MA 589 | End-Semester Examination
Statistical Foundation for Data Science | 9:00–12:00 IST
IIT GUWAHATI | November 18, 2023

Model Answers of End-Semester Examination (Points:40)

1. Let $(X_1,\, X_2)$ be a bivariate normal random vector with $E(X_1) = E(X_2) = 0$, $Var(X_1) = Var(X_2) = 1$ and $Cov(X_1,\, X_2) = 0$. Let U be a $U(0,\, 1)$ random variable, which is independent of $(X_1,\, X_2)$.

   (a) (2 points) Show that $Z_u = \dfrac{uX_1 + X_2}{\sqrt{u^2+1}} \sim N(0,\, 1)$ for all $u \in \mathbb{R}$.

   Solution: For fixed $u \in \mathbb{R}$,

$$Z_u = \frac{u}{\sqrt{u^2+1}}X_1 + \frac{1}{\sqrt{u^2+1}} = \begin{pmatrix} \frac{u}{\sqrt{u^2+1}} & \frac{1}{\sqrt{u^2+1}} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

   Thus, Z_u is a linear combination of $(X_1,\, X_2)$, which has a bivariate normal distribution. Therefore, Z_u as a normal distribution with mean

$$E(Z_u) = \begin{pmatrix} \frac{u}{\sqrt{u^2+1}} & \frac{1}{\sqrt{u^2+1}} \end{pmatrix} \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix} = 0.$$

   and variance

$$Var(Z_u) = \begin{pmatrix} \frac{u}{\sqrt{u^2+1}} & \frac{1}{\sqrt{u^2+1}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{u}{\sqrt{u^2+1}} \\ \frac{1}{\sqrt{u^2+1}} \end{pmatrix} = 1.$$

   Therefore, $Z_u \sim N(0,\, 1)$ for all $u \in (0,\, 1)$.

   (b) (4 points) Find $P\left(Z \le 0\right)$, where $Z = \dfrac{UX_1 + X_2}{\sqrt{U^2+1}}$. [Hint: You may use part (a)]

   Solution:

$$\begin{aligned}
P\left(Z \le 0\right) &= P\left(\frac{UX_1 + X_2}{\sqrt{U^2+1}} \le 0\right) \\
&= \int_0^1 P\left(\frac{UX_1 + X_2}{\sqrt{U^2+1}} \le 0 \middle| U = u,\right) du \\
&= \int_0^1 P\left(\frac{uX_1 + X_2}{\sqrt{u^2+1}} \le 0 \middle| U = u\right) du \\
&= \int_0^1 P\left(\frac{uX_1 + X_2}{\sqrt{u^2+1}} \le 0\right) du \\
&= \int_0^1 P\left(Z_u \le 0\right) du \\
&= \int_0^1 \frac{1}{2} du \\
&= \frac{1}{2}.
\end{aligned}$$

   The second equality is obtained using conditional argument, where $U \sim U(0,\, 1)$. The fourth equality follows as U and (X_1, X_2) are independent. The sixth equality follows as standard normal distribution is symmetric about zero.

2. (3 points) Suppose that the observed value of a random sample of size 10 drawn from a population with the probability density function

$$\frac{1}{2}\left(\frac{1}{\theta}e^{-x/\theta}+\frac{1}{10}e^{-x/10}\right),\ \ 0<x<\infty$$

are 8.59, 22.30, 29.10, 26.14, 1.70, 16.31, 1.34, 2.64, 5.23, 0.58. Find the estimate of θ based on the given observed sample using the method of moments.

Solution: The expectation corresponding to the given probability density function is

$$\int_0^\infty \frac{x}{2}\left(\frac{1}{\theta}e^{-x/\theta}+\frac{1}{10}e^{-x/10}\right)dx=\frac{\theta+10}{2}.$$

The sample mean is 11.393. Therefore, method of moments estimate of θ, say $\widehat{\theta}$, satisfies

$$\frac{\widehat{\theta}+10}{2}=11.393 \implies \widehat{\theta}=12.786.$$

3. (5 points) Let $\{0,\,1,\,2,\,3\}$ be an observed sample of size 4 from a population having $N(\theta,\,5)$ distribution, where $\theta\geq 2$. Find the maximum likelihood estimate of θ based on the observed sample.

Solution: The likelihood function of θ for given realization is

$$L(\theta)=\left(\frac{1}{10\pi}\right)^2\exp\left[-\frac{1}{10}\left\{\theta^2+(\theta-1)^2+(\theta-2)^2+(\theta-3)^2\right\}\right]$$

for $\theta\geq 2$. Therefore, log-likelihood function is

$$l(\theta)=-2\ln(10\pi)-\frac{1}{10}\left\{\theta^2+(\theta-1)^2+(\theta-2)^2+(\theta-3)^2\right\}$$

for $\theta\geq 2$. Now,

$$\frac{d}{d\theta}l(\theta)=-\frac{1}{5}\left(4\theta-6\right)\leq-\frac{2}{5}<0$$

for all $\theta\geq 2$. Thus, the likelihood function is a decreasing function of θ on the domain $[2,\,\infty)$ and hence, attains its' maximum at $\theta=2$. Thus, the maximum likelihood estimate of θ based on the given sample is 2.

4. (5 points) Let $X_1,\,X_2,\,\ldots,\,X_{15}$ be a random sample form a population having $U(-\theta,\,\theta)$ distribution, where $\theta>0$ is unknown parameter. Derive a 95% symmetric confidence interval using pivotal method. [Hint: You may find the pivot using the statistic $T=\max_{i=1,2,\ldots,15}|X_i|$.]

Solution: The CDF of $W = \frac{T}{\theta}$ is

$$\begin{aligned} F_W(w) &= P\left(\frac{1}{\theta}\max_{i=1,\,2,\,\ldots,\,15}|X_i| \le w\right) \\ &= \left\{P\left(|X_1| \le \theta w\right)\right\}^{15} \\ &= \left\{P\left(-\theta w \le X_1 \le \theta w\right)\right\}^{15} \\ &= \begin{cases} 0 & \text{if } w < 0 \\ w^{15} & \text{if } 0 \le w < 1 \\ 1 & \text{if } w \ge 1. \end{cases} \end{aligned}$$

Thus, W is a pivot. Now, the upper 0.975 point, say w_1, of the distribution of W satisfies

$$F_W(w_1) = 0.025 \implies w_1 = (0.025)^{\frac{1}{15}} \approx 0.782.$$

The upper 0.025 point, say w_2, of the distribution of W satisfies

$$F_W(w_2) = 0.975 \implies w_2 = (0.975)^{\frac{1}{15}} \approx 0.998.$$

Therefore,

$$P\left[(0.025)^{\frac{1}{15}} \le \frac{T}{\theta} \le (0.975)^{\frac{1}{15}}\right] = 0.95 \implies P\left[\frac{T}{(0.975)^{\frac{1}{15}}} \le \theta \le \frac{T}{(0.025)^{\frac{1}{15}}}\right] = 0.95.$$

Thus, 95% symmetric confidence interval for θ is

$$\left[\frac{\max_{i=1,\,2,\,\ldots,\,15}|X_i|}{(0.975)^{\frac{1}{15}}},\ \frac{\max_{i=1,\,2,\,\ldots,\,15}|X_i|}{(0.025)^{\frac{1}{15}}}\right] \approx \left[1.002\max_{i=1,\,2,\,\ldots,\,15}|X_i|,\ 1.279\max_{i=1,\,2,\,\ldots,\,15}|X_i|\right].$$

5. (6 points) Let $X_1,\, X_2,\, \ldots,\, X_n$ be a random sample of size $n\,(n > 1)$ form a population having $N(\mu,\, \sigma^2)$ distribution, where both $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown parameters. Derive (in an implementable form) likelihood ratio level 0.01 test for testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.

Solution: The likelihood function of μ and σ^2 is

$$L\left(\mu,\, \sigma^2\right) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Here $\Theta_0 = \{0\} \times \mathbb{R}^+$ and $\Theta_1 = \mathbb{R} \setminus \{0\} \times \mathbb{R}^+$. Hence $\Theta_0 \cup \Theta_1 = \mathbb{R} \times \mathbb{R}^+$. Now

$$\sup_{\Theta_0} L(\mu,\, \sigma^2) = \sup_{\sigma > 0} L(0,\, \sigma^2) = \left(\frac{2\pi e}{n}\sum_{i=1}^{n} x_i^2\right)^{-n/2},$$

and

$$\sup_{\Theta_0 \cup \Theta_1} L(\mu,\, \sigma^2) = \left(\frac{2\pi e}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{-n/2}.$$

Hence, the likelihood test statistic is

$$\Lambda = \left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} x_i^2} \right)^{n/2} = \left(1 + \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{-n/2}.$$

Note that for $x > 0$, $f(x) = (1+x)^{-n/2}$ is a decreasing function in x. Hence

$$\Lambda < k \iff \frac{n\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} > k_1 \iff \frac{\sqrt{n}\,|\overline{x}|}{s} > k_2,$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$. Therefore, the likelihood ratio test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}|\overline{x} - \mu_0|}{s} > k_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $E_{\Theta_0}\left(\psi(\boldsymbol{X})\right) = \alpha$. Now we know that

$$\frac{\sqrt{n}\,\overline{X}}{S} \sim t_{n-1} \text{ under null hypothesis } H_0 : \mu = 0.$$

Hence $E_{\Theta_0}\left(\psi(\boldsymbol{X})\right) = \alpha \implies k_2 = t_{n-1;\alpha/2}$ and the likelihood ratio test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \dfrac{\sqrt{n}\,|\overline{x}|}{s} > t_{n-1;\alpha/2} \\ 0 & \text{otherwise.} \end{cases}$$

6. (5 points) Let X_i, $i = 1, 2, \ldots$ are independently and identically distributed χ^2 random variables with 2 degrees of freedom. Let $N_n = \#\left\{k : 1 \leq k \leq n, X_k \geq 2\right\}$. Show that $\frac{N_n}{n}$ converges to $\frac{1}{e}$ with probability one. You may use the fact that the probability density function of a χ^2 random variable with 2 degrees of freedom is

$$f(x) = \begin{cases} \frac{1}{2} \exp\left\{-\frac{x}{2}\right\} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Solution: For $k = 1, 2, \ldots$, let us define

$$Y_k = \begin{cases} 1 & \text{if } X_k \geq 2 \\ 0 & \text{otherwise.} \end{cases}$$

Then $\frac{N_n}{n} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \overline{Y}_n$. Moreover, Y_k's are i.i.d. random variables with

$$E(Y_1) = P\left(X_1 \geq 2\right) = \int_2^{\infty} \frac{1}{2} \exp\left\{-\frac{x}{2}\right\} dx = \frac{1}{e}.$$

Thus, using SLLN,

$$\frac{N_n}{n} = \overline{Y}_n \to E(Y_1) = \frac{1}{e}$$

with probability one.

7. A researcher is investigating the use of a windmill to generate electricity. The researcher collected data on the DC output (y) of a windmill and the corresponding average wind velocity (x) in miles per hour for 5 consecutive days. The data is given in the following table. The preliminary aim of the researcher is to fit a linear regression model considering DC output as response and average wind velocity as regressor.

x	y
5.00	1.58
6.00	1.82
3.40	1.05
2.70	0.50
10.00	2.23

For all the parts in this question, please write the steps clearly mentioning statistical modeling and expressions. No need to derive any estimator or tests.

(a) (3 points) Compute the coefficients of a linear regression using the above data.

Solution: Here $\overline{x} = 5.42$, $\overline{y} = 1.436$, $S_{xy} = 7.1244$, and $S_{xx} = 32.968$. Therefore, $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 0.2161$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \approx 0.2647$.

(b) (3 points) Determine the coefficient of determination. Interpret the result.

Solution: Here, $SS_T \approx 1.8237$. Therefore, $R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\widehat{\beta}_1 S_{xy}}{SS_T} \approx 0.8442$. As the value of R^2 is large, it shows that the model fits the data quite well.

(c) (2 points) Test, at level 0.05, the significance of the linear regression, by stating null and alternative hypotheses clearly.

Solution: Here we want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The test statistics is

$$t = \frac{\widehat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}},$$

which follows a t-distribution with $n-2 = 3$ degrees of freedom. Now, the observed value of t is 4.032 and $t_{3;0.025} = 3.18$. Thus, observed value of t is more than $t_{3;0.025}$, and hence, the null hypothesis is rejected. Therefore, this regression is a significant one.

(d) (2 points) Suppose that the weather forecast says that the average wind speed for tomorrow will be 7 miles per hour. Find the 99% prediction interval of DC output for tomorrow.

Solution: A 99% prediction interval for DC output for wind speed 7 miles per hour is

$$\begin{aligned}
&\left[\widehat{y}_0 \mp t_{3;0.005}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}\,\right] \\
=&\left[\widehat{y}_0 \mp t_{3;0.005}\sqrt{MS_{Res}\left(1 + \frac{1}{5} + \frac{(7 - 5.42)^2}{32.968}\right)}\,\right] \\
\approx&[-0.253, 3.808].
\end{aligned}$$

You may take $z_{0.1} = 1.28$, $z_{0.05} = 1.64$, and $z_{0.025} = 1.96$.