MECH60017/MECH70041/MECH96014 Statistics

Coursework

23rd February 2024

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

A team is studying outdoor air-pollution in a town in Europe, specifically they are interested in the amount of carbon monoxide (CO) in the air and how it relates to atmospheric conditions of temperature (T) and relative humidity (RH). They are interested in both explaining the variability of CO with respect to these two easily obtained atmospheric measurements and building models to predict CO concentration.

They have collected daily average values for each of these three quantities over a period of 150 days.

Your task is to model the CO concentration as a function of T and RH. In addition you will investigate including a time index as a covariate.

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 Sub-*

missions; one box is for your report, and the other box is for your code. The deadline for submission is **Friday 8th March 2024, 12 noon 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). 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 three columns labelled: 'CO', 'T' and 'RH'. These are the features you will use in your analysis, they correspond to average daily sensor measurements for Carbon Monoxide concentration measured in mg/m^3, Temperature measured in Celsius, and Relative Humidity as a percentage, respectively. There should be 150 rows of data corresponding to 150 consecutive days of daily average measurements.

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 for each of three features, and comment on these plots. Are the ranges of these values sensible? Hint you may have to search online for what are reasonable values for these measurements in the context provided above.

   (b) Construct a scatterplot for each pair of features, i.e. 'CO', 'T' ; 'CO', 'RH' and 'T', 'RH', and compute each pairwise linear correlation. Comment on the plots in relation to the correlation.

2. **Modelling**

   (a) Fit two simple linear regression models, 'CO' versus 'T' and 'CO' versus 'RH' . Comment on the fit using the coefficient of determination.

   (b) Fit a regression model of 'CO' versus'T' and 'RH'. Comment on the fit using coefficient of determination, is your result consistent with the exploratory analysis you have done?

3. **Prediction**

   We now wish to see how well your models can predict 'CO' concentrations. It is best to make predictions for data that your model was not built with, since this will enable us to see how well your model can generalise to new data. However, since you have no additional data, we shall build models using a subset of your data and use the remaining data to evaluate the predictions.

   Take the first half of your data and call it '`fold1`'. The remaining data, i.e. the last half, we shall call '`fold2`'. The procedure for computing the errors from predictions is as follows.

   (i) Build a model using data from '`fold1`'.

   (ii) Obtain predictions from this model for the data from '`fold2`'.

   (iii) Compute the residuals for these predictions.

   Repeat this process but this time building the model using '`fold2`' and obtaining residuals

for the data in '`fold1`'. You will now have a residual computed for each datum in the whole dataset. The above procedure for computing all the prediction errors is an example of *2-fold cross-validation*.

(a) Perform 2-fold cross-validation as described above for each of your 3 models and assess the normality of the errors.

(b) Compute the sum of squares of the residuals and determine which is the better model in terms of smallest overall error.

4. **Including Time as a Feature**

The data you have been provided with comprises 150 consecutive days of measurements in time order. Construct a time index starting from 1 and include this as the fourth column in your dataset. Call this column 'day'.

(a) Produce a scatterplot of 'CO' versus 'day'. Using the entire dataset, compute the correlation and build a simple linear regression model. Comment on the results.

(b) Investigate whether taking the square root of your time index fits the assumptions of linearity better and produces a better simple linear regression model. Hint you can do this by creating yet another column called sqrtDay, which is the square-root of your time index. Comment on both the fit, the normality of the errors, and sum of squares of the residuals. Compare these results to those using 'day'.

(c) Consider the following three regression models:
'CO' versus 'T' ,'RH' and 'day',
'CO' versus 'T' ,'RH' and 'sqrtDay'
'CO' versus 'T' ,'RH'.
Compare these three models using any of the methods shown in this coursework that you deem necessary. Which model is best and why? Using only the information you have produced in this coursework, could you suggest a better model?