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

Paper CODE	EXAMINER	DEPARTMENT	TEL
DTS206TC		AIAC	

2nd SEMESTER 2021/22 FINAL EXAMINATION
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 candy factory HAPPY wants to make sure that the mean of the sugar content μ of their products does not exceed 30%. To test this hypothesis $H_0 : \mu = 30$ against the alternate one $H_1 : \mu > 30$, they choose a sample of size 20.

(a)[10 marks] Assume the standard deviation of the sugar content is equal to 3.5%**.**

(i) [3.5 marks] What is the probability of type I error if the critical region is defined to be $\mu > 31.5$?
(ii) [3.5 marks] Find the probability of type II error assuming the true mean sugar content is 31.8%.
(iii)[3 marks] Would this probability increase or decrease if the sample size was 40?

(b)[5 marks] The sample mean is found to be $\overline{x} = 31.1\%$, and the sample standard deviation turns out to be $s = 2.5\%$ (they have no independent knowledge of the variance at this point). Please use $\alpha = 0.1$.

(i)[3 marks] Will they reject the null hypothesis given this data?
(ii)[2 marks] Now suppose they know that the true standard deviation is $\sigma = 2.4$. What sample size should they use in order to be able to detect the true mean sugar content of 31.5% with probability 0.95?

Question 2. [25 marks]

The following table is the data regarding the age of the person (x_i) and the number of attempts (y_i) to take the driver license test.

Age x_i	19	50	30	40	65
Number of attempts y_i	2	10	5	7	15

Assume that SLR model is appropriate. *(The result is account to two decimals places)*

(a)[6 marks] Fit the regression line using the method of least squares.

(b)[6 marks] Compute the residual variance.

(c)[6 marks] Compute a 95% confidence interval for the slope of the regression line obtained in (a).

(d)[2 marks] Test the hypothesis that number of attempts required depends linearly on age, using a 0.05 significance level.

(e)[5 marks] In a more general scenario where the distribution of error term is unknown, how would you check whether normally distributed error assumption is met? How do you deal with the case when it is violated?

Question 3. [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 ran the regression process on 16

observations and got the information: $SSR = 946.181$ and $SSE = 49.773$. Please answer the following questions .

(a)[5 marks] What is the value of R^2 ? What can we learn from it?

(b)[5 marks] Compute the F-statistic? What does the F-statistic imply in this case? What is difference between F-test and t-test?

(c)[5 marks] Is the overall regression model significant? Test at $\alpha = 0.05$ level of significance.

(d)[5 marks] Is it possbile to check the Homoscedasticity assumption with F-test? Please write your idea briefly if you think it's possible.

Question 4. [20 marks]

An investigation of whether a health system reform in the USA led to reduced doctor visits. The study was conducted by interviewing some individuals on the frequency of doctoral visits. In this problem, the responses y from these interviewees were dichotomized into unfrequent and frequent visitors: If the number of visits were below 7, $y_i = 0$; otherwise, $y_i = 1$.
We define two explanatory variables x_{i1} and x_{i2}, of which x_{i1} is used to indicate whether the interview is before or after the reform (0=before, 1=after), and x_{i2} is to indicate the health condition of the interviewee(1=poor, 0=good).
We conducted logistic regression in R, with the binary outcomes defined as *I(numvisit>6)*, and x_{i1} and x_{i2} are denoted as *reform* and *badh*.

(a)[6 marks] Write the logistic regression model for the case, and calculate the odds.

(b)[6 marks] Calculate the odd ratio when x_{i1} is changed by one unit and x_{i2} is kept constant. Interpret the result.

(c)[8 marks] Calculate 95%-confidence intervals for exp(β_1) and exp(β_2), and conduct hypothesis tests for $H_{0j} : \beta_j = 0$ versus $H_{1j} : \beta_j \neq 0$ with a 5% significance level. (β_1 : coefficient of x_{i1}, β_2 : coefficient of x_{i2})

```
> fit=glm(I(numvisit>6)~badh+reform,family=binomial,data=drv)
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7180     0.1230 -22.103   <2e-16
badh          2.1995     0.1668  13.189   <2e-16
reform       -0.3382     0.1619  -2.089   0.0367
```

Question 5. [20 marks]

We have the Bayesian linear regression model

$$p(\mathbf{y}|\mathbf{w},\tau) = \prod_{n=1}^{N} \mathcal{N}\left(y_n; \mathbf{w}^T\mathbf{x_n}, \tau^{-1}\right)$$

with the prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mu_{\mathbf{0}}, \mathbf{\Sigma})$$

where τ, μ_0, and $\mathbf{\Sigma}$ are known.
The observations are shown in table below.

y	x_1	x_2
3	2	-2
2	3	4
1	5	3

We would like to learn a linear regression model on the form

$$y = ax_1 + bx_2 + \epsilon$$

where, $\epsilon \sim \mathcal{N}(0, 5)$.

(a)[5 marks] Find $\mathbf{w} = (a, b)^T$ using the maximum likelihood approach.

(The result is account to two decimals places)

(b)[10 marks] Find w using the Bayesian approach, with the prior of

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.25 & 0\\0 & 0.25\end{bmatrix}\right)$$

(The result is account to two decimals places.)

(c)[5 marks] Comments on the results from (a) and (b).