Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8302 COMPUTATIONAL TECHNIQUES IN DATA SCIENCE

DATE: July 22, 2024. Time: 3 Hours

Instructions

1. This examination consists of **FIVE** questions and an appendix to one of the questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1

a) Distinguish between numerical and analytical methods in the solution of mathematical problems.

(3 Marks)

b) Describe the bisection method and explain how it differs from the Newton-Raphson method.

(3 Marks)

c) Given an initial guess $x=3$, find an approximate value of the root of the function $f(x) = x^2 - 2$ using 3 iterations. Provide a sample R code that you would use to solve this problem, stopping after 40 iterations.

(6 Marks)

d) Using the inverse-transform approach to explain (mathematically) how you would generate random numbers from the exponential distribution, $f(y) = \theta \exp - (\theta y)$, and further provide an R code that will be used to generate random numbers from this distribution.

(8 Marks)

Question 2

Starting with the Newton-Raphson formula

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}$$

shown that the order of convergence of the Newton-Raphson method is

$$|\varepsilon^{(n+1)}| = k|\varepsilon^{(n)}|^2$$

as $n \to \infty$, with $k = \frac{1}{2}\left|\frac{f''(r)}{f'(r)}\right|$ provided $|f'(r)| = 0$.

(20 Marks)

Question 3

a) The probability density function of the logistic distribution is given by

$$f(x) = \frac{1}{\beta}\frac{e^{-(x-\mu)/\beta}}{[1+e^{-(x-\mu)/\beta}]^2}$$

Derive expressions for its cumulative distribution and inverse cumulative distribution functions and hence explain how random numbers from this distribution would be generated.

(9 Marks)

b) Suppose that $Y_1, \ldots, Y_n$ random numbers from a Exp(1), and that

$$Z = 2\sum_{i=1}^{n} Y_i \sim \chi^2(2n)$$

Using the moment generating function techniques, show that $Z \sim \chi^2(2n)$ and hence explain how you would generate random numbers from the exponential distribution.

(11 Marks)

Question 4

Consider the general linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0}, \sigma^2\boldsymbol{I}),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is an $(n \times 1)$ vector of response values, $\mathbf{X}$ is an $(n \times k)$ design matrix corresponding to the explanatory variables $X_1, \ldots, X_k$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is the vector of parameters, and σ^2 is the variance.

Using a maximum likelihood approach, clearly showing the likelihood function, the log-likelihood function, and the score-vector, derive the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2.

(20 Marks)

Question 5

a) Show that the probability density function of the Bernoulli distribution, $f(y) = \pi^y(1-\pi)^{1-y}, y = 0, 1$ belongs to the exponential dispersion family

$$f(y;\theta) = \exp\left[\frac{y\theta - b(\theta)}{\phi}\right] + c(y;\phi).$$

(3 Marks)

b) Also show that the mean and the variance of the Bernoulli distribution are equal to $b'(\theta)$ and $b''(\theta)$, respectively.

(3 Marks)

c) Consider the generalized linear model

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is an $(n \times 1)$ vector of response values belonging to the Bernoulli distribution, $\mathbf{X}$ is an $(n \times k)$ design matrix corresponding to the explanatory variables $X_1, \ldots, X_k$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is the vector of parameters. Derive an expression for the estimating equation and Hessian that would be used to estimate the vector of parameters $\boldsymbol{\beta}$.

(14 Marks)

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCE
MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS
END OF SEMESTER EXAMINATION
DSA 8101: ETHICS IN DATA SCIENCE

Date: 29TH September 2023 **Time:**3 hours

Instructions; Answer Question ONE and Any Other Three

Question 1. (40 Marks, 4 marks each)

Give a short explanation of the following terms highlighting the ethical aspects

1. Data Ethics,
2. Natural Moral Law,
3. Moral Dilemmas,
4. Virtues,
5. Character,
6. Algorithm Transparency,'
7. Ethical Algorithm Accountability
8. Data Bias
9. Ethical Data ownership
10. Double Effect Principle

Question 2. (20 marks)

Explain how data analytics can assist to improve the five types of internal goods.

Question 3. (20 marks)

Your immediate manager is asking you to work on Saturday and Sunday this month together with the other members of your team for an important project. This will interfere with the family duties and rest each need. Using what you have learned from Giving Voice to Values, present a draft for each of the steps, to train the other members and have a common voice. (2 marks for each explanation, 1 mark for each plan, and 2 extra marks for the introduction to GVV).

PRINCIPLE	EXPLANATION	PLAN
1.		
2.		
3.		
4.		

5.		
6.		
7.		

Question 4 (20 marks)

The police in some countries is using predictive analytics to following people and organizations. Using the six social ethical principles discuss the positive and negative aspects that such use of data could bring.

Question 5 (20 marks)

The new CEO is very social minded. He has asked you to prepare a draft to help the company live the six social ethical principles in a way that assists the employees and clients. Explain what each ethical principle requires and make a matrix for each indicating the suggested policies for each. In your analysis, suggest policies for each and how to measure effectiveness. (1 marks for each explanation, 1 mark for each policy, 1 mark for each measure and 2 extra marks for the introduction to the Social Ethical Principles).

PRINCIPLE	EXPLANATION	POLICY	MEASURE
1.			
2.			
3.			
4.			
5.			
6.			

Question 6. (20 marks)

"There's a growing body of data that shows that the growth of smartphones has also led to a significant decrease in optimism and life satisfaction, particularly among young people. But it's not the phone that's the problem, it's the way attention is manipulated by the companies that create the apps on our phones." Set Godin

The daily joy one feels is part of the immediate feeling of happiness, but to have total happiness one needs more than software. With a clear idea of total happiness, one chooses what is really important. Explain the four levels of happiness and give an example of IT activities that can encourage each level in an ethical way.

Level	Explanation	Example

***

STRATHMORE UNIVERSITY
STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
iLAB AFRICA
END OF SEMESTER EXAMS FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE
DSA 8105: FUNDAMENTAL CONCEPTS IN FINANCE AND ECONOMICS

Date: 28th September, 2023 **Time: 3 hours**

Instructions
Attempt any **Four** (4) questions

Question 1

A data scientist has many roles. Required:

(a) Identify these roles {5 marks}

(b) What is the nexus between these roles and economics and finance? {5 marks}

(c) Provide concrete examples in the Kenya economy where the nexus in 1(b) is manifest {5 marks}

[15 marks]

Question 2

The informal economy in Kenya is 83% of the economy. Much of the informal economy is comprised of micro, small and medium enterprises (MSMEs). Therefore, MSMEs are central to the growth of the economy and creation of jobs. A common characteristic of the informal economy is constrained space but unlimited users of the spaces. To ensure optimal usage of the spaces an optimal number of persons (workers) is required. You are required to:

(a) Identify an appropriate theory in economics that can address the problem of optimality of the number of persons per each space {2 marks}

(b) State the main thesis of this theory {2 marks}

(c) Create and solve a minimum working numerical example (MWNE) that demonstrates the variables and approach required to solve this problem if the output by MSMEs (Q) is related to the number of workers (l) according to the equation $Q = f(s, l) = 6l^2 - 0.4l^3$. {11 marks}

[15 marks]

Question 3

(a) As a data scientist you have noticed that in the recent past fuel prices have been on an upward trend. In the same period, you have also noticed the sale of fuel efficient plug in hybrid vehicles has been increasing. This is not the expected result since plug-in hybrids and fuel are expected to be complements. Using intuition from the price theory, explain this paradox. {5 marks}

(b) Based on economic theory, goods will flow where they are needed the most and where they fetch the most. However, this is not the practice in Kenya. We have marginalized areas that need goods but cannot get them even when the areas are willing to pay a premium for the goods. Using the various types of economic systems, their strengths and weaknesses, provide plausible explanations to this dichotomy between theory and practice. {5 marks}

(c) Consider the statistics provided in Table 1

Table 1: **Price and quantity of bread and butter produced in country Y**

Year	Quantity of bread	price of bread	Quantity of butter	Price of butter
2009	150	30	20	80
2010	160	40	50	95
2011	165	55	55	105
2012	180	60	59	110
2013	185	64	66	115

Assuming that only bread and butter are produced in this country calculate and interpret (where applicable):

(i) The nominal GDP {1 mark}

(ii) The real GDP if 2009=100 {1 mark}

(iii) The GDP deflators for 2010, 2011, 2012 and 2013 {2 marks}

(iv) Economic growth for the various years {1 mark}

[15 marks]

Question 4

(a) You wish to disabuse the notion that financial markets in Kenya are too small to matter using data. As an initial step, you first wish to establish the connection between financial markets and the economy in general using a list of channels. State any five connections between financial markets and the macroeconomy that would be contained in your list {5 marks}

(b) In any financial markets there are three broad categories of participants. Identify these participants and state their roles in the financial markets {6 marks}

(c) The agricultural sector in Kenya is usually characterized by informality of activities which leads to periods of plenty and periods of acute shortages. This is a problem that has been experienced elsewhere and solved using innovative financial instruments. Identify these financial instruments and explain how they work to solve information problems in the agricultural sector {4 marks}

[15 marks]

Question 5

Consider a $10,000 loan borrowed today. If the loan will be repaid in equal instalments at the end of four years. Required:

(a) Calculate the annual payment if interest is 9% {7 marks}

(b) Construct the amortization table for this loan {8 marks}

[15 marks]

Be More Be Strathmore! ☺

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES (SIMS)
MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
END OF SEMESTER EXAMINATION
DSA 8104: FUNDAMENTAL STATISTICAL ANALYSIS

DATE: 27th September, 2023 TIME: 3 Hours

INSTRUCTIONS

1. This examination consists of **FIVE** questions.
2. Answer Question **ONE** (**COMPULSORY**) and any other **TWO** questions.
3. You may use a **SIMPLE CALCULATOR**. No **MOBILE PHONES** in the exams room.

Question One (30 Marks)

(i) Describe the nature of kurtosis using Karl Pearson's measure of kurtosis (formula): 20 18 22 27 25 12 15 (3 marks)

(ii) The mean height of a class of 28 students is 162 cm. A new student of height 149 cm joins the class. What is the mean height of the class now? (2 marks)

(iii) After taking a refresher course, a salesman found that his sales (in dollars) on 9 random days were 1280, 1250, 990, 1100, 880, 1300, 1100, 950 and 1050. Does the sample indicate that the refresher course had the desired effect, in that his mean sale is now more than 1000 dollars? Assume $\sigma = 100$, and the probability of erroneously saying that the refresher course is beneficial should not exceed 0.01. Also assume that the sales are normally distributed. (4 marks)

(iv) To ride the indoor roller coaster "Space Mountain" at Walt Disney World in Orlando, Florida, guests must be at least 44 inches tall. Suppose that the heights of all visitors to Disney World are normally distributed with a mean of 53 inches and a standard deviation of 6. What fraction of visitors to Disney World can ride Space Mountain? (4 marks)

(v) The random variable X has probability density function

$$f(x) = \begin{cases} k(-x^2 + 5x - 5), & 1 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

(a) Show that $k = \frac{2}{3}$. (2 marks)

(b) Find the cumulative distribution function $F(x)$ for all x. (3 marks)

(c) Evaluate $P(X \leq 2.5)$. (2 marks)

(vi) Describe any two forms of data classification (2 marks)

(vii) Explain the importance of regression analysis in business. (2 marks)

(viii) A gardener buys 10 packets of seeds from two different companies. Each pack contains 20 seeds and he records the number of plants which grow from each pack.

Company A	20	5	20	20	20	6	20	20	20	8
Company B	17	18	15	16	18	18	17	15	17	18

(a) Find the mean, median and mode for each company's seeds. (3 marks)

(b) Which company has more stable seeds? Explain. (3 marks)

Question Two (15 Marks)

(i) The head of the Statistics department of a certain university is interested in the difference in writing scores between freshman Statistics students who are taught by different teachers. The incoming freshmen are randomly assigned to one of two Statistics teachers and are given a standardized writing test after the first semester. We take a sample of eight students from one class and nine from the other class. Is there a difference in achievement on the writing test between these two classes? (5 marks)

Class1	35	51	66	42	37	46	60	55	53
Class2	52	87	76	62	81	71	55	67	

(ii) A university has found over the years that out of all the students who are offered admission, the proportion who accept is 0.70. After a new director of admissions is hired, the university wants to check if the proportion of students accepting has changed significantly. Suppose they offer admission to 1200 students and 888 accept. Is this evidence at the $\alpha = .05$ level that there has been a real change from the status quo? How about at the 0.02 level? (4 marks)

(iii) With the current level of communication resources for an online bookstore and their projected growth over the next 6 months, a company will be able to provide satisfactory service if the average connection time per customer is no more than 13.5 minutes. Based on a random sample of 45 connections yielding s a sample mean of 15.3 minutes with a sample standard deviation of 6.7 minutes, would you recommend that the company upgrades their communication resources? (Perform a one-sided test at a 5% significance level.) (6 marks)

Question Three (15 Marks)

(i) An urn contains 3 red balls and 2 blue balls. A ball is drawn. If the ball is red, it is kept out of the urn and a second ball is drawn from the urn. If the ball is blue, then it is put back in the urn and a red ball is added to the urn. Then a second ball is drawn from the urn.

(a) What is the probability that both balls drawn are red? (3 marks)

(b) If the second drawn ball in the experiment is red, what is the probability that the first drawn ball was blue? (3 marks)

(ii) The number of miles that Anita's motorbike will travel on one gallon of petrol may be modelled by a normal distribution with mean 135 and standard deviation 12. Given that Anita starts a journey with one gallon of petrol in her motorbike's tank, find the probability that, without refueling, she can travel:

(a) more than 111 miles (3 marks)

(b) between 141 and 150 miles (3 marks)

(iii) Consider the density function

$$f(x) = \begin{cases} (p+1)x^p, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where p is greater than -1. Compute the expected value of X. (3 marks)

Question Four (15 Marks)

(i) Fifty male subjects drank a measured amount x (in ounces) of a medication and the concentration y (in percent) in their blood of the active ingredient was measured 30 minutes later. The sample data are summarised by the following information.

$$n = 50, \quad \sum x = 112.5, \quad \sum y = 4.83, \quad \sum xy = 15.255, \quad \sum x^2 = 356.25, \quad \sum y^2 = 0.667$$

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem. (4 marks)

(ii) A tobacco company statistician wishes to know whether heavy smoking is related to longevity.From a sample of recently deceased smokers, the number of cigarettes (estimated on a per day for their last five years after visits with their surviving relatives) is paired with the number of years that they lived.

Cigarettes	25	35	10	40	85	75	60	45	50
Years lived	63	68	72	62	65	46	51	60	55

(a) Present the information using a scatter diagram. (3 marks)

(b) Find the regression equation in the form $y = \beta_0 + \beta_1 x$. (4 marks)

(c) Give a practical interpretation of the slope β_1. (2 marks)

(d) Using your answer in (b), find the number of years lived for someone who smoked 101 cigarettes. (2 marks)

Question Five (15 Marks)

(i) In an experiment in breeding mice, a geneticist has obtained 120 brown mice with pink eyes, 48 brown mice with brown eyes, 36 white mice with pink eyes and 13 white mice with brown eyes. Theory predicts that these types of mice should be obtained in the ratios 9 : 3 : 3 : 1. Test the compatibility of the data with theory, using a 5% critical value. (5 marks)

(ii) A man travels from Nakuru to Kisumu by a car and takes 4 hours to cover the whole distance. In the first hour he travels at a speed of 50 km/hr, in the second hour his speed is 64 km/hr, in third hour his speed is 80 km/hr and in the fourth hour he travels at the speed of 55 km/hr. Find the average speed of the motorist. (3 marks)

(iii) The population in a city increased at the rate of 15% and 25% for two successive years. In the next year it decreased at the rate of 5%. Find the average rate of growth. (3 marks)

(iv) The following data refers to the frequency distribution of time in minutes that students arrived late to STA 1102 lecture 2 classes during the first semester of year 2021-22

Time in Minutes	0-2.5	2.5-5.0	5.0-7.5	7.5-10.0	10.0-12.5	12.5-15.0
No of students	40	25	12	22	13	8

Compute Bowley's coefficient of skewness and comment on the skewness. (4 marks)

END

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES (SIMS)
MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
END OF SEMESTER EXAMINATION
DSA 8205: OPTIMIZATION FOR DATA SCIENCE

DATE: 27th February, 2023 TIME: 3 Hours

INSTRUCTIONS

1. This examination consists of **FOUR** questions.
2. Answer Question **ONE (COMPULSORY)** and any other **TWO** questions.
3. You may use a **SIMPLE CALCULATOR**. No **MOBILE PHONES** in the exams room.

Question One (30 Marks)

(i) Solve the following NLP problem graphically. (5 marks)

$$\begin{aligned} \textbf{Maximize} \quad & Z = 2x_1 + 3x_2 \\ \textbf{Subject to:} \quad & x_1 x_2 \leq 8 \\ & x_1^2 + x_2^2 \leq 20 \\ & x_1 \geq 0; x_2 \geq 0 \end{aligned}$$

(ii) A farmer has recently acquired a 110 hectares piece of land. He has decided to grow Wheat and barley on that land. Due to the quality of the sun and the region's excellent climate, the entire production of Wheat and Barley can be sold. He wants to know how to plant each variety in the 110 hectares, given the costs, net profits and labor requirements according to the data shown below:

Variety	Cost (Price/Hec)	Net Profit (Price/Hec)	Man-days/Hec
Wheat	100	50	10
Barley	200	120	30

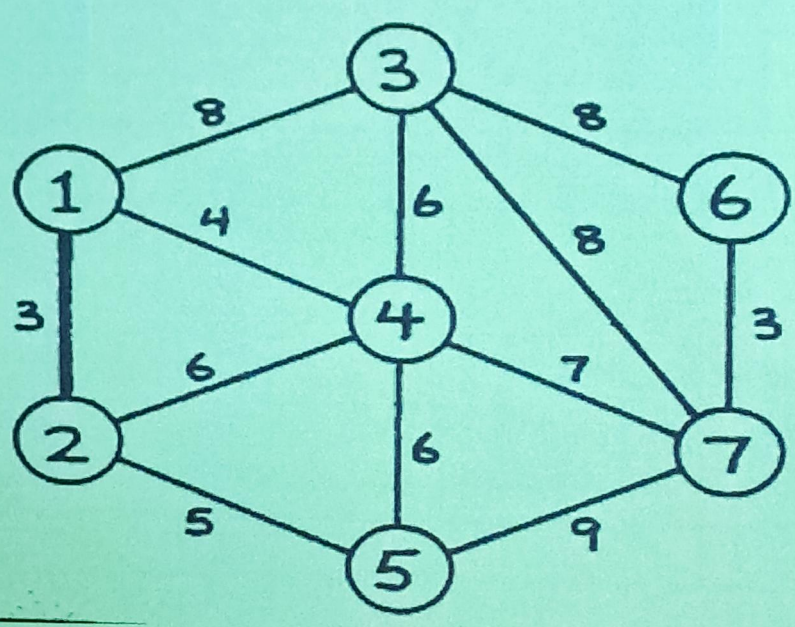
The farmer has a budget of \$10,000 and availability of 1,200 man-days during the planning horizon. Find the optimal solution and the optimal value using the graphical approach. (5 marks)

(iii) Explain five types of nonlinear programming problems. (5 marks)

(iv) Explain the procedure for constructing an initial solution using the North West Corner rule for a transportation problem. (5 marks)

(v) Explain the assumptions involved in the formulation of an assignment problem that must be satisfied to fit its definition. (5 marks)

(vi) The diagram below shows an undirected network, with the labels on the arcs indicating their costs. Find the minimum spanning tree. (5 marks)



Question Two (15 Marks)

(i) A company manufactures three-speed, five-speed, and ten-speed bicycles. Each bicycle passes through three departments; fabrication, painting & plating, and final assembly. The relevant manufacturing data are given in the table below.

	Labour hours per bicycle			Maximum labour hours available per day
	Three-speed	Five-speed	Ten-speed	
Fabrication	3	4	5	120
Painting & plating	5	3	5	130
Final Assembly	4	3	5	120
Profit per bicycle	\$80	\$70	\$100	

(a) How many bicycles of each type should the company manufacture per day in order to maximize its profit? What is the maximum profit? (5 marks)

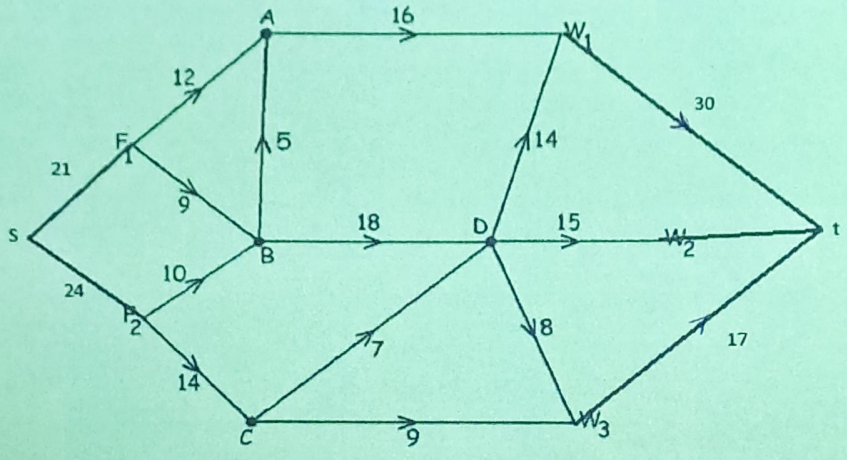
(b) Discuss the effect on the solution to part (a) if the profit on a ten-speed bicycle increases to \$110 and all other data in part (a) remains the same. (3 marks)

(c) Discuss the effect on the solution to part (a) if the profit on a five-speed bicycle increases to \$110 and all other data in part (a) remains the same. (3 marks)

(ii) A steel company has two mills. Mill 1 costs \$70,000 per day to operate, and it can produce 400 tons of high-grade steel, 500 tons of medium-grade steel, and 450 tons of low-grade steel each day. Mill 2 costs \$60,000 per day to operate, and it can produce 350 tons of high-grade steel, 600 tons of medium-grade steel, and 400 tons of low-grade steel each day. The company has orders totaling 100,000 tons of high-grade steel, 150,000 tons of medium-grade steel, and 124,500 tons of low-grade steel. How many days should the company run each mill to minimize its costs and still fill the orders? (4 marks)

Question Three (15 Marks)

(i) An oil company has a underground tank and wishes to transport its product to a tanker truck. The capacities of the various routes through pipes are shown in the diagram below.

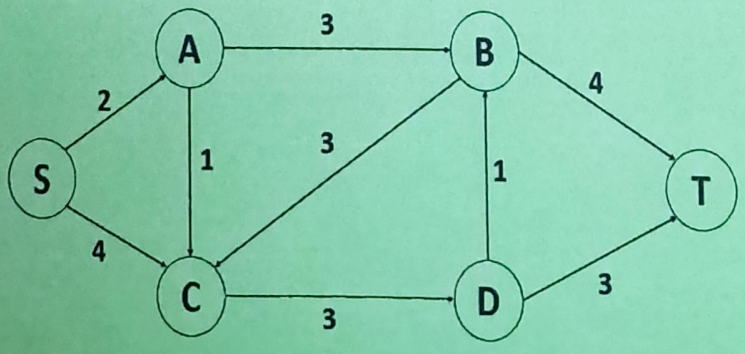


(a) Find the maximum number of units of oil that can be transported through the network and determine if the demand of 40 units can be met. (3 marks)

(b) Verify by finding the minimum cut. (3 marks)

(c) What is the outflow of the source? (2 marks)

(ii) In the flow network illustrated below, each directed edge is labeled with its capacity. We are using the Ford-Fulkerson algorithm to find the maximum flow. The first augmenting path is S-A-C-D-T, and the second augmenting path is S-A-B-C-D-T.



(a) Draw the residual network after we have updated the flow using these two augmenting paths (in the order given). (3 marks)

(b) List all of the augmenting paths that could be chosen for the third augmentation step. (2 marks)

(c) What is the numerical value of the maximum flow? Draw a dotted line through the original graph to represent the minimum cut. (2 marks)

Question Four (15 Marks)

(i) Find the solution to the minimization problem below by solving its dual using the simplex method. Show clearly all the steps. (5 marks)

$$\begin{aligned} &\textbf{Minimize} && Z = 12x_1 + 16x_2 \\ &\textbf{Subject to:} && x_1 + 2x_2 \geq 40 \\ & && x_1 + x_2 \geq 30 \\ & && x_1 \geq 0; x_2 \geq 0 \end{aligned}$$

(ii) Suppose we are required to maximize the function $f(x) = 12x - 3x^4 - 2x^6$, explain how we can use one-dimensional search procedure to solve the problem hence obtain its optimal solution. (5 marks)

(iii) Given the following quadratic programming problem,

$$\begin{aligned} &\textbf{Max} && f(x_1, x_2) = 15x_1 + 30x_2 + 4x_1x_2 - 2x_1^2 - 4x_2^2 \\ &\textbf{Subject to:} && x_1 + 2x_2 \leq 30 \\ & && x_1 \geq 0; x_2 \geq 0 \end{aligned}$$

State the Karush-Kuhn-Tucker (KKT) conditions for the above problem and state the criteria under which the problem is optimal. (5 marks)

END

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8203: PRINCIPLES OF DATA SCIENCE

DATE: 3rd MARCH, 2023 TIME: 3 Hours

INSTRUCTION

There are four questions in this exam, questions one is required. Select two questions from the remaining three. This exam will be taken in person through Google Classroom at Strathmore's ILAB computer labs at 4:00 PM - 7:00 PM on Friday March 3rd Nairobi time.

QUESTIONS:

1. Reflecting on our course and working on your Colab Notebook apply what you learned to this dataset related to the Group Evaluation Questionnaire you completed after group work. (20 points)
    a. Describe at least two code snippets you would use for each of the following activities based on the dataset (include your code and at least one paragraph of explanation for each):
        i. Data Cleaning (2 points)
        ii. Data Analysis (2 points)
        iii. Data Visualization or Machine Learning (2 points)
    b. Based on the dataset carry out some analysis (stating your assumptions) to determine the following (Include your code and an image of the result/visualization):
        i. Best Performing Group (4 points)
        ii. Best Performing Rator (4 points)
        iii. Best Performing Ratee (4 points)
    c. Provide a brief (one paragraph) executive summary to explain what you found and how it may be relevant in evaluating group performance. (2 points)

Below are the documents you need.
i) Questionnaire (For reference) https://forms.gle/bBrAoCzYM4oyKAPKA
ii) Raw Data - https://docs.google.com/spreadsheets/d/1Ao3AjU9nmNm_QAroJrPujfT1c-sFQ8n_bZqGDiZKmW4/edit?usp=sharing
iii) Cleaned Data - https://docs.google.com/spreadsheets/d/1JA0f09cnd8QQYnY-sip2kIFPEXAdk3rQSshqDmJwhws/edit?usp=sharing

Note: If your response is based on the raw data you will be eligible for 100% of the points from the question. If it's from the cleaned data, you will be eligible for 80% of the points from the question.

2. How can Data Science be useful in answering the following questions for either of these organizations? Pick one and include at least three scenarios with specific examples. (20 points)
   a. Lucky's, a chain of petrol stations, wants to know what types of e-commerce opportunities, if any, are relevant to Lucky's? (20 points)
   OR
   b. Plant Away is an International retailer and distributor of trees and shrubs. They have hundreds of smaller nurseries in various countries that grow the plant stock. The majority of their business is conducted online: Consumers purchase typically small quantities of products online and Plant Away coordinates the shipping from the most appropriate nursery.
      i. What unique problems might you anticipate they have in their supply chain? (10 points)
      ii. How can Data Science be useful in addressing these problems? (10 points)

3. Americlinic, a national chain of budget health-care clinics, is creating an information system that will allow patients and doctors at participating franchises to communicate online. The goal of the system is to allow doctors to respond to minor health questions quickly and more efficiently, saving patients unnecessary visits to the clinic. This will be a major procedural change. How can Data Science be useful in evaluating whether the new system is operating as desired. Consider the following questions and use examples to respond. (20 points)
   a. What is regression analysis? What assumptions underpin it.(4 points)
   b. How can regression analysis be useful in assessing the impact of the new system proposed? (4 points)
   c. What is an outlier? (2 points)

d. How can you identify outliers in your dataset? (4 points)
e. What may outliers look like in assessing the results of Americlinic's new system? (6 points)

4. Guaraldi and Associates is a management consulting firm located in Manhattan. The firm is interested in examining the relationship between the amount of vacation that their consultants use per year and how long the consultant has been with the firm. All consultants at Guaraldi and Associates, other than the Managing Partners, are considered either Junior Consultants or Senior Consultants based on their skills and expertise. The file *GuaraldiConsultants* contains data on all Junior and Senior Consultants at Guaraldi and Associates including the years of service the consultant has at the firm and the number of hours that the consultant took as vacation last year.

a. Create a scatter chart and add a trendline to examine the relationship between years of service and amount of vacation time used for all consultants at Guaraldi and Associates. Based on the scatter charts, what appears to be the relationship between years of service and amount of vacation time used? (5 points)

b. Create a scatter chart for the same data, but this time differentiate the points in the scatter chart based on whether the consultant is junior or senior. Do you see the same relationship in this scatter chart individually for junior and senior consultants as you saw in part a for all consultants? (5 points)

c. What additional data may be of value in this analysis? Explain how you would incorporate it in your analysis. (10 points)

plt.scatter(df.X, df.Y,)

plt.

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8203: PRINCIPLES OF DATA SCIENCE

DATE:26TH FEBRUARY, 2024 TIME: 3 Hours

INSTRUCTIONS

There are **four** questions in this exam, **questions one is required**. **Select two questions from the remaining three**. This exam will be taken in person through Google Classroom at Strathmore's ILAB computer labs at 5:00 PM - 8:00 PM on the day of the exam.

QUESTIONS:

1. Reflecting on our course and working on your Colab Notebook apply what you learned to the dataset below related to bird strikes. (20 marks)
    a. Describe at least two code snippets you would use for each of the following activities based on the dataset (include your code and at least one paragraph of explanation for each):
        i. Data Cleaning (2 marks)
        ii. Data Analysis (2 marks)
        iii. Data Visualization or Machine Learning (2 marks)
    b. Based on the dataset carry out some analysis (stating your assumptions) to determine the following (Include your code and an image of the result/visualization):
        i. Best Performing Airline (4 marks)
        ii. Worst Performing Airline (4 marks)
        iii. The most useful visualization from your perspective (4 marks)
    c. Provide a brief (one paragraph) executive summary to explain what you found in part (b) above and how it may be relevant in air travel. (2 marks)

Below are the documents you need.
Bird Strikes Data:
Google Sheets:
https://docs.google.com/spreadsheets/d/1ZmtRxu3Cb4OONE7nTY1g5viaqJd2iVP1rU3ZmW7d5No/edit?usp=sharing

OR

GitHub Raw Data:
https://raw.githubusercontent.com/chrisnjunge/DStest/master/BirdStrikes.xlsx%20-%20Bird%20Strikes.csv

2. How can Data Science be useful in answering the following questions for either of these organizations? Pick one and include at least three scenarios with specific examples. (20 marks)
    a. Recall your analysis of the country you were assigned for the class. Which e-commerce opportunities may exist as a result of the data you were analyzing? (20 marks)
    OR
    b. Plant Away is an International retailer and distributor of trees and shrubs. They have hundreds of smaller nurseries in various countries that grow the plant stock. The majority of their business is conducted online: Consumers purchase typically small quantities of products online and Plant Away coordinates the shipping from the most appropriate nursery.
        i. What unique problems might you anticipate they have in their supply chain? (10 marks)
        ii. How can Data Science be useful in addressing these problems? (10 marks)

3. Ameritalent, an international entertainment talent agency is creating an information system that will allow artists and production companies to communicate online. The goal of the system is to allow production companies to respond to movie or other production projects quickly and more efficiently, saving artists unnecessary visits to the venues. This will be a major procedural change. How can Data Science be useful in evaluating whether the new system is operating as desired. Consider the following questions and use examples to respond. (20 marks)
    a. What is regression analysis? What assumptions underpin it.(4 marks)
    b. How can regression analysis be useful in assessing the impact of the new system proposed? (4 marks)
    c. What is an outlier? (2 marks)
    d. How can you identify outliers in your dataset? (4 marks)

e. What may outliers look like in assessing the results of Ameritalent's new system? (6 marks)

4. Guaraldi and Associates is a management consulting firm located in Manhattan. The firm is interested in examining the relationship between the amount of vacation that their consultants use per year and how long the consultant has been with the firm. All consultants at Guaraldi and Associates, other than the Managing Partners, are considered either Junior Consultants or Senior Consultants based on their skills and expertise. The file *GuaraldiConsultants* contains data on all Junior and Senior Consultants at Guaraldi and Associates including the years of service the consultant has at the firm and the number of hours that the consultant took as vacation last year.

   a. Create a scatter chart and add a trendline to examine the relationship between years of service and amount of vacation time used for all consultants at Guaraldi and Associates. Based on the scatter charts, what appears to be the relationship between years of service and amount of vacation time used? (5 marks)

   b. Create a scatter chart for the same data, but this time differentiate the marks in the scatter chart based on whether the consultant is junior or senior. Do you see the same relationship in this scatter chart individually for junior and senior consultants as you saw in part a for all consultants? (5 marks)

   c. What additional data may be of value in this analysis? Explain how you would incorporate it in your analysis. (5 marks)

   d. How can the CRISP-DM model be applied to Guaraldi and Associates problem described above? (5 marks)

GuaraldiConsultants data –
Google Sheets
https://docs.google.com/spreadsheets/d/1SJd8T5h7Q1Sncf62Crluy6JxqQh-RJMd/edit?usp=sharing&ouid=104253394941702640723&rtpof=true&sd=true

OR

GitHub
https://raw.githubusercontent.com/chrisnjunge/DStest/master/GuaraldiConsultants.xlsx%20-%20Sheet2.csv

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8205: OPTIMIZATION FOR DATA SCIENCE

Date: 27TH FEBRUARY, 2024 **Time: 3 Hours**

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

QUESTION ONE (20 MARKS)

i. Read the Questions keenly, before selecting the appropriate option(s). Do not re-write the question. **[3 marks]**

1. A feasible solution to a linear programming problem
   a. Must satisfy all the constraints of the problem simultaneously
   b. Need not satisfy all the constraints, only some of them.
   c. Must be a corner point of the feasible region.
   d. Must optimize the value of the objective.
2. If any coefficient in the indicator row of the simplex tableau is negative, then the solution is _____.
   a. Infeasible
   b. Feasible
   c. Bounded
   d. No solution
3. Corner points of a feasible region are located at the intersections of the region and coordinate axes.
   a. True
   b. False

ii. The feasible region of a linear programming problem has extreme points: A(0,0), B(1,1), C(0,1), and D(1,0). Identify an optimal solution for the objective function $Minimize\ z = 2x - 2y$ **[2 marks]**

iii. A company has three factories and three distribution centres for a particular product. If the capacities of each factory, the requirements of each centre and the costs of transporting an item from any factory to any centre are known.

	Centres			
	1	2	3	Availability
D	55	30	40	40
E	35	30	100	20
F	40	60	95	40
	50	20	30	100

Requirements

Form a Linear Optimization Model **[3 marks]**

iv. A small toy manufacturing firm has 200 square of felt, 600 ounces of stuffing and 90 feet of trim available to make two types of toys, a small bear and a monkey. The bear requires 1 square of felt and 4 ounces of staffing. The monkey requires 2 squares of felt, 3 ounces of stuffing, and 1 foot of trim. The firm makes \$1 profit on each bear and \$1.50 profit on each monkey. The linear program to maximize profit is

$$\text{Maximize } z = x_1 + 1.5x_2$$

$$\text{Subject to: } \begin{aligned} x_1 + 2x_2 &\leq 200 \\ 4x_1 + 3x_2 &\leq 600 \\ x_2 &\leq 90 \\ x_1,\ x_2 &\geq 0 \end{aligned}$$

What is the corresponding dual problem? **[2 marks]**

v. Three machine shops A, B, C produce three types of products X, Y, Z respectively. Each product involves operation of each of the machine shops. The time required for each operation on various products is given as follows:

	Machine Shops			Profit per unit
Products	A	B	C	
X	10	7	2	\$12
Y	2	3	4	\$3
Z	1	2	1	\$1
Available Hours	100	77	80	

Formulate the LOM and set up the first TWO simplex tableaus that would be use in finding the optimal solution. **[10 marks]**

QUESTION TWO (20 MARKS)

Consider the following integer programming problem (IP):

$$\begin{aligned} Maximize\ z &= 3x_1 - x_2 + 2x_3 \\ & x_1 - x_2 + x_3 \leq 5 \\ Subject\ to: & \ 2x_1 + x_3 \leq 4 \\ & x_1 \leq 3 \\ & x_1, x_2, x_3 \geq 0; \quad x_1, x_2, x_3 \in \mathbb{Z} \end{aligned}$$

a. Write the LP relaxation (P1) of (IP) and explain why the objective value of an optimal solution to (P1) is an upper bound on the value of an optimal solution to (IP). **[5 marks]**

b. Standardize (P1) and set up the first simplex tableau. **[4 marks]**

c. Explain how x_1 may be brought into the basic solution and why this will increase the current objective value. Perform a pivot step that brings x_1 into the basic and explain how you select the variable to leave the basis in that step. **[4 marks]**

d. After several iterations we obtain the following simplex tableau:

Basic solution	z	x_1	x_2	x_3	s_1	s_2	s_3	Solution
z	1	0	0	0	$\frac{5}{3}$	$\frac{1}{3}$	$\frac{4}{3}$	$13\frac{2}{3}$
x_1	0	1	0	0	0	0	1	3
x_2	0	0	1	0	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
x_3	0	0	0	1	$\frac{2}{3}$	$\frac{1}{3}$	$-\frac{2}{3}$	$\frac{8}{3}$

Argue that we have found an optimal solution to (P1). State the solution and its objective value. **[4 marks]**

e. If the dual problem (DP1) of (P1) with variables y_1, y_2, y_3 is:

$$\begin{aligned} Minimize\ z &= 5y_1 + 4y_2 + 3y_3 \\ & y_1 + y_3 \geq -3 \\ Subject\ to: & \ 2y_2 - y_1 \geq 1 \\ & y_1 + y_2 \geq -2 \\ & y_1, y_2, y_3 \geq 0 \end{aligned}$$

Show that $(y_1, y_2, y_3) = \left(\frac{5}{3}, \frac{1}{3}, \frac{4}{3}\right)$ is an optimal solution to (DP1). **[3 marks]**

17 marks.

QUESTION THREE (20 MARKS)

a. Consider the following knapsack problem:

$$\begin{aligned} Maximize\ z &= 50x_1 + 60x_2 + 140x_3 + 40x_4 \\ s.t. \quad & 5x_1 + 10x_2 + 20x_3 + 20x_4 \leq 30 \\ & x_1, x_2, x_3, x_4 \in \{0,1\} \end{aligned}$$

State the LP-relaxation and show how to find an optimal solution to this without using the simplex method. **[6 marks]**

b. Solve the knapsack problem above to optimality using branch and bound. Show the B&B tree in each step and explain briefly (with justification) what you conclude in each step of the algorithm. This includes which nodes in the B&B tree you must continue exploring and which nodes you can finish (and with what conclusion). **[14 marks]**

QUESTION FOUR (20 MARKS)

Apply Gomory's cutting plane method to solve the model below: **[20 marks]**

$$\begin{aligned} Min\ z &= 6x_1 + 8x_2 \\ & 3x_1 + x_2 \geq 4 \\ s.t. \quad & x_1 + 2x_2 \geq 4 \\ & x_1, x_2 \geq 0 \ and\ integer \end{aligned}$$

QUESTION FIVE (20 MARKS)

Joe State lives in Gary, Indiana and owns insurance agencies in Gary, Fort Wayne, Evansville, Terre Haute, and South Bend.

Each December, he visits each of his insurance agencies.

The distance between each agency:

Miles	*G*	*FW*	*E*	*TH*	*SB*
G	0	132	217	164	58
FW	132	0	290	201	79
E	217	290	0	113	303
TH	164	201	113	0	196
SB	58	79	303	196	0

What order of visiting his agencies will minimize the total distance traveled? **[20 marks]**

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

AND ILAB AFRICA

MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS

END OF SEMESTER EXAMINATION

DSA 8202 TIME SERIES ANALYSIS AND FORECASTING

DATE: 28th February 2024 Time: 3 Hours

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and you may choose any other **TWO** additional questions (from Question 2 to 6). Each question is worth 20 marks (This exam has a total of 60 points).
3. Except for Question ONE multiple choice, you must show all of your working for full credit in answering the questions. Correct answers without any working will not earn you any points.

++

Question ONE (COMPULSORY)

1-A. Multiple Choice Questions (10 points total – no need to explain)

1) The concept of "spurious regression" refers to a situation where
a. two or more non-stationary series are regressively analyzed, leading to misleadingly high R-squared values.
b. the regression model mistakenly includes too many lagged terms.
c. the residuals of the model are perfectly autocorrelated.
d. the model parameters are incorrectly identified as significant due to random chance.

2) Which of the following is not a characteristic of a cointegrated time series?
a. The series move together over time and have a long-run equilibrium relationship.
b. The series can be represented by an error correction model.
c. Each series is stationary at their level.
d. Deviations from the long-run equilibrium are temporary.

3) In the context of time series analysis, Granger causality tests are used to determine if

a. one variable can predict another variable.
b. two series have a unit root.
c. the series is stationary.
d. residuals are normally distributed.

4) The presence of seasonality in a time series implies that
a. there is a deterministic trend in the data.
b. the series exhibits regular patterns at specific periodic intervals.
c. the time series is non-stationary.
d. autocorrelation within the series is always positive.

5) When a time series model includes both autoregressive and moving average terms, it is specifically referred to as a(n)
a. AR model.
b. MA model.
c. ARMA model.
d. VAR model.

6) Impulse Response Functions (IRFs) in the context of Vector Autoregression (VAR) models are used to
a. predict the future values of a time series based on past shocks.
b. determine the causality direction between two time series.
c. measure the response of one variable to a one-time shock to another variable while holding everything else constant.
d. estimate the long-run equilibrium relationship between variables in a cointegrated system.

7) Which of the following is not a method for testing the stationarity of a time series?
a. Phillips-Perron test.
b. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.
c. Engle-Granger test.
d. Augmented Dickey-Fuller test.

8) In vector autoregression (VAR) models, the optimal number of lags is often determined using
a. the Cochrane-Orcutt procedure.
b. the Durbin-Watson statistic.
c. information criteria like AIC or BIC.
d. the Phillips-Perron test.

9) Which of the following is true about forecasting with ARIMA models?
a. The differencing order must be determined after estimating the model.
b. The model can only forecast stationary series.
c. The model parameters include the number of AR terms, differencing, and MA terms.
d. Forecast accuracy decreases as the forecast horizon increases.

10) Non-stationary time series data can be transformed into stationary data by
a. applying exponential smoothing.
b. taking the natural logarithm of the series.
c. differencing the series.
d. increasing the sample size.

gretl: ADF test

Augmented Dickey-Fuller test for YNS
including one lag of (1-L)YNS
sample size 298
unit-root null hypothesis: a = 1

test without constant
model: (1-L)y = (a-1)*y(-1) + ... + e
1st-order autocorrelation coeff. for e: -0.207
estimated value of (a - 1): -0.00608704
test statistic: tau_nc(1) = -1.43452
asymptotic p-value 0.1415

Augmented Dickey-Fuller regression
OLS, using observations 3-300 (T = 298)
Dependent variable: d_YNS

	coefficient	std. error	t-ratio	p-value
YNS_1	-0.00608704	0.00424327	-1.435	0.1415
d_YNS_1	0.506102	0.0503010	10.06	1.12e-020 ***

AIC: -447.818 BIC: -440.424 HQC: -444.858

1-B. (Each sub-question is worth 2 points x 5 = 10 points total)

a) Compare and contrast a white noise process with a stationary process.

b) Sketch possible ACF and PACF for an ARMA (1,2) process.

c) Below are results from a unit root test in GRETL.

gretl: ADF test

```
Augmented Dickey-Fuller test for YNS
including one lag of (1-L)YNS
sample size 298
unit-root null hypothesis: a = 1

   test without constant
   model: (1-L)y = (a-1)*y(-1) + ... + e
   1st-order autocorrelation coeff. for e: -0.207
   estimated value of (a - 1): -0.00608704
   test statistic: tau_nc(1) = -1.43452
   asymptotic p-value 0.1415

Augmented Dickey-Fuller regression
OLS, using observations 3-300 (T = 298)
Dependent variable: d_YNS

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  YNS_1      -0.00608704   0.00424327   -1.435    0.1415
  d_YNS_1     0.506102     0.0503010    10.06     1.12e-020 ***

  AIC: -447.818   BIC: -440.424   HQC: -444.858
```

```
Critical Values:
        1%: -3.432
        5%: -2.862
       10%: -2.567
```

Note:

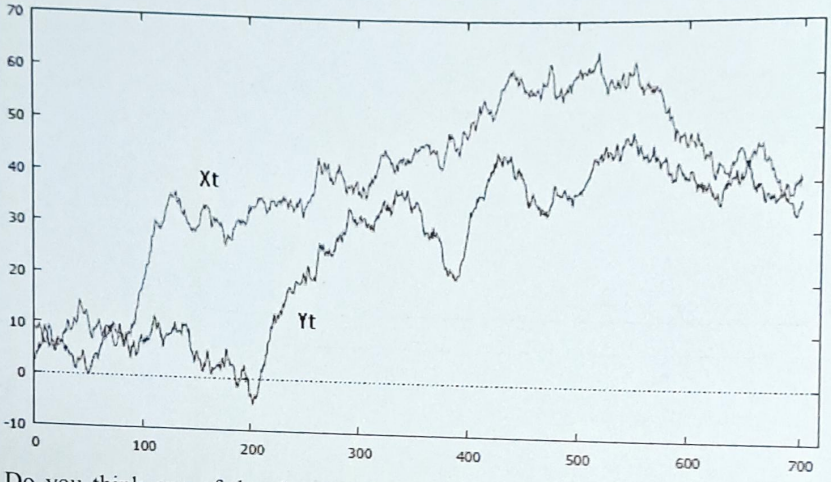
State the null and alternatively clearly for the test. What do you conclude?

d) Can we improve the Augmented Dickey-Fuller test above by adding more lags? Explain your answer.

e) Explain 3 kinds of non-stationarity we may come across in time series analysis (use figures and equations for your answer).

Question TWO (Each sub-question is 5 points x 4 = 20 points total)

Suppose we have two time series X_t and Y_t observed over a period of t=700. A time-series plot is provided below:



(a) Do you think any of the above series are stationary? Are any non-stationary? Which one(s)?

(b) Do you think the two series could be cointegrated? What do you mean by cointegration?

(c) Assuming that Y_t and X_t are I(1) processes, you decide to regress Y_t on X_t by OLS giving:

$$\widehat{Yt} = \underset{(0.87052)}{-7.33993} + \underset{(0.020501)}{0.837184}\,Xt$$

$$T = 700 \quad \bar{R}^2 = 0.7045 \quad F(1, 698) = 1667.6 \quad \hat{\sigma} = 8.5326$$

(standard errors in parentheses)

Would you accept the above results? What would you be worried about? How would you check that your regression is valid?

(d) Alternative specifications to understanding the relationship between Y_t and X_t are:

Dynamic model: $\widehat{d\_Yt} = \underset{(0.037586)}{0.0469393} + \underset{(0.038184)}{0.0279278}\,d\_Xt$

Error Correction Model (ECM): $\widehat{d\_Yt} = \underset{(0.037460)}{0.0468043} + \underset{(0.038056)}{0.0277219}\,d\_Xt - \underset{(0.0043928)}{0.0104865}\,uhat\_1$

where d_Y_t denotes the first difference of Y_t, d_X_t denotes the first difference of X_t, and uhat_1 is the residual from the model in (c) above.

Of the two models (dynamic and ECM), which one would you accept and why?

Question THREE (Each sub-question is 5 points x 4 = 20 points total)

(a) Consider an AR(1) model, $x_t = \varphi x_{t-1} + u_t$. What are the conditions on φ that make x_t stationary? How about non-stationary?

(b) The result below is for some variable x_t.

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.868177	0.334573	2.594884	0.0102
AR(1)	0.975461	0.019471	50.09854	0.0000
MA(1)	-0.909851	0.039596	-22.97840	0.0000

Write down the ARMA(p,q) model in full.

(c) What are the conditions for the ARMA model in (b) to be valid for forecasting?

(d) Given the above results, forecast the one-period-ahead value for x_{t+1}, given $x_t = 100$ and $e_t = 5$

Question FOUR (Each sub-question is 5 points x 4 = 20 points total)

a) Explain how the concepts of overfitting and underfitting are intrinsically related forecasting performance?

b) Explain the difference between static and dynamic forecasting.

c) Here are actual and forecasted values for t = 10.

T	1	2	3	4	5	6	7	8	9	10
Actual	250	110	500	200	330	490	670	210	435	375
Forecast	265	140	480	215	290	515	750	210	420	285

Find the Mean Absolute Error (MAE), Mean Square Error (MSE) and the Root Mean Square Forecasting Error (RMSFE).

d) Compare and comment on the values of the RMSE and MAE above.

Question FIVE (Each sub-question is 5 points x 4 = 20 points total)

a) Derive the mean and variance of an AR(1) process. What are the conditions for the AR(1) process to be stationary?

b) In time series, what is the difference between autocorrelation and partial autocorrelation?

c) Given a time series, explain how you could check for the presence of an ARCH effect.

d) Clearly explain the ARCH and GARCH models. What are the advantages of GARCH over ARCH model?

Question SIX (Each sub-question is 5 points x 4 = 20 points total)

Suppose that industrial production (IP), money supply (M) and trade balance (tb) are jointly determined by a VAR(2,2) model (with a constant be the only exogenous variable)

(a) Write out the VAR(2,2) model in equation form

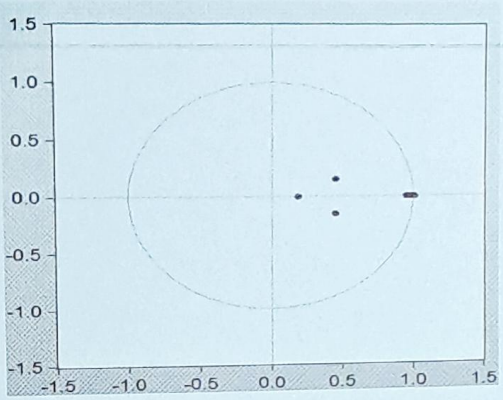
(b) Suppose you ask GRETL to calculate some information criteria as follows:

VAR Lag Order Selection Criteria
Endogenous variables: IP M1 TB3
Exogenous variables: C
Date: 06/02/15 Time: 13:07
Sample: 1959M02 1995M04
Included observations: 428

Lag	LogL	LR	FPE	AIC	SC	HQ
0	NA	NA	22944378	25.46221	25.49067	25.47345
1	NA	7306.121	0.786018	8.272854	8.386661	8.317801
2	NA	264.4260	0.437446	7.686819	7.885982*	7.765478
3	NA	50.97121	0.403867	7.606935	7.891453	7.719304
4	NA	38.17713	0.384206	7.556998	7.926871	7.703078*
5	NA	8.899167	0.392165	7.577454	8.032683	7.757244
6	NA	27.72597	0.382229	7.551720	8.092305	7.765221
7	NA	39.14914*	0.362037*	7.497350*	8.123290	7.744562
8	NA	11.50713	0.367005	7.510852	8.222147	7.791775

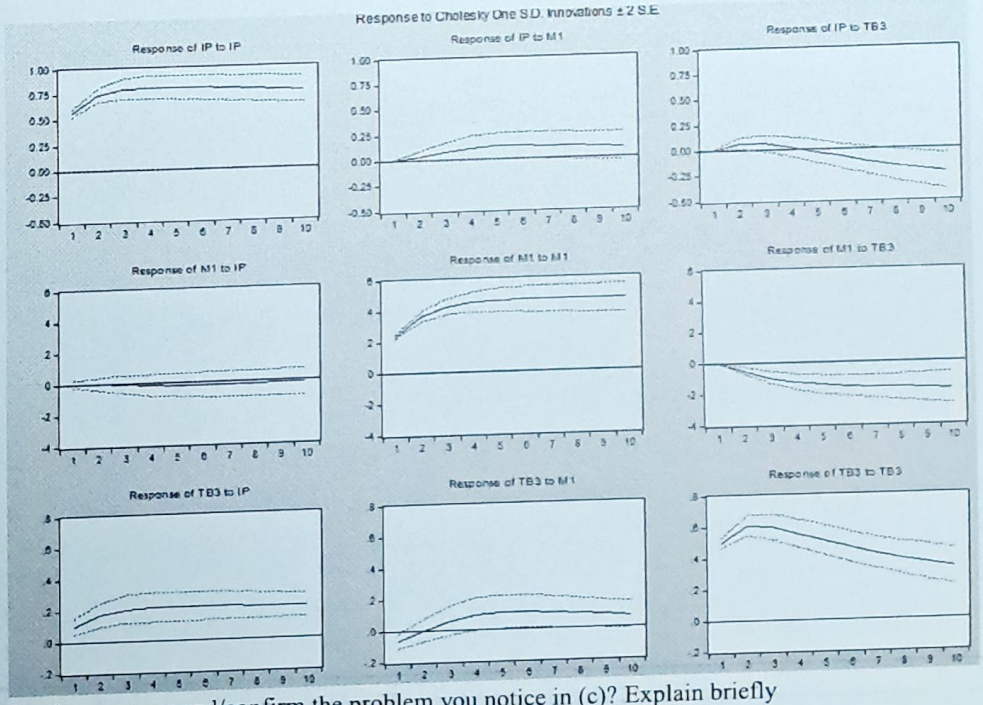
What do you conclude?

(c) However, the inverse roots of the AR characteristic polynomial as follows



What seems to be the problem?

(d) Your impulse response functions are as follows:



Do these correspond/confirm the problem you notice in (c)? Explain briefly

(e) What would you do to solve the problem? What kind of model would you consider running? Briefly explain.

Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
AND ILAB AFRICA
MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS
END OF SEMESTER EXAMINATION
DSA 8202: TIME SERIES ANALYSIS AND FORECASTING

DATE: 1ST March 2023 Time: 3 Hours

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions (from Question 2 to 6). Each question is worth 20 marks (This exam has total 60 points).
3. Except for Question ONE multiple choice, you must show all of your working for full credit in answering the questions. Correct answers without any working will not earn you any points.

++

Question ONE (COMPULSORY)

1-A. Multiple Choice Questions (10 points – no need to explain)

1) In time series econometrics, pseudo out-of-sample forecasting can be used for the following reasons with the exception of
   a. giving the forecaster a sense of how well the model forecasts at the end of the sample.
   b. estimating the RMSFE (root mean square forecasting error).
   c. analyzing whether or not a time series contains a unit root.
   d. evaluating the relative forecasting performance of two or more forecasting models.

2) Autoregressive models include
   a. current and lagged values of the error term.
   b. lags of the dependent variable, and lagged values of additional predictor or explanatory variables.
   c. current and lagged values of the residuals.
   d. lags and leads of the dependent variable.

3) Negative autocorrelation in the change of a variable implies that
   a. the variable contains only negative values.
   b. the series is not stable.
   c. an increase in the variable in one period is, on average, associated with a decrease in the next.
   d. the data are negatively trended.

4) Stationarity means that the
   a. error terms are not correlated.
   b. probability distribution of the time series variable does not change over time.
   c. time series has a unit root.
   d. forecasts remain within 1.96 standard deviation outside the sample period.

5) To choose the number of lags in either an autoregression or a time series regression model with multiple predictors, you can use any of the following test statistics with the exception of the
   a. *F*-statistic.
   b. Akaike information criterion.
   c. Bayes information criterion.
   d. augmented Dickey-Fuller test.

6) The random walk model is an example of a
   a. deterministic trend model.
   b. binomial model.
   c. stochastic trend model.
   d. stationary model.

7) Heteroskedasticity- and autocorrelation-consistent standard errors
   a. result in the OLS estimator being BLUE.
   b. should be used when errors are autocorrelated.
   c. are calculated when using the Cochrane-Orcutt iterative procedure.
   d. have the same formula as the heteroskedasticity robust standard errors in cross-sections.

8) The following is not a consequence of X_t and Y_t being cointegrated:
   a. X_t and Y_t are both I(1), then for some θ, $Y_t - \theta X_t$ is I(0).
   b. X_t and Y_t have the same stochastic trend.
   c. in the expression $Y_t - \theta X_t$, θ is called the cointegrating coefficient.
   d. integrating one of the variables gives you the same result as integrating the other.

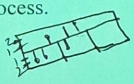
9) The order of integration
   a. can never be zero.
   b. is the number of times that the series needs to be differenced for it to be stationary.
   c. is the value of ϕ_1 in the quasi difference ($\Delta Y_t - \phi_1 Y_{t-1}$).
   d. depends on the number of lags in the VAR specification.

10) To test the null hypothesis of a unit root, the ADF test
   a. has higher power than the so-called DF-GLS test.
   b. uses complicated iterative techniques.
   c. cannot be calculated if the variable is integrated of order two or higher.
   d. uses a test-statistic and a special critical value.

1-B. (10 points)

a) What is a white noise process? What is a stationary process? What is the difference between a white noise process and a stationary process?

b) Sketch possible ACF and PACF for an ARMA(2,1) process.



c) Below are results from a unit root test.

Null Hypothesis: TBILL has a unit root
Exogenous: Constant
Bandwidth: 3.82 (Andrews using Bartlett kernel)

		Adj. t-Stat	Prob.*
Phillips-Perron test statistic		-1.519035	0.5223
Test critical values:	1% level	-3.459898	
	5% level	-2.874435	
	10% level	-2.573719	

*MacKinnon (1996) one-sided p-values.

Residual variance (no correction)	0.141569
HAC corrected variance (Bartlett kernel)	0.107615

State the null for the test. What do you conclude?

d) What does it mean to de-trend a series? Explain why de-trending is useful in time series analysis.

e) Briefly explain what use Theil (U2) statistics is? And what is it used for?

++

Question TWO

Suppose we have two time series X_t and Y_t observed over a period of t=700. In fact the specific equations (data generating process - DGP) are $X_t = X_{t-1} + e_t$ and $Y_t = Y_{t-1} + e_t$, where $e_t \sim N(0,1)$.

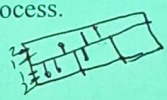
10) To test the null hypothesis of a unit root, the ADF test

a. has higher power than the so-called DF-GLS test.
b. uses complicated iterative techniques.
c. cannot be calculated if the variable is integrated of order two or higher.
d. uses a test-statistic and a special critical value.

1-B. (10 points)

a) What is a white noise process? What is a stationary process? What is the difference between a white noise process and a stationary process? white noise: ts noise with no pattern ≡ variance mean no autocor

b) Sketch possible ACF and PACF for an ARMA(2,1) process.



c) Below are results from a unit root test.

Null Hypothesis: TBILL has a unit root
Exogenous: Constant
Bandwidth: 3.82 (Andrews using Bartlett kernel)

		Adj. t-Stat	Prob.*
Phillips-Perron test statistic		-1.519035	0.5223
Test critical values:	1% level	-3.459898	
	5% level	-2.874435	
	10% level	-2.573719	

*MacKinnon (1996) one-sided p-values.

Residual variance (no correction)	0.141569
HAC corrected variance (Bartlett kernel)	0.107615

p-f = high critical value less non-stationary

State the null for the test. What do you conclude?

d) What does it mean to de-trend a series? Explain why de-trending is useful in time series analysis. remove long term isolate short term fluctu'

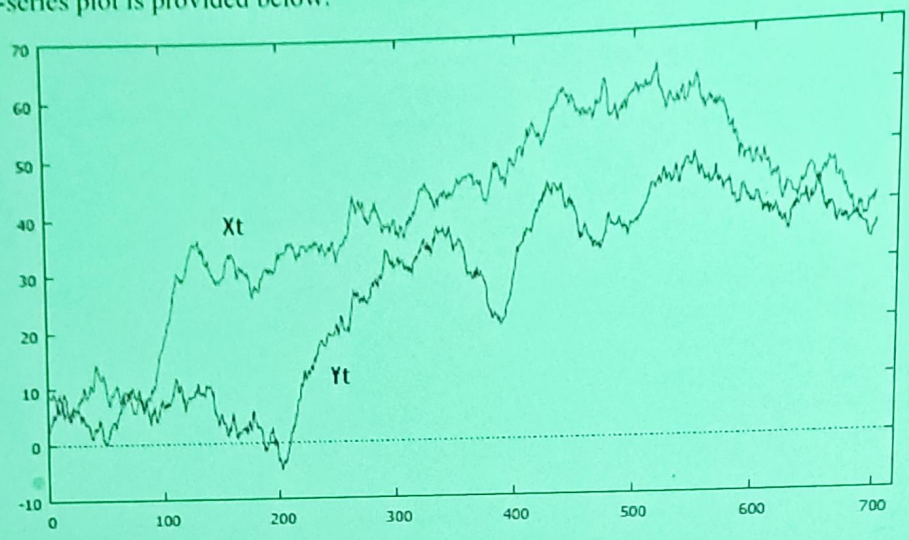
e) Briefly explain what use Theil (U2) statistics is? And what is it used for?

++

Question TWO

Suppose we have two time series X_t and Y_t observed over a period of t=700. In fact the specific equations (data generating process - DGP) are $X_t = X_{t-1} + e_t$ and $Y_t = Y_{t-1} + e_t$, where $e_t \sim N(0,1)$.

A time-series plot is provided below:



(a) Do you think the series are stationary or non-stationary? If they are non-stationary, what kind of non-stationarity would describe the above series best?

(b) To test for stationary, you decide to run Y_t on its lag, Y_{t-1}. Estimation by OLS gives the following:

$$\widehat{Yt} = \underset{(0.071910)}{0.131553} + \underset{(0.0023905)}{0.996752}\ Yt\_1$$

$$T = 699 \quad \bar{R}^2 = 0.9960 \quad F(1, 697) = 1.7386e{+}005 \quad \hat{\sigma} = 0.99174$$

(standard errors in parentheses)

Does it make sense to test the null hypothesis H_0: $\beta = 1$ to establish a unit root or non-stationary? Explain why or why not. no - assume

(c) You try a different specification, running the first difference of Y_t or d_Y_t on the lag of Y_t. Estimation by OLS gives:

$$\widehat{d\_Yt} = \underset{(0.071910)}{0.131553} - \underset{(0.0023905)}{0.00324768}\ Yt\_1$$

$$T = 699 \quad \bar{R}^2 = 0.0012 \quad F(1, 697) = 1.8458 \quad \hat{\sigma} = 0.99174$$

(standard errors in parentheses)

9.45997
7.76357
0.99600
0.99174
0.00426

Based on the above results, is Y_t stationary or non-stationary? State clearly the null and alternative hypothesis. (Use critical Dickey-Fuller statistic of -2.87 for t > 500 and 5% level of significance.)

(d) Assume Y_t and X_t are non-stationary or more specifically I(1). To understand the relationship between Y_t and X_t we regress one on the other by OLS giving:

$$\widehat{Yt} = \underset{(0.87052)}{-7.33993} + \underset{(0.020501)}{0.837184}\,Xt$$

$$T = 700 \quad \bar{R}^2 = 0.7045 \quad F(1,698) = 1667.6 \quad \hat{\sigma} = 8.5326$$

(standard errors in parentheses)

Would you accept the above results? What in particular would you be worried about? How would you check that your regression is valid?

(e) Alternative specifications to understanding the relationship between Y_t and X_t are:

Dynamic model: $$\widehat{d\_Yt} = \underset{(0.037586)}{0.0469393} + \underset{(0.038184)}{0.0279278}\,d\_Xt$$

Error Correction Model (ECM): $$\widehat{d\_Yt} = \underset{(0.037460)}{0.0468043} + \underset{(0.038056)}{0.0277219}\,d\_Xt - \underset{(0.0043928)}{0.0104865}\,uhat\_1$$

where d_Y_t denotes the first difference of Y_t, d_X_t denotes the first difference of X_t, and uhat_1 is the residual from the model in (d) above.

What does the dynamic model tell you about the relationship between Y_t and X_t ? How about the ECM model? Of the two models (dynamic and ECM), which one would you accept and why?

Question THREE

The following least squares residuals come from a sample of T=10.

T	1	2	3	4	5	6	7	8	9	10
$\hat{e}_t$	0.28	-0.31	-0.09	0.03	-0.37	-0.17	-0.39	-0.03	0.03	1.02

(a) Calculate the Durbin-Watson statistic, $DW = \dfrac{\sum_{t=2}^{T}(\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T}\hat{e}_t^2}$. Interpret your results.

(b) Consider an AR(1) model, $x_t = \varphi x_{t-1} + u_t$: What do you think are the conditions on φ that make x_t stationary? How about non-stationary?

(c) The result below is for some variable x_t.

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.868177	0.334573	2.594884	0.0102
AR(1)	0.975461	0.019471	50.09854	0.0000
MA(1)	-0.909851	0.039596	-22.97840	0.0000

Write down the ARMA(p,q) model in full.

(d) What are the conditions for the ARMA model in (c) to be valid for forecasting?

(e) Given the above results, forecast the one-period-ahead value for x_{t+1}, given $x_t = 100$ and $e_t = 5$

Question FOUR

a) Explain the difference between static and dynamic forecasting.

The series below is on Thailand's actual monthly inflation and forecasted inflation by AR(2) by dynamic forecasting method for the year 2014.

date	Actual inflation (π)	Forecast 1 (f1)
2013 - Dec	3.63	
2014 - Jan	3.39	3.60
2014 - Feb	3.23	3.58
2014 - Mar	2.69	3.56
2014 - Apr	2.42	3.54
2014 - May	2.27	3.52
2014 - Jun	2.25	3.50
2014 - Jul	2.00	3.48
2014 - Aug	1.59	3.47
2014 - Sep	1.42	3.45
2014 - Oct	1.46	3.44
2014 - Nov	1.92	3.42
2014 - Dec	1.67	3.41

Evaluate the forecasting performance of forecast 1 (generated by an AR(2) model) with a naïve forecast model (Random Walk) using

b) Root mean square forecasting error (RMSFE) and Mean absolute error (MAE).

c) Bias, Variance and Covariance Proportion

d) What do you conclude from the above? Which is better, AR(2) or Random Walk for forecasting?

Question FIVE

a) Derive the mean and variance of an AR(1) process. What are the conditions for the AR(1) process to be stationary?

b) In time series, what is the difference between autocorrelation and partial autocorrelation?

c) Given a time series, explain how you could check for the presence of an ARCH effect.

d) Clearly explain the ARCH and GARCH models. What are the advantage of GARCH over ARCH model?

Question SIX

Suppose that industrial production (IP), money supply (M) and trade balance (tb) are jointly determined by a VAR(2) model (with a constant be the only exogenous variable)

(a) Write out the model in equation form

(b) What is the main advantage of VAR models over simpler models like the ARMA?

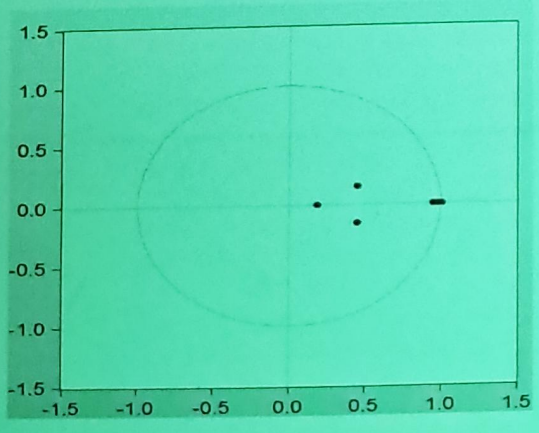
(c) Suppose you ask GRETL to calculate some information criteria as follows:

VAR Lag Order Selection Criteria
Endogenous variables: IP M1 TB3
Exogenous variables: C
Date: 06/02/15 Time: 13:07
Sample: 1959M02 1995M04
Included observations: 428

Lag	LogL	LR	FPE	AIC	SC	HQ
0	NA	NA	22944378	25.46221	25.49067	25.47345
1	NA	7306.121	0.786018	8.272854	8.386661	8.317801
2	NA	264.4260	0.437446	7.686819	7.885982*	7.765478
3	NA	50.97121	0.403867	7.606935	7.891453	7.719304
4	NA	38.17713	0.384206	7.556998	7.926871	7.703078*
5	NA	8.899167	0.392165	7.577454	8.032683	7.757244
6	NA	27.72597	0.382229	7.551720	8.092305	7.765221
7	NA	39.14914*	0.362037*	7.497350*	8.123290	7.744562
8	NA	11.50713	0.367005	7.510852	8.222147	7.791775

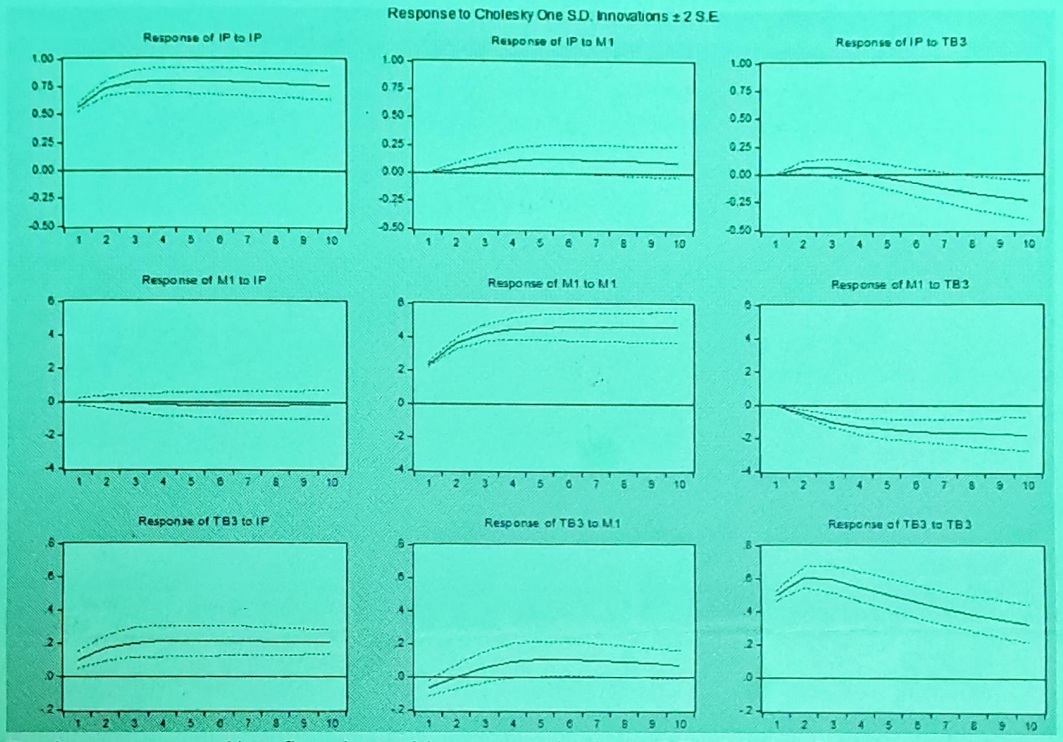
What do you conclude?

(d) However, the inverse roots of the AR characteristic polynomial in EViews is as follows



What seems to be the problem?

(e) Your impulse response functions are as follows:



Do these correspond/confirm the problem you notice in (c) ? Explain briefly

(f) What would you do to solve the problem? What kind of model would you consider running? Briefly explain.

(g) Mention other things you would check to see whether your model fits the data well.

(h) Explain briefly the intuition of a reduced-VAR and a structural VAR? When would you run either a reduced or structural VAR?

(i) Now if you were consider running a VECM model here, what would you have to consider?

(j) Write down a possible VECM model for industrial production (IP), money supply (M) and trade balance (tb) based on all you have done above.