《统计方法与机器学习》（第六周：随机模拟）

1.（上机练习）设计一个模拟：实现欠拟合和过拟合对多元线性回归模型的影响。
回顾：给定 $\boldsymbol{x}_0$，真实值为

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0,$$

其期望为 $E(y_0) = \boldsymbol{x}_0'\boldsymbol{\beta}$。通过线性回归模型得到估计 $\hat{\boldsymbol{\beta}}$，则 y_0 的预测值为

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}.$$

在上一章，我们知道一个结论：

$$\begin{aligned} E(\hat{y}_0 - y_0)^2 &= (E(\hat{y}_0) - E(y_0))^2 + E(\hat{y}_0 - E(\hat{y}_0))^2 + E(\varepsilon_0)^2 \\ &= \text{Bias}^2(\hat{y}_0) + \text{Var}(\hat{y}_0) + \sigma^2, \end{aligned}$$

而我们通常定义均方误差为

$$\text{MSE}(\hat{y}_0) = E(\hat{y}_0 - E(y_0))^2 = \text{Bias}^2(\hat{y}_0) + \text{Var}(\hat{y}_0).$$

在模拟中，我们可以定义

$$\begin{aligned} \text{Bias}_k^2 &= \left(\frac{1}{M}\sum_{m=1}^{M} \hat{y}_{0,m}^{(k)} - \boldsymbol{x}_0'\boldsymbol{\beta}\right)^2 \\ \text{Var}_k &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_{0,m}^{(k)} - \frac{1}{M}\sum_{m=1}^{M}\hat{y}_{0,m}^{(k)}\right)^2 \\ \text{MSE}_k &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_{0,m}^{(k)} - \boldsymbol{x}_0'\boldsymbol{\beta}\right)^2 \end{aligned}$$

分别为第 k 个线性回归模型的偏差平方、方差和均方误差。

(a) 在同一张图上采用三种颜色绘制 Bias_k^2、Var_k 和 MSE_k 的三条曲线。

(b) 标示出 MSE 最小所对应的自变量个数。

2.（作业）按以下步骤，实现一个模拟：

(a) 构造 $n \times (p+1)$ 维自变量矩阵

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{x}_1' \\ 1 & \boldsymbol{x}_2' \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n' \end{pmatrix},$$

其中，$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 是独立同分布的 p 维随机变量，且服从多元正态分布，即

$$\boldsymbol{x} \sim N_p(\mathbf{0}_p, \Sigma_x), \quad \Sigma_x = \sigma_x^2 \cdot \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}$$

(b) 构造因变量 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，其中，$\boldsymbol{\beta} = (\mathbf{1}_{1+p_1}', \mathbf{0}_{p-p_1}')'$。这表明 $\boldsymbol{X}$ 中的前 $(p_1 + 1)$ 列的变量（包括常数项）对因变量有影响。同时，$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$，其中 ε_i 独立同分布于正态分布 $N(0, \sigma_y^2), i = 1, 2, \cdots, n$。

(c) 给定 $\boldsymbol{x}_0$，$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0$ 的期望为 $\boldsymbol{x}_0'\boldsymbol{\beta}$。

(d) 给定训练集 $\boldsymbol{y}$ 以及 $\boldsymbol{X}$，建立第 k 个模型，即

$$y = \beta_0 + \sum_{j=1}^{k} x_j \beta_j + \varepsilon$$

由此，我们得到其最小二乘估计 $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k, \mathbf{0}_{p-k}')$。此时，我们得到 y_0 的预测值为 $\hat{y}_0^{(k)} = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}^{(k)}$。

(e) 重复 (a)-(d) 步 M 次，可以得到 M 个不同预测值，分别记为 $\hat{y}_{0,m}^{(k)}, m = 1, 2, \cdots, M$。

(f) 参数设置如下：

   i. 样本量 $n = 300$
   ii. 变量维度 $(p, p_1) = (20, 10)$
   iii. 自变量的波动 $\sigma_x = 0.2$
   iv. 自变量的相依程度 $\rho_x = 0$
   v. 误差的波动 $\sigma_y = 3$
   vi. 预测点的位置 $\boldsymbol{x}_0 = (1, \mathbf{0.05}_{20}')'$;
   vii. 重复次数 $M = 5000$