MATH50010 coursework 2023-24

30/11/2023

This coursework is due at 1pm on Friday 9th December. Please submit it via the turnitin link on blackboard. Your submission should contain your CID but not your name.

The Task

In this coursework, we will analyse the amount of Nitrogen Dioxide (NO_2) in the environment. The dataset 'tower_bridge.csv' contains the daily average NO_2 level on Tower Bridge Road in London for every day in 2022. It contains the following columns:

- Site: the location, should be SK8 for all entries.
- Species: the particle being measured, should be NO2 for all entries.
- ReadingDateTime: the date and time the reading was taken (note that time will be 0 as it is a daily average).
- Value: the average amount of NO_2 recorded.
- Units: the units of the NO_2 reading, should be ug m-3 ($\mu g/m^3$, micorgrams per cubic meter) for all entries.

The dataset is available to download from blackboard. In this coursework we are interested in determining how frequently we have high pollution periods.

The following is a step-by-step workflow to guide you through the task. Your coursework submission should be written using RMarkdown, and compiled to a PDF for submission. All code should be commented clearly. For the highest marks, you should communicate to the marker clearly what you are trying to do, and justify any arbitrary choices. There are a total of 50 marks available for this coursework. 6/50 marks are available for an extension question, you can still get a good mark overall without attempting this question.

(3 marks) Loading and exploration

1. Read the data in to R.
2. (1 marks) We want to split the data into low and high pollution levels. Typically it is assumed that the pollution level is high if the NO_2 level exceeds $40\mu g/m^3$. Create a new variable called 'state' indicating whether the pollution is high (1) or low (0) on each day.
3. (2 marks) Calculate the proportion of days in each of the two states defined above.

(18 marks) A Markov Chain Model

We will now model the data as a Markov Chain.

4. (3 marks) We look at the transitions between states. Count the number of pairs in each of the possible pairs of successive states (0,0), (0,1), (1,0),(1,1). Overlaps are OK, e.g. the sequence 0100 corresponds to one (0,1) transition, one (1,0) transition and one (0,0) transition.
5. (5 marks) Assume that the high/low pollution state forms a two-state time-homogeneous Markov chain. Use the data to estimate the transition matrix of the chain.
6. (3 marks) Write a function that simulates draws of length `m` from a two state Markov chain with states 0 and 1.

7. (7 marks) Use your function to simulate n independent 'years' of daily high/low classifications of NO_2 using the transition probabilities from the data. For each of the n realizations of the chain, compute the estimates of the transition probabilities. Show that the estimators are approximately unbiased. Are the estimates of different transition probabilities correlated?

(14 marks) Testing the Markov Model

We now need to test whether the pollution level does in fact depend on the pollution level of the previous day.

8. (3 marks) Write down a formal hypothesis test in terms of the transition probabilities to test whether the probability of the pollution level being high is independent of the pollution level of the previous day.

If we want to test whether two sampled data sets x_A and x_B of sizes n_A,n_B come from the same Bernoulli distribution, we can perform a hypothesis test using a Hypergeometric distribution. In particular, let p_A and p_B be the population positivity probabilities of the two data sets, then we want to test $H_0 : p_A = p_B$ vs $H_1 : p_A \neq p_B$. Note that under H_0, both data sets x_A and x_B are from the same distribution. Therefore the number of positive samples in x_A follows a Hypergeometric($n_A + n_B, n_A, s_A + s_B$) distribution where s_A and s_B are the number of positive samples in x_A and x_B respectively (see e.g. Section 8.6.1 of Ross (2020)). The Hypergeometric(N, K, n) distribution for $n \leq N$ has the density:

$$P(X = x|N, K, n) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \qquad \text{for } x \in \{0, 1, \dots, K\}, \text{ and } n - x \in \{0, 1, \dots, N - K\}.$$

To perform the hypothesis test, we use the p-value given by,

$$p = 2 \min\{P(X \leq s_A), P(X \geq s_A)\} \qquad \text{for } X \sim \text{Hypergeometric}(n_A + n_B, n_A, s_A + s_B).$$

This test is sometimes known as the Fisher-Irwin test.

9. (6 marks) Calculate the p-value for your hypothesis test.
10. (2 marks) Conduct the hypothesis test at the 5% significance level.
11. (3 marks) Perform a similar hypothesis test to determine whether the probability of the pollution level being low is independent of the previous days level.

(12 marks) Investigating the Number of Consecutive Low Pollution Days

We now want to investigate the likelihood of consecutive low pollution days.

12. (3 marks) Using the estimated Markov model, calculate the probability that after any high pollution day, it is over a week until the next high pollution day, i.e. calculate $P(X_1 = 0, .., X_7 = 0|X_0 = 1)$.
13. (6 marks, Extension) Using the estimated Markov model, plot the probability mass function of the number of consecutive low pollution days immediately after any high pollution day, i.e. let M be the number of consecutive low pollution days starting from day 1, plot $P(M = m|X_0 = 1)$ for all m.
14. (3 marks) Using the data directly, calculate the average number of consecutive low pollution days between high pollution days.

(3 marks) Conclusion

15. (3 marks) Comment on any limitations of your study.

Academic integrity

You are welcome to use any sources (websites, books, etc), but you should cite them. If you make use of chunks of code that you have found, you should also cite the source. You should write your own submission, including all code. Failure to do so will be considered misconduct.

Useful functions

Scatter plot of x_t against x_{t+1}

```
x<-rnorm(300)
x<-x + c(0,x[-300]) #make correlated data
lag.plot(x) # plot
```

Make a vector into a single string.

```
paste(c(0,1,1,0),collapse="")
```

Count instances of a single letter

```
library(stringr)
str_count("101010",c("0","1"))
```

Count (overlapping) instances of a double letter

```
str_count("00010100111",paste0("(?=",c("00","01","10","11"),")"))
```

Count the lengths of consecutive runs of 1's or 0's

```
vec<-c(0,0,0,1,0,0,1,1,0,0,0,0,1,1,1)
rle(vec)
```

In the output, 'lengths' represent the number of consecutive values and 'values' tells us which value is repeated. We can extract each using e.g. `rle(vec)$values`.

Reference:

Ross, S. M. (2020). Introduction to probability and statistics for engineers and scientists. Academic press.