validation to select the optimal tuning parameter λ.