Practice problems for midterm

April 5, 2024

1. Suppose we have a population $\{1, 2, 3, 4, 5\}$. Suppose that we draw a random sample $X_1, X_2, \cdots, X_n$ from this population with replacement. How much n should be so that the variance of the sample mean is equal to or smaller than 0.1 ?

2. Suppose we want to transform the left dataset(D) into the right dataset($D2$), where the unit of height is not centimeters but meters. **Which matrix** should we multiply to D and on **which side** (left or right) of D should we multiply it?

```
D.head()
```

	height	weight
0	158.64	48.00
1	156.59	45.78
2	172.70	62.65
3	154.18	43.89
4	178.39	66.24

```
D2.head()
```

	height	weight
0	1.5864	48.00
1	1.5659	45.78
2	1.7270	62.65
3	1.5418	43.89
4	1.7839	66.24

3. Suppose we have random variables $X_1, X_2, \cdots, X_n$ which are identically distributed with expectation μ and variance σ^2. ($E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$ and $Var(X_1) = Var(X_2) = \cdots = Var(X_n) = \sigma^2$.) Suppose however that $X_1, X_2, \cdots, X_n$ are *not* independent and $Cov(X_i, X_j) = \rho^2 > 0$ for any $i \neq j$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) What is $E(\bar{X})$? Show your derivation.

(b) What is $Var(\bar{X})$? Show your derivation.

(c) What is the meaning of (b) ?

4. In the following code,

(a) express the expected outcome using the cumulative distribution function of the standard normal distribution, $\Phi()$.

(b) explain how `sum(reject)/len(reject)` will change according to the value of `sigma` and `n`.

In [1]:
```
import numpy as np
import scipy.stats as stats
```

In [2]:
```
sigma=1
n=100
```

In [3]:
```
reject=[]

for i in range(100000):
    samp=np.random.normal(loc=-0.1,scale=sigma,size=n)

    Z=(np.mean(samp)-0)/sigma*np.sqrt(len(samp))

    reject.append(Z<stats.norm.ppf(0.05))
```

In [4]:
```
sum(reject)/len(reject)
```

5. Consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \cdots, N)$$

where ε_i's are i.i.d. with mean 0. Consider the following, slightly different model

$$y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + \varepsilon_i \quad (i = 1, 2, \cdots, N)$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$. Suppose $\hat{\beta}$ and $\hat{\alpha}$ are least-squares estimates. Prove that $\hat{\alpha}_1 = \hat{\beta}_1$.

6. Suppose we have data $(x_1, y_1), (x_2, y_2), \cdots, (x_{10}, y_{10})$ which satisfy

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (1 \le i \le 10)$$

where $\varepsilon_1, \cdots, \varepsilon_{10}$ are independent with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = 2$ for every i. Suppose the values of x_i can be selected anywhere between 0 and 2. What is the smallest value of $Var(\hat{\beta}_1)$ possible? ($\hat{\beta}_1$ is the estimate of β_1 obtained by Least-Square Method.) Explain the response.

7. Suppose we conduct linear regression on outcome variable (y) and explanatory variable (x). We posit the following model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

where ε_i's are i.i.d. $\mathcal{N}(0, \sigma^2)$.

Consider the following, slightly different model

$$y_i = \alpha_0 + \alpha_1 c x_i + \epsilon_i \tag{2}$$

where ϵ_i's are i.i.d. $\mathcal{N}(0, \sigma^2)$.

In (2), we scaled the variable x_i by $c(\neq 0)$. (For example, when x_i is the height in cm, we can change it to height in meter by multiplying $c = 0.01$.)

Prove that the test results for

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

and

$$H_0 : \alpha_1 = 0 \qquad vs \qquad H_1 : \alpha_1 \neq 0$$

are identical. (Assume we know the value of σ^2.)

8. Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix}$, $\mathcal{E} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$. Suppose we have

$$Y = X\beta + \mathcal{E}$$

where $\varepsilon_1, \cdots, \varepsilon_N$ are independent with $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$ for every i. Also, suppose $X^T X = \begin{bmatrix} 4 & -2 & 6 \\ -2 & 2 & -5 \\ 6 & -5 & 29 \end{bmatrix}$ and $X^T Y = \begin{bmatrix} 22 \\ -14 \\ 71 \end{bmatrix}$.

(a) Conduct Cholesky decomposition of $X^T X$.

(b) What is the least-square estimate of β? (Do not use the inverse formula of 3×3 matrix.)

9. Let $M = \begin{bmatrix} y_1 & x_{11} & x_{21} \\ y_2 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ y_N & x_{1N} & x_{2N} \end{bmatrix}$. We posit the following model

$$y_i = \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \varepsilon_i$$

where ε_i's are i.i.d. with mean 0 and $\bar{x}_1 = \frac{1}{N}\sum_{i=1}^N x_{1i}$, $\bar{x}_2 = \frac{1}{N}\sum_{i=1}^N x_{2i}$. Suppose $N = 15$, $\sum_{i=1}^N y_i = 45$, and suppose we have

$$M^T \left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T \right) M = \begin{bmatrix} 5 & 30 & -30 \\ 30 & 10 & 0 \\ -30 & 0 & 15 \end{bmatrix},$$

where I is $N \times N$ identity matrix and $\mathbf{1}$ is a $N \times 1$ vector with all elements equal to 1. What are the least-squares estimates of β_0, β_1 and β_2?

10. Suppose we conduct linear regression on outcome variable (y) and explanatory variables (x_1, x_2, $\cdots$, x_p). We posit the following model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where ε_i's are i.i.d. with mean 0.

Consider the following, slightly different model

$$y_i = \alpha_0 + \alpha_1 \underline{c_1 x_{i1}} + \alpha_2 \underline{c_2 x_{i2}} + \cdots + \alpha_p \underline{c_p x_{ip}} + \varepsilon_i$$

where $c_1, c_2, \cdots, c_p \neq 0$. (For example, when x_{i1} is the height in cm, we can change it to height in meter by multiplying $c_1 = 0.01$.)

Suppose $\hat{\beta}$ and $\hat{\alpha}$ are least-squares estimates. Using matrix operations, show that $\hat{\alpha}_j = \frac{1}{c_j}\hat{\beta}_j$ for $j = 1, 2, \cdots, p$. (*Hint: see Problem 2*)