BT307 Biological Data Analysis

Total Marks: 40 | **End-semester Examination** | **Duration: 3 hr**
Jan-May 2022

Instructions: In all questions, you **MUST** show all the calculation steps. Mark the final answer clearly.

Section A: Eight questions with two marks each

Q1. For the given data, calculate the covariance of P and Q. Use appropriate correction in your calculation.

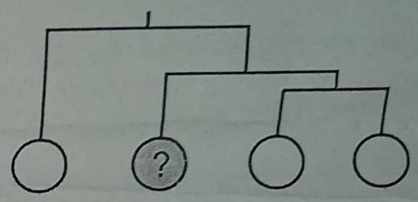
P	2	2	3	4	5
Q	1	1	2	3	4

Q2. For the given data, our regression equation is $G = b_1 S + b_2 M + a$. To perform the linear regression, we formulated a system of equations of the form $\mathbf{G} = \mathbf{PB}$. Here, $\mathbf{B} = \begin{bmatrix} b_1 & b_2 & a \end{bmatrix}^T$. Calculate **P** and **G**.

Salinity (S)	Moisture (M)	Growth (G)
1	0.2	3
1.5	0.2	2.5
2	0.34	2
2	0.4	1.6
3	0.46	1

Q3. We have four data points D1, D2, D3, and D4. The following dendrogram is obtained by hierarchical clustering of this data using single linkage. The distance matrix for the data points is given. Identify the data point marked by '?' in the dendrogram.

	D1	D2	D3	D4
D1	0	3	6	1
D2	3	0	8	8
D3	6	8	0	7
D4	1	8	7	0



Q4. A microarray experiment has been performed in three experimental conditions (E1, E2, and E3). The normalized fold-changes in the expression of two genes (X and Y) in these conditions are given. Calculate the Pearson Correlation Coefficient between these two genes.

	E1	E2	E3
X	0	1	2
Y	5	2	2

Q5. We are performing a sequence of statistical tests using R. At one step of the analysis, we have used the following code: `TukeyHSD(mod, conf.level=.95)`

What is the purpose of using the function `TukeyHSD()`?

Q6. For the same data set d, we performed two modeling using R:

```
reg1 <- lm(y ~ x + 0, data = d)
reg2 <- lm(y ~ x, data = d)
```

What is the difference between these two models?

Q7. We want to project data on new coordinates or vectors in principal component analysis. Like standard coordinate systems, these new coordinates should be orthogonal. How do we assure that the new coordinates or principal components would be orthogonal?

Q8. We have performed linear regression considering X as a predictor for the following data. Calculate TSS. Calculate the degrees of freedom of TSS.

X	3	3	4	5	5	6	7	7
Y	3	2	4	5	4.5	4	6	5

Section B: Six questions with four marks each

Q9. **X** and **Y** are two data vectors. Prove that when the $\text{var}(\mathbf{X}) = \text{var}(\mathbf{Y}) = 1$ and $\text{cov}(\mathbf{X}, \mathbf{Y}) = 0$, the Euclidean distance between **X** and **Y** is equal to the Mahalanobis distance between **X** and **Y**.

Q10. X is a discrete random variable. Prove that $\text{var}(X) = E(X^2) - [E(X)]^2$.

Q11. In a city, it rains on one-third of the days. The local evening newspaper attempts to predict rain on the following day. Three-quarters of rainy days and three-fifths of dry days are correctly predicted by the previous evening's paper. Given that this evening's paper predicts rain, what is the probability that it will rain tomorrow?

Q12. We are performing linear spline regression with two knots using linear algebra. The data is provided in the table. *y* is the dependent variable. The knots are at $x = 4$ and $x = 8$. What is the sum of all elements in the last row of the **X** matrix (matrix of predictor)?

x	1	2	3	4	5	6	7	8	9	10	11	12
y	3.3	5.7	7.8	9.2	9.1	9.7	10	9.7	7.9	5.6	3.9	1.3

Q13. We performed PCA for the data shown here. The loading matrix is given. Project the data on Principal Components 1 and 2 (PC1 and PC2). Calculate the Euclidian distance between samples 1 and 2 in the PC1-PC2 space.

Data:

	Drug 1	Drug 2	Drug 3
Sample 1	1	2	3
Sample 2	4	5	6
Sample 3	1	2	3
Sample 4	2	4	6

Loading matrix:

$$\begin{bmatrix} 1 & 1 & 3 \\ 2 & 1 & 5 \\ 3 & 1 & 8 \end{bmatrix}$$

Q14. **X** is a data matrix with *n* number of samples and *v* number of variables. **S** is its covariance matrix. **u** is a unit vector. Prove that the variance of the data projected on **u** will be maximum if **u** is an eigenvector of **S**.