Xi'an Jiaotong-Liverpool University
西交利物浦大学

Paper CODE	EXAMINER	DEPARTMENT	TEL
DTS206TC	Shanshan Zhao	AIAC	88180410

2nd SEMESTER 2021/22 FINAL EXAMINATION(resit)
Undergraduate - Year 3

APPLIED LINEAR STATISTICAL MODELS

TIME ALLOWED: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. This is an open book online examination, in which a calculator or computer may be used for calculation.
2. Screenshot from books or resources cannot be used in the answer of any questions. Such behaviour will be deemed academic misconduct and will be dealt with accordingly.
3. Total marks available are 100.
4. This exam consists of two parts:
    - Part I consists of 3 short answer questions for a total of 30 marks.
    - Part II consists of 3 independent questions for a total of 70 marks.
5. Answer all questions. There is NO penalty for providing a wrong answer.
6. Answers can be written on white A4 paper (or *Ipad*), and should be scanned or converted into pdf format before being submitted. Alternatively, answers can be written in the (latex) answer sheet template which is provided.
7. Only English solutions are accepted.
8. Answer sheet of *pdf* format must be uploaded to the LMO upon completion of the exam. Failure to do so will be taken as absence of the exam.

Part A. [30 marks]

Instructions: Please fill in the blanks with final results.
Calculation procedure is not needed and will not be marked for questions A1,A2,A3.

Question A1 [5 marks]

Let X be normally distributed with mean 36 and variance 9. $P(X \geq 30)$ equals approximately: ________

Question A2 [10 marks]

We have data on the difference between monthly average temperatures in the northern part of Sealand (2011-2014), and historical (1961-1990) correspondingly for the northern part of Sealand (for example, the month of January in 2012 was 2.1 degrees warmer than the average for January in the years 1961-1990). The average of the n=48 observations is $\overline{x} = 1.363$ and the standard deviation is $s = 1.521$. In the following it may be assumed that the observations are independent and comes from a normal distribution $X_i \sim N(\mu, \sigma^2)$.
The following hypothesis should now be tested

$$H_0 : \mu = 0$$

against the two-sided alternative. On level $\alpha = 0.05$, what is the conclusion of this hypothesis test? Give a short reason.

____.

Question A3 [15 marks]

The estimated data are shown in table below.

X	2	4	3	2	4	5
y	20	60	46	27	61	77

Assume that first-order regression model is appropriate,
(1) what is the estimated regression function: ____________________.[3 marks]
(2) What is the point estimate when $X = 5$?____________________. [2 marks]
(3) Estimate the change in y when X increases by one.__________.[2 marks] What is the 90% confidence interval ________. [3 marks]
(4) Conduct a t test to determine whether or not there is a linear association between X and y. (State the alternatives, decision rule and conclusion. [5 marks]

(The result is account to two decimals places.)

Part B. [70 marks]

Question B1 [20 marks]

Suppose we are going to build a linear regression model with the dependent variable y and four explanatory variables x_1, x_2, x_3, x_4. We have n observations and assume SSR and SSE is known.
Please answer the following questions .

(a)[7 marks] How to compute R^2 ? What does it imply?

(b)[8 marks] How to compute and interpret the F-statistic? What is difference between F-test and t-test?

(c)[5 marks] How does F-test work in ANVOA?

Question B2 [20 marks]

There is a study on the student's grade in a secondary school. Some data has been collected, in terms of how the student spend his/her spare time. The data includes the GPA, which ranges from 0 to 340 points and averages around 200 points. The data also contains how much time the student spends on reading, playing computer games, and sports. Each of the mentioned activities has been normalized on the scale of $[-1, 1]$. No reason can be seen to favor any activity in this explanation. As a matter of fact, the researcher finds that it seems rather unlikely if any of these factors explained more than 10 points, with the exception of reading, which the researcher thinks tends to explain up to 20 points. The researcher also implies that these factors should not be expected to explain the GPA perfectly, but other factors that have not been included in this study would be likely to explain up to 20 points at least.
(a)[15 marks] Formulate a probabilistic linear regression model for the problem above, with all distributions specified.
(b)[5 marks] How would the model change if gender were to be included? According to the researcher, the factor of gender would likely to explain no more than 5 points.

Question B3 [30 marks]

In this problem, we will investigate how we we can use the theory for GLM to fit regression models for log-normal responses.

- (a)[10 marks] Assume that Y is a random variable that has log-normal distribution, and a random variable defined by $Z = \log(Y)$ is normal distributed, the mean of which is $E[Z] = \upsilon$ and $var[Z] = \sigma^2$.

  Demonstrate that $var[Y] = \tau\mu^2$, with $\mu = E[Y]$ and the dispersion parameter $\tau = exp(\sigma^2) - 1$.
  (hint: For $Z \sim N(\upsilon, \sigma^2)$, the moment generating function of random variable Z is $m_Z(t) = E[exp(tZ)] = exp(\upsilon t + \frac{1}{2}t^2\sigma^2)$)

- (b)[5 marks] Assume independent variables $Y_1, Y_2..., Y_n$ are log-normally distributed, with expected values $u_i = E[Y_i] = exp(\beta_0 + \beta_1 x_i + \frac{1}{2}\sigma^2)$, where the x_i are explanatory variables, which have the same variance with Z_i, i.e., σ^2.
  Can we obtain estimates of β_0, β_1 and τ, with the simple linear regression for $Z_i = \log(Y_i)$?
  (hint: $E[Z_i] = E[log(Y_i)] = \beta_0 + \beta_1 x_i$)
- (c)[15 marks] When a simple linear regression technique, for instance $\mu_i = \beta_0 + \beta_1 x_i$, cannot be applied to estimate the parameters, explain how GLM-theory can be used.