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

October 30, 2024

1. (1p) Let X_1, ..., X_n be a random sample from a distribution with mean μ and variance σ^2. The central limit theorem implies that the distribution function of

$$Z = \sqrt{n}\left(\bar{X} - \mu\right)$$

converges to the distribution function of $N\left(0, \sigma^2\right)$. Let $g\left(\right)$ be a differentiable function. We are interesting in approximating the distribution of

$$\sqrt{n}\left(g\left(\bar{X}\right) - g\left(\mu\right)\right).$$

Is bootstrap a valid approach to approximate it? Motivate your answer.

2. (2p) In the EM algorithm, if $\theta^{(t)}$ maximizes $Q\left(\theta \mid \theta^{(t-1)}\right)$ as a function of θ, then the observed likelihood increases, i.e.,

$$\log g\left(y \mid \theta^{(t)}\right) \geq \log g\left(y \mid \theta^{(t-1)}\right).$$

Suppose that $\theta^{(t)}$ does not maximizes $Q\left(\theta \mid \theta^{(t-1)}\right)$, but we have $Q\left(\theta^{(t)} \mid \theta^{(t-1)}\right) > Q\left(\theta^{(t-1)} \mid \theta^{(t-1)}\right)$. Does the observed likelihood still increase? Is the EM algorithm still self-consistent, in the sense that the MLE is one of the local maxima in this algorithm?

3. (1p) Suppose that the complete data is $X = (Y, Z)$, but we only observe Y. To approximate

$$Q\left(\theta \mid \hat{\theta}^{(t-1)}\right) = \mathrm{E}\left[\log f\left(x \mid \theta\right) \mid y, \hat{\theta}^{(t-1)}\right]$$

in the MCEM algorithm, our textbook generates z from $f\left(y, z \mid \hat{\theta}^{(t-1)}\right)$. Is $Q\left(\theta \mid \hat{\theta}^{(t-1)}\right)$ correctly approximated using such generated z?

4. (1p) Consider the following parametric logistic model:

$$Y_i \sim \text{Bernoulli}\left(\frac{\exp\left(\beta_0 + \beta_1 V_i + \beta_2 Z_i\right)}{1 + \exp\left(\beta_0 + \beta_1 V_i + \beta_2 Z_i\right)}\right), \quad i = 1, ..., n.$$

We observe (Y_i, V_i, X_i), where $X_i = Z_i + \epsilon_i$ with $\epsilon_i \sim N\left(0, 0.25\right)$. A sample of size 1000 is observed. You can obtain the data set **SIMEXlogit** on Studium. Find the estimate of β_0, β_1, and β_2.