8 Foundations of Data Science (djw1005)

The exam paper at Oxbridge Academy has three questions, and students are asked to choose two questions and answer them. Each question is marked out of 20. The results for four students were

	question 1	question 2	question 3
student 1	13	15	
student 2	14	12	
student 3	18		10
student 4		16	8

The examiners are concerned that question 3 was harder than the other two questions. Consider the model

$$X_{ij} \sim \text{Normal}\big(\alpha_i + \beta_j, \sigma^2\big)$$

where X_{ij} is the mark for student i on question j. Here, α_i represents the ability of student i, and β_j represents the easiness of question j.

(a) Write this 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 your model in a form with linearly independent feature vectors. [*Hint:* You should have 6 linearly independent feature vectors.] [5 marks]

(c) In order to grade the students fairly, the examiners wish to fill in the blanks in the table using predicted marks. Give pseudocode to find $\alpha_1 + \beta_3$, the predicted mark for student 1 on question 3, using your model from Part (b). Describe briefly any standard library routines you use in your answer. [4 marks]

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

(e) To find out whether question 3 is indeed harder, the examiners wish to find a confidence interval for $\beta_3 - (\beta_1 + \beta_2)/2$. Suggest a confidence interval, and give pseudocode to find its error probability. [4 marks]