Assumption University of Thailand
Vincent Mary School of Science and Technology

Midterm Examination (online)
Semester 1/2022

Subject: ITX 2006/ CSX 2006 - Mathematics and Statistics for Data Science (Section 541)
Date: Tuesday, 2 August 2022
Time: 15:00 – 17:00
Lecturer: Dr. Khaing Sandar Htun

Instructions:

1. Read the questions carefully and answer each question **completely, legibly, and concisely**.
2. This examination is **open-book** and the use of books and lecture notes is allowed.
3. Use MS Excel, Minitab or any similar software to perform the analysis.
4. Submit all your work in **one single** PDF or Excel file to **MS Teams (Private Chat)**. Name your file as “**4**”.

Marking Scale:

The total number of marks for the 4 questions on the exam paper is 100 marks.
The total of 100 marks for this examination corresponds to 20% of the final score.

Marks Awarded	1	2	3	4	Total

Student Name: ____ **ID:** ________________________

Total: 2 Pages (excluding this page)

There are 4 questions for the total of 100 marks.

---

1. **(15 marks) Simple Linear Regression Analysis**

   A biologist assumes that there is a linear relationship between the amount of fertilizer supplied to tomato plants and the subsequent yield of tomatoes obtained.

   Eight tomato plants of the same variety were selected at random and treated, weekly, with a solution in which x grams of fertilizer was dissolved in a fixed quality of water. The yield, y kilograms, of tomatoes was recorded.

Plant	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7.0	7.1	7.3	7.7

   a) **(2 marks)** Plot a scatter diagram of yield, y, against amount of fertilizer, x?
   b) **(4 marks)** Obtain the equation of the least squares regression line of y and x. Hence obtain parameter estimates for the intercept (β_0) and the slope (β_1).
   c) **(2 marks)** How large a part of the variation in yield can be explained by the fertilizer?
   d) **(3 marks)** Estimate the yield of a plan treated weekly with 3.2 grams of fertilizer.
   e) **(4 marks)** What is a 95%-confidence interval for the expected yield for 3.2 grams of fertilizer?

2. **(30 marks) Multiple Regression Analysis**

   The Literacy rate is a reflection of the educational facilities and quality of education available in a country, and mass communication plays a large part in the educational process. In an effort to relate the literacy rate of a country to various mass communication outlets, a demographer has proposed to relate literacy rate to the following variables: number of daily newspaper copies (per 1000 population), number of radios (per 1000 population), and number of TV sets (per 1000 population). Here are the data for a sample of 10 countries:

Country	**Newspapers**	**Radios**	**TV sets**	**Literacy rate**
Czech Republic	280	266	228	0.98
Italy	142	230	201	0.93
Kenya	10	114	2	0.25
Norway	391	313	227	0.99
Panama	86	329	82	0.79
Philippines	17	42	11	0.72
Tunisia	21	49	16	0.32
USA	314	1695	472	0.99
Russia	333	430	185	0.99
Venezuela	91	182	89	0.82

   a) **(6 marks)** Develop all possible scatter diagrams between independent variables and dependent variable and interpret the meaning.
   b) **(5 marks)** Do you see any problems with multicollinearity? Explain.
   c) **(4 marks)** Define the best estimated regression equation that can be used to predict the literacy rate.
   d) **(5 marks)** Predict literacy rate for a country that has 200 daily newspaper copies (per 1000 in the population), 800 radios (per 1000 in the population), and 250 TV sets (per 1000 in the population).
   e) **(10 marks)** Is there a relationship between literacy rate and each independent variables and the independent variables combined? Explain and state a conclusion of your analysis.

3. **(20 marks) Multiple Regression Analysis with Categorical variable**

   The following data is obtained to test the usefulness of IQ and Gender in prediction of Test score.

Student	**Test score**	**IQ**	**Gender**
1	93	125	Male
2	86	120	Female

3	96	115	Male
4	81	110	Female
5	92	105	Male
6	75	100	Female
7	84	95	Male
8	77	90	Female
9	73	80	Male
10	74	80	Female

a) (**4 marks)** Express categorial variable as a single dummy variable.

b) **(4 marks)** Define the best estimated regression equation that can be used to predict the Test score using IQ and Gender.

c) **(4 marks)** Is there a relationship between Test score and the independent variables combined? Explain.

d) **(4 marks)** Does IQ, contribute the prediction of Test score? Explain.

e) **(4 marks)** Does Gender, contribute the prediction of Test score? Explain.

4. **(35 marks) Time Series - Multiplicative decomposition**

Three years of monthly data lawn-maintenance expenses for a house in Bangkok is given below.

a) (2 marks) Construct a time series plot. What type of pattern exists in the data?

b) (3 marks) Compute a 12-month moving average.

c) (3 marks) Compute a centered moving average.

d) (14 marks) Compute the monthly seasonal indexes for the three years of lawn-maintenance expenses for a house in Bangkok.

e) (3 marks) Compute the deseasonalized time series.

f) (5 marks) Compute the linear trend equation for the deseasonalized data and compute the deseasonalized trend forecasts for 2022.

g) (5 marks) Compute the monthly forecasts for 2022 based upon both trend and seasonal effects.

Month	**2019**	**2020**	**2021**
January	170	180	195
February	180	205	210
March	205	215	230
April	230	245	280
May	240	265	290
June	315	330	390
July	360	400	420
August	290	335	330
September	240	260	290
October	240	270	295
November	230	255	280
December	195	220	250

End of Examination Paper