5 Data Science (djw1005)

(*a*) We have two groups of paired data, (a_j, b_j) for $j \in \{1, ..., m\}$ and (a'_k, b'_k) for $k \in \{1, ..., m'\}$. Consider the probability model

$$B_j \sim \text{Poisson}(\lambda a_j)\,, \qquad B'_k \sim \text{Poisson}(\mu a'_k)$$

where λ and μ are unknown parameters.

(*i*) Find maximum likelihood estimates for λ and μ. (*ii*) Describe how to test the hypothesis that $\lambda = \mu$. [5 marks]

(*b*) We have a dataset in which each datapoint $i \in \{1, ..., n\}$ consists of an integer response y_i and a collection of m non-negative features $(e_{1,i}, ..., e_{m,i})$. A *Poisson-linear* model is a model for this dataset of the form

$$Y_i \sim \text{Poisson}\Bigl(\sum_{k=1}^{m} \beta_k e_{k,i}\Bigr)$$

where $\beta_1, ..., \beta_m$ are unknown parameters, assumed to be strictly positive.

(*i*) Give pseudocode for fitting a Poisson-linear model. (*ii*) Express the model from part (*a*) as a Poisson-linear model. [4 marks]

(*c*) In a factory, the workers suspect that one particular manager is incompetent, since there tend to be more safety incidents on days when that manager is present. They have assembled a dataset spanning several months, giving for each day the number of safety incidents and the names of the managers present. They believe that each manager has a competence level, and the number of incidents on a given day is related to the mean competence level among the managers present.

(*i*) Suggest a probability model for this data, in the form of a Poisson-linear model. (*ii*) Explain how to test whether the suspect manager is indeed worse than the others. [11 marks]