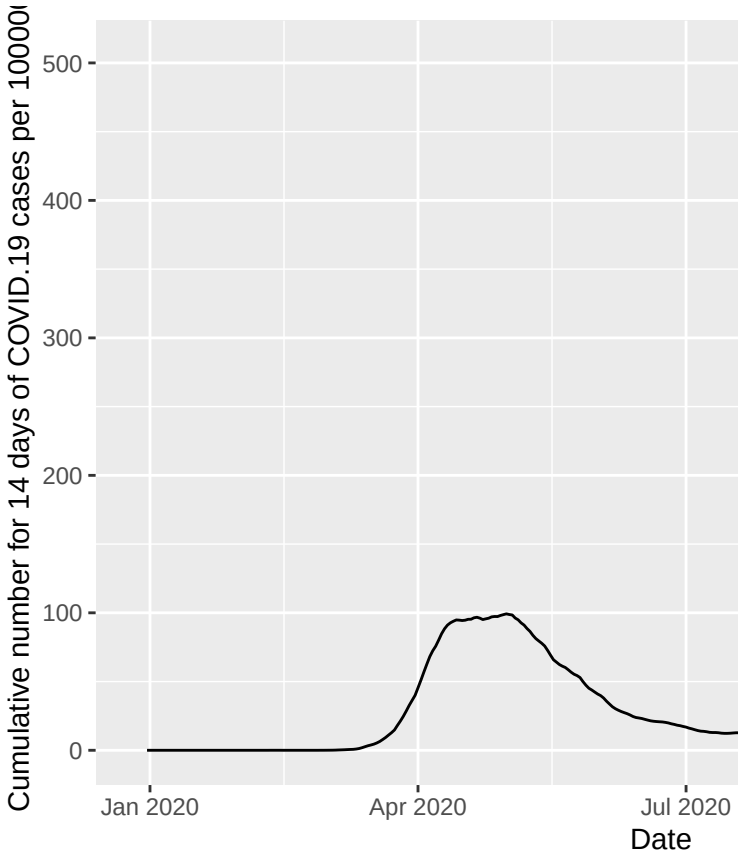
COVID.19 data analysis exercise

Thomas Fulton

08/05/2021

This analysis aims to model the effect of prevention measures taken by the UK on the cumulative number of COVID19 cases over the last 14 days, and determine which prevention measures have been the most effective. This could aid decision on potential measures to include in the case of future lockdowns.



UK cumulative cases over 14 days per 100000 people in 2020:

Data

Two datasets were used: Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country, and data on country response measures to COVID-19.
The datasets were subsetted to get just data for the UK, and combined so that each `Response_measure` was added as a binary variable: either implemented ("1") or not implemented ("0") for every date between 16/03/202 and 15/12/202. An additional lag variable was also included containing the cumulative number of cases over 14 per 100000 people, but 3 weeks prior, for each date. The number of people who could have

spread it at 21 days earlier is autocorrelated with the current number of infected people. The lag is seen roughly 3 weeks later (Roy and Ghosh, 2020. https://doi.org/10.1371/journal.pone.0241165).
Linear Regression
A linear regression model was calculated from the combined dataset. The assumptions of the model were checked, and it was tested by training it on 80% of the data, then using it to predict the cumulative no. cases over 14 days per 100000 for the other 20% of the data.
The model was then improved using a stepwise regression to select factors for a model with lower AIC, retested, and the most influential of the remaining predictive factors were identified with an ANOVA.

Results

The factors were reduced to just 11 factors, 10 of which were highly significant. The cumulative number of cases 3 weeks before ("three_weeks_previous_Cumulative_cases" factor) was the most predicative (F value = 6292), and the most effect Response measures were a Ban on all events, followed by Closing pubs, Paritally closing hotels and other accommodation, and Regional stay at home order.

The improved model had a p-value: < 2.2e-16 and AIC value: 3400.142. Not all the assumptions were properly met: there was still autocorrelation - for future could try and improve by introducing a better lag variable by making multiple models with different lags eg. from 1 week-4 weeks and comparing models. Also the residuals were not normally distributed.

However, the model was highly predicative: the correlation accuracy of the improved model was 0.976.

```
anova_sel_lm_model <- anova(selectedMod)
anova_sel_lm_model
```

```
## Analysis of Variance Table
##
## Response: Cumulative_number_for_14_days_of_COVID.19_cases_per_100000
##                                    Df  Sum Sq Mean Sq   F value    Pr(>F)
## three_weeks_previous_Cumulative_cases   1 5864793 5864793 6292.1748 < 2.2e-16
## BanOnAllEvents                         1 1288578 1288578 1382.4797 < 2.2e-16
## ClosPubAny                             1   94624   94624  101.5190 < 2.2e-16
## ClosPubAnyPartial                      1   12090   12090   12.9711 0.0003638
## EntertainmentVenues                    1   30765   30765   33.0067 2.050e-08
## EntertainmentVenuesPartial             1   61447   61447   65.9250 8.842e-15
## HotelsOtherAccommodationPartial        1   95439   95439  102.3938 < 2.2e-16
## IndoorOver50                           1    6953    6953    7.4596 0.0066409
## PlaceOfWorship                         1   13535   13535   14.5218 0.0001646
## PrivateGatheringRestrictions           1    3293    3293    3.5332 0.0610124
## RegionalStayHomeOrder                  1  156767  156767  168.1906 < 2.2e-16
## Residuals                            338  315042     932
##
## three_weeks_previous_Cumulative_cases ***
## BanOnAllEvents                        ***
## ClosPubAny                            ***
## ClosPubAnyPartial                     ***
## EntertainmentVenues                   ***
## EntertainmentVenuesPartial            ***
## HotelsOtherAccommodationPartial       ***
## IndoorOver50                          **
## PlaceOfWorship                        ***
## PrivateGatheringRestrictions          .
```

```
## RegionalStayHomeOrder                  ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```