Solution to Gate ST 2023 Q 26

Mayank Gupta

Question : Consider the following regression model

$$y_k = \alpha_0 + \alpha_1 \log_e k + \epsilon_k, \qquad k = 1, 2, \ldots, n,$$

where ϵ_k's are independent and identically distributed random variables each having probability density function $f(x) = \frac{1}{2}e^{-|x|}, x \in \mathbb{R}$. Then which one of the following statements is true?

(A) The maximum likelihood estimator of α_0 does not exist
(B) The maximum likelihood estimator of α_1 does not exist
(C) The least squares estimator of α_0 exists and is unique
(D) The least squares estimator of α_1 exists, but it is not unique

Solution:

$$f(\epsilon_k) = \frac{1}{2}e^{-|\epsilon_k|} \tag{1}$$

$$\text{Likelihood function : } f(\epsilon_1\epsilon_2....\epsilon_n) = \prod_{k=1}^{n} f(\epsilon_k) \tag{2}$$

$$L = \prod_{k=1}^{n} \frac{1}{2}e^{-|\epsilon_k|} \tag{3}$$

$$L_1 = \ln L = \ln\left(\prod_{k=1}^{n} \frac{1}{2}e^{-|\epsilon_k|}\right) \tag{4}$$

$$= \sum_{k=1}^{n} \ln\left(\frac{1}{2}e^{-|\epsilon_k|}\right) \tag{5}$$

$$= \sum_{k=1}^{n} \left(-\ln 2 - |y_k - \alpha_0 - \alpha_1 \log_e k|\right) \tag{6}$$

$$= -n\ln 2 - \sum_{k=1}^{n} \left(|y_k - \alpha_0 - \alpha_1 \log_e k|\right) \tag{7}$$

$$L_1 = \text{function of } \alpha_0, \alpha_1 \tag{8}$$

1) Maximum likelihood estimator
   We need to find the value of α_0 and α_1 which will maximise the value of L_1 i.e. the value of α_0 and α_1 which will minimise the value of $\sum_{k=1}^{n} |y_k - \alpha_0 - \alpha_1 \log_e k|$
   a) With respect to α_0
      i) For $y_k - \alpha_0 - \alpha_1 \log_e k > 0$

$$\begin{aligned} \min_{\alpha_0} \quad & y_k - \alpha_0 - \alpha_1 \log_e k \\ \text{s.t.} \quad & \alpha_0 \le y_k - \alpha_1 \log_e k \end{aligned}$$

Using Lagrange multiplier method

$$L(\lambda) = y_k - \alpha_0 - \alpha_1 \log_e k - \lambda(\alpha_0 - y_k + \alpha_1 \log_e k) \tag{9}$$
$$\frac{\partial L}{\partial \alpha_0} = -1 - \lambda = 0 \tag{10}$$
$$\frac{\partial L}{\partial \lambda} = y_k - \alpha_0 - \alpha_1 \log_e k = 0 \tag{11}$$
$$\lambda = -1 \tag{12}$$
$$\alpha_0 = y_k - \alpha_1 \log_e k \tag{13}$$

ii) For $y_k - \alpha_0 - \alpha_1 \log_e k < 0$

$$\min_{\alpha_0} \quad -(y_k - \alpha_0 - \alpha_1 \log_e k)$$
$$\text{s.t.} \quad \alpha_0 \geq y_k - \alpha_1 \log_e k$$

Using Lagrange multiplier method

$$L(\lambda) = -(y_k - \alpha_0 - \alpha_1 \log_e k) - \lambda(\alpha_0 - y_k + \alpha_1 \log_e k) \tag{14}$$
$$\frac{\partial L}{\partial \alpha_0} = 1 - \lambda = 0 \tag{15}$$
$$\frac{\partial L}{\partial \lambda} = y_k - \alpha_0 - \alpha_1 \log_e k = 0 \tag{16}$$
$$\lambda = 1 \tag{17}$$
$$\alpha_0 = y_k - \alpha_1 \log_e k \tag{18}$$

As value of α_0 matches for both cases of modulus
$\therefore$ The maximum likelihood estimator of α_0 exist

b) With respect to α_1

i) For $y_k - \alpha_0 - \alpha_1 \log_e k > 0$

$$\min_{\alpha_1} \quad y_k - \alpha_0 - \alpha_1 \log_e k$$
$$\text{s.t.} \quad \alpha_1 \leq \frac{y_k - \alpha_0}{\log_e k}$$

Using Lagrange multiplier method

$$L(\lambda) = y_k - \alpha_0 - \alpha_1 \log_e k - \lambda\left(\alpha_1 - \frac{y_k - \alpha_0}{\log_e k}\right) \tag{19}$$
$$\frac{\partial L}{\partial \alpha_1} = -\log_e k - \lambda = 0 \tag{20}$$
$$\frac{\partial L}{\partial \lambda} = -\left(\alpha_1 - \frac{y_k - \alpha_0}{\log_e k}\right) = 0 \tag{21}$$
$$\lambda = -\log_e k \tag{22}$$
$$\alpha_1 = \frac{y_k - \alpha_0}{\log_e k} \tag{23}$$

ii) For $y_k - \alpha_0 - \alpha_1 \log_e k < 0$

$$\min_{\alpha_1} \quad -(y_k - \alpha_0 - \alpha_1 \log_e k)$$
$$\text{s.t.} \quad \alpha_1 \geq \frac{y_k - \alpha_0}{\log_e k}$$

Using Lagrange multiplier method

$$L(\lambda) = -(y_k - \alpha_0 - \alpha_1 \log_e k) - \lambda\left(\alpha_1 - \frac{y_k - \alpha_0}{\log_e k}\right) \quad (24)$$

$$\frac{\partial L}{\partial \alpha_1} = \log_e k - \lambda = 0 \quad (25)$$

$$\frac{\partial L}{\partial \lambda} = -\left(\alpha_1 - \frac{y_k - \alpha_0}{\log_e k}\right) = 0 \quad (26)$$

$$\lambda = \log_e k \quad (27)$$

$$\alpha_1 = \frac{y_k - \alpha_0}{\log_e k} \quad (28)$$

As value of α_1 matches for both cases of modulus
$\therefore$ The maximum likelihood estimator of α_1 exist

$\therefore$ Option (A) and (B) are incorrect

c) Least square estimator
The least square estimator of α_0 and α_1 is $\tilde{\alpha_0}$ and $\tilde{\alpha_1}$ which will minimise

parameter	value	description
$\bar{y}$	$\frac{1}{n}\sum_{k=1}^{n} y_k$	Average value of y_k
$\bar{x}$	$\frac{1}{n}\sum_{k=1}^{n} \log_e k$	Average value of $log_e k$

TABLE 1
Variables used

$$Q(\alpha_0, \alpha_1) = \sum_{k=1}^{n} (y_k - \alpha_0 - \alpha_1 \log_e k)^2 \quad (29)$$

$$\frac{\partial Q}{\partial \alpha_0} = -2\sum_{k=1}^{n} (y_k - \alpha_0 - \alpha_1 \log_e k) = 0 \quad (30)$$

$$\sum_{k=1}^{n} (y_k - \alpha_0 - \alpha_1 \log_e k) = 0 \quad (31)$$

$$n\bar{y} - n\alpha_0 - \alpha_1 n\bar{x} = 0 \quad (32)$$

$$\implies \tilde{\alpha_0} = \bar{y} - \tilde{\alpha_1}\bar{x} \quad (33)$$

$$\frac{\partial Q}{\partial \alpha_1} = -2\sum_{k=1}^{n} (y_k - \alpha_0 - \alpha_1 \log_e k)\log_e k = 0 \quad (34)$$

$$\implies \tilde{\alpha_1} = \frac{\sum_{k=1}^{n} (\log_e k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n} (\log_e k - \bar{x})^2} \quad (35)$$

$\therefore$ Least square estimator of α_0 and α_1 exists and are unique
$\therefore$ Option (C) is correct and (D) is incorrect

d) Steps for simulation the given distribution whose probability density function is $f(x) = \frac{1}{2}e^{-|x|}$
  i) Write a function cdf for calculating the cdf of any random variable

$$P_X(x) = \begin{cases} \frac{1}{2}e^{x} & x \leq 0 \\ \frac{1}{2}e^{-x} & x > 0 \end{cases} \quad (36)$$

$$F_X(x) = \begin{cases} \int_{-\infty}^{x} \left(\frac{1}{2}e^{x}\right) dx & x \le 0 \\ \int_{-\infty}^{0} \left(\frac{1}{2}e^{x}\right) dx + \int_{0}^{x} \left(\frac{1}{2}e^{-x}\right) dx & x > 0 \end{cases} \tag{37}$$

$$F_X(x) = \begin{cases} \frac{1}{2}e^{x} & x \le 0 \\ \frac{1}{2}\left(2 - e^{-x}\right) & x > 0 \end{cases} \tag{38}$$

ii) Declare a function inverse cdf ($I(u)$) such that its input is any random number and output is random variable whose cdf equals that of the given distribution
For x $\le$ 0

$$u = \frac{1}{2}e^{x} \tag{39}$$

$$e^{x} = 2u \tag{40}$$

$$x = \ln 2u \tag{41}$$

$$\because x \le 0 \tag{42}$$

$$u \le 0.5 \tag{43}$$

For x $>$ 0

$$u = \frac{1}{2}\left(2 - e^{-x}\right) \tag{44}$$

$$2 - e^{-x} = 2u \tag{45}$$

$$e^{-x} = 2 - 2u \tag{46}$$

$$x = -\ln\left(2 - 2u\right) \tag{47}$$

$$\because x > 0 \tag{48}$$

$$u > 0.5 \tag{49}$$

$$I(u) = \begin{cases} \ln(2u) & u \le 0.5 \\ -\ln(2 - 2u) & u > 0.5 \end{cases} \tag{50}$$

iii) Define three arrays random_vars , cdf_values , theoretical_cdf_values to store random variables, simulated cdf values and theoretical cdf values
iv) Generate random numbers using rand() and calling inverse cdf funtion to generate our random variable
v) Calling cdf function to calculate the cdf of the generated random variable
vi) Storing the random variable,theoretical cdf and generated cdf into their respective arrays
vii) Storing the data of these three array into a .dat file
viii) Plotting these .dat file in python