数据科学导论2024-Project I

假设$(Y, \mathbf{X})$服从线性模型$Y = \mathbf{X}^\top \boldsymbol{\beta} + \epsilon$, 其中$Y$是一元响应变量，$\mathbf{X} \in \mathbb{R}^p$是$p$-维协变量，$\boldsymbol{\beta}$是$p$-维感兴趣的未知参数，$\epsilon \sim (0, \sigma^2)$. 考虑如下的情况: 我们收集到了$\{\mathbf{X}_i\}_{i=1}^N$，其中$\mathbf{X}_i \sim \mathbf{X}$. 而相应的$\{Y_i\}_{i=1}^N$由于成本的约束，只允许在给定其中的一些$\mathbf{X}_i$下，获得其中$n$个响应的观测，即$\{Y_\ell^*, \mathbf{X}_\ell^*\}_{\ell=1}^n$，其中$n \ll N$. 请尝试根据如下的问题撰写报告。

(a) 在这种情况下，我们应该选取什么目标来去从$\{\mathbf{X}_i\}_{i=1}^N$中抽取$\{\mathbf{X}_\ell^*\}_{\ell=1}^n$;

(b) 如何根据你提出的目标来进行抽取呢？

(c) 考虑通过模拟的方法来阐述你提出的方法是否优于简单随机抽样;

(d) 如果Y是分类变量，即观测是1或者0，此时应该如何建立模型？如何进行抽样？

(e) 假设我们还想估计$\mathbb{E}(Y)$，一个自然的估计是$n^{-1}\sum_{\ell=1}^n Y_\ell^*$，可否利用$\{\mathbf{X}_i\}_{i=1}^N$来提高该估计？