



STRATHMORE BUSINESS SCHOOL
BACHELOR OF COMMERCE
SPECIAL EXAMINATION
BFS 4102: ADVANCED BUSINESS DATA ANALYTICS

DATE: Friday, 8th May 2026

TIME: 10:00 – 12:00

Instructions

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

QUESTION ONE

(30 MARKS)

- (a) In Exploratory Data Analysis, data is often classified by type. Define nominal, ordinal, interval, and ratio data, giving an example of each. **(5 marks)**
- (b) The following output is from a simple linear regression predicting Customer Lifetime Value (CLV) using Purchase Frequency:

Coefficients:

	Estimate	Std. Error	t-value	p-value
(Intercept)	52.25	0.90	57.79	<2e-16
Purchase Freq.	19.54	0.17	114.91	<2e-16

$R^2 = 0.9637$, Adjusted $R^2 = 0.9636$, F-statistic = 1.32e+04 (p < 0.001)

Tasks:

- (i) Interpret the meaning of the coefficient of Purchase Frequency. **(1 mark)**
 - (ii) Explain the practical meaning of R^2 . **(1 mark)**
 - (iii) Interpret what the p-value of the F-statistic implies? **(1 mark)**
 - (iv) What does the intercept represent? **(1 mark)**
 - (v) Is this model suitable for business decision-making? Justify your answer. **(1 mark)**
- (c) What is Diagnostic Exploratory Data Analysis (EDA) in the context of statistical tests of hypothesis? Describe **three** diagnostic checks or tests used to validate regression assumptions. **(5 marks)**
- (d) Differentiate between statistical tests of regression, comparison, and correlation, giving one example of each. **(5 marks)**

- (e) Give **two** examples of financial data that are typically analyzed using time series forecasting. Explain why each is suitable for this method. **(5 marks)**
- (f) Discuss one limitation of time series forecasting when applied to financial markets and how it can be addressed. **(5 marks)**

QUESTION TWO (15 MARKS)

- (a) Identify **two** business decisions in banking or insurance that can be improved using time series forecasting. **(5 marks)**
- (b) A Securities Exchange has provided you with 5 years of daily closing prices for Company X. Describe the systematic approach you would use to design an ARIMA model to forecast the next 30 days' closing prices. Include all necessary data transformation and model evaluation steps. **(5 marks)**
- (c) Suppose you are analyzing a country's quarterly GDP data (2016–2025) to forecast future growth. Explain how you would test for and achieve stationarity. **(5 marks)**

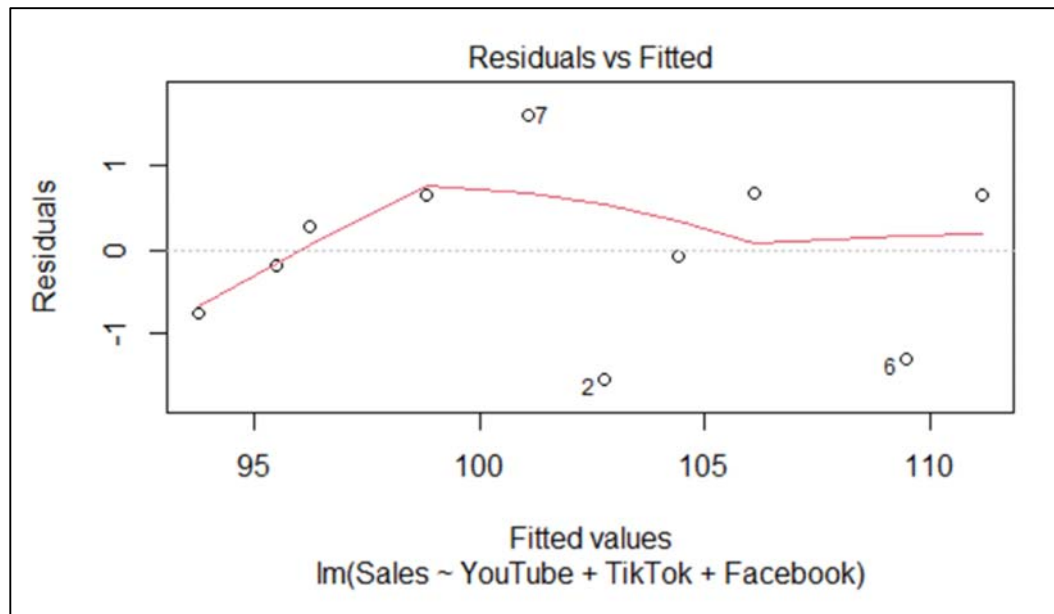
QUESTION THREE (15 MARKS)

- (a) A retail company experiences strong December sales each year. Explain how you would decompose its monthly sales time series into trend, seasonal, and irregular components. Discuss how each component informs managerial decisions. **(5 marks)**
- (b) A supermarket chain is planning its inventory for Week 8. Overstocking increases holding costs, while understocking leads to lost sales and dissatisfied customers.

The weekly sales (in cartons) of a perishable dairy product are shown below:

Week	Sales (Cartons)
1	39
2	44
3	40
4	45
5	38
6	43
7	39
8	?

- (i) Use a 3-Week Simple Moving Average to forecast the sales in cartons for Week 8 **(1 mark)**
- (ii) Calculate the Mean Absolute Percentage Error **(3 marks)**
- (c) Given the “residuals versus fitted” plot below:



Tasks:

- (i) Describe the pattern in the plot. **(1 mark)**
- (ii) What does a plot of the “residuals versus the fitted” values enable us to assess? **(2 marks)**
- (iii) What signifies that the assessment has passed? Has the assessment passed in this case? **(2 marks)**

QUESTION FOUR

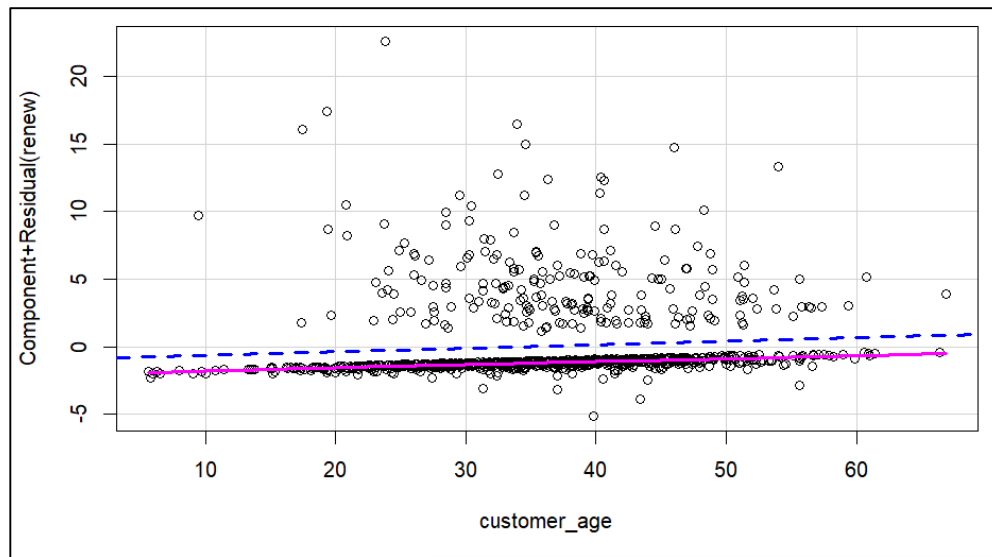
(15 MARKS)

- (a) Differentiate, using business related examples, between the assumptions of “normality of distribution of errors” and “equal variance of errors” applied during statistical tests of regression. **(5 marks)**
- (b) Identify **two** methods used by regularization methods such as Lasso Regression and Ridge Regression reduce the complexity in linear regression models. **(5 marks)**
- (c) A marketing analytics team is using k-Means Clustering to group customers based on purchasing behavior. Explain the purpose of calculating the Within-Cluster Sum of Squares (WCSS) in this context and describe how the team would determine the optimal number of customer segments to target. **(5 marks)**

QUESTION FIVE

(15 MARKS)

- (a) Explain the difference between autocorrelation and multicollinearity in the context of regression analysis. In your answer, specify how each is tested (e.g., using the Durbin-Watson test and the Variance Inflation Factor), what the presence of each indicates about a regression model, and the implications they have on model interpretation and validity. **(5 marks)**
- (b) Given the “component + residual plot” below for a continuous predictor (customer’s age) in a regression model below:



Tasks:

- (i) Describe the pattern in the plot. (1 mark)
 - (ii) What does a plot of the “component + residual plot” enable us to assess? (2 marks)
 - (iii) What signifies that the assessment has passed? Has the assessment passed in this case? (2 marks)
- (c) Given the output of a logistic regression model used to predict whether a customer will renew a subscription (1 = Renew):

	Estimate	Odds Ratio	p-value
Monthly Fee	-0.04	0.96	<0.001
Customer Age	0.03	1.03	0.002
Support Calls	0.70	2.01	<0.001

- (i) Interpret the odds ratio for Support Calls. (1 mark)
- (ii) Why is the Monthly Fee coefficient negative? (1 mark)
- (iii) Calculate the percentage change in odds for Customer Age. (1 mark)
- (iv) What does an AUC of 0.72 indicate about model performance? (1 mark)
- (v) Name one limitation of logistic regression. (1 mark)