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(resit)
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 2 short answer questions ans 4 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.**

Part A. [40 marks]

Short answer questions

Question A1. [5 marks]

What is the purpose of regularization in linear regression?

Question A2. [5 marks]

What is the difference between ordinary least squares (OLS) and maximum likelihood estimation (MLE) in linear regression?

Question A3. [30 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)[20 marks] Assume the standard deviation of the sugar content is equal to 3.5%.

(i) [10 marks] What is the probability of type I error if the critical region is defined to be $\mu > 31.5$?
(ii) [10 marks] Find the probability of type II error assuming the true mean sugar content is 31.8%.

(b)[10 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 $\alpha = 0.1$.

Will they reject the null hypothesis given this data?

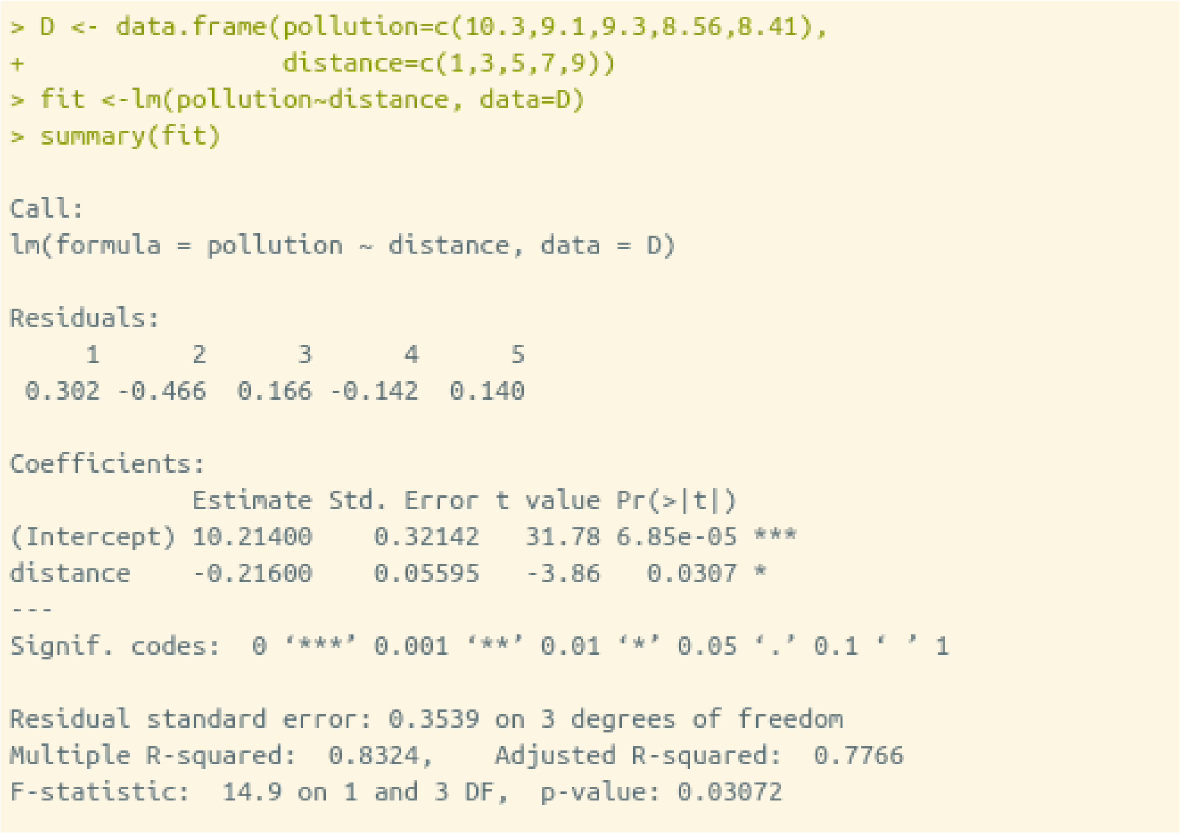
Part B. [60 marks]

Question B1: [25 marks]

In an investigation of pollution in a river, samples are taken from 5 different locations to test the concentration (y) of pollution. The chosen locations are at different distances (x) to the pollution source. These distances and the average pollution concentrations are given in the table below:

x_i (Distance/km)	1	3	5	7	9
y_i Pollution	10.3	9.1	9.3	8.56	8.41

Assume SLR is appropriate. The R code output is given as,



```
> D <- data.frame(pollution=c(10.3,9.1,9.3,8.56,8.41),
+                 distance=c(1,3,5,7,9))
> fit <-lm(pollution~distance, data=D)
> summary(fit)

Call:
lm(formula = pollution ~ distance, data = D)

Residuals:
     1      2      3      4      5 
 0.302 -0.466  0.166 -0.142  0.140 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.21400    0.32142   31.78 6.85e-05 ***
distance    -0.21600    0.05595   -3.86   0.0307 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3539 on 3 degrees of freedom
Multiple R-squared:  0.8324,	Adjusted R-squared:  0.7766 
F-statistic:  14.9 on 1 and 3 DF,  p-value: 0.03072
```

(a)[6 marks] What are the parameter estimates for the three unknown parameters in the usual linear regression model?

1. The intercept β_0
2. the slope β_1
3. error standard deviation σ

(b)[6 marks] How much variability of y is explained by the model?

(c)[13 marks] What is a 95%-confidence interval for the expected pollution concentration 6 km from the pollution source?

Question B2 [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 ANOVA?

Question B3 [15 marks]

For a binary mathced pair data (Y_{i1}, Y_{i2}) with

- explanatory variables , x_{ij}
- random intercept , u_i

the model is written as

$$logit(P(Y_{ij} = 1|x_{ij}, u_i)) = \beta_0 + \beta_1 x_{ij} + \mu_i$$

where $logit(\pi) = \log(\pi/(1-\pi))$.
Assuming that,

- the μ_i's density $f(\mu; \sigma^2)$ is from a $\mathcal{N}(0, \sigma_\mu^2)$ distribution
- Y_{i1} and Y_{i2} independent conditionally on μ_i
- the μ_i's are independent so that the pairs (Y_{i1}, Y_{i2}); $i = 1, \cdots, n$ are independent

(a)[5 marks] Show an expression for the marginal probability $P(Y_{ij} = 1|x_{ij})$.

(b)[5 marks] Demonstrate that Y_{i1} and Y_{i2} are marginally dependent.

**(c)[5 marks] Otherwise, the μ_i's can be considered as fixed effects and β_1 can be estimated by conditioning on $Y_{i1} + Y_{i2}$.
We have**

$$P(Y_{i1} = 1|Y_{i1} + Y_{i2} = 1) = \frac{exp(\beta_1(x_{i1} - x_{i2}))}{1 + exp(\beta_1(x_{i1} - x_{i2}))}$$

and

$$P(Y_{i1} = 1|Y_{i1} + Y_{i2} = 2) = 1 = P(Y_{i1} = 0|Y_{i1} + Y_{i2} = 0).$$

Explain how can we set up logistic regression to get an estimate of β_1.