COMPUTER SCIENCE TRIPOS Part IB – mock – Paper 6

4 Foundations of Data Science (DJW)

A group of three friends want to find out how well they know each other. They agree on a questionnaire about tastes (favourite film, favourite food, favourite mustelid, etc.) and then each tries to guess what each other will answer. Let x_{ij} be the number of correct answers when i tries to guess j's answers, and consider the model

$$X_{ij} \sim \alpha_i + \beta_j + \text{Normal}(0, \sigma^2). \tag{1}$$

Here α_i represents the perceptiveness of i, and β_j represents the openness of j.

(*a*) Write the model for X_{ij} as a linear model, and identify the feature vectors. [3 marks]

(*b*) Are the feature vectors in your model linearly independent? Justify your answer. If they are not independent, rewrite the model in a form with linearly independent feature vectors. [6 marks]

(*c*) Explain what is meant by *residuals* in a linear model. Give pseudocode to compute the residuals for this model. [3 marks]

(*d*) Assuming the model (1) is correct, explain how to compute the distribution of $T = \sum_{i \neq j} \varepsilon_{ij}^2$ where ε_{ij} is the residual for X_{ij}. [4 marks]

(*e*) It may be that some pairs of friends know each other particularly well, or particularly badly, in which case the model (1) might be inaccurate. Describe how to test the hypothesis that the model is accurate, using T from part (*d*) as your test statistic. [4 marks]

5 Foundations of Data Science (DJW)

We are given a dataset with a real-valued response variable $y \in \mathbb{R}$ and an integer covariate $e \in \{1, \ldots, E\}$. Consider the model

$$Y_i \sim \text{Normal}(\beta_{e_i}, \sigma^2)$$

where the parameters $\beta = (\beta_1, \ldots, \beta_E)$ and σ are unknown, and where $i \in \{1, \ldots, n\}$ indexes the records in the dataset.

(*a*) Write the model for Y_i as a linear model. Identify the feature vectors, and explain why they are linearly independent. [5 marks]

(*b*) What is meant by *parametric resampling*? Explain how to use parametric resampling to generate a resampled version of the dataset. [4 marks]

(*c*) Let δ_e be the difference between group e and the average,

$$\delta_e = \beta_e - \frac{\beta_1 + \cdots + \beta_E}{E}. \tag{1}$$

We would like to produce a simultaneous confidence interval for every δ_e, of the form

$$-c\,\hat{\delta}_{\max} \leq \delta_e \leq c\,\hat{\delta}_{\max} \quad \text{for all } e \in \{1, \ldots, E\}, \quad \text{where } \hat{\delta}_{\max} = \max_e |\hat{\delta}_e|$$

where c is given, and $\hat{\delta}_e$ is obtained by plugging in the maximum likelihood estimators $\hat{\beta}$ into equation (1). Give pseudocode to compute the error probability of this confidence interval. Comment your code appropriately. [7 marks]

(*d*) An engineer friend suggests it would be easier to rewrite the model as

$$Y_i \sim \text{Normal}(\alpha + \delta_{e_i}, \sigma^2),$$

fit it, and then just read off the estimated $\hat{\delta}_e$. Explain what you would expect to see, if you tried to use numerical optimization to compute the maximum likelihood estimates of $(\alpha, \delta_1, \ldots, \delta_E)$. [4 marks]