**Predicting trends of COVID-19 using the frequency of symptom searches**

**Introduction and Motivation:**

COVID-19 has impacted all aspects of society over the 18 months. Our ability to promptly detect changes in the occurrence of COVID-19 is primarily informed by testing and reporting through the NHS and private laboratory systems. Unfortunately, the information, despite considerable improvement in the last year, is slow with trends often recognised long after the opportunity to intervene has passed. Predictive models offer a chance to use other sources of data to detect areas where COVID-19 infection is rising. Based on previous published work(1) by my supervisors, the aim of this project is to explore a multiple linear regression model to predict the occurrence of COVID-19 cases using the online search data. The key motivation for the project are:

To understand the statistical approaches used in predictive modelling

To improve my ability to use Python to query databases

The project consists of the following the development of Python code to extract symptom search data from publicly available sources, downloads of individual case reports of confirmed COVID-19 cases and predictive modelling to estimate parameters that define which combination of symptom searches can correctly predict COVID-19 trends.

**Methodology:**

***Data sources and processing***

Symptom search frequency data were obtained from the publicly available Google Symptom database(2). Using Python code, I extracted symptom data from this repository. The data downloaded included information on the daily frequency of the search terms for the UK and sub geographical areas between 1st January 2021 and 15th August 2021 that matched a list of relevant COVID search queries: chest pain, cough, diarrhoea, fatigue, fever, headache, migraine, nasal congestion, nausea, pneumonia, shortness of breath, sore throat and vomiting.

Similarly, data on daily individual case reports of confirmed COVID-19 cases for the UK and sub geographical areas for the same time period were downloaded from the Public Health England website(3).

Using Python, I aligned the daily data from both sources into a single dataset for each geographical area and utilised national (UK-wide) data for training.

***Analysis***

Using the correlate function, relevant the Pearson correlation coefficient in each geographical area between each symptom frequency and the number of daily reported cases were calculated. To visually examine the relationship between symptom frequency and case reports, scatter plots of each symptom with the daily new case number as well as line plots of each symptom and the cases over the time period were produced. Based on the observed patterns and a priori knowledge of the incubation period and reporting delays, I added a time lag of 14 days between symptom onset/report and case reporting. Following this adjustment, correlation coefficients were recalculated, sequentially removing the symptoms which were not relevant to the model upon reviewing the information gathered, these were: chest pain, fatigue, headache, migraine, nasal congestion, nausea, pneumonia and shortness of breath.

To visually illustrate correlation clusters and check for multicollinearity, a Pearson correlation matrix between the symptoms was plotted. The worse performing features of those that had a correlation coefficient over an arbitrary value of 0.75 were subsequently removed from the model.

Further to this, linear regression was run on the training data. To do this, X, y and theta matrices were created to store the features, target and coefficient values respectively.

Where:

X is an n + 1 x m dimensional matrix, where m is the number of training examples.

is a n + 1 x 1 dimensional matrix.

is the predicted value

The data was normalised by dividing the case number by the population size resulting in cases per 100,000 (.

Where: p is the population of the area and is the case number

The cost and gradient descent functions as well as the learning rate and maximum iterations were defined as follows.

Cost Function J:

Gradient Descent Algorithm:

Where , is the learning rate

Upon running the simple linear regression on the dataset, the theta values were returned.

To assess multiple variables in regression model, a multiple linear regression model was developed. This model utilised the selected symptom search frequencies stored in a NumPy array of the training predictions and allowed the calculation of the root mean squared error as well as the final value of the cost function.

Assessment of Model Accuracy: To assess the accuracy of the model I plotted the predicted values against the actual case data for training and testing dataset.

**Results**

Table 1 shows the correlation coefficient for all included symptoms demonstrating considerable variation ranging from a -0.107595 inverse relationship for fatigue to a strong positive correlation at 0.721270 for fever.

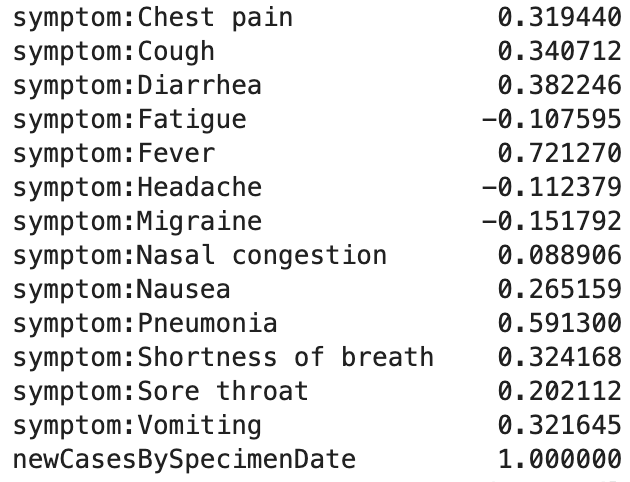
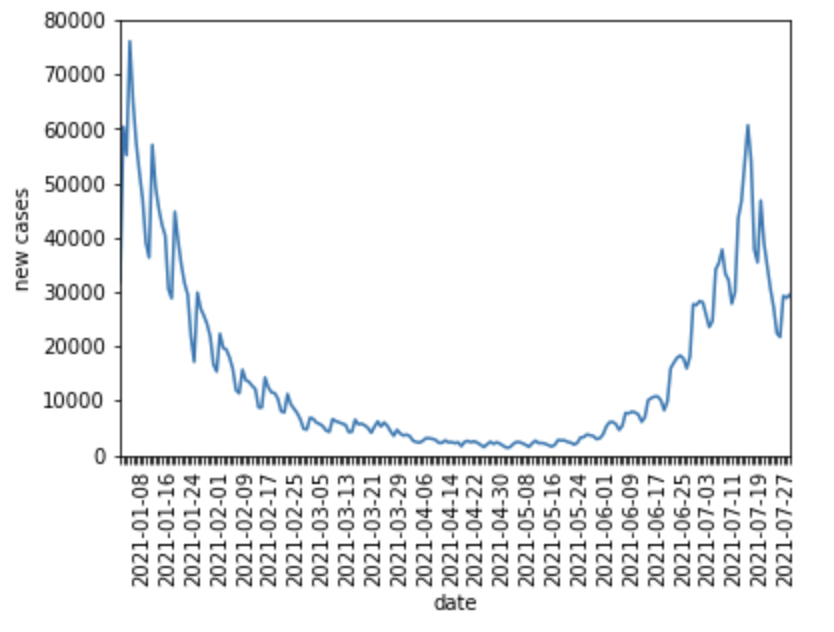


Table 1. Pearson correlation coefficient between symptoms and daily UK cases before shifting time series by 14 days

Figure 1 shows a broadly similar pattern with both variables peaking in January and July. However, there are other peaks and greater variability in the fever data.



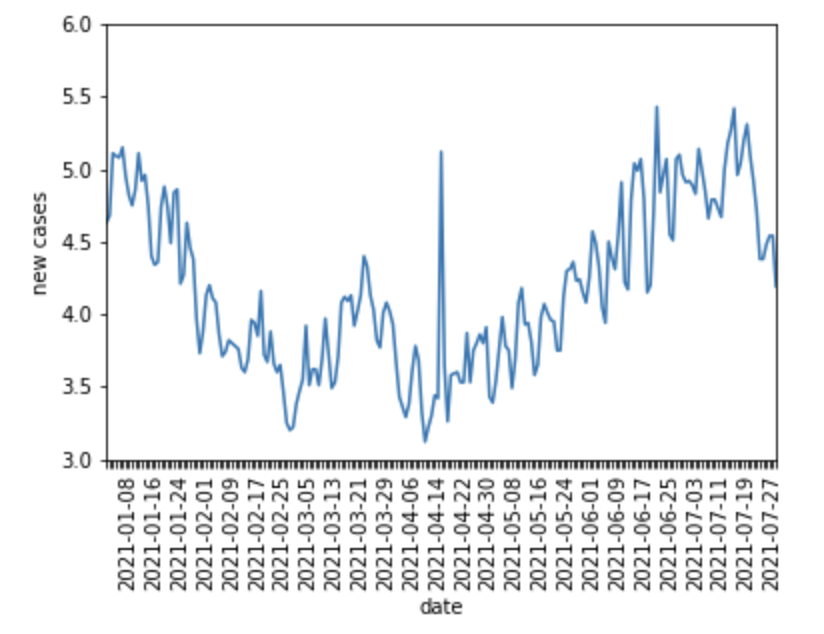


Figure 1. Line plot of time series data for new cases and fever search frequency

Figure 2 and 3 demonstrate an improvement in the correlation between cases and fever by shifting the time frame of the data by 14 days.

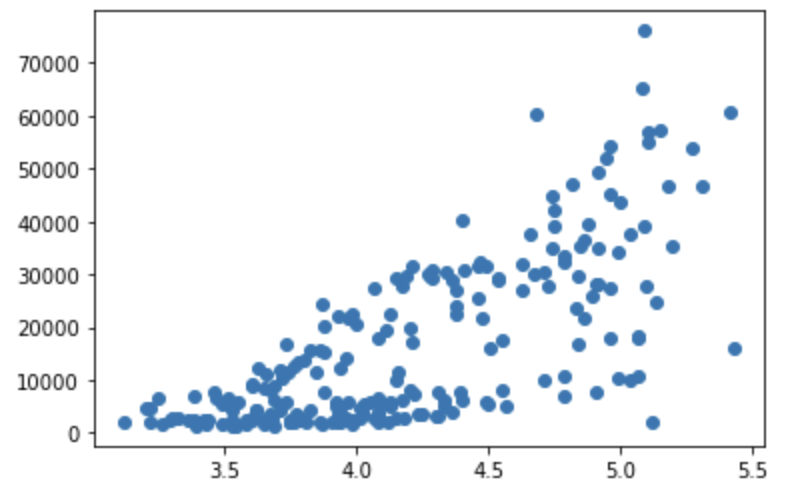


Figure 2. Scatter plot of new cases to fever search frequency before time shift of 14 days

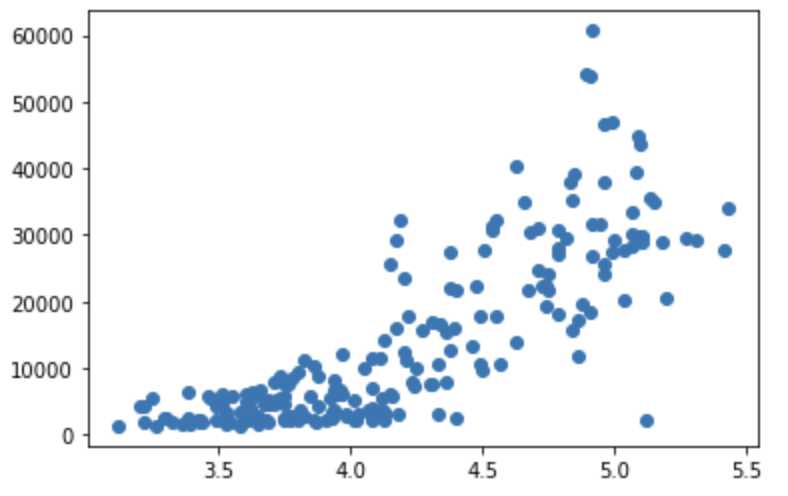


Figure 3. Scatter plot of new cases to fever search frequency after time shift of 14 days

Recalculating the correlation coefficients and sequentially removing the symptoms which were not relevant to the model showed an improvement in correlation for fever, cough, diarrhoea, sore throat and vomiting (table 2).

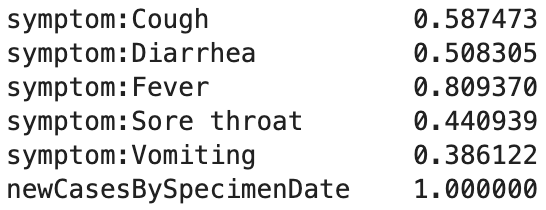


Table 2: Pearson correlation coefficient between symptoms and cases after shifting time series by 14 days and dropping features

The Pearson correlation matrix between the symptoms to assess multicollinearity showed the variables which explained the same variation (sore throat and cough, vomiting and diarrhoea)

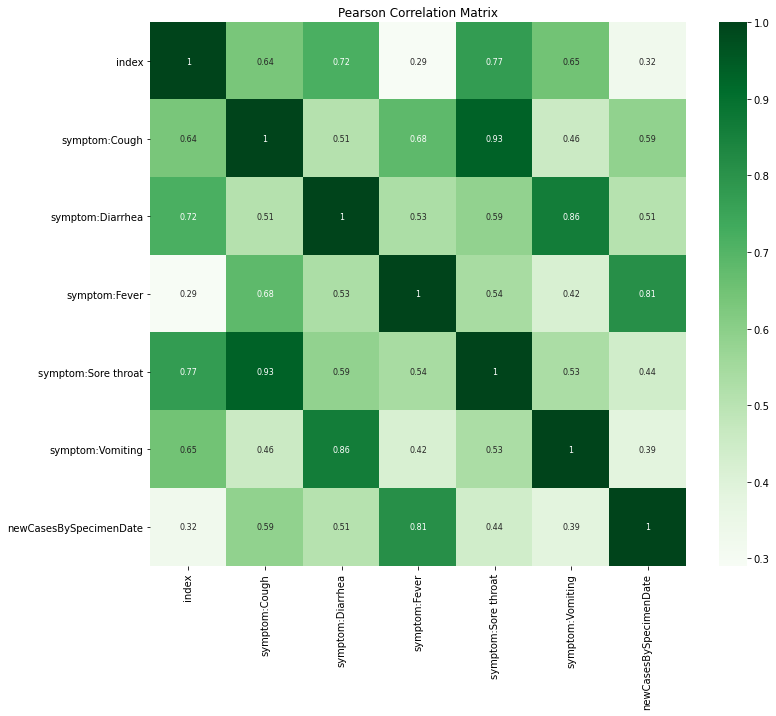


Figure 5. Pearson correlation matrix to check for multicollinearity

Plotting the value of the cost function over the number of iterations allowed me to adjust the maximum iterations and learning rate to suitable levels to ensure the algorithm converges.

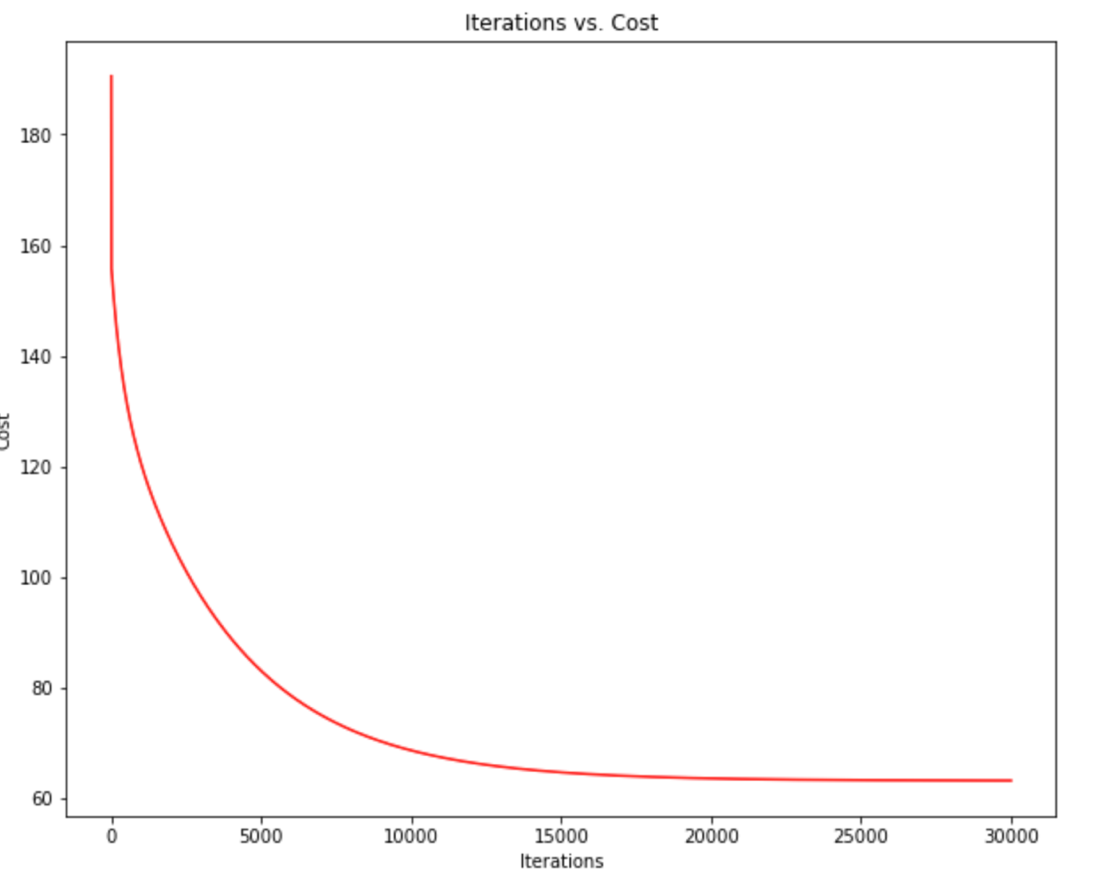


Figure 6. Plot iterations vs cost to see if the gradient descent algorithm converges

I selected a sub region of the UK, Aberdeen City, as a testing dataset and made similar calculations after using the model – plotting the actual cases and predicted cases for both the training and testing datasets as well as the RMSE and final residuals

Figure 7 shows that the general trend to be accurate however there was a large amount of variance in the predicted values especially for the testing graph. The model also struggles to predict large spikes in cases as seen in both graphs.

Figure 7. Plot predicted values against actual case data for training dataset

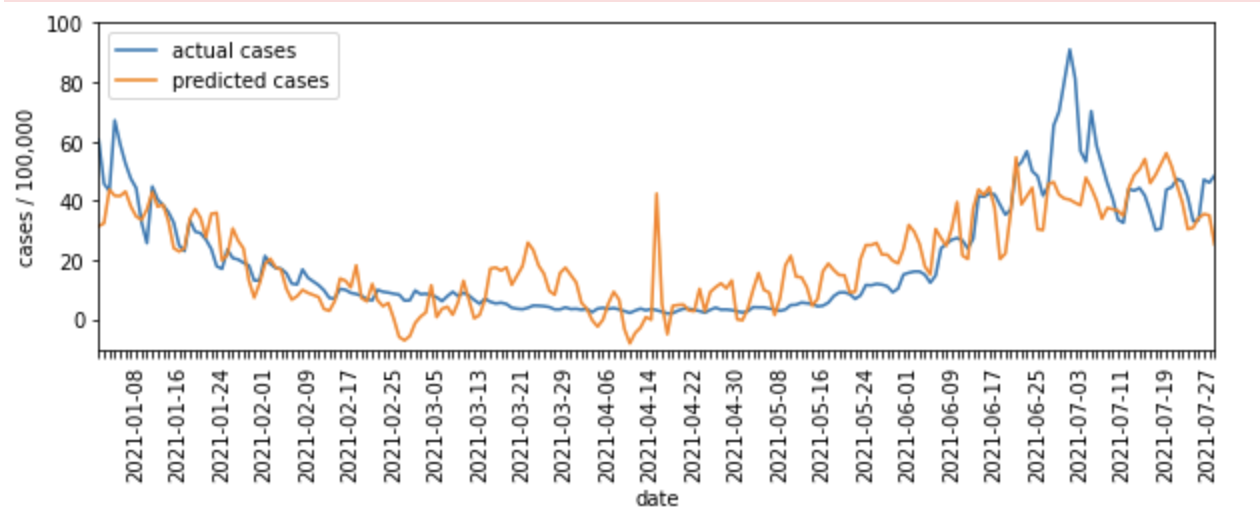


Figure 8. Plot predicted values against case data for training dataset

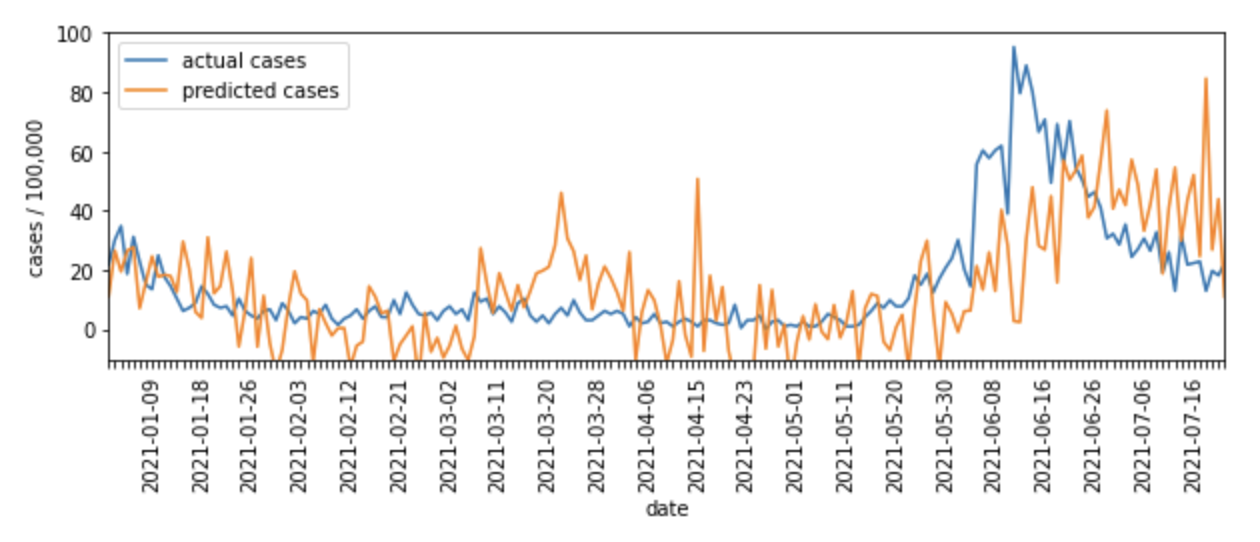
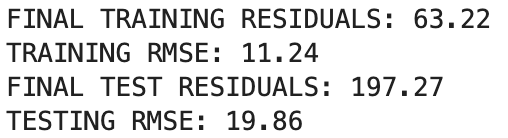


Table 3. Training and testing residuals and RMSE



**Discussion:**

The multiple linear regression model used was able to produce predictions for cases using the online search frequencies of selected symptoms. The model was able to predict the general trend of cases however had very high variance and struggled to predict spikes. I believe the high variance was due to the over representation of fever search frequency in the model, using other search terms outside of symptoms may have reduced this variation. The variation and struggle to predict spikes in cases may have been due to observations in figure 3, as there was an overabundance of points with low case totals and varying frequency within that time period added noise to the predicted values. The work is based on published work(1) which provides insight that minimising the effect of news media and trying other unsupervised and supervised models may yield improved results.

This work has a number of limitations. First, the analysis is limited by the inability to adjust for the effect of news coverage on searches. I intended to use Global news corpus for news media adjustment which was not undertaken due to the limited time available for this project. Previous research suggests that adjusting for news items considerably improves the predictive accuracy of models. Second, unfortunately I was also unable to find data on excess deaths; which would have provided a further outcome measure to correlate with symptoms. Third, I would have been keen to explore time series models of varying complexity to attempt to improve predictions which was not possible in the time available. Nevertheless, the analysis presented demonstrates the feasibility of undertaking this work and the potential for future analysis to predict COVID-19 trends.

Public Health England currently utilises a similar model, adjusted for news items as one of the approaches in their report to explore current COVID-19 trends(4). A potential exploration of the current analysis will be to develop the model presented in this paper further and present predictions for multiple geographical areas that may be used to inform local policy and planning.

**Conclusion:**

This project has presented an initial exploration of the potential of predictive regression models using COVID-19 symptom search data to predict the occurrence of cases. Despite multiple limitations, further work could allow the use of these models to inform local COVID-19 control policy.

**Bibliography:**

1. Tracking COVID-19 using online search [Internet]. [cited 2021 Oct 23]. Available from: https://www.nature.com/articles/s41746-021-00384-w.pdf

2. Explore COVID-19 Symptoms Search Trends [Internet]. [cited 2021 Oct 23]. Available from: https://pair-code.github.io/covid19\_symptom\_dataset

3. National flu and COVID-19 surveillance reports [Internet]. GOV.UK. [cited 2021 Oct 23]. Available from: https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports

4. Weekly\_Flu\_and\_COVID-19\_report\_w42\_v2.pdf [Internet]. [cited 2021 Oct 23]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/1027644/Weekly\_Flu\_and\_COVID-19\_report\_w42\_v2.pdf