MECH60017/MECH70041/MECH96014 Statistics

Coursework

27 February 2023

The purpose of this coursework is to develop your practical skills in statistical modelling, using data extracted from a real engineering application.

A team of engineers is studying the structural characteristics of a steel railway bridge. They have fitted fibre-optic sensors to the bridge. When a train travels over the bridge, the sensors are deformed, and the wavelength of the light emerging from the sensors changes.

Your task is to model the wavelength emerging from one sensor as a train passes over the bridge, as a function of time-index.

This is an individual assessment and so you must work on the coursework on your own, and abide by the College's policy regarding collaboration on assessed work.

The coursework will be marked out of 20 (16 marks for answers and 4 marks for presentation) and it is worth 10% of the final mark.

Instructions

Submission and deadline

There are two submission boxes in the Blackboard folder *Coursework Information and Submissions*; one box is for your report, and the other box is for your code. The deadline for submission is **Monday 13 March 2023, 11am UK time**. Once the files are uploaded there is option for re-uploading, but if possible, avoid last minute uploads as the system can crash if it simultaneously receives too many requests.

Report

The report must be typed up; handwritten reports will not be accepted. You can use any text processing system to produce your report, e.g. Microsoft Word or LaTeX, but the submitted document should be a single PDF file. The report should not exceed five A4 pages. You may include a cover page, table of contents and appendices and these are not included in the five-page limit; references (if any) should be included in the five pages.

Code

You can use MATLAB, Python or R (up to you). At the very beginning of your code, you must set the seed to create reproducible results; for MATLAB use `rng(CID)`, for Python use `random.seed(CID)`, and for R use `set.seed(CID)`, where `CID` is your College ID number. For polynomial regression tasks use the following functions: for MATLAB use `fitlm`, for Python use `PolynomialFeatures`, and for R use `lm`. I should be able to execute your code, without making any modifications, except a first line to read in the data. Badly formatted or unclear code will be penalised. Accepted document file types are: .m, .ipynb, .py, .r.

Data

You are provided with your own, unique datasets. To access your datasets go to the folder *Datasets* within the Blackboard folder *Coursework Information and Submissions*.
The datasets have been named after your username, i.e. if your username is `aa15514`, download the file `aa15514` and save it as csv file in a suitable location. You can then import the file as usual into MATLAB/Python/R.

You should see two columns: the first column, labelled 'X', is the time index and the second column, labelled 'Y', is the wavelength, measured in nanometres.

If you have difficulty downloading your dataset, please email me. For any other issues, please use Blackboard's discussion board.

Questions

1. **Exploratory data analysis**

   (a) Construct a histogram and a boxplot of the wavelength, and comment on the plots.

   (b) Compute the mean, 10% trimmed mean, median, standard deviation, and interquartile range of the wavelength and present these in a table. Which of these measurements would you use to best describe the data? Justify your answer.

   (c) Construct a scatterplot of wavelength versus time index.

2. **Modelling**

   (a) Fit a simple linear regression model for wavelength versus time index. Plot the linear fit on your scatterplot in question 1(c). Comment on the appropriateness of this model for your data.

   (b) An assumption of linear regression is that the response is linear in the parameters. Therefore the response can be modelled as a linear function of polynomials in the predictor variable. Fit a model for wavelength including both a linear and a quadratic term in 'X', and plot the resulting quadratic fit on your scatterplot (from part 1(c)).

   (c) You can now proceed to fit higher order polynomials, plotting the model fit on the question 1(c) scatterplot in each case. Note that a polynomial model of order k should include all lower terms. In polynomial regression you can avoid some numerical problems if you use the standardised 'X' variable. Standardise 'X' by subtracting the sample mean of 'X' and dividing by the sample standard deviation of 'X'. Can you explain why it is better to use the standardised 'X' variable in polynomial regression models? Continue fitting higher order polynomials for as long as you judge to be reasonable (the next question should help you judge).

   (d) Model comparison may be made by Akaike's Information Criterion (AIC), which is equal to $2 * q - 2 * l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, where $l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ is the log-likelihood of the model at the maximum, and q is the number of parameters estimated in the model. Briefly explain what the AIC tells you, and how it can be used to select between models. Assuming the errors are normally distributed in order to calculate the log-likelihood, produce

a table of AIC for each model you have fitted. Which model do you select? Justify your answer.

(e) For your chosen model from question 2(c), calculate the residuals and construct plots of the residuals to check your model assumptions. State clearly which assumptions you are checking with each plot, and your conclusions.

(f) Now suppose you have only been given the responses for time indices equal to $10, 20, 30, \ldots, 850$. Write code to extract the relevant sample from your data set, and fit the same model you chose in question 2(c) to the sample (i.e. if you chose a linear model in question 2(c), fit a linear model in the sample). Make a scatterplot including the model fit in the sample.

3. **Bootstrapping**

Using the sample from question 2(f), you are asked to calculate a 95% pointwise confidence band for the expected wavelength as a function of time index. It is pointwise as the band is constructed from a set of confidence intervals, one at each point (value of 'X'). You decide to bootstrap.

(a) Calculate the residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$, where $\hat{y}_i$ denote the predicted values from the model in question 2(f), i.e. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_k x_i^k, i = 1, \ldots, 85$. Then code the following algorithm:

- Resample from the vector of residuals with replacement, to get a boostrapped sample $\epsilon_1^*, \ldots, \epsilon_{85}^*$.
- Calculate a new response variable, $y_i^* = \hat{y}_i + \epsilon_i^*$.
- Fit a polynomial regression model (same order as in question 2(f)) for the new response values y_i^* – note the predictor variables are the same as in question 2(f).
- Save the predicted values from the polynomial regression model.

Repeat this many times, then calculate across all repetitions the 0.025 and 0.975 quantiles of the predicted frequencies for each value of 'X' – these are the upper and lower limits of the confidence intervals at each time point, which may be joined together to form a 95% confidence band. Add the bootstrapped 95% confidence band

to your question 2(f) scatterplot.

(b) Show graphically that as the number of bootstrapped samples increases, the bootstrapped confidence band converges to the actual confidence band.