CHIC402/CHIC602 Coursework 2 - Variant Analysis

Barry Rowlingson

1 Notes

For this task you should start with the empty R Markdown file `variant_task.Rmd` included in the materials. The final output should be a PDF which also shows the code.

R code should be included within the document and should be neatly formatted and commented, so pay attention to consistent spacing and indenting of blocks of code. Use meaningful names for your objects and make comments that help with the understanding of the code. Put code into functions where appropriate. Write explanatory text chunks to create a narrative as if this was a laboratory notebook. Divide into sections where appropriate for clarity.

Submission should include your `variant_task.Rmd` file and its `variant_task.pdf` version.

2 Data

The data are from genetically-typed SARS-CoV-2 ("Covid") infections in the UK. Each sample has been classified into one of the named variants (with a Greek letter), a "Variant Under Investigation" (VUI), a "Variant of Concern" (VOC), or "Other".

The data are further broken down into cases with or without known travel links.

The first few records of the named variants and the "Other" proportion look like this:

week	travel	Alpha	Beta	Delta	Eta	Gamma	Kappa	Theta	Zeta	Lambda	Mu	Other
2021-01-31	None	0.95	0.00	0	0.00	0	0	0	0.00	0	0	0.05
2021-01-31	Travel	0.62	0.08	0	0.08	0	0	0	0.00	0	0	0.19
2021-02-07	None	0.96	0.00	0	0.00	0	0	0	0.00	0	0	0.03
2021-02-07	Travel	0.47	0.19	0	0.09	0	0	0	0.00	0	0	0.21
2021-02-14	None	0.97	0.00	0	0.00	0	0	0	0.00	0	0	0.02
2021-02-14	Travel	0.60	0.09	0	0.06	0	0	0	0.02	0	0	0.20

and the first VUIs and VOC records look like this:

VUI.21FEB.01	VOC.21FEB.02	VUI.21FEB.04	VUI.21APR.03	VUI.21MAY.01	VUI.21MAY.02	VUI.21OCT.01
0	0	0.00	0	0	0	0
0	0	0.04	0	0	0	0
0	0	0.00	0	0	0	0
0	0	0.04	0	0	0	0
0	0	0.00	0	0	0	0
0	0	0.03	0	0	0	0

Each value is a fraction of total classifications, so that the rows sum to one.

The data is in a CSV file which should read into R in this form. Make sure the week column is a date column, and convert using `as.Date` if it isn't. You may need to use this data in this "wide" format or converted to "long" format with columns for week, travel-status, variant, and proportion.

3 Non Travel-Related Data

Create a subset of the data for the **non-travel-related cases only** and work on it for this section.

3.1 Variant Peaks

Plot a graph of the Delta variant proportion over time. Choose appropriate form, style, colours, labels etc. Include a one-line title that succinctly describes the curve.

The peak of the Delta variant occurred at 2021-07-04 with a value of 0.991. Construct a table of the weeks in which peak occurrence of each variant occurred with the fraction at the peak. Sorted by proportion, the first rows should look like this (but with all the dates and proportions complete for all variants).

variant	week	p
Delta	2021-07-04	0.991
Alpha	yyyy-mm-dd	0.xxx
VUI.21OCT.01	yyyy-mm-dd	0.xxx
Other	yyyy-mm-dd	0.xxx
Kappa	yyyy-mm-dd	0.xxx
..	..	..

4 Graphs of Variant for Travel/non-Travel

Find the five variants (not including "Other") with the largest proportion in any week over the full data set (i.e. including Travel and non-Travel related cases).

4.1 Line Plot

Plot the proportions of each of these variants over time as lines on a single plot. Use a colour to show the variant, and a line style to show the travel/non-travel group.

4.2 Stacked Bars

Plot two stacked bar charts – one for travel-related and one for non-travel related cases – of these five variants, including a bar labelled "Other" which will be the sum of the original "Other" proportion plus the proportions of the variants not in the top five. The bars should sum to one.

4.3 Description

Write a short paragraph (max 100 words) describing what the two stacked bar plots tell us about the changes of the variant mixtures in both travel and non-travel related cases.

5 New Variant Increase

The `VUI.21OCT.01` variant started to increase in proportion in summer 2021 and can be seen to be continually increasing towards the end of the data set period. For this section only use the **non-travel** related cases - subset the data and work with this reduced set.

Plot a line of just this variant over time for **non-travel** cases only.

An exponential model can be fitted to this growth. The resulting fitted model is:

$$p = \exp(d\beta + \alpha)$$

where β is 0.0277 and α is -526.0385, and d is the date, as stored by R in the usual way as the number of days since 1970-01-01.

Write a function to return p given values of d, α and β. You should find that the proportion of this variant on the first of September 2021 is 0.037.

Plot a graph of the proportion of the variant over time together with the fitted model.

What does the model predict for the proportion on the first day of 2022? Comment on this.

6 Submission

Remember to include your R Markdown, generated PDF, and any R code file if you have one. Make sure your student ID is in the R Markdown metadata and appears in the PDF (you can also put it in a comment in the R code too).