STAT0023 ICA 2

Group G: 19000151 19005054

Contents

1 Introduction to the research question and data

1.1 Background

Coronavirus has been rampaging the world since October 2019. Millions of lives suffered and are still suffering. As a result, this report aims to study factors associated with variation in numbers of Covid-19 deaths in some areas of England and Wales during one major period from March to July 2020. Hopefully, an effective statistical model could be built to estimate numbers of deaths for other areas of England and Wales and demonstrate crucial risk factors.

1.2 Data structure

The data include death numbers between March and July 2020, and demographic information for areas mentioned above.

To be specific, the dataset has 7201 observations of "Middle Layer Super Output Areas" (MSOAs) in total (death numbers missing for 1800 of them), and it has over 80 social & demographic characteristics of MSOAs.

- The target variable is the number of deaths.
- All other variables are numeric except `Region` and `RUCode` which are categorical.
- The variables could be categorized into 11 groups:
  - Population
  - Household information
  - Age profile
  - Ethnicity and immigration
  - Unpaid carers
  - Household accommodation
  - People living in communal establishments
  - Employment / occupation
  - Social grade
  - Public transport use
  - Education and qualifications

1.3 Data cleaning and processing

In terms of data cleaning, there is no missing value except for the 1800 numbers of deaths. In addition, there are no conspicuously strange or erroneous data spotted.

Hence, it seems proper to move to the data processing stage. As is suggested, there are several variables already proven to be major risk factors such as age, population density, ethnicity, socioeconomic deprivation, gender, pre-existing health conditions [1]. Apart from those, we decided to look at each set of variables to identify potentially influential factors.

1.3.1 Overall information about each MOSA and its population

- `RUCode` originally has 8 levels and we grouped them into 3 levels (`new.RU`) to represent major city (U1), urban city and town(U2), and rural (R).
- `PopTot` will be an offset as we decided to use death rate as our target variable.
- `PopM` and `PopF` are highly co-linear and correlated from the matrix plot and correlation, but we kept them anyway because gender is known to be a major factor. **[Figure 1]**
- `PopComm` is kept due to its non-linear relationship with the rest.
- `PopDens` also has low linear relationship and correlation with others.

1.3.2 Household information for each MSOA

- `HH` is the sum of the other three types of household and should be deleted due to linear dependence. We decided to keep `HH_1Pers`, `HH_1Fam`, `HH_Oth` as literature says household type might play a big role. [1]
- `HH_HealthPrb`: unhealthy people are supposed to suffer higher risk.
- `HHNoCH`: household without central heating might be safer as virus cannot spread through ventilation.
- `HHRooms` and `HHBedrooms` are highly co-linear. We chose to keep the more general `HHRooms` as they contain roughly same information - whether the house is capable of self-isolation.
- Matrix plot and PCA of dimension of deprivation led to summation of `HHDepriv3` and `HHDepriv4` (`HHDepriv34`).
- We did not figure out how the language speaking could possibly relate with death. For completeness and by PCA, we summed them all in case (`HHLang`).

1.3.3 Age profile for each MSOA

- `MeanAge` and `MedianAge` are extremely co-linear. We kept `MedianAge` as median is more robust to outliers.
- As for the 16 age groups, the matrix plot shows age 0-17 and age 60-90 are clearly co-linear within their group, but age 18-60 is hard to determine. Here is valid to apply PCA to reduce dimensions because here all variables are about age groups. That is to say, after dimension reduction the interpretability of PCA will not be damaged. Eventually, the results imply it is suitable to combine them into 4 age groups: `Age0.17`, `Age18.29`, `Age30.59`, `Age60.90..`

1.3.4 Ethnicity and immigration

According to literature, there is increasing evidence that some racial and ethnic minority groups are being disproportionately affected by COVID-19. [2]

- For example, data shows the deaths of black or African American people are one time higher than that of Asian people. [3]
- Matrix plot here is very complicated and no obvious pattern could be observed. Therefore, we kept all ethnicity variables as they were.

1.3.5 Unpaid carers

It is natural, by its definition, to relate unpaid carers people with health problems (`HH_HealthPrb`) and the old age.

- The number of carers could be an indicator of whether that MSOA has enough helpers to take care of old and unhealthy people.
- The matrix plot indicates that all 3 variables are, to different extents, linear with each other.
- We tended to think `CarersMid` and `CarersHi` should be put together and performed PCA to confirm it. Hence, we took the sum of `CarersMid` and `CarersHi` as `CarersMidHi`.

1.3.6 Household accommodation

The dwelling information is very similar to the household ones.

- `Dwell` is highly co-linear with `HH`. And since they represent pretty much the same aspect, it is valid to be deleted.
- The major values of `DwellShared2` and `DwellShared3.` are quite small. Due to PCA, it is possible to sum them (`DwellShared`).

1.3.7 People living in communal establishments

Communal establishments include hospitals and care homes. From the literature, hospitals and care rooms are shown to be under a stark impact of the virus. [4]

- It is not hard to understand the fact that the distribution of medical resource would influence the death rate i.e. the communal establishments per person, calculated by `CommEstab` / `PopTot` (`CommPerPerson`).
- Matrix plot of the three types of care rooms has no obvious pattern, but from the same literature above, deaths in care rooms are rather more out of control from hospitals due to reasons like less strict protections and no professional medical treatment once infected. We decide to consider 2 covariates, namely, the care room population by summing `LACare`, `PrivCareNurs` and `PrivCareNoNurs`; The hospital population calculated by subtracting the care room population from `PopComm` (`Hospital`).

1.3.8 Employment / occupation & social grade

From the definition of social grade and matrix plot, we can see that each grade represents one or more corresponding employment and occupations. Therefore, social grade and occupations stand for the almost the same information. We decide to drop social grade as employment and occupation are more specific so as to be classified more freely.

The reference gives a clear way of grouping occupations by death rates. Elementary, caring & machine (`WrkC`) leads to overall highest death rate, followed by skilled, sales & admin (`WrkB`). Mgr, ProfTech, Prof (`WrkA`) lead to the lowest death rate. [5]

1.3.9 Public transport

Public transport is an important concept affecting the deaths of Covid [1]. From the matrix plot and also correlation coefficients between the three types of transportations, there is no obvious pattern. Hence, we chose to leave them as they are for the moment. **[Figure 2]**

1.3.10 Education and qualifications

In terms of education, we suspected it to be related with people's ideology. In other words, people with higher education might be better at protecting themselves by listening to advice.

- From the matrix plot and correlation, several variables are quite linear with each other, indicating some potential groups could be formed. The results of PCA show that `NoQual`, `Qual1`, `Qual2` , and `QualApp` can be summed (`QualLow`), and `Qual3`, `Qual4.`, `Stud18.` can be combined (`QualHigh`).

1.4 Possible transformation and interaction

After defining and selecting variables, we ended up with 46 covariates in total. Next, we sought to detect suitable transformations and interactions.

1.4.1 Transformation

By plotting histograms of every potential covariates, it was clear to see that some had skewed distribution and a square-root transformation could work. (`PopComm`, `PopDens`, `HHDepriv34`, `Age18.29`, all variables about ethnicity, dwelling, communal establishments, and public transport) [**Figure 3**]

1.4.2 Interaction

We utilized faceting to identify possible interactions. After checking all combinations between categorical variables (`Region` and `new.RU`) and other continuous variables, we selected 9 pairs for our initial model. In particular, they are all interactions with `new.RU`. (`PopDens`, `HHRooms`, `Age60.90.`, `EthAsian`, `EthBlack`, `CarersMidHi`, `CommPerPerson`, `MetroUsers`, `QualHigh`) [**Figure 4**]

1.5 Starting model specification

First, considering the fact that deaths can only be non-negative and integers, using a normal distribution in the model doesn't seem a good idea. Given also that the response variable - Death is a count, it is natural to consider using a Poisson distribution.

Second, the assumption of constant variance is not valid as more deaths are likely to occur when there are more people.

We also preferred GLM rather than GAM because from the scatterplots y vs. x, some patterns did look like a poisson PMF and most were not weird. [**Figure 5 & 6**]

Finally, we chose log link function to ensure modelled means are positive.

2 Model building and checking

2.1 Model 1

Therefore, a glm with poisson distribution and log link function would be a good starting point. For the first model, we decided to throw in all the 46 covariates, together with the 9 interaction terms. It is worth mentioning that since we chose a poisson glm model, we decided to use deaths as the response variable and add on `log(PopTot)` as an offset.

For diagnostic plots [**Figure 7**]:

- The residual plot shows a pattern of larger residual variance for larger predicted values, which is consistent with our poisson assumption. Also, there are quite various residuals outside the range (-2,2). The calculated variance of the Pearson residuals is about 2.76, far from 1. This suggested using quasipoisson distribution to account for overdispersion.
- The normal QQ plot appears heavy-tailed, looking like poisson. The Cook's distance imply several outliers.
- It is not hard to understand as the Covid dataset is a huge dataset and every data is rather important.

2.2 Model 2

We replaced poisson with quasipoisson. Notice that we couldn't compare them through ANOVA, because it only works for distribution of the same family or type.

Also, the diagnostic plots are the same as those of model 1. From the p-values of the summary, there are some covariates already shown to be statistical significant, in particular, `sqrt(PopComm)`, `HHNoCH`, several `Household` covariates about

deprivation, several `Communal` covariates and some more. However, there are still plenty covariates seemingly not explaining the death well.

2.3 Model 3

For this final model, we didn't achieve this model directly, it is actually a result of trying to delete several insignificant groups of covariates one by one to see we should keep them or not.

Considering both the summary and ANOVA chi-square test, we took turns to delete the following groups of covariates:

- Education: all with interaction
- Dwelling: `DwellShared`
- Unpaid carers: `CarersMidHi` with interaction
- Communal establishments: `CommPerPerson` with interaction, `LACare`
- Household: `HHRooms` with interaction, `HH_Oth`, `HHLang`
- Population: `PopF`
- Ethnicity: `BornIreland`, `BornEU`, `BornNonEU`

As mentioned previously in step 2 above, we tried repeating these processes in turn for every main groups of covariates. By continuously updating and running ANOVA for the consecutive two new models. As a result of repetitive optimization, model 3 is chosen to be our final model.

In this final model, most covariates have significant p-values. We also kept some promising ones with higher p-values. Take `Age0.17` as an example, p-value (>0.1) isn't everything. Youngsters should be a susceptible group given less mature immune systems.

The diagnostic plots of model 3 are the same as previous ones, indicating our model assumptions hold. [**Figure 8**]

3 Model interpretation and limitations

3.1 Model interpretation

Model 3 is a glm with log link function and quasipoisson distribution. It contains 36 covariates, including 5 interaction terms.

The following are the risk factors indicated by our model:

- Offset: `log(PopTot)`
- MOSA: `Region`, `new.RU`
- Population: `PopM`, `sqrt(PopDens)` with interaction
- Household: `HH_1Pers`, `HH_HealthPrb`, `HHNoCH`, `HHDepriv1`, `HHDepriv2`, `sqrt(HHDepriv34)`
- Age: `Age0.17`, `sqrt(Age18.29)`, `Age30.59`, `Age60.90.` with interaction, `MedianAge`
- Ethnicity: `sqrt(EthWhite)`, `sqrt(EthMixed)`, `sqrt(EthAsian)` with interaction, `sqrt(EthBlack)` with interaction, `sqrt(EthOther)`
- Unpaid carers: `sqrt(CarersLo)`
- Communal establishments: `sqrt(PrivCareNurs)`, `sqrt(PrivCareNoNurs)`, `sqrt(Hospital)`, `sqrt(PopComm)`
- Occupation: `WrkA`, `WrkB`, `WrkC`
- Public transport: `sqrt(MetroUsers)` with interaction, `sqrt(TrainUsers)`, `sqrt(BusUsers)`

There are some interesting insights drawn from the summary:

1. Population density in rural areas only is positively related to death.
2. An increase in the number of households without central heating leads to a decrease in death rate. This might be because central heating makes it easier for virus to spread throughout the house.
3. Poverty will also boost the probability of death.
4. Youngsters and senior citizens are more susceptible to virus, probably due to their weaker immune systems. On the contrary, people aged from 18 to 59 are less likely to die.
5. White people are on an edge, whereas Asians, black people, and other minorities suffer a higher risk.
6. Hospitals do have positive impact on saving people while private care homes even make it worse.
7. All occupations are dangerous.
8. All public transports are risky, with metro being the most dangerous.

3.2 Limitations

One conspicuous disadvantage of using quasipoisson model is that we could only use p-values and chi-square test to improve our models. Therefore, although being extremely careful, We could potentially delete insignificant covariates that are indeed important in predicting death rate.

Since this is a huge dataset, there are some 'outliers' or 'leverage points' that might influence the quality of model and hence prediction. However, considering Covid death is a very serious social problem that we should treat them rigorously. We did not try to change them.

4 Appendix: graphs and tables

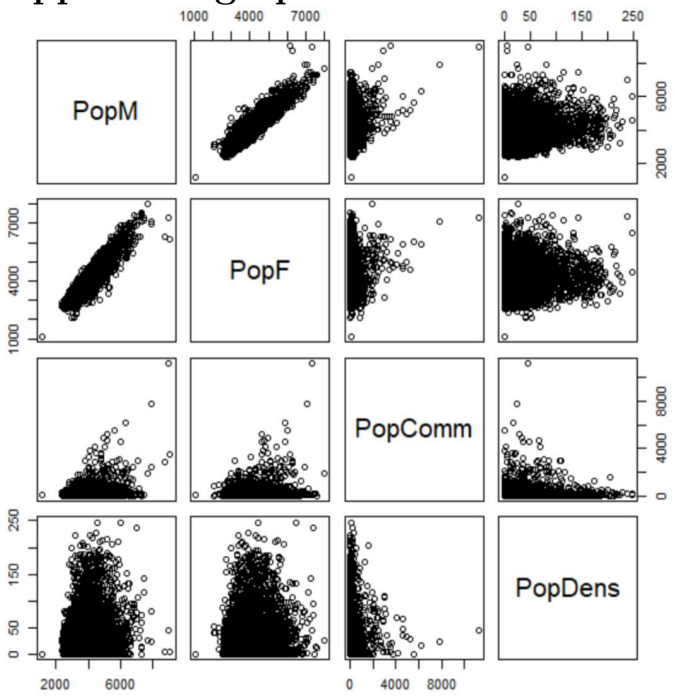

Figure 1. Matrix plot of population

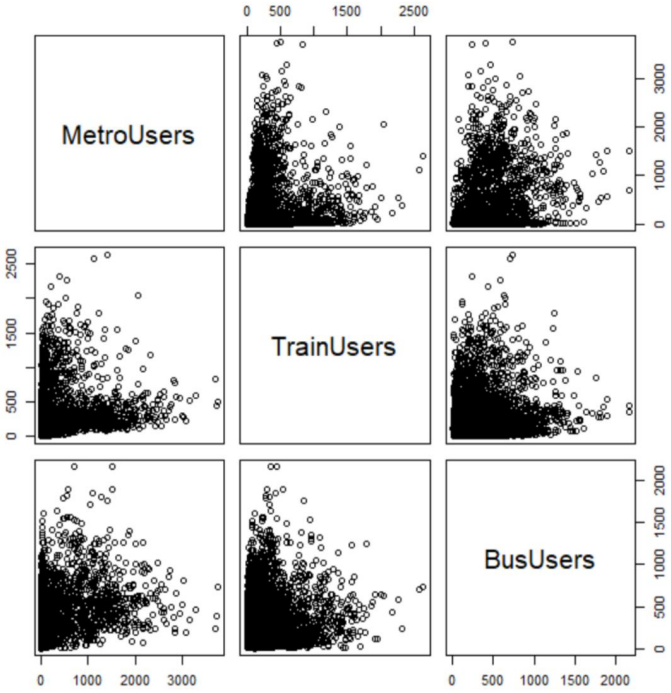

Figure 2. Matrix plot of public transport

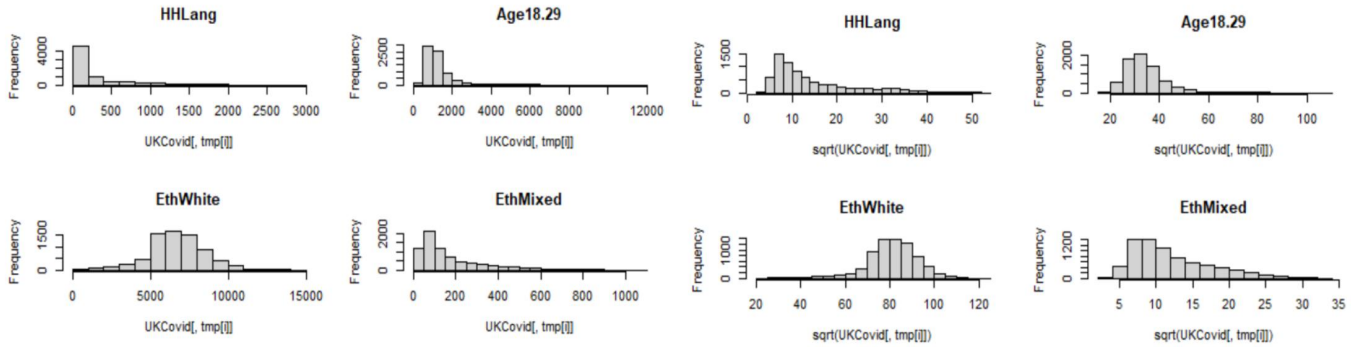

Figure 3. Histogram before and after square root transformation

Figure 4. Interaction between population density and new.RU by faceting

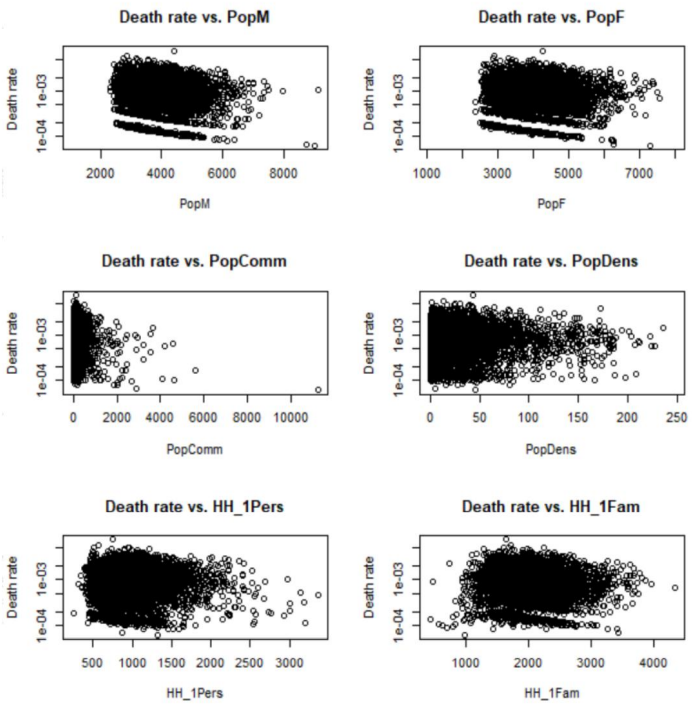

Figure 5. Scatterplot of death rate vs. covariates

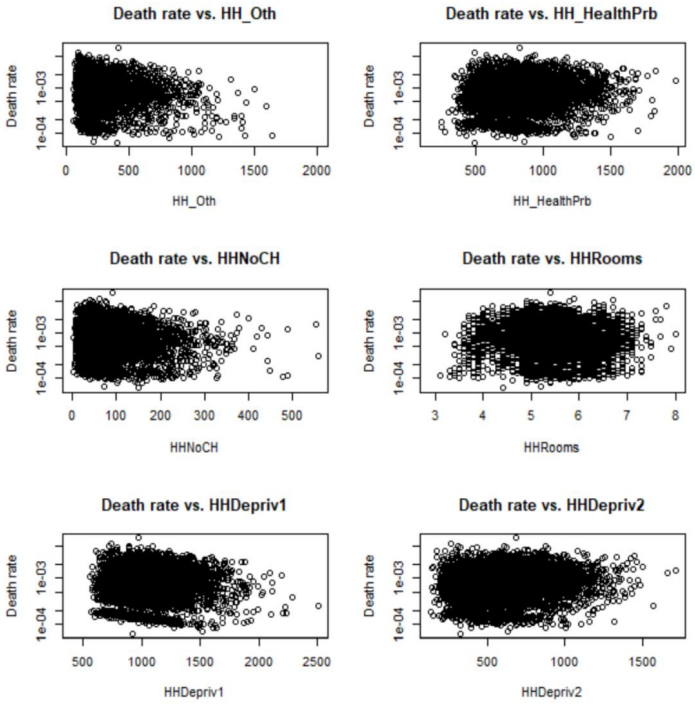

Figure 6. Scatterplot of death rate vs. covariates

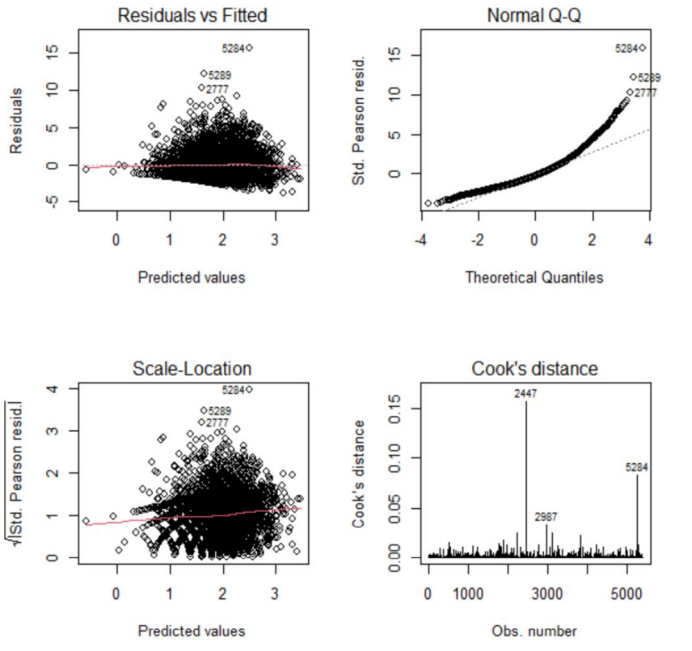

Figure 7. Diagnostic plots of model 1

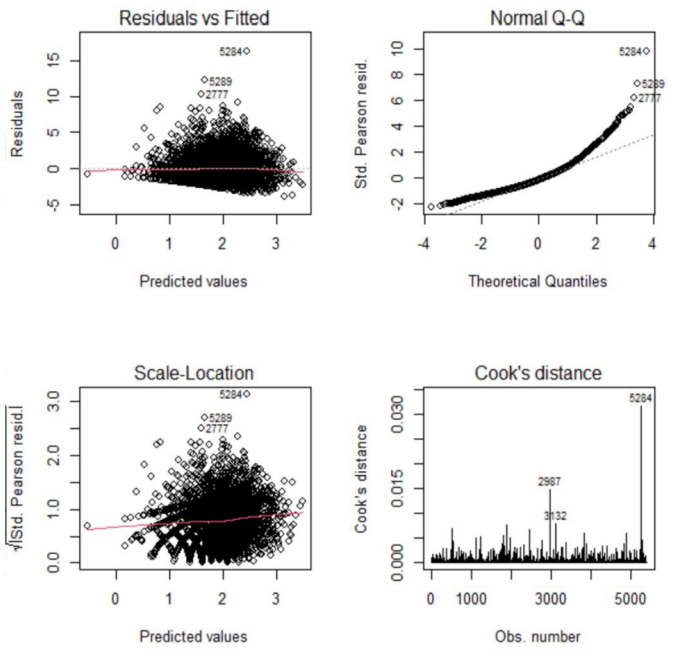

Figure 8. Diagnostic plots of model 3

5 Reference

1. Ons.gov.uk. 2021. Analysis of geographic concentrations of COVID-19 mortality over time, England and Wales - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/analysisofgeographicconcentrationsofcovid19mortalityovertimeenglandandwales/deathsoccurringbetween22februaryand28august2020#toc [Accessed 5 April 2021].

2. Stokes, E., Zambrano, L., Anderson, K., Marder, E., Raz, K., El Burai Felix, S., Tie, Y. and Fullerton, K., 2021. Coronavirus Disease 2019 Case Surveillance — United States, January 22–May 30, 2020.

3. Centers for Disease Control and Prevention. 2021. Cases, Data, and Surveillance. [online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html#footnote01 [Accessed 7 April 2021].

4. The Health Foundation. 2021. What has been the impact of COVID-19 on care homes and the social care workforce? | The Health Foundation. [online] Available at: https://www.health.org.uk/news-and-comment/charts-and-infographics/what-has-been-the-impact-of-covid-19-on-care-homes-and-social-care-workforce [Accessed 18 April 2021].

5. Ons.gov.uk. 2021. Coronavirus (COVID-19) related deaths by occupation, England and Wales - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/bulletins/coronaviruscovid19relateddeathsbyoccupationenglandandwales/deathsregisteredbetween9marchand28december2020 [Accessed 18 April 2021].

6 Contribution

Both members contributed equally.