BStat (Hons) I Year
2024-25 Second Semester
Statistical Methods II

Duration: 2 hours | 27 February 2025 | Maximum marks: 100

This examination is closed book, closed notes. Calculators are NOT allowed during the examination. Please answer all the questions.

1. Write an R code to simulate 10 rolls of a fair die and to calculate the maximum of the numbers obtained. [15]

2. For a data set `X` of four variables `x1`, `x2`, `x3` and `x4`, the R code
   `cor(X)`
   produces the following output.

```
            x1          x2         x3         x4
x1  1.00000000 -0.03057823 0.83495178 0.03799898
x2 -0.03057823  1.00000000 0.03497818 0.95391431
x3  0.83495178  0.03497818 1.00000000 0.11653807
x4  0.03799898  0.95391431 0.11653807 1.00000000
```

   State which ordering of the variables you will use for a parallel lines plot, and why. [15]

3. The R code

```
airpollt <- read.table(file = "https://people.stat.sc.edu/Hitchcock/airpoll.txt",
            header = TRUE)
dim(airpollt)
head(airpollt)
```

   produces the following output.

```
[1] 60  8
  City     Rainfall Education Popden Nonwhite NOX SO2 Mortality
1 akronOH        36      11.4   3243      8.8  15  59     921.9
2 albanyNY       35      11.0   4281      3.5  10  39     997.9
3 allenPA        44       9.8   4260      0.8   6  33     962.4
4 atlantGA       47      11.1   3125     27.1   8  24     982.3
5 baltimMD       43       9.6   6441     24.4  38 206    1071.0
6 birmhmAL       53      10.2   3325     38.5  32  72    1030.0
```

   (a) Describe the plot that you will get after running the R code
   `plot(airpollt[,-1])`

   (b) Suppose the R object `indx` contains the vector of the unique integers in your BStat roll number (excluding 0, 1 and 9) in increasing order. Describe the plot that you will get after running the R code
   `stars(airpollt[,indx])`

   In your descriptions, please remember to relate the plots to the dimensions of the data matrix and the names of the variables. [10+10=20]

(*Please turn over*)

4. The table given below shows the midyear population of the world for the years 1981-2000 (source: U.S. Census Bureau, International Data Base).

Year	Population (billion)	Year	Population (billion)
1981	4.533	1991	5.367
1982	4.613	1992	5.450
1983	4.694	1993	5.531
1984	4.774	1994	5.611
1985	4.855	1995	5.691
1986	4.938	1996	5.769
1987	5.024	1997	5.847
1988	5.110	1998	5.925
1989	5.196	1999	6.003
1990	5.284	2000	6.080

(a) Consider expressing the midyear population y in the year x through the piecewise linear (broken line) regression model, with zero-mean errors,

$$y = \begin{cases} \alpha_0 + \alpha_1 x + \varepsilon & \text{if } x \leq 1990, \\ \beta_0 + \beta_1 x + \varepsilon & \text{if } x > 1990. \end{cases}$$

Show that this model can be written as a linear regression model with a suitable choice of explanatory variables.

(b) According to the model of part (a), $E(y|x)$ may be discontinuous at $x = 1990$, unless the parameters of the model satisfy a certain restriction. Can you rewrite this model with an equivalent set of fewer parameters, so that $E(y|x)$ is continuous for all x?

(c) For the data given in the above table, write the model of part (b) in the matrix-vector form, while identifying the matrices and vectors precisely.

(d) Express the 'reduction in the rate of annual increase of world population from the 1980s to the 1990s' in terms of the parameters of the model of part (b).

(e) Can $\alpha_1 - \beta_1$ in the model of part (b) be interpreted as the 'reduction in the rate of annual increase of world population from the 1980s to the 1990s'? Explain.

[7+7+7+5+4=30]

5. Suppose a random variable Y has to be approximated by a linear function of k random variables $X_1, X_2, \ldots, X_k$, having the form $\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$. The random variables may be assumed to have finite means (not necessarily 0) and finite variances and covariances.

(a) Write the mean squared error of approximation $E[\{Y - (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)\}^2]$, in terms of the means, variances and covariances.

(b) Find an explicit expression for the optimum values of the coefficients $\beta_1, \beta_2, \ldots, \beta_k$ that minimizes the mean squared error.

(c) Does the linear function with coefficients optimized as in part (b) have the largest possible correlation with Y, among all the linear functions? Explain. [6+9+5=20]