

# STAT0023 Computing for Practical Statistics

## In-course assessment 2, take-home component (2020–21 session)

### Table of Contents

Rubric.....	2
Background and overview .....	3
Detailed instructions.....	4
Marking criteria.....	5
Hints on tackling the assessment.....	7
Appendix: the UKCovidWave1.csv dataset.....	10
Data sources and pre-processing.....	10
Description of variables.....	12
Overall information about each MSOA and its population .....	12
Household information for each MSOA.....	13
Age profile for each MSOA: variables Age0-4, Age5-7, ..., Age90+ .....	13
Ethnicity and immigration .....	13
Unpaid carers.....	14
Household accommodation .....	14
People living in communal establishments.....	14
Employment / occupation .....	14
Social grade: variables GradeAB, GradeC1, GradeC2 and GradeDE .....	15
Public transport use: variables MetroUsers, TrainUsers and BusUsers .....	15
Education and qualifications.....	15

---

## Rubric

- Your solutions should be your own work and are to be submitted electronically to the course Moodle page by **12 noon on MONDAY, 26TH APRIL 2021**.
  - You can work either alone or in pairs for this assessment. It is up to you to form your own pairs. *You MUST register your choices on Moodle by 12 noon on MONDAY, 29TH MARCH 2021, even if you choose to work alone.*
  - If you choose to work in a pair, you will be jointly responsible for the work that is submitted and you will be awarded the same mark.
  - Ensure that you electronically 'sign' the plagiarism declaration on the Moodle page when submitting your work. If you choose to work in a pair, both of you should check what has been submitted before signing this declaration: if any plagiarism or collusion is identified with anyone outside your pair, you will share responsibility for it.
  - Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation and notified within one week of the deadline above. Penalties, and the procedure in case of extenuating circumstances, are set out in the latest editions of the Statistical Science Department student handbooks which are available from the departmental web pages.
  - Failure to submit this in-course assessment will mean that your overall examination mark is recorded as "non-complete", i.e. you will not obtain a pass for the course.
  - Submitted work that exceeds the specified word count will be penalized. The penalties are described in the detailed instructions below.
  - Your solutions should be your own work. When uploading your scripts, you will be required to electronically sign a statement confirming this, and that you have read the Statistical Science department's guidelines on plagiarism and collusion (see below).
  - Any plagiarism or collusion can lead to serious penalties for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the departmental student handbooks: the relevant extract is provided on the 'In-course assessment 2' tab on the STAT0023 Moodle page. The Turn-It-In plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
  - You will receive feedback on your work via Moodle, and you will receive a provisional grade. *Grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2021.*
-

## Background and overview

When the Covid-19 pandemic was first recognised in early 2020, it quickly became apparent that age was the main risk factor for becoming seriously ill or dying from the disease. Researchers have also identified other risk factors including gender, social deprivation, pre-existing health conditions and ethnicity.<sup>1</sup> Understanding these risk factors can potentially help to develop strategies for reducing deaths, for example by targeting appropriate healthcare resources in areas that need them the most.<sup>2</sup>

In the UK, the Office for National Statistics (ONS) publishes a variety of information on Covid. An ONS report from August 2020<sup>3</sup> produced a simple analysis of Covid death rates across England and Wales, between March and July 2020. In this assessment we will examine more closely the data used in that report and try to understand why some areas have more deaths than others, by linking to UK Census data on the socio-economic characteristics of the different areas.

We will use data consisting of the total numbers of reported deaths in the period March–July 2020, where Covid-19 was given as the cause of death, for each of 7201 “Middle Layer Super Output Areas” (MSOAs) in England and Wales. According to the ONS report cited above, Super Output Areas are “small-area statistical geographies covering England and Wales”, each of which has a similarly sized population and remains stable over time. These data are from the [ONS web site](#).<sup>4</sup> They have been combined with demographic and socioeconomic data from the most recent UK Census in 2011, obtained by querying datasets at the [Nomis Labour Market Statistics service](#), and also with some geographic information from the UK’s [Open Geography Portal](#).

The data are provided in the file `UKCovidWave1.csv`, available from the ‘In-course assessment 2’ tab of the STAT0023 Moodle page. This contains an anonymised version of the original data. Full details, including the anonymisation procedure (which includes rounding of most variables) can be found in the Appendix to these instructions. The first 5 401 rows are complete, i.e., contain all values of the death count and covariates. The last 1 800 rows contain all values of the covariates, but -1 for the death counts.

Your task in this assessment is to use the data from the first 5 401 records, to build a statistical model that will help you to:

- Understand the social, demographic and economic factors associated with variation between MSOAs in numbers of Covid deaths during the period March–July 2020; and
- Estimate the numbers of deaths for each of the 1 800 records where you don’t have this information.

---

<sup>1</sup> See, for example, Williamson *et al.* (2020): “Factors associated with COVID-19-related death using OpenSAFELY” ([Nature 584, pp. 430–436](#)).

<sup>2</sup> For a more general overview of the key role that statistics has to play in responding to crises, see the Royal Statistical Society’s [Ten recommendations on better use of stats and data in a pandemic](#), released on 8<sup>th</sup> March 2021.

<sup>3</sup> ONS Statistical Bulletin “[Deaths involving COVID-19 by local area and socioeconomic deprivation: deaths occurring between 1 March and 31 July 2020](#)”, published August 2020.

<sup>4</sup> Here and elsewhere, clicking on the blue text will take you to the relevant web site.

## Detailed instructions

You may use either R or SAS for this assessment.

1. Read the data into your chosen software package and carry out any necessary recoding (e.g. to deal with the fact that -1 represents a missing value).
2. Carry out an exploratory analysis that will help you to start building a sensible statistical model to understand and predict the numbers of Covid deaths in each MSOA. This analysis should aim to identify an appropriate set of candidate variables to take into the subsequent modelling exercise, as well as to identify any important features of the data that may have some implications for the modelling. You will need to consider the context of the problem to guide your choice of exploratory analysis. See the 'Hints' below for some ideas.
3. Using your exploratory analysis as a starting point, develop a statistical model that enables you to *predict* the number of Covid deaths for each MSOA based on (a subset of) the other variables in the dataset, and also to *understand* the variation in deaths between different MSOAs. To be convincing, you will need to consider a range of models and to use an appropriate suite of diagnostics to assess them. Ultimately however, you are required to recommend a single model that is suitable for interpretation, and to justify your recommendation. Your chosen model should be either a linear model, a generalized linear model or a generalized additive model.
4. Use your chosen model to predict the number of Covid deaths for each MSOA where this information is missing, and also to estimate the standard deviation of your prediction errors.

Submission for this assessment is electronic, via the STAT0023 Moodle page. You are required to submit three files, as follows:

- A report on your analysis, not exceeding 2 500 words of text plus two pages of graphs and / or tables. The word count includes titles, footnotes, appendices, references etc. – in fact it includes everything except the two pages of graphs / tables and, if present, the separate page describing the contribution of each pair member (see below). Your report should be in three sections, as follows:

**Section I:** Describe briefly what aspects of the problem context you considered at the outset, how you used these to start your exploratory analysis, and what were the important points to emerge from this exploratory analysis.

**Section II:** Describe briefly (without too many technical details) what models you considered in step (3) above, and why you chose the model that you did.

**Section III:** State your final model clearly, summarise what your model tells you about the factors associated with variation of death counts in each MSOA, and discuss any potential limitations of the model.

Your report should not include any computer code. It should include some graphs and / or tables, but only those that support your main points. **Graphs and tables must appear on separate pages**, or they will be included in the word count.

In addition to your data analysis, **if you are working as a pair then you must include an additional page at the end of their report where each pair member briefly describes their contribution to the project.** You will need to agree this in your pairs before

submitting the report. If both pair members agree that they contributed equally then it is sufficient to write a single sentence to that effect, or alternatively you are very welcome to describe your own personal contribution to the project. Note that this page will not be marked and does not contribute to the word count; nor will different marks be allocated to different pair members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of group-work.

**Your report should be submitted as a PDF file named as #####\_rpt.pdf, where ##### is your group ID, with any spaces replaced by underscores (IMPORTANT!!!).** For example, if your group ID is 'ICA2Group C', your report should be named ICA2Group\_C\_rpt.pdf.

- An R script or SAS program corresponding to your analysis and predictions. Your script/program should run *without user intervention* on any computer with R or SAS installed, providing the file UKCovidWave1.csv is present in the current working directory / current folder. When run, it should produce any results that are mentioned in your report, together with the predictions and the associated standard deviations. **The script / program should be named #####.r or #####.sas as appropriate, where ##### is your group ID with underscores instead of spaces.** For example, if your group ID is 'ICA2Group C' and you use R, your should be named ICA2Group\_C.r.

You may not create any additional input files that can be referenced by your script; nor should you write any code that requires access to the internet in order to run it. If you use R however, you may use the following additional libraries if you wish (together with other libraries that are loaded automatically by these): mgcv, ggplot2, grDevices, RColorBrewer, lattice and MASS. You may not use any other add-on libraries: for present purposes, an "add-on library" is one that requires a library() or require() command or equivalent (e.g. the package::command syntax) before it can be used, if your R system is installed using default settings.

- A text file containing your predictions for the 1 800 observations with missing counts. **This file should be named #####\_pred.dat, where ##### is your group ID with underscores instead of spaces.** The file should contain three columns, separated by spaces and with *no header*. The first column should be the record identifier (corresponding to variable ID in file UKCovidWave1.csv); the second should be the corresponding count prediction, and the third should be the standard deviation of your prediction error.
- **NOTE:** if you work in pairs, **both members of a pair must confirm their submission on Moodle before the submission deadline.**

## Marking criteria

There are 75 marks for this exercise. These are broken down as follows:

- **Report: 40 marks.** The marks here are for: displaying awareness of the context for the problem and using this to inform the statistical analysis; good judgement in the choice of exploratory analysis and in the model-building process; a clear and well-justified argument; clear conclusions that are supported by the analysis; and appropriate choice and presentation of graphs and / or tables. The mark breakdown is as follows:

- *Awareness of context: 5 marks.*
- *Exploratory analysis: 10 marks.* These marks are for (a) tackling the problem in a sensible way that is justified by the context (b) carrying out analyses that are designed to inform the subsequent modelling.
- *Model-building: 10 marks.* The marks are for (a) starting in a sensible place that is justified from the exploratory analysis (b) appropriate use of model output and diagnostics to identify potential areas for improvement (c) awareness of different modelling options and their advantages and disadvantages (d) consideration of the social, economic and demographic context during the model-building process.
- *Quality of argument: 5 marks.* The marks are for assembling a coherent 'narrative', for example by drawing together the results of the exploratory analysis so as to provide a clear starting point for model development, presenting the model-building exercise in a structured and systematic way and, at each stage, linking the development to what has gone before.
- *Clarity and validity of conclusions: 5 marks.* These marks are for stating clearly what you have learned about how and why the numbers of deaths vary between MSOAs, and for ensuring that this is supported by your analysis and modelling.
- *Graphs and / or tables: 5 marks.* Graphs and / or tables need to be relevant, clear and well presented (for example, with appropriate choices of symbols, line types, captions, axis labels and so forth). There is a one-slide guide to 'Using graphics effectively' in the slides / handouts for the Week 1 videos for the course. **Note** that you will only receive credit for the graphs in your report if your submitted script / program generates and automatically saves all of these graphs when it is run.

**Note** that you will be penalised if your report exceeds EITHER the specified 2 500-word limit or the number of pages of graphs and / or tables. Following [UCL guidelines](#), the maximum penalty is 7 marks, and no penalty will be imposed that takes the final mark below 30/75 if it was originally higher. Subject to these conditions, penalties are as follows:

- *More than two pages of graphs and / or tables: zero marks for graphs and / or tables, in the marking scheme given above.*
- *Exceeding the word count by 10% or less: mark reduced by 4.*
- *Exceeding the word count by more than 10%: mark reduced by 7.*

*In the event of disagreement between reported word counts on different software systems, the count used will be that from the examiner's system. The examiners will use an R function called `PDFcount` to obtain the word count in your PDF report: this function is available from the Moodle page in file `PDFcount.r`.*

- **Coding: 15 marks.** There are 3 marks here for reading the data, preprocessing and setting up variable names correctly and efficiently; 7 marks for effective use of your chosen software in the exploratory analysis and modelling (e.g. programming efficiently and correctly); and 5 marks for clarity of your code – commenting, layout, choice of variable / object names and so forth.
- **Prediction quality: 20 marks.** The remaining 20 marks are for the quality of your predictions. **Note**, however, that you will only receive credit for your predictions if your

*submitted#####\_pred.dat* file is identical to that produced by your script / program when it is run: if this is not the case, your predictions will earn zero marks.

For these marks, you are competing against each other. Your predictions will be assessed using the following score:

$$S = \sum_{i=1}^{1800} \left[ \log \sigma_i + \frac{(Y_i - \hat{\mu}_i)^2}{2\sigma_i^2} \right],$$

where:

- $Y_i$  is the actual number of deaths (which the examiners know) for the  $i$ th prediction;
- $\hat{\mu}_i = \mathbb{E}(Y_i)$  is your corresponding prediction;
- $\sigma_i$  is your quoted standard deviation for the prediction error.

The score  $S$  is an approximate version of a *proper scoring rule*, which is designed to reward predictions that are close to the actual observation and are also accompanied by an accurate assessment of uncertainty (this was discussed during the Week 10 lecture, along with the rationale for using this score for the assessment). Low values are better. The scores of all of the students in the class (and the lecturer) will be compared: students with the lowest scores will receive all 20 marks, whereas those with the highest scores will receive fewer marks. The precise allocation of marks will depend on the distribution of scores in the class.

If you don't supply standard deviations for your prediction errors, the values of the  $\{\sigma_i\}$  will be taken as zero: this means that your score will be  $-\infty$  if you predict every value perfectly (this is the smallest possible score, so you'll get 20 marks in this case), and  $+\infty$  otherwise (this will earn you zero marks).

## Hints on tackling the assessment

1. There is not a single 'right' answer to this assignment. There is a huge range of options available to you, and many of them will be sensible.
2. You are being assessed not only on your computing skills, but also on your ability to carry out an informed statistical analysis: material from other statistics courses (in particular STAT0006, for students who have taken it) will be relevant here. To earn high marks, you need to take a structured and critical approach to the analysis and to demonstrate appropriate judgement in your choice of material to present.
3. At first sight, the task will appear challenging. However, there is a lot of information that can guide you: look at some of the web links earlier in these instructions, and at other commentaries on Covid, to gain some understanding of what kinds of relationships you might look for in the data.
4. When building your model, you have two main decisions to make. The first is: should it be a linear, generalized linear or generalized additive model? The second is: which covariates should you include? You might consider the following points:
  - **Linear, generalized linear or generalized additive?** This is best broken down into two further questions, as follows:

- *Conditional on the covariates, can the response variable be assumed to follow a normal distribution with constant variance?* In this assignment, the response variable cannot be negative and it is an integer. Therefore, it cannot have exactly a normal distribution. However, you may find that the residuals from a linear regression model are *approximately* normal – and you may judge that the approximation is adequate for your purposes. The ‘constant variance’ assumption may also be suspect: for positive-valued quantities, it is common for the variability to increase with the mean. If this is the case here, you need to decide whether it varies enough to matter: you need to think about whether the effect is big enough that you can improve your predictions (and hence your score!) by accounting for it e.g. using a GLM. You might consider using your exploratory analysis to gain some preliminary insights into this point.
- *Are the covariate effects best represented parametrically or nonparametrically?* Again, your exploratory analysis can be used to gain some preliminary insights into this. You may want to look at the material from week 6, for examples of situations where a nonparametric approach is needed.
- **Which covariates?** The data file contains a lot of potential covariates, some of which are more important than others. You have many choices here, and you will need to take a structured approach to the problem in order to avoid running into difficulties. The following are some potentially useful ideas:
  - *Look at other literature on risk factors for death from Covid.* What are considered to be the most important risk factors for serious illness or death from Covid? Can these be linked to covariates for which you have information? Obviously, if you do this then you will need to acknowledge your sources in your report.
  - *Define new variables based on the correlations between the existing variables, and work with these.* If several continuous variables are highly correlated, then it is difficult to disentangle their effects and it may be preferable to work with a single ‘index’ that combines all of them. This is the basis of techniques such as Principal Components Analysis, that were discussed during the Week 10 lecture (along with how to implement them in R and SAS).

You should not start to build any models until you have formed a fairly clear strategy for how to proceed. Your decisions should be guided by your exploratory analysis, as well as your understanding of the context.

5. Don’t forget to look for interactions! For example, in rural areas it may take longer to access emergency hospital treatment than in urban areas: delays in obtaining help may further increase the risk of death for those who are already in the most vulnerable groups, although the risk may not change so much for those who are less vulnerable.
6. You probably won’t find a perfect model in which all the assumptions are satisfied: models are just models. Moreover, you should not necessarily expect that your model will have much predictive power: maybe the covariates in the data set just don’t provide very much



useful information. You should focus on finding the best model that you can, therefore – and acknowledge any deficiencies in your discussion.

7. To obtain the standard deviations of your prediction errors, you need to do some calculations. Specifically:
  - i. Suppose  $\hat{\mu}_i = \hat{\mathbb{E}}(Y_i)$  is your  $i$ th predicted death count and that  $Y_i$  is the corresponding actual value.
  - ii. Then your prediction error will be  $Y_i - \hat{\mu}_i$ .
  - iii.  $Y_i$  and  $\hat{\mu}_i$  are independent, because  $\hat{\mu}_i$  is computed using only information from the first 5 401 records and  $Y_i$  relates to one of the 'new' records.
  - iv. The *variance* of your prediction error is thus equal to  $\text{Var}(Y_i) + \text{Var}(\hat{\mu}_i)$ .
  - v. You can calculate the standard error of  $\hat{\mu}_i$  in both R and SAS, when making predictions for new observations – see the materials from Weeks 6 and 9. Squaring this standard error gives you  $\text{Var}(\hat{\mu}_i)$ .
  - vi. You can estimate  $\text{Var}(Y_i)$  by plugging in the appropriate formula for your chosen distribution – for example, if you're using a linear model then this is just the error variance estimate, whereas if you're using a Poisson distribution (which is a possibility when the response variable is a count) then  $\widehat{\text{Var}}(Y_i) = \hat{\mu}_i$ .
  - vii. Hence you can estimate the standard deviation of your prediction error as  $\hat{\sigma}_i = \sqrt{\widehat{\text{Var}}(Y_i) + \text{Var}(\hat{\mu}_i)}$ . In fact, for the case of linear models this is exactly the calculation that is used in the construction of prediction intervals (see your STAT0006 notes or equivalent).
8. Larger MSOAs will tend to have more deaths simply because they have bigger populations. You may therefore think that it is more sensible to model the effects of covariates upon the death *rates* (i.e. the numbers of deaths as a proportion of the population) instead of the actual counts. However, the assignment instructions tell you to model and predict the *numbers* of deaths. One option here would be to fit models to the rates and then to derive the corresponding expressions for the counts (since the count is equal to the rate times the population size). If you use a Poisson or quasiPoisson GLM or GAM however, there is a more elegant approach. To illustrate, suppose you use a GLM in which the number of deaths  $Y_i$  has a Poisson distribution – with mean  $\mu_i$  say, where  $\mu_i$  potentially depends on the covariates. If the corresponding population size (i.e. value of PopTot – see Appendix) is  $P_i$  then the death rate is  $Y_i/P_i$ , which has mean  $\mu_i/P_i = \lambda_i$  say and which does *not* have a Poisson distribution (notice, for example, that it can take non-integer values). Notice, however, that  $\mu_i = P_i \lambda_i$  so that

$$\log \mu_i = \log P_i + \log \lambda_i . \quad (1)$$

If you fit a GLM or a GAM to the counts  $\{Y_i\}$  therefore, then you could use a log link function and include a fixed term  $\log P_i$  in the linear predictor. In the GLM case, this gives a model of the form

$$\log \mu_i = \log P_i + \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} . \quad (2)$$

Comparing (2) with (1), you can now see that fitting the Poisson model (2) to the counts  $\{Y_i\}$  is equivalent to fitting the model

$$\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

to the death rates  $\{Y_i/P_i\}$  – so this gives you a way to model the effects of covariates on these rates directly. The term  $\log P_i$  in (2) is called an *offset*. Both R and SAS allow the specification of offsets when fitting GLMs – look at the help system for details.

---

## Appendix: the UKCovidWave1.csv dataset

### Data sources and pre-processing

The data provided for the analysis are from three sources, as follows:

- **Data source “ONS”: numbers of Covid deaths in each MSOA**, from the [ONS web site](#). The dataset was downloaded on 3<sup>rd</sup> March 2021 and contains information on deaths between 1 March 2020 to 31 July 2020. The documentation states that “to protect confidentiality, a small number of deaths have been reallocated between neighbouring areas. Due to the method used for this, figures for some areas may be different to previously published data”.
- **Data source “Nomis”: demographic and socioeconomic data from the 2011 UK Census**, obtained by querying datasets at the [Nomis Labour Market Statistics service](#). The data provided are from some of the Nomis “Key Statistics” and “Quick Statistics” datasets, accessed between 5<sup>th</sup> and 8<sup>th</sup> March 2021. The specific datasets used were:
  - KS101EW: Usual resident population
  - KS102EW: Age structure
  - KS105EW: Household composition
  - KS106EW: Adults not in employment and dependent children and persons with long-term health problems or disability for all households
  - KS201EW: Ethnic group
  - KS206EW: Household language
  - KS301EW: Health and provision of unpaid care
  - KS401EW: Dwellings, household spaces and accommodation type
  - KS403EW: Rooms, bedrooms and central heating
  - KS405EW: Communal establishment residents
  - KS501EW: Qualifications and students
  - KS608EW to KS610EW: Occupation by sex
  - QS119EW: Households by deprivation dimensions
  - QS611EW: Approximated Social Grade
  - QS701EW: Method of travel to work

The Nomis documentation states “In order to protect against disclosure of personal information from the 2011 Census, there has been swapping of records in the Census database between different geographic areas, and so some counts will be affected. In the main, the greatest effects will be at the lowest geographies, since the record swapping is

targeted towards those households with unusual characteristics in small areas". More details can be found [here](#).

- **Data source "OGP": geographic information** from the UK's [Open Geography Portal](#). These data provide information on where each MSOA is located, and whether it is urban or rural. Specifically:
  - The OGP "[Output Area to LSOA to MSOA to Local Authority District \(December 2017\) Lookup with Area Classifications in Great Britain](#)" (RGC) was used to identify the region of England and Wales to which each MSOA belongs. The data provided were downloaded on 4<sup>th</sup> March 2021.
  - The OGP "[Rural Urban Classification \(2011\) of Middle Layer Super Output Areas in England and Wales](#)" was used to obtain an urban-rural classification of each MSOA (more details below). The data were downloaded on 10<sup>th</sup> March 2021.

The data were merged and preprocessed, to prevent you from being able to identify the actual values that you're supposed to be predicting from information available online. The preprocessing involved some anonymisation and transformations, that will have a negligible effect on any models that are fitted. Specifically:

1. The 18 original datasets were merged using the nine-digit MSOA identification code (identified as "MSOA code" in data source ONS).
2. The rows of the data table were randomly shuffled, so that the order of MSOAs no longer corresponds to that in any of the data sources. This was done in order to prevent 'cheating' when making predictions.
3. A sample of roughly 75% of the records was selected for use in the 'model building' part of the assessment (this will be referred to as 'Group 1' below), with the remaining 25% used for 'prediction' ('Group 2'). This was done in such a way that the two samples were non-overlapping but had very similar distributions for all numeric covariates. Specifically:
  - i. For each region, 75% of the observations were sampled at random, without replacement, as candidates to use in Group 1; and the remaining 25% were allocated to Group 2.
  - ii. For each of the numeric covariates in the data set, a Kolmogorov-Smirnov test was performed to test the null hypothesis that the underlying distributions in Groups 1 and 2 are the same.
  - iii. The samples were accepted only if the  $p$ -values for *all* of the Kolmogorov-Smirnov tests were greater than 0.25. Otherwise, a new candidate sample was drawn in step (i) and the procedure was repeated.

The Kolmogorov-Smirnov test is used here as a convenient way to measure whether two distributions are roughly similar. Note, however, that the death counts were *not* included in this balancing exercise: this is because the performance of predictions would be artificially enhanced if they were included (for example, we would know that the mean number of deaths for Group 2 is similar to that for Group 1). Note also that no attempt has been made to balance the groups in terms of *combinations* of the covariates.

4. The data table was sorted by group, and within that by region; the death counts for the 'Group 2' records were set to -1; and the original MSOA code was replaced with a new ID variable taking values from 1 to 7 201.

5. Each of the numeric covariates was rounded, to a different resolution depending on the original value. The changes due to the rounding are within  $\pm 0.5\%$  in all cases.

## Description of variables

This section gives a brief description of each of the variables in UKCovidWave1.csv, and an indication of which data source it came from. Descriptions are provided on the basis of information provided in the original data sources (links given above). For convenience, the variables have been grouped into broad categories in the descriptions below – although in practice, some variables may be considered to belong to more than one category.

### Overall information about each MSOA and its population

#### Variable

name	Source	Description
ID	–	MSOA identifier, assigned during step 4 of the preprocessing described above
Region	OGP	Name of the geographic region containing the MSOA
RUCode	OGP	Rural-urban categorisation of the MSOA. See below for an explanation of the codes
Deaths	ONS	Number of deaths from Covid, during the period from 1 <sup>st</sup> March to 31 <sup>st</sup> July 2020
PopTot	Nomis	Total population, according to the 2011 Census
PopM	Nomis	# males in the population <sup>5</sup>
PopF	Nomis	# females in the population
PopComm	Nomis	# people living in a communal establishment
PopDens	Nomis	Population density (individuals per hectare)

#### *Key to rural-urban classification*

The RUCode variable is coded as follows:

- A1: Urban major conurbation
- B1: Urban minor conurbation
- C1: Urban city and town
- C2: Urban city and town in a sparse setting
- D1: Rural town and fringe
- D2: Rural town and fringe in a sparse setting
- E1: Rural village and dispersed
- E2: Rural village and dispersed in a sparse setting

All of the remaining variables are from the Nomis data source.

---

<sup>5</sup> Here and elsewhere, '#' means 'number of'.

## Household information for each MSOA

Variable name	Description
HH	Total # households in the MSOA
HH_1Pers	# single-person households
HH_1Fam	# single-family households
HH_Oth	# "other" households
HH_HealthPrb	# households where at least one person has a long-term health problem or disability
HHNoCH	# households without central heating
HHRooms	Average # rooms per household
HHBedrooms	Average # bedrooms per household
HHDepriv1, HHDepriv2, HHDepriv3, HHDepriv4	# households deprived in 1, 2, 3 or 4 dimensions according to the 2011 census definition (described in Part 4 of the "Variables and classifications" section of the <a href="#">2011 Census user guide</a> )
HHAdultUKLang	# households where at least one but not all people aged 16 and over in household has English (or Welsh in Wales) as a main language
HHChildUKLang	# households where no people aged 16 and over has English (or Welsh in Wales) as a main language, but at least one person aged 3 to 15 does.
HHNoUKLang	# households where nobody has English (or Welsh in Wales) as a main language

## Age profile for each MSOA: variables Age0-4, Age5-7, ..., Age90+

These variables give the numbers of people in the specified age ranges. MeanAge and MedianAge are the mean and median age, in years.

## Ethnicity and immigration

Variable name	Description
EthWhite	# individuals self-identifying as "White"
EthMixed	# individuals self-identifying as of "mixed" ethnicity or "multiple ethnic groups"
EthAsian	# individuals self-identifying as "Asian" or "Asian British"
EthBlack	# individuals self-identifying as "Black", "African", "Caribbean" or "Black British"
EthOther	# individuals self-identifying as being from another ethnic group
BornIreland	# individuals born in the Republic of Ireland (RoI)
BornEU	# individuals born in the European Union (excluding UK and the RoI)
BornNonEU	# individuals born elsewhere in the world

## Unpaid carers

The UK census documentation states that “a person is a provider of unpaid care if they look after or give help or support to family members, friends, neighbours or others because of long-term physical or mental ill health or disability, or problems related to old age”, and are not paid for it.

Variable name	Description
CarersLo	# individuals providing between 1 and 19 hours of unpaid care per week
CarersMid	# individuals providing between 20 and 49 hours of unpaid care per week
CarersHi	# individuals providing 50 or more hours of unpaid care per week

## Household accommodation

These variables summarise the ‘dwellings’ in each MSOA. A dwelling (e.g. an address) is shared if it is used by more than one household, and if it contains shared facilities (e.g. kitchen and bathroom) that are used by more than one household.

Variable name	Description
Dwell	Total # dwellings in the MSOA
DwellShared2	# dwellings shared by two households
DwellShared3+	# dwellings shared by three or more households

## People living in communal establishments

“Communal establishments” include hospitals and care homes.

Variable name	Description
CommEstab	# communal establishments in the MSOA
LACare	Total # residents in a local authority or other care home
PrivCareNurs	Total # residents in a private care home with nursing
PrivCareNoNurs	Total # of residents in a private care home without nursing

## Employment / occupation

Occupations are classified according to the 2010 ONS [Standard Occupational Classification](#). The data represent numbers of individuals aged from 16 to 74, who were in employment at the time of the census.

Variable name	Description
WrkMgr	# individuals working as “managers, directors and senior officials”
WrkProf	# individuals working in “professional occupations”
WrkProfTech	# individuals working in “associate professional and technical occupations”
WrkAdmin	# individuals working in “administrative and secretarial occupations”
WrkSkilled	# individuals working in “skilled trades occupations”
WrkCaring	# individuals working in “caring, leisure and other service occupations”

WrkSales	# individuals working in "sales and customer service occupations"
WrkMachine	# of individuals working as "process plant and machine operatives"
WrkElementary	# of individuals in "elementary occupations"

### **Social grade: variables GradeAB, GradeC1, GradeC2 and GradeDE**

The Nomis documentation states that the "social grade" data relate to the number of "household reference persons" (HRPs) aged from 16 to 64 on the date of the census; and that "social grade is the socio-economic classification used by the Market Research and Marketing Industries, most often in the analysis of spending habits and consumer attitudes. Although it is not possible to allocate social grade precisely from information collected by the 2011 Census, the Market Research Society has developed a method for using census information to provide a good approximation of social grade".

Variables GradeAB, GradeC1, GradeC2 and GradeDE are respectively the number of HRPs in social grades A or B, C1, C2, and D or E. Social grade definitions are given by the [Market Research Society](#), and are roughly as follows:

- A or B: professionals and senior managers, middle-management executives, small business owners
- C1: Supervisory, clerical and junior managerial, administrative, professional occupations
- C2: skilled manual occupations
- D or E: Semi-skilled and unskilled manual occupations, unemployed and lowest grade occupations

### **Public transport use: variables MetroUsers, TrainUsers and BusUsers**

These variables represent the numbers of individuals using the respective method of travel "for the longest part, by distance, of the usual journey to work." The data are restricted to the usual residents of an MSOA, who were aged from 16 to 74 and who were in work during the week before the census date. "Metro" includes underground, metro, light rail and tram.

### **Education and qualifications**

Variable name	Description
NoQual	# individuals aged 16 and over, with no academic or professional qualifications
Qual1, Qual2, Qual3, Qual14+	# individuals aged 16 and over, whose highest level of qualification is 1, 2, 3 or "4 and above". The levels are described in Part 4 of the "Variables and classifications" section of the <a href="#">2011 Census user guide</a> : for example, Qual14+ is the number of people educated to at least degree level.
QualApp	# individuals aged 16 and over, whose highest level of qualification is an apprenticeship
QualOther	# individuals aged 16 and over, whose highest level of qualification does not fall in any of the categories above
Stud18+	# of schoolchildren and full-time students aged 18 and over