BIOS -900 Homework #2

Date assigned: 09/25/2017
Due Date: 10/09/2017 by 11:59 pm (Blackboard clock time);

Instructions:

1. To receive full credit, show all work. Please make your work legible.
2. Total points for this homework are 100.
3. Do not forget to write your name on the homework.
4. Insert page numbers on all pages and also total # of pages submitted.
5. Homework can be typed or hand-written. Provide SAS code wherever necessary.
6. Use the BLACKBOARD drop box to turn in the homework (preferably as pdf) or bring it to class on 10/09/2017.

Question # 1: **16 points**

Refer to hand out on 'Sample Geometry and Random Sampling' given to you in the class

[A]
{i} Prove Equation 3.21 from this hand out.

{ii}Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be two 2x1 vectors and let $\mathbf{v}_1 = \mathbf{A}\mathbf{u}_1$ and $\mathbf{v}_2 = \mathbf{A}\mathbf{u}_2$ for a 2x2 matrix $\mathbf{A}$. Show that: (Area generated by $\mathbf{v}_1$ and $\mathbf{v}_2$) = $|\mathbf{A}|$.(Area generated by $\mathbf{u}_1$ and $\mathbf{u}_2$)

[B]
When the generalized variance is 0, it is the columns of the deviation matrix that are linearly dependent, and not necessarily those of the data matrix itself. Given the data

$$\begin{bmatrix} 3 & 1 & 0 \\ 6 & 4 & 6 \\ 4 & 2 & 2 \\ 7 & 0 & 3 \\ 5 & 3 & 4 \end{bmatrix}$$

{i} Obtain the deviation matrix, and verify that the columns are linearly dependent. Specify an $\mathbf{a}^{\boldsymbol{T}} = [a_1, a_2, a_3]$ vector that establishes the dependence.

{ii} Obtain the sample covariance matrix **S** and verify the generalized variance is 0.

{iii} Show that the columns of the data matrix are linearly independent.

Question # 2: **16 points**

[A]
Exercise 2.55 in your textbook asks you to use Theorem 2.9c and Corollary 1 of Theorem 2.9b to prove Theorem 2.9b. But on Page #39 of the textbook, Corollary 1 is itself obtained from Theorem 2.9b. Thus in this question, you are asked to do the following:

{i} Prove that $\begin{vmatrix} \mathbf{A}_{11} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}|.|\mathbf{A}_{22}|$ by using Theorem 2.9c alone.

{ii} Prove Corollary 1 by using Theorem 2.9c alone. You may use the result from {i} above.
Hint: $\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{O} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{C_1}||\mathbf{C_2}|$ That is, find $\mathbf{C_1}$ and $\mathbf{C_2}$ such that the result from {i} above can be used to show the required result.

[B]
Using your results from part [A] above, prove Equation 2.71 and Equation 2.72 for a square matrix **A**.
Hint: Partition **A** and verify that

$$\begin{bmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O}^T & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{O} \\ \mathbf{O}^T & \mathbf{A}_{22} \end{bmatrix}$$

[C] Using result from Part [B], show that;

$$\mathbf{A^{-1}} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{bmatrix}\begin{bmatrix} (\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & \mathbf{O} \\ \mathbf{O}^T & \mathbf{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O}^T & \mathbf{I} \end{bmatrix}$$

Question # 3: **25 points**

[A] Solve Exercise 4.16 but with $\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 4 \\ -1 \\ -2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 3 & -4 & 2 & 1 \\ -4 & 4 & -3 & -1 \\ 2 & -3 & 2 & 6 \\ 1 & -1 & 6 & 5 \end{bmatrix}$

[B] Find the maximum likelihood estimates of the 2x1 mean vector $\boldsymbol{\mu}$ and the 2x2 covariance matrix $\boldsymbol{\Sigma}$ based on the random sample:

$$\mathbf{X} = \begin{bmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{bmatrix}$$

[C] Research at least two different ways of testing bivariate normality for bivariate data. Write a two page summary (one page for each method) explaining the methods that you learned.

Question # 4: **25 points**

[A]
The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. (Data from "The SENIC Project", *American Journal of Epidemiology* 1980, 111:465-653). Dataset: ***SENIC.csv***

Each line of the data set has an identification number and provides information on 11 other variables for a single hospital. The data presented here are for the 1975-76 study period. The 12 variables are:

Variable Name	**Description**
Identification Number	1 – 113
Length of stay	Average length of stay of all patients in hospital (in days)
Age	Average age of patients (in years)
Infection risk	Average estimated probability of acquiring infection in hospital (in percent)
Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
Number of beds	Average number of beds in hospital during study period
Medical School affiliation	1 = yes, 2 = no
Region	Geographic region; 1 = NE, 2 = NC, 3 = S, 4 = W
Average daily census	Average number of patients in hospital per day during study period
Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number of full time plus one half the number of part time)
Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

Read in this dataset into SAS and make the following objects: Y = average length of stay in a hospital, and X = [1 | X1 | X2 | X3| X4] be a matrix with variables AGE (X1), INFECTION RISK (X2), AVAILABLE FACILITIES & SERVICES (X3) and ROUTINE CHEST X-RAY RATIO (X4).

i. Make a scatterplot matrix for X1, X2, X3, X4 and Y.
ii. Using matrix operations in R, compute the OLS estimate b for the parameters in the linear model $Y_i = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \beta_3(X_{i3}) + \beta_4(X_{i4}) + \varepsilon_i$ (model 1). Do not use any SAS PROC to compute the solution (i.e., DO NOT USE *PROC REG or PROC GLM).* Note: you need to make your design matrix X for this model using PROC IML.
iii. What do the parameters in the model in (b) mean? (i.e., interpret the parameter estimates in the context of this model.)

iv. Find the eigenvalues and eigenvectors for $(X^TX)^{-1}$. What do these quantities tell you about the data?
v. Find the rank of X. How does this value related to your eigenvalues?
vi. Compute the estimated means $\hat{Y} = Xb$ and the vector of residuals $e = Y - \hat{Y}$. Plot the residuals against the estimated means – what does this plot reveal?
vii. Create a normal probability plot from the values in the residual vector. What does this plot reveal?

[B]
Using the same data from question #1, answer the following questions.

i. Using matrix operations in SAS, compute the OLS estimate b for the parameters in the linear model $Y_i = \beta_1(X_{i1}) + \beta_2(X_{i2}) + \beta_3(X_{i3}) + \beta_4(X_{i4}) + \varepsilon_i$ (model 2). How does this model differ from model 1?
ii. What is the rank of design matrix X for model 2?
iii. How do the plot of the residuals and scatterplot of predicted means vs residuals change from the old model (model 1) to this new model (model 2)?
iv. Is $E(Y) = X\mathrm{B}$ estimable? Explain how you reached this conclusion?

Note: You may use R instead of SAS to solve this problem if R is the software of your choice. In that case also, you cannot use any R functions and have to solve everything using Matrix Algebra.

Question # 5: **18 points**

Let **X** be a nxp matrix of constants of rank = k. Partition such that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ is of size nxq, 0 < q < p. Let the nx1 random vector $\mathbf{Y} \sim \mathbb{N}(\mathbf{0}, \mathbf{I})$ and consider the following equation:

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{Y}^T\left(\mathbf{I} - \mathbf{X}\mathbf{X}^{(-)}\right)\mathbf{Y} + \mathbf{Y}^T\left(\mathbf{X}\mathbf{X}^{(-)} - \mathbf{X}_2\mathbf{X}_2^{(-)}\right)\mathbf{Y} + \mathbf{Y}^T\left(\mathbf{X}_2\mathbf{X}_2^{(-)}\right)\mathbf{Y}$$
$$= Q_1 + Q_2 + Q_3$$

where $\mathbf{X}\mathbf{X}^{(-)}$ and $\mathbf{X}_2\mathbf{X}_2^{(-)}$ are assumed to be symmetric. Rank of $\mathbf{X}_2$= m.

[A] Prove the following:

{i} $Q_1 \sim \chi^2_{n-k}$
{ii} $Q_2 \sim \chi^2_{k-m}$
{iii} $Q_3 \sim \chi^2_m$
{iv} Show that Q_1, Q_2, Q_3 are pairwise independent.

[B] Solve Exercise 5.20 from your textbook.

Question # 6 (Bonus) **5 points**

{i} Read the 3 pages posted as a pdf on Blackboard on the topic of "Singular Value Decomposition" of a matrix. Confirm that you read and understood this topic. You may read extra material by searching on the internet about this topic.

{ii} Read the following webpage about "why determinant of a 2x2 matrix is equal to the area of a parallelogram.
https://math.stackexchange.com/questions/29128/why-determinant-of-a-2-by-2-matrix-is-the-area-of-a-parallelogram
Confirm that you read and understood at least three different proofs. Did you enjoy reading the different proofs? Which proof did you like the best?

GOOD LUCK ☺☺