**NANYANG TECHNOLOGICAL UNIVERSITY**

**SEMESTER 1 EXAMINATION 2022-2023**

**MH3510 – Regression Analysis**

Nov/Dec 2022 Time Allowed: 2 hours

---

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FOUR (4)** questions and comprises **THREE (3)** printed pages.

2. Answer **ALL** questions. The marks for each question are indicated at the beginning of each question.

3. Answer each question beginning on a **FRESH** page of the answer book.

4. This is a **RESTRICTED OPEN BOOK** exam. You are only allowed to bring in **ONE DOUBLE-SIDED A4-SIZE REFERENCE SHEET WITH TEXTS HANDWRITTEN ON THE A4 PAPER** (no sticky notes/post-it notes on the reference sheet).

5. Candidates may use calculators. However, they should write down systematically the steps in the workings.

---

**Question 1 (30 marks)** David would like to investigate the relation between the number of times a carton is transferred ($X$) and the number of broken ampules (Y). He collected the data in the following table. He starts with a linear regression model to fit the data, i.e., uses $y = \beta_0 + \beta_1 x + \epsilon$ to fit the data.

| X | Y |
|---|---|
| 3 | 20 |
| 8 | 150 |
| 4 | 50 |
| 6 | 80 |
| 5 | 70 |
| 3 | 15 |
| 7 | 120 |
| 5 | 60 |
| 2 | 16 |
| 6 | 85 |

(1) Find the least squares estimates of $\beta_0$ and $\beta_1$.

(2) Set up the ANOVA table and conduct an F test to see whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the $\alpha$ risk at 0.05. State the hypothesis, decision rule, and conclusion.

(3) He would like to examine whether a quadratic function is a better fit for the data. Please help him conduct an F test to answer whether it is significant to apply a quadratic function for the data.

(4) Prove that the residual sum of square (SSE) of a least squares estimator from a polynomial regression is lower than that from a linear regression. Some useful statistics: $F^{0.05}_{1,6} = 5.9874$, $F^{0.05}_{1,7} = 5.5914$, $F^{0.05}_{1,8} = 5.3177$, $F^{0.05}_{2,6} = 5.1433$, $F^{0.05}_{2,7} = 4.7374$, $F^{0.05}_{2,8} = 4.4590$.

**Question 2 (20 marks)**

Suppose that $X_1$, $X_2$ and $X_3$ are independent random variables from the following population: $X_1 \sim N(\mu, 2\mu^2), X_2 \sim Uniform(0, \mu), X_3 \sim Exponential(1/\mu)$. Let $\hat{\mu} = \frac{\delta}{3}X_1 + \frac{2\delta}{3}X_2 + (1 - \frac{2}{3}\delta)X_3$, where $\delta$ and $\mu$ are unknown parameters.

(1) Show that $E[\hat{\mu}] = \mu$.

(2) Among all the estimators of the form $\frac{\delta}{3}X_1 + \frac{2\delta}{3}X_2 + (1 - \frac{2}{3}\delta)X_3$, find $\delta$, such that it is the Minimum Variance Unbiased Estimator (MVUE) of $\mu$.

**Question 3 (30 marks)**

Consider a multiple linear regression (MLR) model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ over 30 past observations. Mary conducted an F test to check the significance of the MLR model. The obtained F statistics is $6.899 > F_{2,27}^{0.05} = 3.35$, which implies the MLR model is significant. (Keep 4 decimal places)

(1) Mary tries to check the significance of the reduced model $y_i = \beta_0 + \beta_1 x_{i1}$. She obtained the sum of squares including SSR and SSEs for both reduced and full models but she forgot to label them. She only has a record of those numbers: 2.6562, 5.6611, 11.0812 and 14.0861. Get $SSR$ and $SSE$ for both reduced and full models.

(2) Write down your hypothesis. Complete the ANOVA table and conclude your test results.

(3) Calculate $R^2$ for both reduced and full models.

(4) Mary obtained the 95% prediction interval for a new observation of the predictor $X_0 = [1, 6, 3]$ as follows: $(2.762, 8.965)$. Write down the 95% prediction interval for its mean response.

(5) Mary further applied $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ to fit the data and got the extra sums of squares by adding $x_{i3}$ given $x_{i1}$ and $x_{i2}$ in the model as 1.5023. Please construct the ANOVA table and use F test to conclude whether this model is significant or not for the data.

Some useful statistics: $F_{1,26}^{0.05} = 4.2252$, $F_{2,26}^{0.05} = 3.3690$, $F_{3,26}^{0.05} = 2.9752$, $F_{1,27}^{0.05} = 4.2100$, $F_{2,27}^{0.05} = 3.3541$, $F_{3,26}^{0.05} = 2.9604$, $t_{26}^{0.025} = 2.0555$, $t_{27}^{0.025} = 2.0518$, $t_{28}^{0.025} = 2.0484$

**Question 4 (20 marks)**

To examine the effectiveness of Covid-19 vaccines, the government collected data on the effective level for each type of vaccine and provided their sample means and sample variances in the following table.

| | Pfizer-BioNTech | Moderna | Sinovac |
|---|---|---|---|
| Sample mean | 112 | 108 | 126 |
| Sample variance | 12.3 | 11.6 | 10.8 |
| Sample size | 10 | 8 | 12 |

(1) Construct the ANOVA table to see whether there is significant difference among the three vaccines.

(2) Note that Pfizer-BioNTech and Moderna are based on the mRNA technique, and Sinovac is an inactivated virus COVID-19 vaccine. The government would like to see whether there is a significant difference between these two techniques. Can you help them test the significance using a $t$ test?

**END OF PAPER**

**MH3510 REGRESSION ANALYSIS**

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.