**Southwest University**

∣

∣

∣

∣

∣

∣

∣

∣

∣

密

∣

∣

∣

∣

∣

∣

∣

∣

∣

∣∣

封

∣

∣

∣

∣

∣

∣

∣

∣

∣

∣

∣

线

∣

∣

∣

∣∣

∣

∣

∣∣

∣

学院 专业 年级 姓名 学号

**College of Computer & Information Science**

**《 Data Analysis 》Exam Paper 【A】**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year 2019～2021 Semester** **2** | | | | | | | | | **Final Exam** | | | |
| **Timing** | **120 Mins** | | **Assessment Method** | | **Closed-Book**  **Exam** | | **Student category** | | **Under-**  **graduate** | **Number of Students** | | **121** |
| **Profession** | | | **Computer Science** | | | | | | **Enrolled Year** | **2019** | | |
| **No.** | | **One** | | **Two** | | **Three** | | **Four** | **Five** | | **Total** | |
| **Score** | |  | |  | |  | |  |  | |  | |
| **Sign** | |  | |  | |  | |  |  | |  | |

**Marking notes: Score with a red ink pen, scoring with Arabic numerals written before the title of each question, with positive points, writing 0 before the title while no points; score in the corresponding score box for each section; unified exam should be collective marking, flow process; review after marking, errors should be corrected in time; the modified person must sign after modify the record.**

**Special reminder: students must abide by the curriculum assessment discipline, violators will be dealt with severely.**

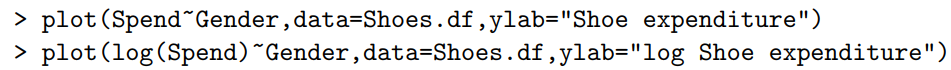
**INSTRUCTIONS**

**• Attempt ALL questions.**

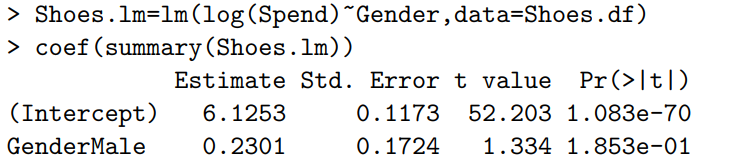
**• This exam has a total of 100 marks.**

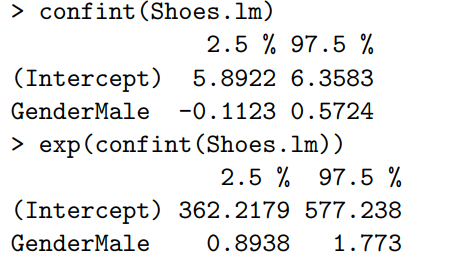
1. A student conducted a small survey on 95 fellow students because she had read that men now spend more than women on shoes. She was curious to see how much her fellow students spent on shoes, and also to see if men really did spend more.

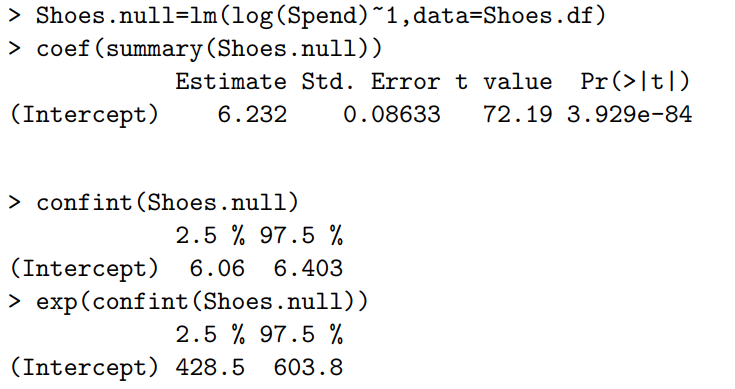
The dataframe Shoes.df contains the variables Gender (”Female” or ”Male”) and Spend (Amount spent on shoes in the last 12 months, $).











You may assume that all model assumptions are adequately satisfied. Please answer the questions on the next page. [Total marks: 16]

a. Comment on the two boxplots. [5 marks]

b. Write down the equation for the model Shoes.lm. [5 marks]

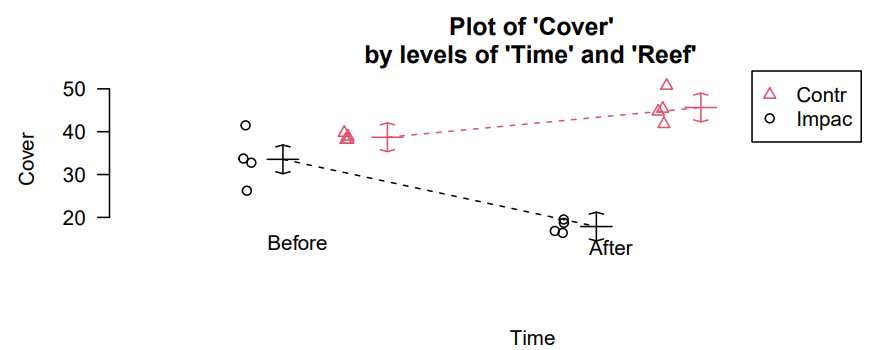
c. Write down the equation for the model Shoes.null. [2 marks]

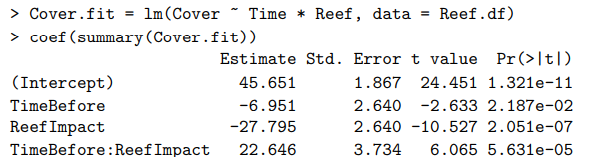
d. Write a brief Executive Summary, being sure to address the questions of interest. (Be sure to use Occam’s razor.) [4 marks]

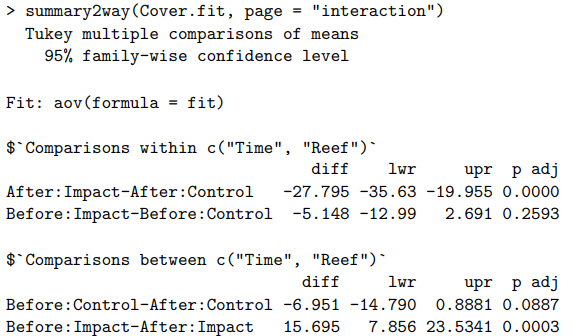
2. Underwater video was used to assess the health of a coral reef, measured as the percent coverage of hard coral. Two surveys were done at two reefs. The surveys are labelled ”Before” and ”After”, since one was done before a strong tropical cyclone, and the second was done just after the storm. Four samples were taken at each survey. Dataframe Reef.df contains the three variables Cover (% cover of hard coral), Reef (”Control” and ”Impact”) and Time (”Before” and ”After”)

It is of particular interest to understand how the cyclone affected the reefs, and whether the effect was the same at each reef.









You may assume that all model assumptions are adequately satisfied. Please answer the questions on the next page. [Total marks: 17]

a. Comment on the interaction plot. [5 marks]

b. What is the fitted value of coral coverage at the Impact site before the cyclone?

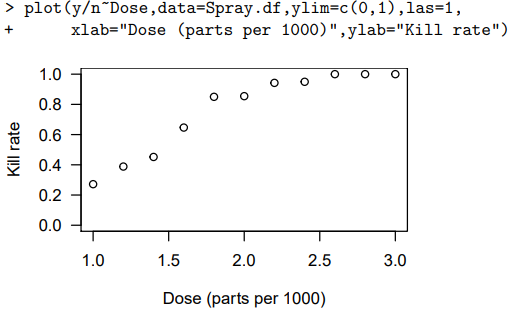
[4 marks]

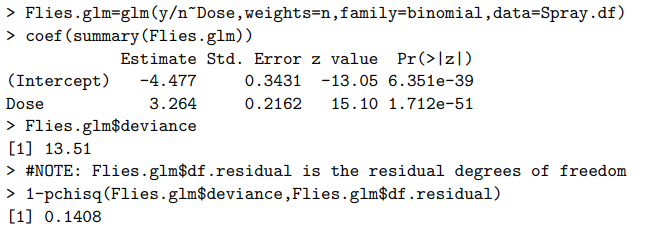
c. What is the fitted value of coral coverage at the Impact site after the cyclone?

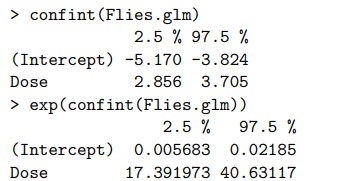
[3 marks]

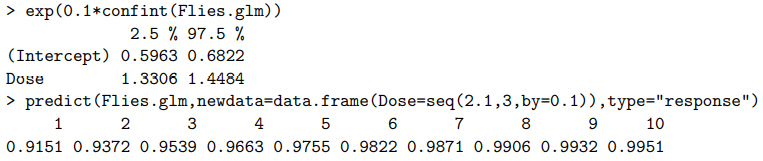
d. Write a brief Executive Summary, being sure to address the questions of interest. [5 marks]

3. An experiment was conducted to determine the concentration of phenothrin required for an insect spray to have a high kill rate. Insect sprays with different concentration of phenothrin were used on independent batches of approximately one hundred flies each. The number killed within 30 seconds was recorded. Data frame Spray.df contains variables Dose, n and y, corresponding to the concentration (parts per 1000), number of flies in the batch, and the number that were killed within 30 s.









Please answer the questions on the next page. [Total marks: 16]

a. Explain the purpose of the 1-pchisq(Flies.glm$deviance,Flies.glm$df.residual)

command, and what it is that we conclude from its output. [5 marks]

b. Write down the equation for model Flies.glm. [4 marks]

c. What is the effect on fly mortality of increasing the concentration of phenothrin by 0.1 (parts per 1000)? [4 marks]

d. According to model Flies.glm, what is the minimum concentration required to achieve at least 99% fly mortality. [3 marks]

4. Short-Answer Questions.

1) A power law model  was fitted to some data. The 95% confidence interval for β1 was (0, 2). What is the effect on y of increasing x by 20%? Be sure to write you answer in a complete sentence, as you would for an Executive Summary. [5 marks]

2) The weather forecast gives a 50% probability of sun on Saturday, and a 80% probability on Sunday. What is the odds ratio for a sunny day on Sunday relative to Saturday? [4 marks]

5. Single-Choice Questions.

The remaining 12 questions are all **single-choice** questions. Answer ALL questions. Each question is worth 3.5 marks. [Total: 42 marks]

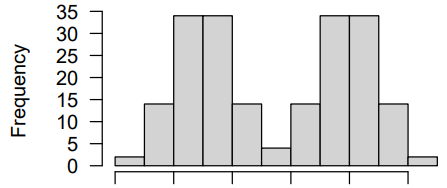
•For each question, select the **ONE** answer by circling the number it is next to.

• If you give more than one answer to any question, you will receive zero marks for that question.

• If you wish to change your answer, make it very obvious what your ﬁnal answer is.

• Each question has a single correct answer.

1) 200 observations of some variable x are plotted in the histogram below. Which statement most accurately describes the relationship between the mean, , and the median, ?



a.  < 

b.  and  are both undeﬁned.

c.  > 

d.  ≈ 

2) For a particular linear model, the sum of the squared residuals of the null model was 1000. For the chosen model, the sum of the squared residuals was 900. What is the value of the chosen model?

a. 0.9

b. 0.81

c. 1.9

d. 0.1

3) Assuming model assumptions are satisﬁed, which of the following statements about the residuals, , from a ﬁtted linear regression model is TRUE?

a. The absolute value of  is always smaller than the absolute value of .

b. The sum of the absolute values of the residuals is maximized.c. The sum of the squared values of the residuals is maximized.d. The residuals have expected value of zero.

4) Consider a linear model which has one numeric explanatory variable x and a factor explanatory variable z with three levels. How many regression coefficients are there in the interaction model with formula ?

a. 3

b. 4

c. 5

d. 6

5) Consider a linear model that suffers from multicollinearity. Which of the following is a potential remedy?

a. Log-transform the response variable.

b. Fit a generalised linear model instead.

c. Remove an explanatory variable from the model.

d. Add a quadratic term to the model.

6) Which of the following is the most reasonable to assume about the distribution of the number of children that a mother has given birth to.

a. Poisson.

b. Binomial.

c. Normal.

d. None of the above because this number can not be zero (since she is a mother).

7) Which of the following is **NOT** an assumption of a generalised linear model with a Poisson response distribution? Let  be the expectation of the response variable for the *i*th observation.

a. The variance of the response for the *i*th observation is .

b. The observations are independent of one another.

c. Each observation’s response variable has a Poisson distribution.

d. is a linear combination of the explanatory variables.

8) A generalised linear model was ﬁtted with the glm() function in R, using the argument family = "quasibinomial". Which of the following statements **IS** an assumption of the model? Let  be the number of successful trials for the *i*th observation, out of a total of trials. The expected number of successful trials is.

a. .

b. Constant variance of the response across all observations.

c. .

d. The observations are independent.

9) When we ﬁt a Poisson regression model with a single explanatory term,  , we often assume that



where  is the expected value of observation i.

Which of the following is a good reason for using the log link function?

a. It ensures that the variance of the response is the same for all observations.

b. It allows us to model nonconstant variance in the response variable.

c. We can’t use more than one explanatory term unless we use this link function.

d. It ensures that the expected value is not negative.

10) Suppose that a Poisson GLM model ﬁtted to some count data is found to fail the residual deviance test, and that a quasi-poisson is then ﬁtted. Which of the following is **FALSE**?

a. The estimated regression co-eﬃcients of both models are the same.

b. The estimated standard errors of the regression co-eﬃcients will be higher for the quasi-poisson model.

c. The t-statistic values for the regression co-eﬃcients will be bigger in magnitude for the quasi-poisson model.

d. The p-values for the regression co-eﬃcients will be bigger for the quasi-poisson model.

11) A Poisson GLM model was fitted to a two-by-two contingency table. The model included the row and column effects and their interaction. The summary table output showed that the degrees of freedom of the fitted model was zero. Which of the following statements is TRUE?

a. The model has too many parameters and the results can not be trusted.

b. The row and column effects are multicollinear.

c. A quasi-poisson model should be used instead.

d. There is nothing to worry about. This is to be expected, since there are as many parameters as there are observations.

12) Which of the following statements about the R function relevel is TRUE?

a. It changes the levels of a factor variable so that the level specified by ref= is first.

b. It changes the coverage level of a confidence interval.

c. It changes the grouping in a histogram to make the bars more level.

d. It changes the performance level of the CPU running the R session.