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 (mock)
Undergraduate - Year 3

APPLIED LINEAR STATISTICAL MODELS

TIME ALLOWED: 2 Hours

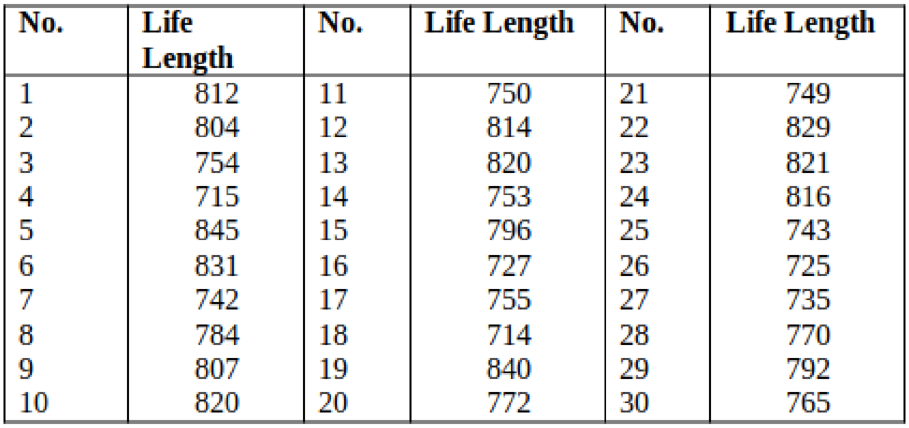
INSTRUCTIONS TO CANDIDATES

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 5 independent questions.
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.

Question 1. [15 marks]

A factory INFINITY wants to conduct interval estimate for their light bulbs. Suppose that the life length of an INFINITY light bulb follows the normal distribution. A random sample of $n = 30$ bulbs is selected for the test and the light bulb life length data is as shown in TABLE 1.

TABLE I



No.	Life Length	No.	Life Length	No.	Life Length
1	812	11	750	21	749
2	804	12	814	22	829
3	754	13	820	23	821
4	715	14	753	24	816
5	845	15	796	25	743
6	831	16	727	26	725
7	742	17	755	27	735
8	784	18	714	28	770
9	807	19	840	29	792
10	820	20	772	30	765

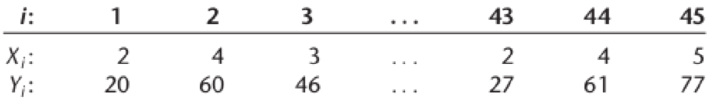
(a)[5 marks] Compute the sample mean life length $\overline{X}$.

(b)[5 marks] Assume that the variance σ^2 is unknown and the random variable has a t distribution with $n - 1$ degrees of freedom. Let $S = 40$ be the sample variance. Construct a 95% two-sided confidence interval on the mean life length of an INFINITY light bulb.

(c)[5 marks] When the number of degrees of freedom of the t distribution is very large, what distribution does it become?

Question 2. [25 marks]

Copier maintenance. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, x is the number of copiers serviced and y is the total number of minutes spent by the service person. Assume that SLR model is appropriate.



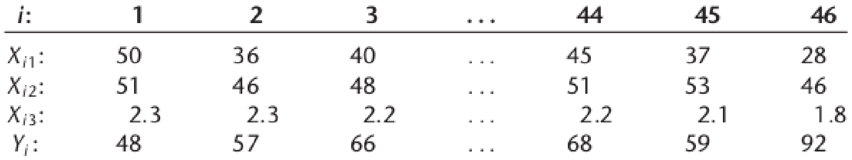
i:	1	2	3	...	43	44	45
X_i:	2	4	3	...	2	4	5
Y_i:	20	60	46	...	27	61	77

(a)[10 marks] Obtain the estimated regression function with the method of least squares.

(b)[15 marks] Conduct a t test to determine whether or not there is a linear association between x and y here; control the α risk at 0.10. State the alternatives, decision rule, and conclusion. What is the p-value of your test?

Question 3. [20 marks]

A hospital administrator wished to study the relation between patient satisfaction (y) and patient's age (x_1 , in years), severity of illness (x_2 , an index), and anxiety level (x_3 , an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of y , x_2 , and x_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.



i:	1	2	3	...	44	45	46
X_{i1}:	50	36	40	...	45	37	28
X_{i2}:	51	46	48	...	51	53	46
X_{i3}:	2.3	2.3	2.2	...	2.2	2.1	1.8
Y_i:	48	57	66	...	68	59	92

The estimated MLR model is $\hat{y} = 158.491 - 1.1416x_1 - 0.4420x_2 - 13.4702x_3$. We got $SSR = 9,120.46, SSTO = 13,369.3$.

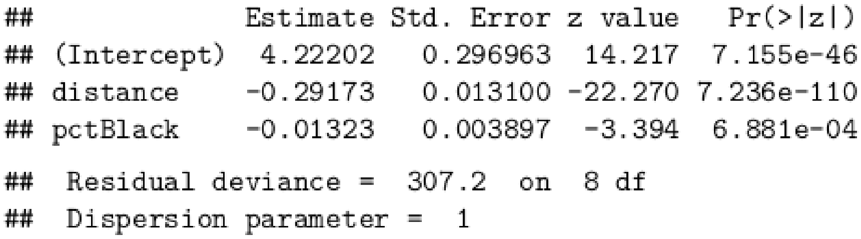
(a)[5 marks] Compute the F-statistic.

(b)[15 marks] Test whether there is a regression relation; use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2 , and β_3 ? What is the p-value of the test?

Question 4. [20 marks]

This case study demonstrates how wide-ranging applications of statistics can be. Many would not associate statistics with historical research, but this case study shows that it can be done. U.S. Census data from 1870 helped historian Michael Fitzgerald of St. Olaf College gain insight into important questions about how railroads were supported during the Reconstruction Era.
In a paper entitled "Reconstructing Alabama: Reconstruction Era Demographic and Statistical Research," Ben Bayer performs an analysis of data from 1870 to explain factors that influence voting on referendums related to railroad subsidies [Bayer and Fitzgerald, 2011]. Positive votes are hypothesized to be inversely proportional to the distance a voter is from the proposed railroad, but the racial composition of a community (as measured by the percentage of blacks) is hypothesized to be associated with voting behavior as well. Separate analyses of three counties in Alabama—Hale, Clarke, and Dallas—were performed; we discuss Hale County here. This example differs from the soccer example in that it includes continuous covariates. Was voting on railroad referenda related to distance from the proposed railroad line and the racial composition of a community?

Assume a logistic model is appropriate. The following *R* output of the fitted model is shown below. State the model fitted here, and interpret the race and gender effects. Construct confidence intervals for the effects.



```
##              Estimate Std. Error z value   Pr(>|z|)
## (Intercept)  4.22202   0.296963  14.217  7.155e-46
## distance    -0.29173   0.013100 -22.270 7.236e-110
## pctBlack    -0.01323   0.003897  -3.394  6.881e-04

##  Residual deviance =  307.2  on  8 df
##  Dispersion parameter =  1
```

Question 5. [20 marks]

Refer to data in Question 3.

(a)[5 marks] Show how to find the coefficients using the maximum likelihood approach.

(b)[15 marks] Formulate the problem with Bayesian linear regression model. Define a prior you like, show how to find the coefficients of the model.