MA30280 Applied data science - Coursework 2

Due: Friday 26 April 2024, 23.59

See the document MA30280_CW2_2024_instructions for further instructions on this coursework.

[Total marks for this coursework: 50]

Part I

For the following scenarios, specify a GAM that you think may be appropriate to use in each case. That is, specify the response distribution and an equation for the mean including your choice of basis, for example,

“$Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \beta_0 + f(x_i)$ where f is a smooth function modelled as a cubic regression spline.”

You do not need to specify basis size or smoothing parameter estimation method.

- Water surface temperature has been measured at 100 locations across a lake. If t_i denotes the temperature measurement at location (x_i, y_i) for $i = 1, \ldots, 100$, write down a model that could be used to predict water surface temperature in locations of the lake that were not measured.
- A marketing company is interested in assessing what factors might influence the performance of an ad placed on a particular website. The performance of an ad is measured by the number of clicks C received within a month. The company has collected data on 300 ads placed on the website in the past. For each ad i, the data set includes the position (x_i, y_i) of the ad on the page, the size A_i of the ad and the corresponding performance C_i. Write down a model that can be used to assess how ad performance is affected by page position and ad size.
- A company has developed an environmentally friendly alternative to pesticides that can be sprayed on plants to prevent a certain plant disease. To assess the effectiveness of the spray at different concentrations the company conducted a test on 100 plants. The plants were divided into 5 equally sized groups, each of which had the spray applied at different concentrations. After 30 days they recorded how many plants in each group were infected with the disease. Let I_i denote the number of infected plants in group i and c_i the concentration used for group i. Write down a model that could be used to estimate the effect of concentration on the probability of infection.
- A Scottish manufacturer produces warm woolen hats. The number of hats they need to produce in a month is strongly related to how cold it is. Write down a model that estimates the total cost c_i (in GBP) of materials used for production by the manufacturer in month i from the average monthly temperature t_i (in degrees Celsius) and the average monthly cost of wool w_i (in GBP per skein).
- Satellite images are used to detect elephants across a large area that has been divided into 1×1 km grid cells. For a selected number of grid cells with locations (x_i, y_i), the presence or otherwise of elephants in the cell is recorded. Write down a model that can be used to estimate the probability of finding elephants in any given grid cell.

[Total marks for Part I: 5]

Part II

Download the data file MA30280_cwdat_2024.RData and load using the command `load("MA30280_cwdat_2024.RData")`.

The data includes weekly visitor numbers `visitors` to the made up (and surprisingly popular) Museum of Modelling (MoMO) during the period 1 January 2011 - 1 January 2019. The museum is open all weeks of the year with exhibitions on anything related to data modelling. There are also special exhibitions and events through the year, in particular, during school holidays.

The data also includes the variables `RPuptake` and `hotel.prices` that you will use in 4. below.

1. Briefly examine and comment on the data using some simple summaries/plots.

[5 marks]

2. Fit a model with a normal response distribution and response variable $\sqrt{\texttt{visitors}}$ that estimates how visitor numbers are changing over time.

- Explain your model specification choices and summarise what you conclude from the output.
- What might be the reason for modelling $\sqrt{\texttt{visitors}}$ rather than `visitors`?
- Create a line plot of the data overlaid by the predictions from your model.

[10 marks]

3. There are some weeks during the period where the visitor numbers were not recorded.

- Use your model to estimate these missing values.
- Create a line plot of the data showing the predicted points.
- Comment on the appropriateness of using the model for this purpose.

[5 marks]

4. In January 2011 the city council introduced a residents' pass that gives local residents heavily discounted entry to museums, including MoMO. The variable `RPuptake` records the weekly uptake of this pass. You have also been given the data `hotel.prices` which records the average price per night of a hotel room in the city.

- Can these variables be used to improve your model from 2. above?
- Describe and discuss the differences between different modelling approaches.

[15 marks]

5. For the final 10 marks, choose either 5a or 5b below.

5a. The museum director would like to be able to forecast how many visitors to expect next week, next month or even next year. Discuss whether your models can be used for this purpose.

OR

5b. Investigate other model specifications for the response variable in 2. above that could be used to achieve a similar fit. Compare and contrast a number of different such choices.

[10 marks]

[Total marks for Part II: 45]