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

Paper CODE	EXAMINER	DEPARTMENT	TEL
DTS206TC		AIAC	

2nd SEMESTER 2022/23 FINAL EXAMINATION
Undergraduate - Year 3

APPLIED LINEAR STATISTICAL MODELS
TIME ALLOWED: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. **This is a closed-book examination, which is to be written without books or notes.**
2. **Total marks available are 100.**
3. **This exam consists of 5 independent questions.**
4. **Answer all questions. There is NO penalty for providing a wrong answer.**
5. **Answer should be written in the answer booklet(s) provided.**
6. **Only English solutions are accepted.**
7. **All materials must be returned to the exam supervisor upon completion of the exam. Failure to do so will be deemed academic misconduct and will be dealt with accordingly.**

1	2	3	4	5	Σ

Question 1. [10 marks]

If $\overline{X}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, show that the $100(1-\alpha)\%$ Confidence Interval on the population mean μ is given by

$$\overline{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.

Question 2. [25 marks]

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature in $^\circ C$ (x)	0	25	50	75	100
Yield in grams (y)	14	38	54	76	95

The average and standard deviation of temperature and yield are
$\overline{x} = 50, s_x = 39.52847, \overline{y} = 55.4, s_y = 31.66702$
The usual linear regression model is used.
By running the regression in R, we can get the output,

```
D <- data.frame(x=c(0,25,50,75,100),
                y=c(14,38,54,76,95))
fit <- lm(y ~ x, data=D)
summary(fit)


Call:
lm(formula = y ~ x, data = D)

Residuals:
   1    2    3    4    5 
-1.4  2.6 -1.4  0.6 -0.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.4000     1.4967    10.3    0.002 **
x             0.8000     0.0244    32.7 0.000063 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.93 on 3 degrees of freedom
Multiple R-squared:  0.997, Adjusted R-squared:  0.996
F-statistic: 1.07e+03 on 1 and 3 DF,  p-value: 0.0000627
```

(a)[6 marks] Can a significant relationship between *yield* and *temperature* be documented on the usual significance level $\alpha = 0.05$?

(b)[9 marks] Compute a 95% confidence interval for the slope of the regression line obtained in (a).

(c)[10 marks] Give the 95% confidence interval of the expected yield at a temperature of $x_{new} = 80$ °C.

Question 3. [25 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 ANOVA?

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

Question 4. [25 marks]

Suppose that 3000 high students are enrolled in a study and the outcome is the occurrence of depression cases. Possible predictors of depression include academic pressure, social isolation, financial stress, family problems and sleep disturbances.

(a)[5 marks] Formulate the problem of selecting the important predictors of depression in a generalized linear model (GLM) framework.

(b)[10 marks] Show the components of the GLM, including the link function and distribution (in exponential family form).

(c)[10 marks] Describe (briefly) how estimation and inference could proceed via a frequentist approach.

Question 5. [15 marks]

Consider the following Gaussian data distribution.

$$p(y_i|w) = \mathcal{N}(y_i; x_i^T w, \sigma^2)$$

We are interested in a so-called *maximum a posteriori estimate* of w,

$$\hat{w} = argmax_w p(w|y)$$

(a)[9 marks] Show that $\hat{w}$ is the solution to linear regression with ridge regression,

$$\hat{w} = argmin_w \left\{ \sum_{i=1}^{N} (y - x_i^T w)^2 + \lambda \sum_{j=1}^{p} |w_j|^2 \right\} \tag{1}$$

with prior of

$$p(w) = \prod_{j=1}^{p} p(w_j) = \prod_{j=1}^{p} \mathcal{N}(w_j; 0, \alpha^2)$$

(b)[6 marks] Idetify the value of α that makes equation (1) correct.