Assignment 4 (three pages)

Statistics 32950-24620 (Spring 2023)

Due 9 am Tuesday, April 18. (Reminder: In-class midterm April 20.)

Requirements

- Your answers should be typed or clearly written, started with your name, Assignment 4, STAT 24620 or 32950; saved as `LastnameFirstnamePset4.pdf`. Make sure to upload to Gradescope under the correct section: 246Pset4 or 329Pset4, and tag pages.
- When you use R (or other software) to solve problems, select only relevant parts of the output, edit, then insert in your writing.
- You may discuss approaches with others. However the assignment should be devised and written by yourself. Capturing contents from other sources then pasting as your answers are not allowed.

References: Chapters 5 (up to 5.5 for this assignment), 6 (up to 6.6), 7 (up to 7.3), 12.2, 12.6, and 12.7 in Johnson & Wichern. Section 14.8 in *The Elements of Statistical Learning* (2009) by Hastie, Tibshirani and Friedman.

Problem assignments:

1. (*Simple* MANOVA *layout*)

   Observations on two responses $\boldsymbol{x}' = [x_1\ x_2] = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}'$ are collected for three treatments.

   Treatment 1 observations: $[6\ 7], [5\ 9], [8\ 6], [4\ 9], [7\ 9]$;
   Treatment 2 observations: $[3\ 3], [1\ 6], [2\ 3]$;
   Treatment 3 observations: $[2\ 3], [5\ 1], [3\ 1], [2\ 3]$.

   (a) Breakup the observations (indexed by j) into mean, treatment (indexed by t), and residual components

   $$\underset{\text{observation}}{\boldsymbol{x}_{tj}} = \underset{\text{overall mean}}{\bar{\boldsymbol{x}}} + \underset{\text{treatment effect}}{(\bar{\boldsymbol{x}}_t - \bar{\boldsymbol{x}})} + \underset{\text{residual}}{(\boldsymbol{x}_{tj} - \bar{\boldsymbol{x}}_t)}$$

   by constructing the data arrays for each of the two component variables. For example, for the first variable,

   $$\begin{bmatrix} 6\ 5\ 8\ 4\ 7 \\ 3\ 1\ 2 \\ 2\ 5\ 3\ 2 \end{bmatrix} = \begin{bmatrix} ?\ ?\ ?\ ?\ ? \\ ?\ ?\ ? \\ ?\ ?\ ?\ ? \end{bmatrix} + \cdots + \cdots$$

   (b) Using the information in Part (a) to construct the one-way MANOVA table.

   (c) Evaluate $|\boldsymbol{W}|, |\boldsymbol{B}|$, and Wilks' lambda $\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$. $(|\boldsymbol{W}| = det(\boldsymbol{W}))$

   (d) Using Barlett's approximation $-\ln(\Lambda^*)\left(n - 1 - \frac{p+g}{2}\right) \approx \chi^2_{p(g-1)}$ (under H_0) to find the observed p-value from the data.

   (e) Write the hypothesis test (H_0, H_a) that the p-value in (d) refers to. Describe the hypotheses in words.

   (Note: If you want to check your calculation in R, remember to code treatment index as factor instead of numeric.)

2. (MANOVA *and confidence region using R*)

   Researchers have suggested that a change in skull size over time is evidence of the interbreeding of a resident population with immigrant populations. Four measurements were made of male Egyptian skulls for three different time periods: period 1 is 4000 B.C., period 2 is 3300 B.C., and period 3 is 1850 B.C. The measured variables are

   X_1 = maximum breath of skull (mm)
   X_2 = basibregmatic height of skull (mm)
   X_3 = basialveolar length of skull (mm)
   X_4 = nasal height of skull (mm)

   The data are in https://www.stat.uchicago.edu/~meiwang/courses/s23-mva/T6-13.DAT.

```
skull=read.table("T6-13.DAT")
colnames(skull)=c("x1","x2","x3","x4","period")
```

(a) Conduct a one-way MANOVA (periods as "treatments") of the Egyptian skull data , use $\alpha = 0.05$.

(b) Apply Hotelling's T^2 to determine which pair of time periods differ, treating each pair of time periods as two independent samples of equal covariance structure.

(c) If you will construct confidence intervals for pairwise differences of component means **simultaneously** for all component pairs and all time periods,

i. How many simultaneous confidence intervals do you need to construct? (Note: just the number, so do not construct.)

ii. In this part you are asked to provide a formula for a subset of the simultaneous confidence intervals. (Note: just the formula with details, not the C.I.'s with numerical endpoints)

Let $\bar{x}_{ki}$ denote the sample mean of the ith component variable (X_i) during period k.

For component i, write the formula of the 85% Bonferroni simultaneous confidence interval for the difference of the mean between samples of periods 1 and 2.
Provide the details (including formula for the estimated variance, formula and numerical value of the multiplier).

Note: In R, the command for the upper (1-a)100% quantile of t-distribution with degrees of freedom = m is

```
qt(1-a,df=m)
```

3. (*Simple multivariate linear regression exercise*)

The data https://www.stat.uchicago.edu/~meiwang/courses/s23-mva/basketball.csv (download may start at click) contains measurements on each of the 54 basketball players. The data can be input in R by the command

```
basket=read.table("basketball.csv",header=T,sep=",")
```

The five variables are

```
   Height = height in feet
   Weight = weight in pounds
    Field = percent of successful field goals (out of 100 attempted)
Freethrow = percent of successful free throws (out of 100 attempted)
    Avgpt = average points scored per game
```

Use `(Field, Freethrow, Avgpt)` jointly as the response vector.

(a) Fit multivariate linear regression models, with and without interaction. Check the residuals. The residuals of which two response variables are mostly correlated?

(b) Conduct a sequential analysis of variance. Construct two MANOVA tables with different orders of the explanatory variables. Which variable or variables are important? (Hint: Check the correlations of the explanatory variables.)

4. (*Multidimensional Scaling*)

The table in https://www.stat.uchicago.edu/~meiwang/courses/s23-mva/T12-13.DAT gives the "distances" between certain archaeological sites from different periods, based upon the frequencies of different types of potsherds found at the sites. The dates of the period 1, 2, ..., 9 corresponding to A.D. years 918, 1131, 960, 987, 1024, 1005, 945, 1137, and 1062 respectively.

The data table can be input into R by the following commands.

```
mat=matrix(0,9,9)
mat[row(mat)<=col(mat)] = scan("T12-13.DAT")
X = t(mat)
```

(a) Given these distances, determine the coordinates of the sites in $q = 3, 4$ and 5 dimensions using (classical metric) multidimensional scaling.

(b) The $Stress(q)$ can be calculated as

$$Stress(q) = \left[\frac{\sum_{j<i} \left(x_{ij} - d_{ij}^{(q)} \right)^2}{\sum_{j<i} x_{ij}^2} \right]^{1/2}$$

where $d_{ij}^{(q)}$ is the distance between sites i and j in q diminutional representation by using (classical metric) multidimensional scaling method, and x_{ij} is the corresponding distance matrix X given by the data.
Plot (the minimum) $Stress(q)$ versus q and interpret the graph.

(c) Plot the nine sites (treated as variables) in two dimensions using the first two coordinates for the $q = 5$ dimensional solutions. Noting the periods associated with the sites.
Archeologists would say that the two dimensional configuration contains a time trend or movement pattern. Do you see it (admittedly very vague)?

5. (*Correspondence Analysis*)

A sample of 592 students is cross-classified according to hair colors and eye colors in the table in https://www.stat.uchicago.edu/~meiwang/courses/s23-mva/HairEyeAll.txt.
The data can be input by R commands

```
data = read.table("HairEyeAll.txt")
rownames(data) = c("Black","Brown", "Red", "Blond")   # for variable Hair
colnames(data)= c("Brown","Blue", "Hazel", "Green")   # for variable Eye
```

(a) Obtain the tables of cell percentages ($p_{ij} = x_{ij}/n$), row percentages, and column percentages.

(b) Obtain the table of expected cell counts (i.e. expected cell frequency) E_{ij} if the variables *Hair* and *Eye* were independent (for the given data).

(c) Obtain the table of cell *mass* $(x_{ij} - E_{ij})^2/E_{ij}$. Conduct a Peason's Chi-square test (What is the hypothesis? What is the degrees of freedom of the χ^2?) and check that the χ^2-statistic/n = total inertia.

(d) Perform a correspondence analysis of the data. What percentage of variation in the data is captured in the 2-dimensional CA plot? Are there any association of Eye category and Hair category reflected in the plot?

6. (*Conditional multivariate normal*)

Suppose random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are jointly multivariate normal with mean and covariance matrix given by

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \mathbf{0} \\ \Sigma_{31} & \mathbf{0} & \Sigma_{33} \end{bmatrix}$$

where Σ_{ii} are positive definite and $\mathbf{0}$ denotes a matrix with all entries 0.

Note: Your results below should be in terms of $\Sigma_{ij}, \Sigma_{ij}^{-1}$ and $\boldsymbol{\mu}_i$ (not matrices containing them as parts). Show your steps.

(a) i. Derive the conditional expectation $\mathbb{E}(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3)$.
ii. Derive the conditional variance $Var(\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2, \boldsymbol{X}_3 = \boldsymbol{x}_3)$.

(b) (Required for 32950. Optional for 24620.)
Derive the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 + \boldsymbol{X}_3 = \boldsymbol{x}_0$. Start by finding the distribution of $\boldsymbol{X}_2 + \boldsymbol{X}_3$.